Causality Pursuit from Heterogeneous Environments via Neural Adversarial Invariance Learning

Yihong Gu[1] Cong Fang[2] Peter Bühlmann[3] Jianqing Fan[1]

[1]Department of Operations Research and Financial Engineering, Princeton University
[2]School of Intelligence Science and Technology, Peking University
[3]Seminar for Statistics, ETH Zürich

This version: June 30, 2024

Abstract

Pursuing causality from data is a fundamental problem in scientific discovery, treatment intervention, and transfer learning. This paper introduces a novel algorithmic method for addressing nonparametric invariance and causality learning in regression models across multiple environments, where the joint distribution of response variables and covariates varies, but the conditional expectations of outcome given an unknown set of quasi-causal variables are invariant. The challenge of finding such an unknown set of quasi-causal or invariant variables is compounded by the presence of endogenous variables that have heterogeneous effects across different environments, including even one of them in the regression would make the estimation inconsistent. The proposed Focused Adversial Invariant Regularization (FAIR) framework utilizes an innovative minimax optimization approach that breaks down the barriers, driving regression models toward prediction-invariant solutions through adversarial testing. Leveraging the representation power of neural networks, FAIR neural networks (FAIR-NN) are introduced for causality pursuit. It is shown that FAIR-NN can find the invariant variables and quasi-causal variables under a minimal identification condition and that the resulting procedure is adaptive to low-dimensional composition structures in a non-asymptotic analysis. Under a structural causal model, variables identified by FAIR-NN represent pragmatic causality and provably align with exact causal mechanisms under conditions of sufficient heterogeneity. Computationally, FAIR-NN employs a novel Gumbel approximation with decreased temperature and stochastic gradient descent ascent algorithm. The procedures are convincingly demonstrated using simulated and real-data examples.

Keywords: Adversarial Estimation, Causal Discovery, Conditional Moment Restriction, Gumbel Approximation, Invariance, Neural Networks.

1 Introduction

A fundamental problem in statistics and machine learning is to predict the response variable Y based on explanatory covariates denoted as $X \in \mathbb{R}^d$ using collected data. The objective often centers on estimating the regression function $m_0(x) = \mathbb{E}[Y|X = x]$, which minimizes the population L_2 risk $\mathsf{R}(m) = \int |y - m(x)|^2 \mu_0(dx, dy)$, starting from the pioneering work of least squares in Legendre (1805); Gauss (1809). In the age of data, the problem of achieving sample-efficient estimation of m_0 was extensively studied. There are a lot of structural methods attempting to exploit the low-dimensional structure such as sparsity, low-rankness and additivity, and design corresponding optimal methods tailored to that assumed structure (Hastie et al., 2009; Wainwright, 2019; Fan et al., 2020). However, these methods lack scalable applicability and suffer from model misspecification due to their reliance on imposed structures. As an alternative, algorithmic methods (Breiman, 2001) like neural networks can be adaptive to the low-dimensional structure efficiently (Schmidt-Hieber, 2020; Fan & Gu, 2024) with no supervision of function structure. This nature endows them with universal applicability across various tasks and data.

Despite many celebrated efforts in the efficient estimation of m_0 or its variants like quantile function, the ultimate goal is to utilize observations to fit a model capable of making decent predictions on unseen data, elucidating the causal relationships among variables, and guiding decision-making in real-world scenarios. We instinctively regard m_0 as such a target function for achieving decent prediction and causal attribution. However, this can be flawed: m_0 can produce unstable predictions on unseen data and risk false scientific conclusions in numerous cases. Consider a simple thought experiment where we aim to classify an object in a picture as either a cow ($Y = 1$) or a camel ($Y = 0$) using two provided features X_1 (body shape) and X_2 (background color). In the data we collected from μ_0, the cows usually appear on green grass, while camels often stay on yellow sand. Consequently, the conditional expectation $m_0(x_1, x_2) = \mathbb{E}_{\mu_0}[Y|X_1 = x_1, X_2 = x_2]$ would be heavily dependent on x_2. Such a model is problematic both for prediction and attribution. Its application in a setting with a different background such as zoos would lead to unreliable predictions. Furthermore, attributing the determination of an object to the background surrounding it also contradicts our understanding of causality. In the above case, we may prefer $m_\star(x) = \mathbb{E}[Y|X_1 = x_1]$ for prediction and attribution as we know the causal mechanisms.

We refer to the above problem as the "*curse of endogeneity*" in that the conditional expectation of the residual for the "potential" interested (causal) $m_\star$ is not zero given all the explanatory variables, i.e., $\mathbb{E}[Y - m_\star(X)|X] \neq 0$, leading to a misalignment between m_0 and $m_\star$, i.e., $m_0(X) - m_\star(X) \neq 0$. Hence traditional regression techniques for estimating m_0 will result in an unsatisfactory solution.

Causal inference methods offer structural remedies to the curse of endogeneity. These methods are structural in that they are tailored to pre-assumed, task-specific, and untestable identification conditions or causal-effect knowledge. This prior knowledge can be formally encoded in the potential outcome (Rubin, 1974), or structural causal model (Glymour et al., 2016) framework and fully shaped the "causality skeleton" that exactly determines the causal estimand of interests by some statistical estimand, and the "association flesh" of the latter can be further estimated via structural or algorithmic regression techniques. Examples include estimating the average treatment effect (Robins et al., 1994) and conditional average treatment effects (Athey et al., 2019; Kennedy et al., 2024) under the unconfoundedness condition. These methods' reliance on prior knowledge limits their scalable use, exposes them to severe model misspecification, and prevents their drawn conclusion from going beyond hindsight because it is impossible to falsify (Popper, 2005) these assumptions using data.

This paper aims to answer the following fundamental question:

> *Can we design methods that can algorithmically circumvent the "curse of endogeneity" without the supervision of cause-effect knowledge?* (Q)

Without prior causal structural knowledge, we leverage the principle of how humans understand causality: the causal association consistently occurs in the past, now, and (potentially) future, or more broadly, in diverse environments. In other words, we pursue certain data-driven or data-shaped causality that is invariant across diverse environments, this is essentially what one can pursue based only on observed data without prior knowledge. Hence we do not differentiate the concepts of invariance and (data-driven) causality in this paper. Levering the invariance principle, we propose a unified and algorithmic framework for causality pursuit that is robust to model misspecification based on data from multiple environments. Though the proposed data-driven causality is conceptually different from previous knowledge-based causality that pre-assumes the ground truth, these two types of causality can coincide when the heterogeneity of environments is sufficient.

1.1 The Canonical Model under Study

Let us revisit the thought experiment from the perspective of a hyper-intelligent alien, Alice. Alice knows nothing about cows and camels except for 1000 images with annotated labels highly associated with the background, for example, $r = 90\%$ cows/camels on grass/sand. It's impossible for her to know that the background cannot determine the object given this limited information. In other words, both X_1 and X_2 can be regarded as causality out of pragmatic considerations. However, if she receives another set of 1000 labeled images, where $r = 70\%$ cows/camels on grass/sand, she might begin to question the causality role of the background: the emerging evidence of the varying associations between X_2 and Y falsify the hypothesis that X_2 is causality if she believes that causality persists across diverse environments.

When there is no supervision of the cause-effect relationship, the observation from heterogeneous sources is essential. We consider the following multi-environment regression problem that mimics human

causality learning. Let $\mathcal{E}$ be the set of sources/environments. For each environment $e \in \mathcal{E}$, we observe n i.i.d. data $(X_1^{(e)}, Y_1^{(e)}), \ldots, (X_n^{(e)}, Y_n^{(e)}) \sim \mu^{(e)}$, where $\mu^{(e)}$, the joint distribution of $(X^{(e)}, Y^{(e)})$, satisfies

$$Y^{(e)} = m^\star(X_{S^\star}^{(e)}) + \varepsilon^{(e)} \qquad \text{with} \qquad \mathbb{E}[\varepsilon^{(e)} | X_{S^\star}^{(e)}] \equiv 0. \tag{1.1}$$

Here $S^\star$, the unknown true important variable set, and $m^\star : \mathbb{R}^{|S^\star|} \to \mathbb{R}$, the target regression function, are both *invariant* across different environments; but the joint distributions $\mu^{(e)}$ can vary. We aim to learn the set of quasi-causal variables $S^\star$ and estimate the *invariant regression function* $m^\star$ using data $\{\{(X_i^{(e)}, Y_i^{(e)})\}_{i=1}^n\}_{e\in\mathcal{E}}$ from $|\mathcal{E}|$ heterogeneous environments. The same n in the problem formulation is just for expository simplicity, the extension to varying $n^{(e)}$ is straightforward. We refer to the above problem as *nonparametric invariance pursuit* or *nonparametric causality pursuit* exchangeably, as based on the data alone, without prior knowledge, we can not differentiate these concepts.

Here, we temporarily refrain from causal discussions. Under particular scenarios, such a problem can be instantiated to causal discovery in the Structural Causal Model (SCM) framework (Peters et al., 2016) and transfer learning with a more realistic assumption (Rojas-Carulla et al., 2018); see the details in Appendix A.1. We offer in Section 3 a rigorous and comprehensive interpretation of what $S^\star$ is in the SCM with interventions on X. It is also notable to mention that model (1.1) only requires invariance in the first moment instead of full distributional invariance, i.e., $\varepsilon^{(e)} \sim F_\varepsilon$ and independent of $X_{S^\star}^{(e)}$, as typically required for causal discovery (Peters et al., 2016). It is more realistic and allows for between-environment heteroscedastic errors.

It is important to note that the standard nonparametric regression generally diverges from our target $m^\star$, i.e., $\mathbb{E}[Y^{(e)}|X^{(e)} = x] \neq m^\star(x_{S^\star})$. This discrepancy arises because $\mathbb{E}[\varepsilon^{(e)}|X^{(e)}] \neq 0$. Such a "curse of endogeneity" problem is the main challenge we need to address. Including even one of endogenous spurious variables, for example, X_2 background color in the above thought experiment, in the regression function will create an inconsistent estimation of $m^\star$. Thus, it is essential to design an algorithm to eliminate all endogenous spurious variables.

1.2 Our Algorithmic Remedy: FAIR Estimation

This paper proposes a unified estimation framework – the *Focused Adversarial Invariance Regularized (FAIR)* estimator. It regularizes the user-specified risk loss $\ell(y, v)$ by a novel regularizer. Specifically, the FAIR estimator is the solution of the following minimax optimization program

$$\min_{g\in\mathcal{G}} \max_{f^{(e)}\in\mathcal{F}_{S_g}, \forall e\in\mathcal{E}} \underbrace{\sum_{e\in\mathcal{E}} \mathbb{E}_{\mu^{(e)}}\left[\ell(Y, g(X))\right]}_{\mathsf{R}(g)} + \gamma \underbrace{\sum_{e\in\mathcal{E}} \mathbb{E}_{\mu^{(e)}}\left[\{Y - g(X)\}f^{(e)}(X) - \{f^{(e)}(X)\}^2/2\right]}_{\mathsf{J}(g, \{f^{(e)}\}_{e\in\mathcal{E}})}. \tag{1.2}$$

Here $\ell(\cdot,\cdot)$ is a loss whose population solution leads to the conditional expectation, $\gamma > 0$ is the regularization hyper-parameter to be determined, $(\mathcal{G}, \mathcal{F})$ are the function classes to be specified by the user satisfying $\mathcal{G} \subseteq \mathcal{F}$. The first part is the risk minimization, and the second component is the test of exogeneity of the variables $S_g = \text{supp}(g)$ used by the regression function g, where $\mathcal{F}_{S_g} = \{f \in \mathcal{F} : f(x) = h(x_{S_g})$ for some $h : \mathbb{R}^{|S_g|} \to \mathbb{R}\}$ is the testing function class for the prediction functions in $\mathcal{G}$ that only "focuses" on the variables S_g that g used. Two useful classes of functions are linear and square-integrable classes for $(\mathcal{G}, \mathcal{F})$, which correspond respectively to linear models and nonparametric regression models; see Section 4.1 for additional details. Note that the second component is nonnegative after maximization by comparing with $f^{(e)} = 0$ so that the penalty is nonnegative. For the empirical counterpart, we solve a similar minimax optimization program that substitutes $\mathbb{E}_{\mu^{(e)}}[\cdot]$ with the corresponding sample means.

To see why such a FAIR penalty works, let us consider the nonparametric regression setting in which $\mathcal{F} = \{f : \mathbb{E}_{\mu^{(e)}}[f^2(X_{S_g})] < \infty\}$. By conditioning on X_{S_g}, for $f^{(e)} \in \mathcal{F}_{S_g}$, we have

$$\mathbb{E}_{\mu^{(e)}}\left[\{Y - g(X)\}f^{(e)}(X)\right] = \mathbb{E}_{\mu^{(e)}}\left[\{\mathbb{E}_{\mu^{(e)}}[Y|X_{S_g}] - g(X)\}\, f^{(e)}(X)\right].$$

Then, the supremum in (1.2) can be explicitly found and the objective now becomes

$$\min_{g\in\mathcal{G}} \mathsf{R}(g) + \gamma \cdot \mathsf{J}^\star(g) \qquad \text{with} \qquad \mathsf{J}^\star(g) = \frac{1}{2}\sum_{e\in\mathcal{E}} \mathbb{E}_{\mu^{(e)}}\left[\left|g(X) - \mathbb{E}_{\mu^{(e)}}[Y|X_{S_g}]\right|^2\right]. \tag{1.3}$$

Therefore, $g(X) = m^\star(X_{S^\star})$ is a minimax solution.

To motivate (1.2), let us first consider the additional constraint $\mathbb{E}_{\mu^{(e)}}[f^{(e)}(X_{S_g})^2] = 1$ so that the first part of the second component in (1.2) is basically the maximal correlation between the residual $\{Y - g(X_{S_g})\}$ and testing functions $f^{(e)}(X_{S_g})$. Hence, the criterion (1.2) is to find a set of variables X_{S_g} as exogenous (weakly correlated) with the residuals as possible for all testing functions in $\mathcal{F}_{S_g}$. By the Lagrange multiplier method, the constrained maximization problem can be written as

$$\max_{f^{(e)} \in \mathcal{F}_{S_g}} \mathbb{E}_{\mu^{(e)}} \left[\{Y - g(X)\} f^{(e)}(X) - \lambda \{f^{(e)}(X)\}^2 \right].$$

Choosing the multiplier $\lambda = 1/2$ (justified in the above paragraph) gives rise to the object function (1.2).

FAIR penalty screens out all endogenous spurious variables when γ is sufficiently large. This is easily seen when the penalty in (1.2) is not zero, such a g is dominated by $g = m^\star$ when γ is sufficiently large. After endogenous spurious variables, we can apply the commonly-used statistical variable selection methods (Hastie et al., 2009; Wainwright, 2019; Fan et al., 2020) to further eliminate exogenous spurious or weak causal variables. In addition, we will show that under the SCM with *arbitrary* and *nondegenerate* interventions on X, our proposed FAIR estimator can unveil $S^\star$ being precisely expressed by the graph structure of the SCM, which can be interpreted as the "pragmatic" direct cause of the response Y in general and will coincide with the direct causes if all the root children are intervened. The obtained result is clearly distinguished from what least squares, or even its worst-case variants like distribution robust optimization (Duchi & Namkoong, 2021), Maximin (Meinshausen & Bühlmann, 2015), can obtain. Our method indeed learns certain data-driven causality, while others cannot go beyond learning associations.

1.3 New Contributions

We propose a unified, algorithmic, and sample-efficient methodological framework that can discover the invariant regression function, i.e. to solve a generalized version of the problem in Section 1.1. The method is simple, universal, fully algorithmic, and sample-efficient: It is just one optimization objective (1.2) complemented by one extra hyper-parameter γ; it accommodates many losses and can be seamlessly integrated by various machine learning algorithms; it does not require any prior structural knowledge, and it is almost as statistically efficient as standard regression under various cases.

As a special instance in our framework, the FAIR neural network (FAIR-NN) estimator is proposed for which $\mathcal{G}$ and $\mathcal{F}$ are neural networks to unveil $m^\star$ in (1.1). It is the *first* theoretically guaranteed estimator that can *efficiently* recover $m^\star$ under a single *general* and *minimal* identification condition associated with the heterogeneity of the environments. Its sample efficiency can be understood in several notable aspects: it requires the minimal identification condition, leading to fewer required environments; it exhibits the same L_2 error rate as if directly regressing Y on known $X_{S^\star}$, regardless of the complexity of spurious associations; and it adapts to the unknown low-dimension structure of the invariant association $m^\star$ in a same manner as Kohler & Langer (2021). In summary, the FAIR-NN estimator circumvents the "curse of dimensionality" and "curse of endogeneity" simultaneously in a fully algorithmic manner, which does not rely on the prior knowledge of $m^\star$ structure or cause-effect relationships among variables.

While the complicated combinatorial constraint and minimax optimization are introduced in (1.2), we show that a variant of gradient descent – gradient descent ascent with Gumbel approximation to handle the combinatorial-nature "focused" constraint $f \in \mathcal{F}_{S_g}$ – continues to apply to our specifically designed algorithm and neural network estimators with no curse-of-dimension in implementation. Numerical results in Section 5 support this.

Though our framework is designed for algorithmic learning, it is versatile in that the user can also incorporate their strong prior structural knowledge such as linearity or additivity of $m^\star$ into the FAIR estimation. This can be realized by restricting the function class $\mathcal{G}$ within this known structure and designating $\mathcal{F}$ as a more expansive class. We demonstrate that harnessing such strong structural knowledge can relax the condition for identification. It is worth pointing out that identification is viable even when $|\mathcal{E}| = 1$ corresponding to observational data; see examples in Appendix B.6. At the methodology level, our method bridges the invariance principle (Peters et al., 2016) and asymmetry principle (Janzing et al., 2016) for observational data into a unified framework.

1.4 Related Works and Comparisons

Starting from the pioneering work of Peters et al. (2016), there is considerable literature proposing methods to estimate $m^\star$ in (1.1), predominantly when $m^\star$ is linear. These methods broadly fall into two categories: hypothesis test-based methods and optimization-based methods. For the hypothesis test-based methods (Peters et al., 2016; Heinze-Deml et al., 2018; Pfister et al., 2019), the Type-I error is controlled for an estimator $\widehat{S}$ with $\mathbb{P}(\widehat{S} \subseteq S^\star) \geq 1-\alpha$. Nonetheless, these procedures may result in missing important variables or conservative solutions like $\widehat{S} = \emptyset$ due to the inherent worst-case construction in the algorithm. Additionally, the introduction of hypothesis tests also hinders its seamless integration by machine learning algorithms, limiting their scalability. On the other hand, some optimization-based methods (Ghassami et al., 2017; Rothenhäusler et al., 2019, 2021) focus on linear $m^\star$ and tackle the problem under additional structures such as linear SCMs with additive interventions (Rothenhäusler et al., 2019). This limitation curtails its applicability to a broader nonparametric setting. Some optimization-based methods (Pfister et al., 2021; Yin et al., 2021) designed for linear models are heuristic and lack finite sample guarantees. In summary, there is still a crucial gap towards efficiently estimating $m^\star$ without additional assumptions on the underlying model. Although Fan et al. (2023) recently bridged this gap for linear $m^\star$ through an optimization-based method, it is still unclear under the general nonparametric setting. This paper is the first to attain sample-efficient estimation for the general model with non-asymptotic guarantees in terms of both $|\mathcal{E}|$ and n.

Arjovsky et al. (2019) considers a general task, which aims to search for a data representation such that the optimal solution given that representation is optimal across diverse environments. They propose an optimization-based approach called invariant risk minimization (IRM), with many subsequent variants proposed later. However, their method comes with no statistical guarantees and requires at least d environments even for the linear model, and the improvement over standard empirical risk minimization is not clear (Rosenfeld et al., 2021; Kamath et al., 2021). Our paper is the first to offer a comprehensive theoretical analysis of general invariance learning when the representation class is $\{(x_1, \ldots, x_d) \to (a_1 x_1, \ldots, a_d x_d) : a_1, \ldots, a_d \in \{0, 1\}\}$ and to show that sample efficient estimation is in general viable even when $|\mathcal{E}| = 2$. The main reason why this is attainable is due to the *exact* invariance pursued by our FAIR penalty and its "focused" nature, see the discussion in Appendix A.2.

Under the SCM framework, there is considerable literature on causal discovery using observational data (Spirtes et al., 2000; Richardson, 1996; Chickering, 2002; Hyttinen et al., 2013, 2014), but they cannot go beyond Markov equivalent class (Geiger & Pearl, 1990) and thus fail to establish the exact cause-effect direction in general. Such a problem can be resolved by imposing additional assumptions under the circumstances that the algorithm can only passively observe data rather than performing intervention actively. These methods can be divided into two categories – one based on the invariance principle and the other based on the asymmetry principle. The invariance-based approaches (Peters et al., 2016) use samples from multiple experiments where some unknown intervention may apply to the variables other than Y. It leverages the idea that the cause-effect mechanism will remain constant while the reverse effect-cause association may vary. On the other hand, the asymmetry-based approaches (Shimizu et al., 2006; Hoyer et al., 2008; Zhang & Hyvärinen, 2009; Janzing et al., 2012; Peters et al., 2014) only observe one sample of observational data and use the idea that the cause-effect mechanism admits a simple prior known structure, whereas its inverse does not, example includes the additive noise structure (Hoyer et al., 2008). These two principles for causal discovery seem to have been orthogonal before. Our estimation framework is the first to offer a unified methodological perspective on these two principles with theoretical guarantees. It demonstrates the ability to simultaneously leverage both principles for identification and estimation.

Adversarial estimation is introduced in Goodfellow et al. (2014) for generative modeling. Its application in the statistics spans distribution estimation (Liang, 2021), instrumental variable regression (Dikkala et al., 2020), estimating the (implicit) influence function (Chernozhukov et al., 2020; Hirshberg & Wager, 2021), and so on. The idea of minimizing the worst-case reward among diverse environments can also be considered as "an algorithmic remedy" for out-of-distribution generalization. There are different considerations of the "reward" such as risk (Sagawa et al., 2020), excess risk (Agarwal & Zhang, 2022), and the negative of the explained variance (Meinshausen & Bühlmann, 2015). However, these methods are conceptually similar to running least squares in regression and thus cannot go beyond just learning associations. We adopted the adversarial estimation in our estimation from two novel aspects. Firstly, it allows us to use a simple objective function that homogenizes different tasks and prediction

models for estimation. Moreover, such a minimax optimization objective and the Gumbel approximation in the implementation jointly relax the combinatorial nature in (1.3) and make a variant of gradient descents continue to work numerically.

1.5 Organization

This paper is structured as follows. We first provide the proposed method with non-asymptotic theoretical analysis, and causal interpretations for our canonical *nonparametric causality (invariance) pursuit* problem in Sections 2–3, respectively. Such a special instance of our framework also helps to illustrate the main idea and philosophy of our general invariance pursuit problem and FAIR estimation framework, which will be formally presented in Section 4. In the main text, we provide a sketch of the abstract unified result, from which all non-asymptotic results are derived as corollaries, along with its other applications in Section 4.3 and defer the detailed statements to the Appendix. We provide a computationally efficient implementation using variants of gradient descent and Gumbel approximation, followed by its application to the simulation and real data analysis in Section 5. All the proofs are collected in the supplemental material.

1.6 Notations

We use upper case (X, Y, Z) to represent random variables/vectors and denote their instances as (x, y, z). Define $[n] = \{1, \ldots, n\}$. For a vector $x = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$, we let $\|x\|_2 = (\sum_{j=1}^d x_j^2)^{1/2}$. For given index set $S = \{j_1, \ldots, j_{|S|}\} \subseteq [d]$ with $j_1 < \cdots < j_{|S|}$, we denote $[x]_S = (x_{j_1}, \ldots, x_{j_{|S|}})^\top \in \mathbb{R}^{|S|}$ and abbreviate it as x_S if there is no ambiguity. We let $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. We use $a(n) \lesssim b(n)$, $b(n) \gtrsim a(n)$, or $a(n) = O(b(n))$ if there exists some constant $C > 0$ such that $a(n) \le Cb(n)$ for any $n \ge 3$. Denote $a(n) \asymp b(n)$ if $a(n) \lesssim b(n)$ and $a(n) \gtrsim b(n)$. In the theorem statement and proof, we will use C to represent the universal constants that may vary from line to line and will use $\widetilde{C}, \widetilde{C}_1, \ldots$ to represent the constant that may depend on the other constants defined in the paper.

In the context of the multi-environment setup, consider the following notations. For each $e \in \mathcal{E}$, let $\Theta^{(e)} = L_2(\mu_x^{(e)}) := \{f : \int f^2(x)\mu_x^{(e)}(dx) < \infty\}$, and denote $\|f\|_{2,e} = \{\int f^2(x)\mu_x^{(e)}(dx)\}^{1/2}$. Given n observations $\{(X_i^{(e)}, Y_i^{(e)})\}_{i=1}^n \subseteq \mathbb{R}^d \times \mathbb{R}$ drawn i.i.d. from $\mu^{(e)}$, we define $\mathbb{E}[f(X^{(e)}, Y^{(e)})] = \int f(x, y)\mu^{(e)}(dx, dy)$ and $\widehat{\mathbb{E}}[f(X^{(e)}, Y^{(e)})] = \frac{1}{n}\sum_{i=1}^n f(X_i^{(e)}, Y_i^{(e)})$ for any $f \in \Theta^{(e)}$. We assume $\mathbb{E}[|Y^{(e)}|^2] < \infty$. Let $\bar{\mu} = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \mu^{(e)}$, and $\Theta = L_2(\bar{\mu}_x)$ equipped with the norm $\|\cdot\|_2 = \{\int f^2(x)\bar{\mu}_x(dx)\}^{1/2}$. It is easy to verify that $\Theta = \bigcap_{e\in\mathcal{E}} \Theta^{(e)}$.

Let $S \subseteq [d]$ be any index set. Given a function class $\mathcal{H} \subseteq \{h : \mathbb{R}^d \to \mathbb{R}\}$, we define $\mathcal{H}_S$ be the class of functions in $\mathcal{H}$ that only depend on variables x_S, i.e., $\mathcal{H}_S = \{h \in \mathcal{H}, h(x) \equiv u(x_S)$ for some $u : \mathbb{R}^{|S|} \to \mathbb{R}$ $\mu^{(e)}$-$a.s. \forall e \in \mathcal{E}\}$. We sometimes also write $h(x_S)$ instead of $h(x)$ for $h \in \mathcal{H}_S$ since h only depends on x_S. For any $h \in \mathcal{H}$, we use $S_h \subseteq [d]$ to represent the index set of the variables h depends on. We let $\{\mathcal{H}\}^k = \{(h_1, \ldots, h_k) : h_i \in \mathcal{H}\ \forall i \in [k]\}$. For any (X, Y)'s joint distribution ν, we use ν_x to denote the marginal distribution of X, and $\nu_{x,S}$ to denote the marginal distribution of X_S.

Neural Networks. We use neural networks as a scalable nonparametric technique: we adopt the fully connected deep neural network with ReLU activation $\sigma(\cdot) = \max\{0, \cdot\}$, and call it *deep ReLU network* for short. Let L, N be any positive integer, a *deep ReLU network with depth L width N* admits the form of

$$g(x) = T_{L+1} \circ \bar{\sigma}_L \circ T_L \circ \bar{\sigma} \circ \cdots \circ T_2 \circ \bar{\sigma}_1 \circ T_1(x). \tag{1.4}$$

Here $T_l(z) = W_l z + b_l : \mathbb{R}^{d_l} \to \mathbb{R}^{d_{l+1}}$ is a linear map with weight matrix $W_l \in \mathbb{R}^{d_l \times d_{l-1}}$ and bias vector $b_l \in \mathbb{R}^{d_l}$, where $(d_0, d_1 \ldots, d_L, d_{L+1}) = (d, N, \ldots, N, 1)$, and $\bar{\sigma}_l : \mathbb{R}^{d_l} \to \mathbb{R}^{d_l}$ applies the ReLU activation $\sigma(\cdot)$ to each entry of a d_l-dimensional vector. Here the equal width is for presentation simplicity.

Definition 1 (Deep ReLU network class)**.** *Define the family of deep ReLU networks taking d-dimensional vector as input with depth L, width N, truncated by B as $\mathcal{H}_{\mathtt{nn}}(d, L, N, B) = \{\widetilde{g}(x) = \mathrm{Tc}_B(g(x)) : g(x)$ in (1.4)$\}$, where $\mathrm{Tc}_B : \mathbb{R} \to \mathbb{R}$ is the truncation operator defined as $\mathrm{Tc}_B(z) = \max\{|z|, B\} \cdot \mathrm{sign}(z)$.*

2 FAIR Least Squares Estimator Using Neural Networks

In this section, we show that one can use the FAIR-NN least squares estimator, a realization of the FAIR estimator by setting $\ell(y,v)=\frac{1}{2}(y-v)^2$ and specifying both $(\mathcal{G},\mathcal{F})$ as neural networks, to attain sample-efficient estimation in nonparametric causality pursuit.

The main messages of this section are two-fold. From a theoretical perspective, it shows that sample-efficient estimation (in both n and $|\mathcal{E}|$) in the general nonparametric causality pursuit problem is viable under a minimal identification condition related to the heterogeneity of the environments. From a methodological perspective, it demonstrates one key feature of our proposed framework: one can seamlessly integrate black-box machine learning models (e.g. neural networks) into it and fully exploit these models' sample efficiency and capability in being adaptive to low-dimension structures.

2.1 Setup

We introduce some notations. Recall that $\mu^{(e)}$ is the joint distribution of (X,Y) in environment e. Let $m^{(e,S)}(x):=\mathbb{E}[Y^{(e)}|X_S^{(e)}=x_S]$ be the conditional expectation of Y given X_S in environment e. Recall that $\nu_{x,S}$ is the marginal distribution of X_S for $(X,Y)\sim\nu$. It is easy to see that $\mu_{x,S}^{(e)}$ is absolutely continuous with respect to $\bar{\mu}_{x,S}=[\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\mu^{(e)}]_{x,S}$ for any $S\subseteq[d]$ hence $\rho_S^{(e)}$, the Radon–Nikodym derivative of $\mu_{x,S}^{(e)}$ with respect to $\bar{\mu}_{x,S}$, is well defined. We define $\bar{m}^{(S)}(x)=\sum_{e\in\mathcal{E}}\rho_S^{(e)}(x_S)m^{(e,S)}(x)$, which can be interpreted as the population-level least squares that regress Y on X_S using all the data in $\mathcal{E}$.

Condition 1 (Model and Regularity Conditions)**.** *There exists some positive constants* $(C_0,s_{\min})$ *such that the following conditions hold.*

(a) *Data Generating Process* *We collect data from* $|\mathcal{E}|\in\mathbb{N}^+$ *environments with* $|\mathcal{E}|\le n^{C_0}$*. For each environment* $e\in\mathcal{E}$*, we observe* $\{(X_i^{(e)},Y_i^{(e)})\}_{i=1}^n\overset{i.i.d.}{\sim}\mu^{(e)}$.

(b) *Invariance Structure:* *There exists some set* $S^\star$ *and* $m^\star:\mathbb{R}^{|S^\star|}\to\mathbb{R}$ *such that* $m^{(e,S^\star)}(x)\equiv m^\star(x_{S^\star})$ *for any* $e\in\mathcal{E}$.

(c) *Sub-Gaussian Response:* *For any* $e\in\mathcal{E}$ *and* $t\ge 0$, $\mathbb{P}\left[|Y^{(e)}|\ge t\right]\le C_0e^{-t^2/(2C_0)}$.

(d) *Boundedness:* $X\in[-C_0,C_0]^d$ $\bar{\mu}$*-a.s. and* $\|m^{(e,S)}\|_\infty\le C_0$ *for any* $S\subseteq[d]$ *and* $e\in\mathcal{E}$.

(e) *Nondegenerate Covariate:* *For any* $S\subseteq[d]$ *with* $S^\star\setminus S\neq\emptyset$, $\inf_{m\in\Theta_S}\|m-m^\star\|_2^2\ge s_{\min}>0$.

Condition 1 (a)–(b) is just a restatement of (1.1) together with i.i.d. data within each environment; data across different environments may be dependent. (c)–(d) are standard in nonparametric regression. (e) rules out some degenerate cases, for example, $m^\star(x_1)=x_1^2$ with $S^\star=\{1\}$ and $X_2=X_1^4$, or $m^\star(x_1,x_2)=f(x_1)$ with $S^\star=\{1,2\}$, and is imposed for technical convenience. The target (invariant) regression function in nonparametric causality pursuit is $m^\star$.

2.2 Proposed FAIR-NN Least Squares Estimator

Given all the data $\{\{(X_i^{(e)},Y_i^{(e)})\}_{i=1}^n\}_{e\in\mathcal{E}}$ from heterogeneous environments, we consider using the following FAIR-NN least squares estimator to learn $m^\star$ in (1.1). Specifically, the FAIR-NN least squares estimator is the solution to the subsequent minimax optimization objective

$$\widehat{g}\in\operatorname*{argmin}_{g\in\mathcal{G}}\sup_{f^{\mathcal{E}}\in\{\mathcal{F}_{S_g}\}^{|\mathcal{E}|}}\frac{1}{|\mathcal{E}|\cdot n}\sum_{e\in\mathcal{E},i\in[n]}\left\{Y_i^{(e)}-g(X_i^{(e)})\right\}^2+\gamma\widehat{\mathsf{J}}(g,f^{\mathcal{E}}). \tag{2.1}$$

where the first part of the objective $\widehat{\mathsf{Q}}_\gamma(g,f^{\mathcal{E}})$ is the pooled least squares loss preventing the estimator from collapsing to conservative solutions, γ is the hyper-parameter to be determined, and $\widehat{\mathsf{J}}(g,f^{\mathcal{E}})$ is the empirical counterpart of the focused adversarial invariance regularizer defined as

$$\widehat{\mathsf{J}}(g,f^{\mathcal{E}})=\frac{1}{|\mathcal{E}|\cdot n}\sum_{e\in\mathcal{E},i\in[n]}\left[\left\{Y_i^{(e)}-g(X_i^{(e)})\right\}f^{(e)}(X_i^{(e)})-\frac{1}{2}\left\{f^{(e)}(X_i^{(e)})\right\}^2\right]. \tag{2.2}$$

The minimax program (2.1) is the empirical version of (1.2) via setting $\ell(y,v) = \frac{1}{2}(y-v)^2$. Here we specify the predictor function class $\mathcal{G}$ and testing (discriminator) function class $\mathcal{F}$ as

$$\mathcal{G} = \mathcal{H}_{\texttt{nn}}(d, L, N, B) \qquad \text{and} \qquad \mathcal{F} = \mathcal{H}_{\texttt{nn}}(d, L+2, 2N, 2B) \tag{2.3}$$

for neural network architecture hyper-parameters N, L and truncation parameter $B = C_0$. Here B can be larger than C_0 but should satisfies $B = O(1)$. A larger width, depth, and truncation parameter can also be adopted for $\mathcal{F}$. Our specification of (N, L, B) for $\mathcal{F}$ here is for technical purposes, that is, any $m^{(e,S)} - g$ for $g \in \mathcal{G}$ can be well approximated by some $f \in \mathcal{F}$.

2.3 Non-Asymptotic Result for FAIR-NN

Condition 2 (Identification for Nonparametric Causality Pursuit). *For any $S \subseteq [d]$ such that $\bar{\mu}(\{m^\star \neq \bar{m}^{(S\cup S^\star)}\}) > 0$, there exists some $e, e' \in \mathcal{E}$ such that $\min\{\mu^{(e)}, \mu^{(e')}\}(\{m^{(e,S)} \neq m^{(e',S)}\})\} > 0$.*

Remark 1 (Minimal Heterogeneity Condition for Identification). *The above identification condition necessitates that whenever a bias emerges when regressing Y on $X_{S\cup S^\star}$ using least squares, there should be noticeable shifts in conditional expectation $m^{(e,S)}$ across environments. In other words, $S^\star$ the maximum set that preserves the invariant associations. This condition is minimal. If it is violated, it would imply*

$$\exists \widetilde{S} \subseteq [d] \text{ with } \widetilde{S} \setminus S^\star \neq \emptyset \qquad s.t. \qquad \forall e \in \mathcal{E} \;\; \mathbb{E}[Y^{(e)}|X^{(e)}_{\widetilde{S}}] \equiv g(X^{(e)}_{\widetilde{S}}) \;\; \mu^{(e)}\text{-a.s. for some } g : \mathbb{R}^{|S|} \to \mathbb{R},$$

in which both set $S^\star$ and $\widetilde{S}$ embody the invariant conditional expectation structure, thus more environments are needed in this case to pinpoint $S^\star$. Such a minimal identification condition underscores that our proposed FAIR-NN estimator is "sample efficient" regarding the number of environments $|\mathcal{E}|$ required; see the discussions in Section 3. Notably, such an identification condition relaxes those employed in approaches using intersections like ICP (Peters et al., 2016; Heinze-Deml et al., 2018). These approaches require the shifts of conditional distributions for all the S with $\bar{m}^{(S)} \neq m^\star$ for identifying $S^\star$.

Remark 2 (Relaxing Condition 2). *We claim that Condition 2 can be slightly relaxed given our algorithm searches for the most predictive variable set that preserves the invariance structure. But it is of a technical style and lacks semantic meaning; see discussions in Appendix A.4.*

The following theorem provides an oracle-type inequality for the FAIR-NN least squares estimator in a structure-agnostic manner. The first term is the maximum approximation bias of neural networks across environments and the second term is related to the complexity of the neural networks used in the fitting. It implies that when the FAIR-NN penalty parameter γ is large enough, all endogenous spurious variables can be surely screened Fan & Lv (2008) when n is large enough, thus $m^\star$ can be estimated as well as if the invariant quasi-causal set of variables $S^\star$ is known. In addition, the theorem quantifies the amount of penalty needed, which is related to the signal-to-noise ratio of the problem.

Theorem 1 (Oracle-type Inequality for FAIR-NN Least Squares Estimator). *Assume Condition 1 and Condition 2 hold. Then $\gamma^\star_{\texttt{NN}} = \sup_{S\subseteq[d]:\mathsf{b}_{\texttt{NN}}(S)>0}(\mathsf{b}_{\texttt{NN}}(S)/\bar{\mathsf{d}}_{\texttt{NN}}(S)) < \infty$, where*

$$\mathsf{b}_{\texttt{NN}}(S) = \|m^\star - \bar{m}^{(S\cup S^\star)}\|_2^2 \qquad \text{and} \qquad \bar{\mathsf{d}}_{\texttt{NN}}(S) = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \|m^{(e,S)} - \bar{m}^{(S)}\|_{2,e}^2. \tag{2.4}$$

Consider the estimator that solves (2.1) using $\gamma \geq 8\gamma^\star_{\texttt{NN}}$ and function classes (2.3) with L, N satisfying $NL \leq n$ and $N \geq 4$. Then, there exists some constant $\widetilde{C}$ depending on (d, C_0) such that for any $n \geq 3$,

$$\frac{\|\widehat{g} - m^\star\|_2}{\widetilde{C}} \leq \max_{e\in\mathcal{E}} \inf_{h\in\mathcal{G}_{S^\star}} \|m^\star - h\|_{2,e} + \frac{NL\log^{3/2} n}{\sqrt{n}} + 1_{\{\delta_{\texttt{NN},1} > \mathsf{s}\}} \cdot (\gamma\delta_{\texttt{NN},1})$$

occurs with probability at least $1 - \widetilde{C}n^{-100}$. Here $\delta_{\texttt{NN},1} = \max_{e\in\mathcal{E}, S\subseteq[d]} \inf_{h\in\mathcal{G}_S} \|m^{(e,S)} - h\|_{2,e} + \frac{NL\log^{3/2} n}{\sqrt{n}}$ and $\mathsf{s} = \widetilde{C}^{-1}[1 \wedge s_{\min} \wedge \{\gamma \inf_{S:\bar{\mathsf{d}}_{\texttt{NN}}(S)>0} \bar{\mathsf{d}}_{\texttt{NN}}(S)\}]/(1+\gamma)$, where $s_{\min}$ is defined in Condition 1(5).

As our result is non-asymptotic, for a given n, we may not be able to eliminate all endogenous spurious variables. The third term in Theorem 1 reflects this when the signal is not sufficiently large. It is more explicitly given in Corollary 1.

Remark 3 (Interpretation of $\mathsf{b}_{\mathtt{NN}}(S)$ and $\bar{\mathsf{d}}_{\mathtt{NN}}(S)$). *We refer to $\mathsf{b}_{\mathtt{NN}}(S)$ as bias mean since it exactly characterizes the bias of the least squares estimator in the presence of endogenous spurious variables like the background color in the thought experiment. In particular, letting $\widehat{g}_{\mathtt{LSE}(S)}$ be the least squares estimator that regresses Y on X_S using all the data, namely, the FAIR-NN estimator with $\gamma = 0$, Proposition 6 implies*

$$\left| \frac{\|\widehat{g}_{\mathtt{LSE}(S)} - m^\star\|_2^2}{\mathsf{b}_{\mathtt{NN}}(S)} - 1 \right| = o_{\mathbb{P}}(1) \qquad \text{if } \ S^\star \subseteq S \text{ and } \mathsf{b}_{\mathtt{NN}}(S) > 0.$$

We refer to $\bar{\mathsf{d}}_{\mathtt{NN}}(S)$ as the bias variance because it measures the variations of bias across environments. Specifically, when $S^\star \subseteq S$, the bias in environment e is $(m^{(e,S)} - m^\star)$, and $\bar{\mathsf{d}}_{\mathtt{NN}}(S)$ can be viewed as the variance of the bias concerning the uniform distribution on $\mathcal{E}$ since $\bar{\mathsf{d}}_{\mathtt{NN}}(S) = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \|(m^{(e,S)} - m^\star) - (\bar{m}^{(S)} - m^\star)\|_{2,e}^2$. We have $\bar{\mathsf{d}}_{\mathtt{NN}}(S^\star) = 0$ by the invariance structure in Condition 1(b).

Remark 4 (Identification). *Theorem 1 combines the identification result, which characterizes when it is possible to consistently estimate $m^\star$, and the finite-sample estimation error result, which characterizes how accurately we can estimate $m^\star$. The main identification message disentangled from the above theorem is that if the minimal heterogeneity condition Condition 2 holds, then one can consistently estimate $m^\star$ provided γ is larger than some threshold $8\gamma^\star_{\mathtt{NN}}$ that is independent of n.*

2.4 Adapting to the Low-dimensional Structures Algorithmically

To present the explicit L_2 error rate under a specific nonparametric setup, we first introduce the concept of (β, C)-smooth function.

Definition 2 ((β, C)-smooth Function). *Let $\beta = r+s$ for some nonnegative integer $r \geq 0$ and $0 < s \leq 1$, and $C > 0$. A d-variate function f is (β, C)-smooth if for every non-negative sequence $\alpha \in \mathbb{N}^d$ such that $\sum_{j=1}^d \alpha_j = r$, the partial derivative $\partial^\alpha f = (\partial f)/(\partial x_1^{\alpha_1} \cdots x_d^{\alpha_d})$ exists and satisfies $|\partial^\alpha f(x) - \partial^\alpha f(z)| \leq C\|x - z\|_2^s$. We use $\mathcal{H}_{\mathtt{HS}}(d, \beta, C)$ to denote the set of all the d-variate (β, C)-smooth functions.*

One significant advantage of neural networks over traditional nonparametric methods is their intrinsic capability for algorithmic nonparametric regression. This enables them to learn low-dimensional structures with little or no explicit guidance regarding the forms of functions (Bauer & Kohler, 2019; Schmidt-Hieber, 2020; Kohler & Langer, 2021; Fan & Gu, 2024). We begin by elucidating the concept of the Hierarchical Composition Model (HCM), which is basically the compositions of t-variate functions with (β, C)-smooth l times for (t, β) in a certain set $\mathcal{O}$.

Definition 3 (Hierarchical Composition Model $\mathcal{H}_{\mathtt{HCM}}(d, l, \mathcal{O}, C)$). *We define function class of hierarchical composition model $\mathcal{H}_{\mathtt{HCM}}(d, l, \mathcal{O}, C)$ (Kohler & Langer, 2021) with $l, d \in \mathbb{N}^+$, $C \in \mathbb{R}^+$, and $\mathcal{O}$, a subset of $[1, \infty) \times \mathbb{N}^+$, in a recursive way as follows. Let $\mathcal{H}_{\mathtt{HCM}}(d, 0, \mathcal{O}, C) = \{h(x) = x_j, j \in [d]\}$, and for each $l \geq 1$,*

$$\mathcal{H}_{\mathtt{HCM}}(d, l, \mathcal{O}, C) = \big\{h : \mathbb{R}^d \to \mathbb{R} : h(x) = g(f_1(x), ..., f_t(x)), \text{ where } g \in \mathcal{H}_{\mathtt{HS}}(t, \beta, C) \text{ with } (\beta, t) \in \mathcal{O} \text{ and } f_i \in \mathcal{H}_{\mathtt{HCM}}(d, l-1, \mathcal{O}, C)\big\}.$$

Following Kohler & Langer (2021), we assume all the compositions are at least Lipschitz functions to simplify the presentation. The minimax optimal L_2 estimation risk over $\mathcal{H}(d, l, \mathcal{O}, C_h)$ is $n^{-\alpha^\star/(2\alpha^\star+1)}$, where $\alpha^\star = \min_{(\beta,t)\in\mathcal{O}}(\beta/t)$ is the smallest dimensionality-adjusted degree of smoothness that represents the hardest component in the composition. For example, if $m^\star(x) = f_1(x_1) + f_2(f_3(x_2, x_3), f_4(x_4, x_5)) + f_5(x_1, x_3, x_5)$ and all functions have a bounded second derivative, then the hardest component is the last one, and the dimensionality-adjusted degree of smoothness is $\alpha^* = 2/3$.

Condition 3 (Function Complexity and Neural Network Architecture). *The following holds:*
(a) $m^{(e,S)} \in \mathcal{H}_{\mathtt{HCM}}(|S|, l, \mathcal{O}, C_h)$ for any $e \in \mathcal{E}$ and $S \subseteq [d]$ with $\alpha_0 = \inf_{(\beta,t)\in\mathcal{O}}(\beta/t)$.
(b) $m^\star \in \mathcal{H}_{\mathtt{HCM}}(|S^\star|, l, \mathcal{O}^\star, C_h)$ with $\alpha^\star = \inf_{(\beta,t)\in\mathcal{O}^\star}(\beta/t)$.
(c) We choose N, L satisfying $LN \asymp \{n(\log n)^{8\alpha^\star-3}\}^{\frac{1}{2(2\alpha^\star+1)}}$ and $(\log n)/(N \wedge L) = o(1)$.
(d) $\max\{C_0, d, l, C_h, \sup_{(\beta,t)\in\mathcal{O}}(\beta \vee t), \sup_{(\beta,t)\in\mathcal{O}^\star}(\beta \vee t)\} \leq C_1$ for some constant $C_1 > 1$.

Corollary 1 (Optimal Rate for FAIR-NN). *Under the setting of Theorem 1, assume further that Condition 3 holds. Then, for any $n \geq 3$, with probability at least $1 - \widetilde{C}n^{-100}$, the following holds*

$$\frac{\|\widehat{g} - m^\star\|_2}{\widetilde{C}\log^7(n)} \leq n^{-\alpha^\star/(2\alpha^\star+1)} + 1_{\{n<n_0\}}\gamma \cdot n^{-\alpha_0/(2\alpha^\star+1)}, \tag{2.5}$$

where n_0 depends on $(C_1, \gamma, s_{\min}, \inf_{S:\bar{\mathsf{d}}_{\mathtt{NN}}(S)>0} \bar{\mathsf{d}}_{\mathtt{NN}}(S))$, and $\widetilde{C}$ is a constant dependent only on C_1.

From Corollary 1, we can get (up to logarithmic factors) minimax convergence rate $n^{-\alpha^\star/(2\alpha^\star+1)}$, which is independent of both α_0 and γ, when n is larger than some constant n_0. Utilizing neural networks in predictor and discriminator function classes allows the estimator to adapt to the invariant regression function $m^\star$ efficiently from two crucial perspectives. Firstly, similar to using neural networks in nonparametric regression (Schmidt-Hieber, 2020; Kohler & Langer, 2021; Fan & Gu, 2024), adopting neural networks in $\mathcal{G}$ endows the estimator with the capability of being adaptive to the low-dimensional hierarchical structure algorithmically. Secondly, the choice of model parameter (N, L), and the convergence rate depends only on $m^\star$. The (spurious) conditional expectations $m^{(e,S)}$ can be much more complex than $m^\star$. Notably, this complexity will not affect the convergence rate. This can be credited to the scalability of neural networks used as discriminators, i.e., their adaptivity capability in the regularization part of FAIR.

Remark 5 (Guaranteed for All n). *The error bound (2.5) is applicable for any $n \geq 3$, even when it selects the wrong variables. Notably, the error bound will not inflate if the invariant signal $s_{\min}$ and the heterogeneity signal $\inf_{S\subseteq[d]:\bar{\mathsf{d}}_{\mathtt{NN}}(S)>0} \bar{\mathsf{d}}_{\mathtt{NN}}(S)$ is small. Though the error bound scales linearly with γ, the estimator we propose is not vulnerable to "weak spurious" variables, e.g., x_j with $\sup_{e\in\mathcal{E}} \|m^{(e,S^\star\cup\{j\})} - m^\star\|_{2,e} \leq \epsilon$, provided all the ratio of the bias $\mathsf{b}_{\mathtt{NN}}(S)$ to heterogeneity $\bar{\mathsf{d}}_{\mathtt{NN}}(S)$ gets controlled.*

Remark 6 (Choice of the Hyper-parameter γ). *Though we have to choose a hyper-parameter γ larger than a certain threshold to attain such a rate, the convergence rate is independent of γ. This implies that when the sample size n is large, we do not need to tune the hyper-parameter γ for optimal performance. Instead, we can choose some conservative (large) γ such that the lower bound $\gamma \geq 8\gamma^\star_{\mathtt{NN}}$ is guaranteed.*

3 Nonparametric Invariance Pursuit under SCMs

The results in Section 2 are for the problem *nonparametric invariance pursuit* itself. In a population-level view, it pursues "maximum invariant set" $S^\star$ satisfying

$$m^{(e,S^\star)} \equiv \bar{m}^{(S^\star)} \text{ (invariant)} \quad \text{and} \quad \forall S \subseteq [d],\ m^{(e,S)} \equiv \bar{m}^{(S)} \Longrightarrow \bar{m}^{(S\cup S^\star)} = \bar{m}^{(S^\star)} \text{ (maximum)}. \tag{3.1}$$

Section 2 shows the FAIR-NN estimator can estimate such a $S^\star$ ($m^\star$) efficiently. It is natural to ask

Does such a maximum invariant set $S^\star$ exist? What's the semantic meaning of it?

We offer a clean yet general answer to the question under the SCM with arbitrary interventions (on X) setting. The short answer is: Yes, and it can be interpreted as the "pragmatic direct causes".

3.1 Structural Causal Model with Interventions on Covariates

We first introduce the concept of the structural causal model (Glymour et al., 2016). See Fig. 1 for examples of SCM. It says that each variable in the directed graph is a function of its parents (if any) and an independent innovation or noise.

Definition 4 (Structural Causal Model). *A structural causal model $M = (\mathcal{S}, \nu)$ on p variables $Z_1, \ldots, Z_p$ can be described using p assignment functions $\{f_1, \ldots, f_p\} = \mathcal{S}$:*

$$Z_j \leftarrow f_j(Z_{\mathtt{pa}(j)}, U_j) \qquad j = 1, \ldots, p,$$

where $\mathtt{pa}(j) \subseteq \{1, \ldots, p\}$ is the set of parents, or the direct causes, of the variable Z_j, and the joint distribution $\nu(du) = \prod_{j=1}^p \nu_j(du_j)$ over p independent exogenous variables $(U_1, \ldots, U_p)$. For a given

model M, there is an associated directed graph $G(M) = (V, E)$ that describes the causal relationships among variables, where $V = [p]$ is the set of nodes, E is the edge set such that $(i, j) \in E$ if and only if $i \in \mathtt{pa}(j)$. $G(M)$ is acyclic if there is no sequence $(v_1, \ldots, v_k)$ with $k \geq 2$ such that $v_1 = v_k$ and $(v_i, v_{i+1}) \in E$ for any $i \in [k-1]$.

As in Peters et al. (2016), we consider the following data-generating process in $|\mathcal{E}|$ environments. For each $e \in \mathcal{E}$, the process governing $p = d+1$ random variables $Z^{(e)} = (Z_1^{(e)}, \ldots, Z_{d+1}^{(e)}) = (X_1^{(e)}, \ldots, X_d^{(e)}, Y^{(e)})$ is derived from an SCM $M^{(e)}(\mathcal{S}^{(e)}, \nu)$, whose induced graph $G(M^{(e)})$ is acyclic, and assignments as

$$\begin{aligned} X_j^{(e)} &\leftarrow f_j^{(e)}(Z_{\mathtt{pa}(j)}^{(e)}, U_j), \qquad\qquad j = 1, \ldots, d \\ Y^{(e)} &\leftarrow f_{d+1}(X_{\mathtt{pa}(d+1)}^{(e)}, U_{d+1}). \end{aligned} \tag{3.2}$$

Here the distribution of exogenous variables $(U_1, \ldots, U_{d+1})$, the cause-effect relationship graph G, and the structural assignment f_{d+1} are *invariant* across $e \in \mathcal{E}$, while the structural assignments for X may vary among $e \in \mathcal{E}$. We use superscript (e) to highlight this heterogeneity. This heterogeneity may arise from performing arbitrary interventions on the variables X. We use $Z_{\mathtt{pa}(j)}$ to emphasize that Y can be the direct cause of some variables in the covariate vector. See an example in Fig. 1 (a).

To present the result, we consider an augmented SCM that incorporates the environment label e as a variable E. We consider the case where $\mathcal{E} = \{0, \ldots, |\mathcal{E}| - 1\}$. We let 0 be the observational environment, and the rest are the interventional environments where some *unknown*, *arbitrary* interventions are applied to the variables in some given set $I \subseteq [d]$ defined as $I := \{j : \exists e \in \mathcal{E} \text{ s.t. } f_j^{(e)} \neq f_j^{(0)}\}$. The interventions can be arbitrary: it can be a "hard" do-intervention via set X_j being v_j, or soft intervention that slightly perturbs the association, e.g., replace $X_j \leftarrow 2X_k + U_j$ by $X_j \leftarrow 1.5X_k + U_j$. The shared cause-effect relationships in all the environments are encoded by G, or $\{\mathtt{pa}(j)\}_{j=1}^{d+1}$.

The following SCM $\widetilde{M} = (\widetilde{\mathcal{S}}, \widetilde{\nu})$ on $d+2$ variables $Z = (Z_1, \ldots, Z_d, Z_{d+1}, Z_{d+2}) = (X_1, \ldots, X_d, Y, E)$ encodes all the information of $|\mathcal{E}|$ models $\{M^{(e)}(\mathcal{S}^{(e)}, \nu)\}_{e \in \mathcal{E}}$ in (3.2). Denote $\nu_b \sim \text{Uniform}(\mathcal{E})$. Here $\widetilde{\nu}(du_1, \ldots, du_{d+2}) = \nu(du_1, \ldots, du_{d+1})\nu_b(du_{d+2})$, and the assignments $\widetilde{\mathcal{S}} = \{\widetilde{f}_1, \ldots, \widetilde{f}_{d+2}\}$ are defined as

$$\begin{aligned} E &\leftarrow \widetilde{f}_{d+2}(U_{d+2}) := U_{d+2} \\ X_j &\leftarrow \begin{cases} \widetilde{f}_j(Z_{\mathtt{pa}(j)}, U_j) := f_j^{(0)}(Z_{\mathtt{pa}(j)}, U_j) & \forall j \in [d] \setminus I \\ \widetilde{f}_j(Z_{\mathtt{pa}(j)}, E, U_j) := f_j^{(E)}(Z_{\mathtt{pa}(j)}, U_j) & \forall j \in I \end{cases} \\ Y &\leftarrow \widetilde{f}_{d+1}(X_{\mathtt{pa}}(d+1), U_{d+1}) := f_{d+1}(X_{\mathtt{pa(d+1)}}, U_{d+1}), \end{aligned} \tag{3.3}$$

where I is the set of all intervention variables in $\mathcal{E}$. It should be noted that throughout this section, the direct cause map $\mathtt{pa} : [d+1] \to [d+1]$ matches the causal relationship G instead of $\widetilde{G} = G(\widetilde{M})$. See a graphical illustration of the construction in Fig. 1 (b).

We summarize the above construction as a condition.

Condition 4 (SCM with Interventions on X)**.** *Suppose $M^{(0)}, \ldots, M^{(|\mathcal{E}|-1)}$ are defined by (3.2), and G is acyclic. Let $\widetilde{M}$ be the model constructed as (3.3) by $\{M^{(e)}\}_{e \in \mathcal{E}}$ with I be given set of variables intervened.*

3.2 Maximum Invariant Set as the Pragmatic Direct Causes

We characterize what $S^\star$ would satisfy (3.1) given a fixed intervention set I, and how large I should be to recover the Y's direct causes under arbitrary types of interventions. We define $\mathtt{ch}(k) := \{j : k \in \mathtt{pa}(j)\}$ as the set of children of variable k and $\mathtt{at}(k)$ as the set of all the ancestors of the variable Z_k, defined recursively as $\mathtt{at}(k) = \mathtt{pa}(k) \cup \cup_{j \in \mathtt{pa}(k)} \mathtt{at}(j)$ in the topological order of G. The following condition rules out some degenerated cases.

Condition 5 (Nondegenerate Interventions)**.** *The following holds for $\widetilde{M}$: (a) $\forall S \subseteq [d]$ containing Y's descendants, if $E \not\perp\!\!\!\perp_{\widetilde{M}} Y | X_S$, then there exists some $e, e' \in \mathcal{E}$ such that $(\mu^{(e)} \wedge \mu^{(e')})(\{m^{(e,S)} \neq m^{(e',S)}\}) > 0$; (b) $\widetilde{M}$ is faithful, that is,*

$$\forall \text{ Disjoint } A, B, C \subseteq [d+2], \qquad Z_A \perp\!\!\!\perp Z_B | Z_C \implies Z_A \perp\!\!\!\perp_{\widetilde{G}} Z_B | Z_C,$$

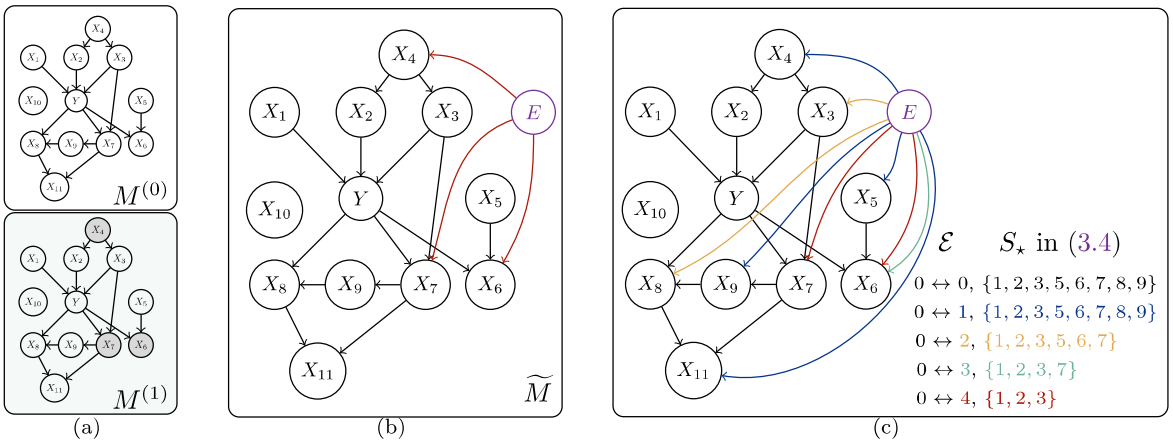


Figure 1: (a) is an illustration of the two-environment model, the SCMs in the two environments share the same associated graph: $M^{(0)}$ is an observational environment, and $M^{(1)}$ is an intervention environment where some unknown intervention is applied to (X_4, X_6, X_7), where $M^{(0)}$ and $M^{(1)}$ are defined as (3.2). (b) visualizes $\widetilde{G}$, the associated graph of $\widetilde{M}$ constructed based on $(M^{(0)}, M^{(1)})$ and (3.3), which is another plot of the environments in (a). (c) An illustration of Theorem 2 by showing how $S_\star$ therein will change as we see more and more environments: the arrow from E to X_j with color e means X_j is intervened in $e \in \{1,2,3,4\}$. For example, $0 \leftrightarrow 3$ means with interventions in environments 1, 2, and 3, the invariant variable set is $\{1,2,3,7\}$. Although X_7 and is reverse causal and hence related to Y, we do not know this based only on the given environments.

where $Z_A \perp\!\!\!\perp_{\widetilde{G}} Z_B | Z_C$ means the node set A and B are d-separated by C in the graph $\widetilde{G}$; see Definition 2.4.1 in Glymour et al. (2016) for a formal definition of d-separation.

The condition (b), faithfulness on the graph $\widetilde{G}$ constraining that the graph $\widetilde{G}$ truly depicts all the conditional independence relationships, is widely used in the causal discovery literature. Condition (a) is further imposed since we only leverage the information of conditional expectations instead of conditional distributions. We impose Condition 5 such that the dependence on E in conditional expectation of Y given X_S with any $S \subseteq [d]$ can be represented by the graph $\widetilde{G}$ itself. The imposed Condition 5 rules out the possibility of some degenerated cases; see the justifications for Condition 5 and some degenerated examples in Appendix A.5. It should be noted that our general results in Theorem 2 and Proposition 1 apply to arbitrary forms of interventions under Condition 5, which is a mild condition as the violation of faithfulness in Condition 5 occurs with probability zero under some suitable measure on the model (Spirtes et al., 2000).

Theorem 2 (General Identification under SCM with Interventions on X)**.** *Under Condition 4, for*

$$S_\star = \mathtt{pa}(d+1) \cup A(I) \cup \bigcup_{j \in A(I)} (\mathtt{pa}(j) \setminus \{d+1\}) \tag{3.4}$$

with $A(I) = \{j : j \in \mathtt{ch}(d+1), \mathtt{at}(j) \cap \mathtt{ch}(d+1) \cap I = \emptyset\}$, we have the invariance $m^{(e,S_\star)} \equiv \bar{m}^{(S_\star)} := m_\star$. Suppose further Condition 5 holds, then Condition 2 holds with $S^\star = S_\star$ and $m^\star = m_\star$.

Theorem 2 exactly characterizes what $S^\star$ is in our nonparametric invariant pursuit under the SCM with interventions on X – it doesn't require intervention to be "sufficient". Firstly, such a $S^\star$ is well-defined in that there exists one maximum set $S_\star$ satisfying the invariant condition (1.1) and heterogeneity condition Condition 2 simultaneously. Secondly, in the SCM setting, such a $S^\star = S_\star$ can be represented in a simple way in (3.4), which lies in between the Markov blanket of the variable Y and the set of Y's direct causes. Note that $A(I)$ can be interpreted as the "unaffected" children of Y from the interventions I. Theorem 2 states explicitly that the pursued set of invariant variables $S^\star$ is the union of (1) parents of Y, (2) unaffected children of Y; and (3) parents of these unaffected children. The size of that set $S^\star$ will keep decreasing when I enlarges. It will finally recover the set of direct causes of Y when I includes "root children set" $I^\star$ as stated in the following Proposition 1. See an illustration in Fig. 1 (c).

Proposition 1 (Direct Cause Recovery)**.** *(Sufficiency) Under Condition 4, define $I^\star = \{j : j \in \mathtt{ch}(d+1), \mathtt{at}(j) \cap \mathtt{ch}(d+1) = \emptyset\}$. If Condition 5 holds and $I \supseteq I^\star$, then Condition 2 holds with $S^\star = \mathtt{pa}(d+1)$.*

(Necessity) Moreover, if $\bar{m}^{(S^\star \cup S)} \neq m^\star$ *for any* S *with* $\mathtt{ch}(d+1) \cap S \neq \emptyset$*, i.e.,* Y *does not have degenerated children, then Condition 2 holds only if* $I \supseteq I^\star$*.*

We refer to $I^\star$ as the *minimal intervention set* because it is the exact minimal set of variables that should be intervened on for exact direct cause recovery in general, nondegenerated cases. The set $I^\star$ is determined by the cause-effect relationship graph G. In particular, $I^\star$ is $\{6, 7\}$ for the example in Fig. 1. Notably, X_8 does not require intervention, as X_7, one of its ancestors, is included in $I^\star$.

Unfortunately, $S_\star \supsetneq \mathtt{pa}(d+1)$ when $I^\star \not\subseteq I$ in general. This is due to a lack of evidence in environments to falsify that some variables in $S^\star$ are not direct causes. Nevertheless, $S_\star$ in this setup can still be interpreted as the "contemporary direct causes" or "pragmatic direct causes" of Y based on the observed environments. If the future interventions are made within the set I, then $S_\star$ can be regarded as the direct causes since the conditional expectation of Y given $X_{S_\star}$ will remain invariant in a new environment t. Moreover, one can deploy such a predictor in unseen environments because it depicts the most predictive one among all the associations in environment 0 that remains in environment t. This can be formally stated in Proposition 2.

Proposition 2 (Robust Transfer Learning)**.** *Under Condition 4, for a new environment* t *with SCM* $M^{(t)} = \{\mathcal{S}^{(t)}, \nu\}$ *satisfying* $f_j^{(t)} \equiv f_j^{(0)}$ *for any* $j \in [d+1] \setminus I$*, i.e., only* X_I *is intervened, we have* $\mathbb{E}[Y^{(t)}|X_{S_\star}^{(t)}] \equiv \mathbb{E}[Y^{(0)}|X_{S_\star}^{(0)}]$ *with* $S_\star$ *in (3.4). If Condition 5 holds and* $M^{(t)}$ *satisfies a condition akin to Condition 5 (see Appendix A.6), then* $S_\star$ *is the maximum set whose conditional expectation is transferable in that for any* $S \subseteq [d]$ *such that* $\mathbb{E}[Y^{(t)}|X_{S_\star \cup S}^{(t)}] \neq \mathbb{E}[Y^{(t)}|X_{S_\star}^{(t)}]$*, one has* $\mathbb{E}[Y^{(t)}|X_S^{(t)}] \neq \mathbb{E}[Y^{(0)}|X_S^{(0)}]$*.*

4 A Unified Framework

The proposed FAIR-NN least squares is a special instance of our generic FAIR estimation framework, which homogenizes different risk loss and prediction models. Moreover, our framework also allows the user to incorporate additional structural knowledge into estimation such that identification is sometimes viable when $|\mathcal{E}| = 1$. The invariance pursuit problem, the estimation method, and the non-asymptotic results will be presented in a unified manner in this section.

4.1 General Invariance Pursuit from Heterogeneous Environments

In this section, we formalize the problem of *invariance pursuit* using data from multiple environments, which admits the canonical *nonparametric invariance pursuit* in Section 1.1 as a special case.

Let $Y \in \mathbb{R}$ be the response variable and $X \in \mathbb{R}^d$ be the explanatory variable. We consider the general setting in which we have collected data from multiple environments $\mathcal{E} = \{e_1, \ldots, e_{|\mathcal{E}|}\}$, where $\mathcal{E}$ is the set of a finite number of environments. In each environment $e \in \mathcal{E}$, we observe n i.i.d. observations $\{(X_i^{(e)}, Y_i^{(e)})\}_{i=1}^n$ that follow from some distribution $\mu^{(e)}$. Let $\Theta_g, \Theta_f \subseteq \Theta$ be the class of prediction functions and testing functions, respectively. Our goal is to estimate the underlying *invariant regression function* $g^\star \in \Theta_g$ satisfying the invariance structure

$$\forall e \in \mathcal{E} \qquad \mathbb{E}\left[\left(Y^{(e)} - g^\star(X_{S^\star}^{(e)})\right) f(X_{S^\star}^{(e)})\right] = 0 \qquad \forall f \in [\Theta_f]_{S^\star}, \tag{4.1}$$

where $S^\star$ is the *unknown* set of true important variables. We refer to the above problem as *invariance pursuit* or *causal pursuit* exchangeably, as no evidence against casualty with the available experiments.

The problem of estimating $g^\star$ in (4.1) is a generalized version of the canonical *nonparametric invariance pursuit* with $g^\star = m^\star$ in (1.1) and $\Theta_f = \Theta_g = \Theta$. It depicts a general form and unifies several problems of interest in predecessors. For example, when Θ_g and Θ_f are all linear function classes, it reduces to the *linear invariance pursuit* problem, i.e., estimating $g^\star(x) = (\beta^\star)^\top x = (\beta_{S^\star}^\star)^\top x_{S^\star}$ with $\beta^\star \in \mathbb{R}^d$ satisfying $\mathrm{supp}(\beta^\star) = S^\star$ in the multi-environment linear regression (Fan et al., 2023) with linear invariance structure

$$\mathbb{E}\left[\left(Y^{(e)} - (\beta_{S^\star}^\star)^\top X_{S^\star}^{(e)}\right) X_j^{(e)}\right] = 0 \qquad \forall e \in \mathcal{E}, j \in S^\star. \tag{4.2}$$

Another example is the *augmented linear invariance pursuit* where Θ_g is linear and $\Theta_f = \{f(x) = \sum_{j=1}^d \beta_{0,j} x_j + \beta_{1,j}\phi(x_j)\}$ with some transform function $\phi : \mathbb{R} \to \mathbb{R}$. This can further generalize this to

multiple transformed testing functions such as $\phi_1(x_j) = x_j^2$ and $\phi_2(x_j) = |x_j|$ but we keep one here for simplicity. The augmented linear invariance structure that realizes (4.1) in this case is

$$\mathbb{E}\left[\left(Y^{(e)} - (\beta^\star_{S^\star})^\top X^{(e)}_{S^\star}\right) X^{(e)}_j\right] = \mathbb{E}\left[\left(Y^{(e)} - (\beta^\star_{S^\star})^\top X^{(e)}_{S^\star}\right) \phi(X^{(e)}_j)\right] = 0 \qquad \forall e \in \mathcal{E}, j \in S^\star \tag{4.3}$$

It coincides with the problem considered by Fan & Liao (2014) when $|\mathcal{E}| = 1$ and our method reduces to the FGMM method therein. The *augmented linear invariance pursuit* leverages further a part of the structural knowledge that $\mathbb{E}[Y^{(e)}|X^{(e)}_{S^\star}] = (\beta^\star_{S^\star})^\top X^{(e)}_{S^\star}$, which is much weaker than the assumption $\mathbb{E}[Y^{(e)}|X^{(e)}] = (\beta^\star_{S^\star})^\top X^{(e)}_{S^\star}$ in the sparse linear regression. Identification is possible in this case even when $|\mathcal{E}| = 1$. This is important for most biological medical studies, where data are usually collected in similar settings. In this case, the FAIR penalty eliminates endogenous spurious variables, making traditional variable selection methods applicable.

Remark 7. *We point out here that there are two kinds of spurious variables. One is endogenous spurious variables such as X_2 = background color, and the other is exogenous spurious variables such as X_3 = the time the photo was taken or the types of camera used. The former is harmful, and the latter is nearly harmless in statistical prediction, transfer learning, and even statistical attribution or causality, thinking of X_3 as a weak causal variable. The introduction of our FAIR method is to surely screen the endogenous spurious variables (Fan & Lv, 2008). Exogenous spurious variables can be reduced by using commonly used statistical variable selection methods.*

Similar to the discussion in Section 1.1, the main challenge here is the curse of endogeneity. To address this issue, we will harness the insight that the distributions of (X, Y) across diverse environments capture the invariance structure (4.1). The central idea of this paper is to exploit both the heterogeneity among different environments, i.e., the shifts in population distributions $\mu^{(e)}$, in conjunction with the above invariance structure (4.1) to pinpoint the invariant regression function $g^\star$.

It should be noted that both $g^\star$ and $S^\star$ are determined by (Θ_g, Θ_f) and $\mathcal{E}$ through the structure (4.1). It is required that $\partial\Theta_g = \{g - g' : g, g' \in \Theta_g\} \subseteq \Theta_f$. In the case of $\Theta_f = \partial\Theta_g$, one uses only heterogeneity among different environments, or the "invariance principle", to identify the invariant regression function $g^\star$, as in (4.2). Heterogeneous environments are essential in this case. By choosing substantially large $\Theta_f \supsetneq \partial\Theta_g$, one further injects the strong structural assumption that the invariant regression function lies in the class Θ_g rather than $\Theta_f \setminus \Theta_g$ as in (4.3). In this case, one leverages both heterogeneity among environments, i.e., the "invariance principle", and the mentioned prior structure knowledge, i.e., the "asymmetry principle", to jointly identify $g^\star$. Only one environment may be enough for identifying $g^\star$ when the intersection of both principles gives sufficient conditions.

4.2 General FAIR Estimation Framework

Let $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a user-determined risk loss such that

$$\frac{\partial \ell(y, v)}{\partial v} = (v - y)\psi(v) \qquad \text{and} \qquad \frac{\partial^2 \ell(y, v)}{\partial v^2} > 0, \tag{4.4}$$

which is slightly more general than the quasi-likelihood in the generalized linear model (Nelder & Wedderburn, 1972). The constraints in (4.4) ensure that the conditional expectation aligns with the unique global minima and can be satisfied by various risk losses. Two leading examples are the least square loss $\ell(y, v) = \frac{1}{2}(y - v)^2$ with $\psi(v) = 1$ for regression, and the cross-entropy loss $\ell(y, v) = -\log(1 - v) - y \log\{v/(1 - v)\}$ with $\psi(v) = 1/\{v(1 - v)\}$ for classification.

Given all the data $\{\{(X^{(e)}_i, Y^{(e)}_i)\}_{i=1}^n\}_{e\in\mathcal{E}}$ from heterogeneous environments together with (Θ_g, Θ_f) that may encode part of the prior information when $\Theta_g \neq \Theta$, our proposed focused adversarial invariance regularized estimator (FAIR estimator) is the solution to the subsequent minimax optimization objective

$$\widehat{g} \in \operatorname*{argmin}_{g\in\mathcal{G}} \sup_{f^{\mathcal{E}} \in \{\mathcal{F}_{S_g}\}^{|\mathcal{E}|}} \underbrace{\widehat{\mathsf{R}}(g) + \gamma \widehat{\mathsf{J}}(g, f^{\mathcal{E}})}_{=:\widehat{\mathsf{Q}}_\gamma(g, f^{\mathcal{E}})}. \tag{4.5}$$

where $\mathcal{G} \subseteq \Theta_g$ and $\mathcal{F} \subseteq \Theta_f$ are function classes that approximates Θ_g and Θ_f, respectively. Here $\widehat{\mathsf{R}}(g)$ is the pooled sample mean of the user-specified loss across all the environments $\mathcal{E}$:

$$\widehat{\mathsf{R}}(g) = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \widehat{\mathbb{E}}\left[\ell(Y^{(e)}, g(X^{(e)}))\right] = \frac{1}{|\mathcal{E}|\cdot n}\sum_{e\in\mathcal{E}, i\in[n]} \ell(Y_i^{(e)}, g(X_i^{(e)})), \tag{4.6}$$

γ is the hyper-parameter to be determined, and $\widehat{\mathsf{J}}(g, f^{\mathcal{E}})$ is defined the same as (2.2).

Discussions and Extensions. From a high-level perspective, our proposed FAIR estimator searches for the most predictive variable set S that preserves some invariance structure imposed by the specification of (Θ_g, Θ_f). The FAIR estimation framework presented has several limitations: (1) the loss ℓ has restrictions in that the conditional expectation must uniquely minimize it; (2) the environment label is discrete; and (3) the discussion still lies within the variable selection level invariance rather than general representation level invariance. We will discuss in Appendix A.3 that our entire framework can be easily extended to the cases where (1) and (2) fail to hold. We add some discussions on the rationale, comparison with IRM, and extension on (3) in Appendix A.2.

4.3 Sketch of the Generic Result and Its Applications

The non-asymptotic results in Section 2 can be extended to be the result for the general FAIR estimation framework, formally stated in Theorem 4, which unifies the identification condition and L_2 estimation errors for specific (Θ_g, Θ_f) or $(\mathcal{G}, \mathcal{F})$ under the least squares loss $\ell(y, v) = \frac{1}{2}(y-v)^2$. We sketch the main idea and informal statement here and defer the complete result and applications to Appendix B.

Suppose $[\Theta_g]_S$ and $[\Theta_f]_S$ are closed subspaces of Θ_S for any $S \subseteq [d]$ so that one can define

$$\bar{g}^{(S)}(x) = \operatorname*{argmin}_{g\in[\Theta_g]_S} \|g - \bar{m}^{(S)}\|_2 \qquad \text{and} \qquad f^{(e,S)}(x) = \operatorname*{argmin}_{f\in[\Theta_f]_S} \|f - m^{(e,S)}\|_{2,e}.$$

In this case, the invariant structure and the invariant regression function in (4.1) can be simplified as

$$f^{(e,S^\star)}(x) \equiv \bar{g}^{(S^\star)}(x) := g^\star(x). \tag{4.7}$$

Similar to the nonparametric bias mean and bias variance in Remark 3, we can define the generalized bias mean and bias variance with respect to (Θ_g, Θ_f) as $\mathsf{b}(S) = \|\bar{g}^{(S\cup S^\star)} - g^\star\|_2^2$ and $\bar{\mathsf{d}}(S) = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \|\bar{g}^{(S)} - f^{(e,S)}\|_{2,e}^2$. The general identification condition akin to Condition 2 is

$$\forall\ S \subseteq [d], \qquad \mathsf{b}(S) > 0 \implies \bar{\mathsf{d}}(S) > 0, \tag{4.8}$$

The above condition requires that whenever incorporating more variables in S will lead to better prediction performance, the set S will not satisfy the invariance structure (4.1). Condition 2 instantiates (4.8) by letting $\bar{\mathsf{d}}(S) = \bar{\mathsf{d}}_{\mathtt{NN}}(S)$ and $\mathsf{b}(S) = \mathsf{b}_{\mathtt{NN}}(S)$ with $(\mathsf{b}_{\mathtt{NN}}(S), \bar{\mathsf{d}}_{\mathtt{NN}}(S))$ defined in (2.4).

Theorem 3 (Main Result for FAIR Least Squares Estimator, Informal)**.** *Under (4.7), (4.8) and some regularity conditions in regression, one can consistently estimate $g^\star$ by choosing $\gamma \geq 8\sup_{S:\mathsf{b}(S)>0}\{\mathsf{b}(S)/\bar{\mathsf{d}}(S)\}$. In this case, the FAIR estimator $\widehat{g}$ in (4.5) with $\ell(y, v) = \frac{1}{2}(y-v)^2$ satisfies, for any $n \geq 3$, w.h.p.,*

$$\frac{\|\widehat{g} - g^\star\|_2}{C_1} \leq \delta_{\mathtt{stoc}} + \delta^\star_{\mathtt{approx}} + \gamma(\delta_{\mathtt{stoc}} + \delta_{\mathtt{approx}})1_{\{\delta_{\mathtt{stoc}}+\delta_{\mathtt{approx}} \geq \frac{s}{1+\gamma}\}} \tag{4.9}$$

Here $\delta_{\mathtt{stoc}}$ is the stochastic error characterized by the local Rademacher complexity of $\mathcal{F}, \partial\mathcal{G}$ and n, $\delta^\star_{\mathtt{approx}}$ measures certain approximation error of $(\mathcal{G}, \mathcal{F})$ w.r.t. $g^\star$, and $\delta_{\mathtt{approx}}$ measures the worst case approximation error of $(\mathcal{G}, \mathcal{F})$ w.r.t. all the $\{f^{(e,S)}\}$. The constant $s > 0$ is the signal strength related to $\min_{S:\bar{\mathsf{d}}(S)>0}\bar{\mathsf{d}}(S)$ and $\min_{S:S^\star\setminus S\neq\emptyset}\inf_{g\in[\Theta_g]_S}\|g - g^\star\|_2$, and C_1 is a universal constant independent of the two quantities.

The complete and rigorous statement is deferred to Theorem 4 in Appendix B.1, with more loss function ℓ in Theorem 5. These generic results can characterize several advantages in our FAIR framework's sample efficiency. Firstly, the error (4.9) is structure-agnostic in that it is represented by the

Θ_g	Θ_f	$\mathcal{G}$	$\mathcal{F}$	Priors	$\lvert\mathcal{E}\rvert = 1$ Ident	Result
Linear	Linear	Linear	Linear	None	Impossible	Thm 8
Linear	Linear w/ ϕ	Linear	Linear w/ ϕ	Nearly Linear	Possible	Thm 9
Linear	Θ	Linear	NN	Linear	Possible	Thm 10
Additive	Θ	Additive NN	NN	Additive	Impossible	Thm 7
Θ	Θ	NN	NN	None	Impossible	Thm 1

Table 1: Applications of Theorem 4. Recall that Θ is the set of all $L_2(\bar{\mu}_x)$ functions. For the function classes in columns $\Theta_g, \Theta_f, \mathcal{G}$ and $\mathcal{F}$, "Linear" is $\{f(x) = \sum_{j=1}^d \beta_j x_j\}$, "Linear w/ ϕ" is $\{f(x) = \sum_{j=1}^d \beta_j x_j + \alpha_j \phi(x_j)\}$, "NN" is deep ReLU network class, "Additive" is the additive functions $\{f(x) = \sum_{j=1}^d f_j(x_j)\}$ and "Additive NN" is a structured neural network approximating additive functions. The column "Priors" indicates what prior structure knowledge is injected by the choice of (Θ_g, Θ_f). For the second row, it is "nearly linear" given it only requires that the residual is uncorrelated with all the $\phi(x_j)$ with $j \in S^\star$; the prior for the third row is exactly linear provided $\Theta_f = \Theta$. The column "$|\mathcal{E}| = 1$ Ident" indicates whether identification for $S^\star$ in (1.1) is possible with only one environment.

sum of approximation error and stochastic error, indicating that (1) our framework can fully exploit the capability of $(\mathcal{G}, \mathcal{F})$ in learning low-dimensional structures, and (2) it has almost no additional cost in sample efficiency compared with standard regression. Moreover, the error rate applies to any n, implying the estimation error is guaranteed even when it selects the wrong variable, especially when the signal s is weak. Finally, though a large enough regularization hyper-parameter γ is needed to guarantee consistent estimation, the error will be free of γ when n is large enough. We also apply our unified result to various specifications of $(\mathcal{G}, \mathcal{F})$, including the non-asymptotic results in identification and convergence rate; see a summary in Table 1.

5 Experiments

5.1 An End-to-End Implementation

We realize the minimax optimization using gradient descent ascent, a similar approach adopted in GAN (Goodfellow et al., 2014) training. The main challenge here is how to do "focused regularization" which enforces $f^{(e)} \in \mathcal{F}_{S_g}$. Here we consider a re-parameterization trick that disentangles the function g and the the variable S_g it selects. To start with, we can write $g(x) = g(a \odot x) = g(x_1 a_1, \ldots, x_d a_d)$ with $a \in \{0,1\}^d$ indicating presence and absence of variables. Then the objective (4.5) can be written as

$$(\widehat{g}, \widehat{a}) \in \operatorname*{argmin}_{g \in \mathcal{G}, a \in \{0,1\}^d} \sup_{f^{\mathcal{E}} \in \{\mathcal{F}\}^{|\mathcal{E}|}} \widehat{\mathsf{R}}(g(a \odot \cdot)) + \gamma \widehat{\mathsf{J}}(g(a \odot \cdot), f^{\mathcal{E}}(a \odot \cdot)) \tag{5.1}$$

A naive implementation is to first enumerate all the possible $a \in \{0,1\}^d$ and then do gradient descent ascent for given a, which is computationally inefficient. To avoid this, we first rewrite the optimization as a "continuous" optimization:

$$(\widehat{g}, \widehat{w}) \in \operatorname*{argmin}_{g \in \mathcal{G}, w \in R^d} \sup_{f^{\mathcal{E}} \in \{\mathcal{F}\}^{|\mathcal{E}|}} \mathbb{E}_{B(w)} \left[\widehat{\mathsf{R}}(g(B(w) \odot \cdot)) + \gamma \widehat{\mathsf{J}}(g(B(w) \odot \cdot), f^{\mathcal{E}}(B(w) \odot \cdot)) \right],$$

where the j^{th} component of $B(w) \in \{0,1\}^d$ follows an independent Bernoulli with probability of success $\mathrm{sig}(w_j) = \exp(w_j)/(1 + \exp(w_j))$. This is easily seen by taking $\widehat{w} = \mathrm{logit}(\widehat{a}) = \log(\frac{\widehat{a}}{1-\widehat{a}})$. Note that $B_j(w_j) = I(\mathrm{logit}(U_j) \le w_j)$ is discontinuous in w_j where $U_j \sim$ uniform[0,1], but can be approximated as

$$B_j(w_j) \approx \frac{1}{1 + e^{(\mathrm{logit}(U_j) - w_j))/\tau}} \equiv V_\tau(U_j, w_j) \quad \text{as} \quad \tau \to 0^+, \tag{5.2}$$

for which its gradient can be taken. Let

$$A_\tau(U, w) = (V_\tau(U_1, w_1), \ldots, V_\tau(U_d, w_d))^\top \in \mathbb{R}^d$$

with $\{U_j\}_{j=1}^d$ being i.i.d. uniform random variables. One can approximate of the original objective (5.1) by

$$(\widehat{\theta}, \widehat{w}) \in \operatorname*{argmin}_{\theta \in \mathbb{R}^{N_g}, w \in \mathbb{R}^d} \sup_{\forall e \in \mathcal{E}, \phi^{(e)} \in \mathbb{R}^{N_f}} \underbrace{\mathbb{E}_{A_\tau(U,w)} \left[\widehat{\mathsf{R}}(g(A \odot \cdot; \theta)) + \gamma \widehat{\mathsf{J}}(g(A \odot \cdot), f^{\mathcal{E}}(A \odot \cdot; \{\phi^{(e)}\}_{e \in \mathcal{E}})) \right]}_{\mathbb{E}_U[\widehat{\mathsf{L}}(A_\tau(U,w), \theta, \{\phi^{(e)}\}_{e \in \mathcal{E}})]}, \tag{5.3}$$

where parametrizations of $g \in \mathcal{G}$ and $f^e \in \mathcal{F}$ are used. Since $\text{logit}(U_j) \stackrel{d}{=} U_{j,1} - U_{j,2}$ with $\{U_{j,1}, U_{j,2}\}_{j=1}^d$ being i.i.d. Gumbel(0,1) random variables, the approximation (5.2) is also referred to as the Gumble approximation.

One can use similar implementation tricks widely used in stochastic gradient descent with Gumbel approximation that gradually anneals the Gumbel approximation hyperparameter τ. We defer the formal pseudo-code Algorithm 1 to the Appendix C.1. The code to reproduce the results in this section can be found at https://github.com/wmyw96/FAIR.

5.2 Simulations

In this section, we present the simulation result for the FAIR-Linear estimator and FAIR-NN estimator implemented by the Gumbel approximation trick and gradient descent ascent algorithm.

5.2.1 Finite Performance of FAIR-Linear Estimator

Data Generating Process. We consider the case where $|\mathcal{E}| = 2$ and the data $(X^{(e)}, Y^{(e)})$ in each environment $e \in \{0, 1\}$ are generated from two SCMs sharing the same causal relationship between variables. For each trial, we first generate the parent-children relationship among the variables. We enumerate all the $i \in [d+1]$. For each $i \in [d+1]$, we randomly pick at most 4 parents for the variable Z_i from $\{Z_1, \ldots, Z_{i-1}\}$, this step ensures that the induced graph is a DAG. We use fixed $d = 70$, and let the variable Z_{36} be Y and the rest variables constitute the covariate X, that is, we let $(Z_1, \ldots, Z_{35}, Z_{36}, Z_{37}, \ldots, Z_{71}) = (X_1, \ldots, X_{35}, Y, X_{36}, \ldots, X_{70})$. We also enforce that Y has at least 5 parents and at least 5 children by adding parents and children when needed. The structural assignment for each variable Z_j is defined as

$$Z_j^{(e)} \leftarrow \sum_{k \in \texttt{pa}(j)} C_{j,k}^{(e)} f_{j,k}^{(e)}(Z_k^{(e)}) + C_{j,j}^{(e)} \varepsilon_j$$

where $(\varepsilon_1, \ldots, \varepsilon_{71})$ are independent standard normal random variables. For $j \neq 36$, $f_{j,k}^{(e)}$ are sampled randomly from the candidate functions $\{\cos(x), \sin(x), \sin(\pi x), x, 1/(1+e^{-x})\}$, $C_{j,k}^{(e)}$ are sampled from Uniform$[-1.5, 1.5]$ with $|C_{j,j}^{(e)}| \geq 0.5$. For $j = 36$ and $k < j$, we have $f_{36,k}^{(e)}(x) = x$ and $C_{36,k}^{(0)} \equiv C_{36,k}^{(1)}$ for linearity and invariance. The above data-generating process can be regarded as one observation environment $e = 0$ and an interventional environment $e = 1$ where the random and simultaneous interventions are applied to all the variables other than the variable Y, while the assignment from Y's parent to Y remains and furnishes the target regression function $m^\star(x) = \sum_{k \in \texttt{pa}(36)} C_{36,k}^{(e)} x_k$ in pursuit. In this case, we let $S^\star = \texttt{pa}(36)$ and $\beta^\star$ with support set $S^\star$ be such that $\beta_j^\star = C_{36,k}^{(0)} = C_{36,k}^{(1)}$ for any $k \in S^\star$. We also let the noise variance be different for the two environments, i.e., $C_{36,36}^{(0)} \neq C_{36,36}^{(1)}$. Now, the model only has conditional expectation invariance rather than the full conditional distribution invariance. Fig. 2 (a) visualizes the induced graph in one trial. The complex cause-effect relationships in high-dimensional variables make the problem of causal pursuit and estimating $\beta^\star$ very challenging.

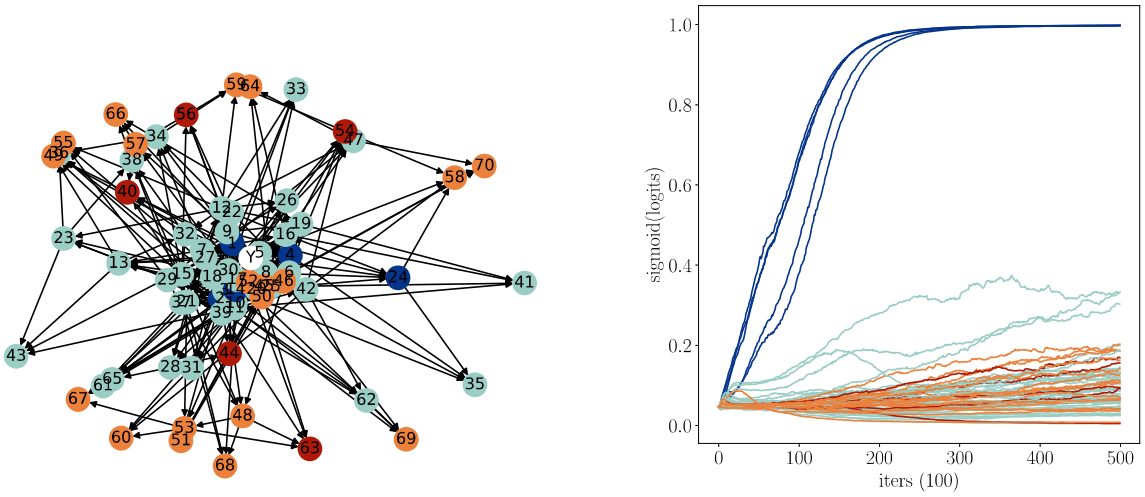
Implementation. For the FAIR-Linear estimator, we realize $\mathcal{G}$ and $\mathcal{F}$ by linear function classes, i.e., $\mathcal{G} = \{g(x) = \beta_g^\top x : \beta_g \in \mathbb{R}^d\}$ and $\mathcal{F} = \{f(x) = \beta_f^\top x : \beta_f \in \mathbb{R}^d\}$, and run gradient descent ascent using Adam optimizer with a learning rate of 1e-3, batch size 64 for $50k$ iterations. In each iteration, one gradient descent update of the parameters of the predictor β_g and Gumbel logits parameters w is followed by the three gradient ascent updates of the discriminators' parameters $(\beta_f^{(1)}, \beta_f^{(2)})$. We adopt a fixed hyper-parameter $\gamma = 36$ and report the performance of the following estimators using the median of the estimation error $\|\widehat{\beta} - \beta^\star\|_2^2$ over 50 replications and varying $n \in \{200, 500, 1000, 2000, 5000\}$.

(1) Pool-LS: it simply runs least squares on the full covariate X using all the data.

(2) FAIR-GB: Our FAIR-Linear estimator with Gumbel approximation that outputs $\beta_g \odot \text{sig}(w)$.

(3) FAIR-RF: it selects the variables x_j with $\text{sig}(w_j) > 0.9$ of the fitted model in (2), i.e., $\widehat{S} = \{j : \text{sig}(w_j) > 0.9\}$, and refits least squares again on $X_{\widehat{S}}$ using all the data.

(4) Oracle: it runs least squares on $X_{S^\star}$ using all the data.



(a) (b)

Figure 2: The visualization of (a) the SCM and (b) the $\mathrm{sig}(w)$ *during training in one trial for the FAIR-Linear estimator. We use different colors to represent the different relationships with* Y*: blue = parent, red = child, orange = offspring, lightblue = other.*

(5) Semi-Oracle: it runs least squares on X_{G^c} using all the data, where G is the set of all the descendants of Y. Compared with the ERM, it manually removes all the variables that will lead to a biased estimation, but it will also keep uncorrelated variables compared with the full Oracle estimation.

Fig. 2 (b) visualizes how the Gumbel gate values for different covariables $\mathrm{sig}(w)$ evolve during training in one trial. We can see that $\mathrm{sig}(w_j)$ for $j \in S^\star$ quickly increases and dominates the values for other variables like children/offspring of Y during the whole training process.

Results. The results are shown in Fig. 3 (a). We can see that the square of the ℓ_2 estimation error $\|\widehat{\beta} - \beta^*\|_2^2$ for the pooled least squares estimator ($\times$) does not decrease and remains to be very large (≈ 1.5) as n increases, indicating that it converges to a biased solution. At the same time, the estimation error for FAIR-GB ($\blacklozenge$) decays as n grows (≈ 0.01 when $n = 1k$) and lies in between that for least squares on X_{G^c} (Semi-Oracle $\blacktriangledown$) and least squares on X_{S^*} (Oracle $\blacktriangle$). This is expected to happen since the FAIR-Linear estimator is not designed to screen out all the exogenous spurious variables: They can be further regularized using the commonly variable selection techniques; see footnote 4. We also observe that the training dynamics of adversarial estimation are highly non-stable: though it can converge to an estimate around $\beta^\star$ when n is very large, it fails to converge to $\beta^\star$ at a comparable rate compared to the standard least squares. The FAIR-RF ($+$) estimator then completes the last step towards attaining better accuracy in this regard: we can see that its performances are very close to that of the Oracle estimator when n is very large ($n = 5000$).

Comparison with Other Methods. We also compare our FAIR-Linear estimator with the cousin estimator EILLS ($\blacktriangleright$) in Fan et al. (2023) and other invariance learning estimators (dotted lines), including invariant causal prediction Peters et al. (2016) (ICP$\blacktriangledown$), invariant risk minimization Arjovsky et al. (2019) (IRM$+$), anchor regression Rothenhäusler et al. (2021) (Anchor$\bullet$) in a similar but smaller dimension setting with $d = 15$, under which ICP and EILLS can be computed within affordable time. For the FAIR-Linear estimator, we report the performance of the FAIR-RF ($\blacklozenge$) and the one with brute force search (FAIR-BF$\blacksquare$). The results are shown in Fig. 3 (b): we can see that the FAIR family estimators ($\blacktriangleright\blacksquare\blacklozenge$ with solid lines) are the only ones attaining consistent estimation among all the invariant learning methods; see a detailed discussion of the data generating process and results in Appendix C.2.1.

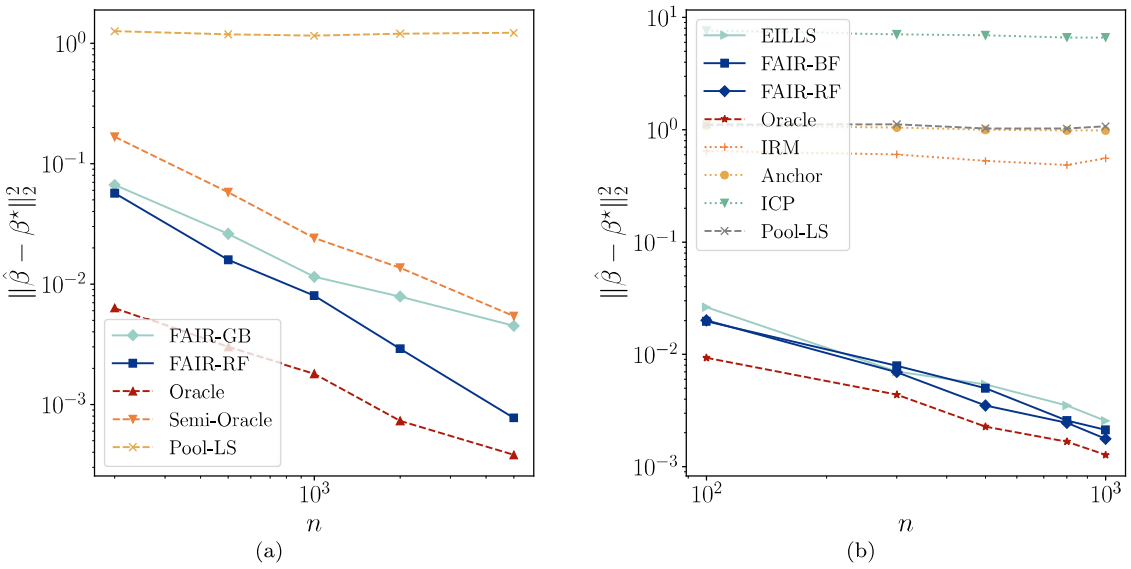


Figure 3: The simulation results for linear models with (a) $d = 70$ and (b) $d = 15$. Both figures depict how the median estimation errors (based on 50 replications, shown in log scale) for different estimators (marked with different shapes and colors) change when n varies in (a) $\{200, 500, 1000, 2000, 5000\}$ and (b) $\{100, 200, 500, 800, 1000\}$, respectively.

5.2.2 Finite Performance of FAIR-NN Estimator

Data Generating Process. We consider the following data generating process with $d = 26$ and $|\mathcal{E}| = 2$ in each trial as

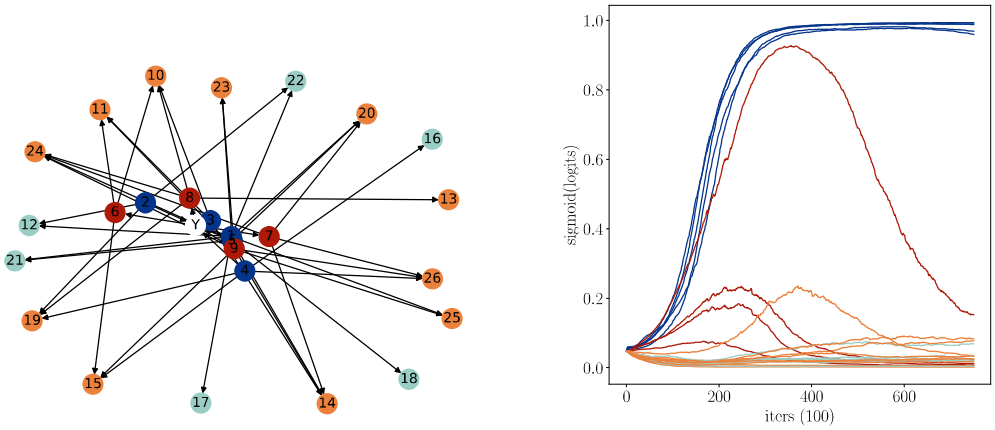
$$X_i^{(e)} \leftarrow \begin{cases} \varepsilon_i^{(e)} & i \le 5 \\ f_{i,0}^{(e)}(Y^{(e)}) + \varepsilon_i^{(e)} & 6 \le i \le 9 \\ \sum_{j \in \mathtt{pa}(i) \subseteq [8]} f_{i,j}^{(e)}(X_j^{(e)}) + \varepsilon_i^{(e)} & 10 \le i \le 26 \end{cases}$$
$$Y^{(e)} \leftarrow m_k^\star(X_1^{(e)}, \ldots, X_5^{(e)}) + \varepsilon_0,$$

where the regression function $m^\star$ is either $m_1^\star(x) = \sum_{k=1}^5 m_{0,j}(x_j)$ with random chosen $m_{0,j}$ or a hierarchical composition model $m_2^\star(x) = x_1 x_2^3 + \log(1 + e^{\tanh(x_3)} + e^{x_4}) + \sin(x_5)$; see detailed model and omitted implementation details in Appendix C.2.2. In the two environments, the cause-effect relationships are shared. The variable Y's parent set is $\{1, 2, 3, 4, 5\}$, its children set is $\{6, 7, 8, 9\}$, and may have potential descendants in $\{9, \ldots, 26\}$. The above data generating process can be regarded as one observation environment $e = 0$ and an interventional environment $e = 1$ where the random and simultaneous interventions are applied to all the variables other than the variable Y, while the assignment from Y's parent to Y remains and furnishes the target regression function $m_k^\star(x)$ with $k \in \{1, 2\}$ in pursuit. Fig. 4 (a) visualizes the induced graph in one trial.

Implementation. We let $\mathcal{G}$ be the class of ReLU neural network with depth 2 and width 128 and $\mathcal{F}$ be the class of ReLU neural network with depth 2 and width 196, and run gradient descent ascent using similar experimental configurations. We use the following empirical mean squared square computed using another $2 \times n_{\text{test}} = 2 \times 30000$ i.i.d. sampled data

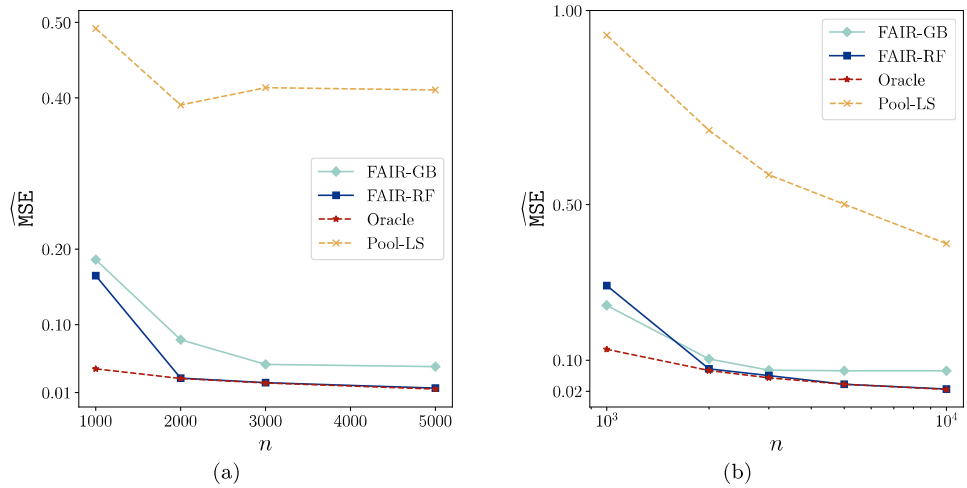
$$\widehat{\mathtt{MSE}} = \frac{1}{2n_{\text{test}}} \sum_{e \in \mathcal{E}} \sum_{i=1}^{n_{\text{test}}} \{m^\star(x_i^{(e)}) - \widehat{m}(x_i^{(e)})\}^2$$

as the evaluation metric. We report the median of $\widehat{\mathtt{MSE}}$ over 100 replications for the estimators (1) – (4) akin to that for the linear model. For (1), (2), and (4), we also use ReLU neural network width depth 2 and width 128 in running least squares. Fig. 4 (b) also visualizes how the Gumbel gate values for different covariables $\mathrm{sig}(w)$ evolve during training in one trial. We can see that the training dynamics for $\mathrm{sig}(w)$ is much more challenging and interesting than that for the linear model depicted in Fig. 2:



(a) (b)

Figure 4: The visualization of (a) the SCM and (b) the $\mathrm{sig}(w)$ *during training in one trial for FAIR-NN estimator when* $k = 1$*. We use different colors to represent the different relationships with* Y*: blue = parent, red = child, orange = offspring, lightblue = other.*



(a) (b)

Figure 5: The simulation results for nonlinear models with (a) $m_1^\star$ *and (b)* $m_2^\star$*. Both figures depict how the median estimation errors (based on 50 replications) for different estimators (marked with different shapes and colors) change when* n *varies in* $\{1000, 2000, 3000, 5000\}$ *for (a) and* $\{1000, 2000, 3000, 5000, 10000\}$ *for (b).*

the weight for some Y's children quickly increases at a comparable rate than the variables in $S^\star$ at the beginning, but such a trend slows down and finally completely reverses in the middle. We leave the rigorous and in-depth analysis behind such dynamics for future studies.

Results. The results are shown in Fig. 5 and the messages are similar to those for FAIR-Linear estimators. The pooled least squares yield biased estimation, while our proposed FAIR-NN estimator can unveil the invariant association $m^\star$ from the two environments. Moreover, the refitted FAIR-NN estimator can obtain a near-oracle performance when n is large.

5.3 Application I: Discovery in Real Physical Systems

We apply our method to perform causal discovery in the light tunnel datasets from Gamella et al. (2024). The data are collected from a real physical device under different manipulation settings. The tunnel device contains a controllable light source at one end and two linear polarizers mounted on rotating

frames. Several sensors are deployed in various positions to measure the light intensity. The causal relationships between the variables of interest is known such that we can get access to the ground-truth cause-effect relationship; see Fig. 2(d) and Fig. 3(a) therein for the device diagram and the cause-effect graphs, respectively. It is worth noticing that the data are collected from a real-world device where the associations between the measurements follow from real-world physical laws. This realistic nature together with the knowledge of ground-truth cause-effect knowledge make it an excellent testbed for causal discovery algorithms.

Using the notation in Gamella et al. (2024), we use the variables $(R, G, B, \theta_1, \theta_2, \widetilde{V}_3, \widetilde{V}_2, \widetilde{V}_1, \widetilde{I}_3, \widetilde{I}_2, \widetilde{I}_1, \widetilde{C})$. Here (R, G, B) is the intensity of the light source at three different wavelengths, $\widetilde{C}$ is the drawn electric current, (θ_1, θ_2) represent the angles of the polarizer frame, and $(\widetilde{V}_3, \widetilde{V}_2, \widetilde{V}_1, \widetilde{I}_3, \widetilde{I}_2, \widetilde{I}_1)$ are the measurement of light-intensity sensors in various positions.

We plan to learn algorithmically the direct cause for $Y = \widetilde{I}_3$, the infrared measurement of the light-intensity sensor after the polarizers, among a subset of manipulable variables and measurement variables $(X_1, \ldots, X_{11}) = (R, G, B, \theta_1, \theta_2, \widetilde{V}_3, \widetilde{V}_2, \widetilde{V}_1, \widetilde{I}_2, \widetilde{I}_1, \widetilde{C})$ under the following two-environment experimental setting: $e = 0$ is the observational environment, $e = 1$ is the interventional environment where the variables $\{\widetilde{V}_j\}_{j=1}^3$ and $\{\widetilde{I}_j\}_{j=1}^2$ are *weakly* intervened on. This leads to the following "equivalent" ground-truth cause-effect relationship among those variables and the effect of "environment intervention" in Fig. 6 (a). In this case, the variables $(R, G, B, \theta_1, \theta_2)$ are the direct causes, i.e., $S^\star = \{1, 2, 3, 4, 5\}$, $\widetilde{V}_3$ are the spurious variables that will lead to biased estimation. The remaining variables are exogenous but have marginal predictive power, i.e., $\mathrm{Var}[Y|X_j] > 0$ for $j \geq 7$.

We will use the following dataset in the experiment: the environment dataset $\mathcal{D}_0$ with size $|\mathcal{D}_0| = 10^4$, the weakly interventional environment dataset $\mathcal{D}_1$ with $|\mathcal{D}_1| = 3000$, and five strongly interventional environment dataset $\mathcal{D}_{2,Z}$ with $Z \in \{\widetilde{V}_1, \widetilde{V}_1, \widetilde{V}_3, \widetilde{I}_1, \widetilde{I}_2\}$ and $|\mathcal{D}_{2,Z}| = 1000$. In each trial, different methods use the same random subsample $\breve{\mathcal{D}} = \{\breve{\mathcal{D}}_0, \breve{\mathcal{D}}_1\}$ with $\breve{\mathcal{D}}_k \subseteq \mathcal{D}_k$ and $|\breve{\mathcal{D}}_k| = n = 1000$ to fit the model. How the fitted model $\widehat{f}$ quantitatively depends on exogenous/endogeneous spurious variable Z is evaluated using the OOS R^2 in corresponding $\mathcal{D}_{2,Z}$ defined as

$$R^2_{\mathrm{OOS},Z} := \frac{\sum_{(X,Y)\in\mathcal{D}_{2,Z}} \{\widehat{f}(X) - Y\}^2}{\sum_{(X,Y)\in\mathcal{D}_{2,Z}} \{Y - \bar{Y}\}^2} \qquad \text{with} \qquad \bar{Y} = \frac{\sum_{(X,Y)\in\breve{\mathcal{D}}_0\cup\breve{\mathcal{D}}_1} Y}{2n}.$$

See the detailed data collection and experimental configuration in Appendix C.3.

The first four rows in Fig. 6 (d) report the variable selection result for several methods over 100 trials. The nonlinear ICP (Heinze-Deml et al., 2018) method does not select any variables because of its conservative nature and stronger heterogeneity condition to recover the direct cause. We can see that FAIR-NN can successfully recover the direct cause $(R, G, B, \theta_1, \theta_2)$ in this case. It exploits neural networks' capability in efficiently detecting the nonlinear associations (the Malus's law, $\widetilde{I}_3 \propto \cos^2(\theta_1 - \theta_2)$ for fixed (R, G, B)), while the linear counterpart FAIR-Linear fails to select the variables (θ_1, θ_2). It is worth pointing out that such a causality recovery cannot be attained by the traditional predictive power and simplicity tradeoff: the variable selection method based on random forest variable importance measures (ForestVarSel) in Heinze-Deml et al. (2018) cannot detect $(G, B, \theta_1, \theta_2)$ and falsely select $(\widetilde{I}_1, \widetilde{I}_2)$. The last three rows in Fig. 6 (d) illustrate how the variable selection rate for the FAIR-NN estimator changes when n grows.

Fig. 6 (b) offers a quantitative illustration by showing the out-of-sample (OOS) R^2 of different estimators under environments with *strong* interventions on $(\widetilde{I}_1, \widetilde{I}_2, \widetilde{V}_1, \widetilde{V}_2, \widetilde{V}_3)$, respectively. The estimator denoted as Oracle-M with $M \in \{\mathrm{Linear}, \mathrm{NN}\}$ referred to the method that runs regress Y on $X_{S^\star}$ using model M. In the spider chart, the red shade represents the out-of-sample R^2 under different interventions for the Oracle-NN estimator that regresses Y on its direct causes. We can see that its performances behave uniformly under various interventions: all the OOS R^2 are approximately equal to 0.91. This is slightly better than that for the linear model (Oracle-Linear) by 0.04. This illustrates the capability of neural networks introduced to detect *weak*, *nonlinear* causal signals from heterogeneous environments. The PoolLS-NN estimator regressing Y on X using neural network and all the data fully exploits the strong spurious association between $\widetilde{V}_3$ and $Y = \widetilde{I}_3$, its heavy reliance on $\widetilde{V}_3$ let it predict better (than the causal model Oracle-NN) when $\widetilde{V}_3$ is not intervened. However, its OOS R^2 significantly decreases by 0.2 when $\widetilde{V}_3$ is strongly intervened hence the spurious association changes. On the contrary, the OOS R^2 for FAIR-NN after refitting (FAIR-NN-RF) behaves almost identical to that for Oracle-NN. This

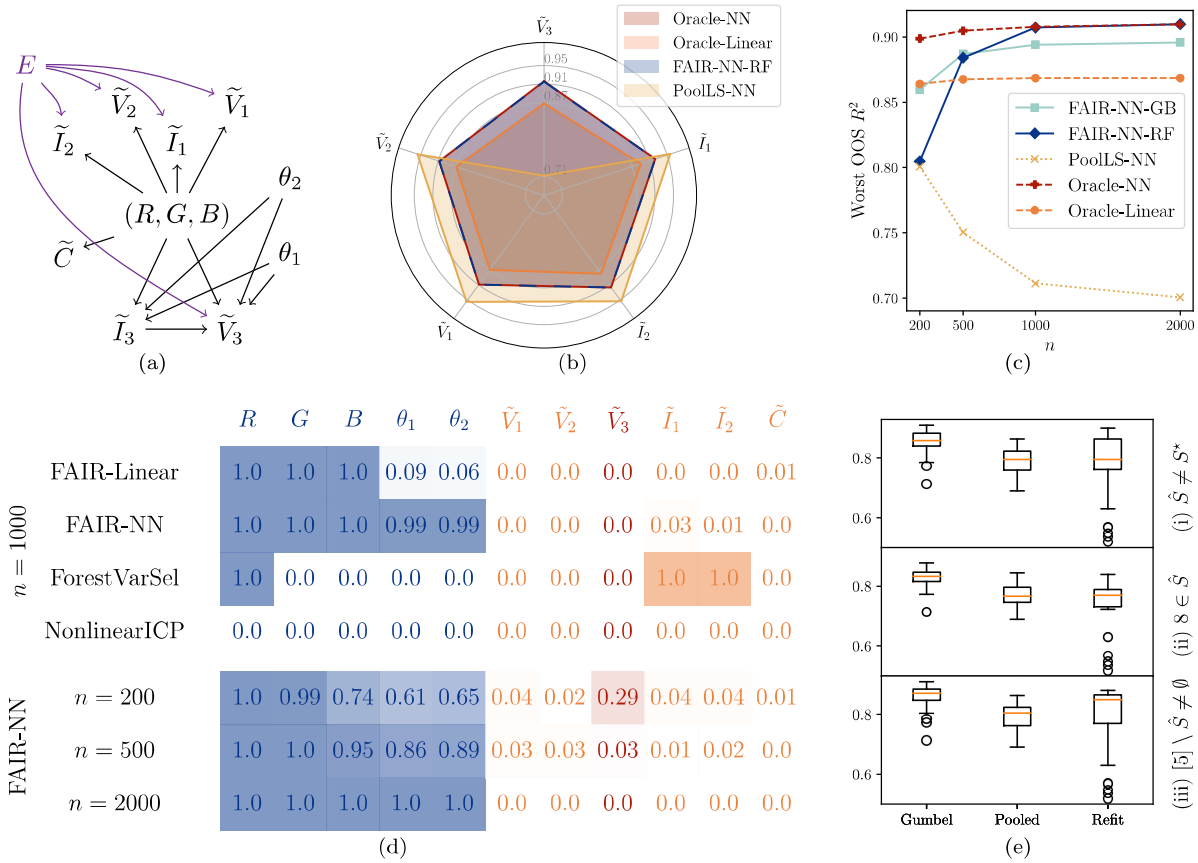


Figure 6: Discovery in Real Physical Systems: (a) the unified cause-effect relationship and interventions similar to Fig. 1 (b). (b) the average out-of-sample R^2 for different estimators using the spider chart: the axis annotated by placeholder variable Z corresponds to the test environment where Z is strongly intervened on. We can see the performance of Oracle-NN and FAIR-NN-RF is almost identical. (c) the average (based on 100 replications) of the worst-case (across 5 environments) of OOS R^2 for different methods as a function of n. (d) the variable selection rate over 100 trials for different methods (top panel) and the variable selection rate for FAIR-NN for various n (bottom panel). We use different colors to represent different relationships with Y: blue=parent, red=child, orange=neither ancestor nor descendants. (e) the distribution of worst-case OOS R^2 (y-axis) for Gumbel-trick optimized FAIR-NN (Gumbel), the follow-up refitted estimator (Refit), and Pooled LS (Pooled) when FAIR-NN selects the wrong variables: the subplots from top to bottom consider the cases of (i) failure in selection consistency (ii) false positive that it falsely selects the child $X_8 = \widetilde{V}_3$ (iii) false negative that it does not select the entire ground-truth $(X_1, \ldots, X_5) = (R, G, B, \theta_1, \theta_2)$.

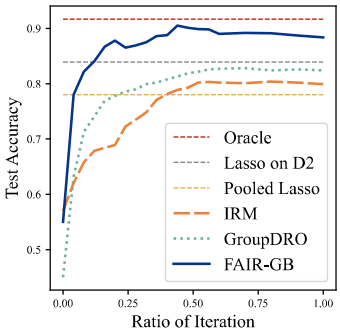
quantitative result illustrates its capability to correct non-trivial and strong bias without no supervision and its efficiency in detecting nonlinear and weak signals.

Fig. 6 (c) shows how the worst-case OOS R^2 among the five, strong intervention environments changes for different estimators when n grows. The performance of the Gumbel-trick optimized FAIR-NN estimator without refitting (FAIR-NN-GB) lies between Oracle-NN and Oracle-Linear and significantly outperforms that of the PoolLS-NN estimator. This suggests that the gradient descent optimized algorithm has already found predictions nearly independent of the spurious variable, and the success of variable selection in Fig. 6 (d) is not because of truncating weak but non-negligible spurious signals. Moreover, as shown in Fig. 6 (e), its performance significantly outperforms the least squares estimator using either the full covariate or the selected covariates when it selects the wrong variable. This further supports the theoretical claims and the advantages of adopting penalized least squares.

5.4 Application II: Prediction Based on Extracted Features

We consider an image object classification task with a spurious background. The target is to classify water birds ($Y = 1$) and land birds ($Y = 0$) (see examples in Fig. 7 (a)) under backgrounds of water or land based on the feature $X \in \mathbb{R}^{500}$ extracted from ResNet pre-trained on ImageNet. We train a linear

Method	Test Accuracy
Oracle	91.06 ± 0.24 %
Lasso on D2	84.57 ± 0.71 %
Pooled Lasso	79.08 ± 0.54 %
IRM	80.32 ± 0.67 %
GroupDRO	82.75 ± 1.10 %
FAIR-GB	89.56 ± 0.53 %



(a) (b) (c)

Figure 7: Prediction Based on Extracted Features: (a) provides two sample images in the dataset: land bird on land (up) and water bird in water (bottom). (b) reports the average ± standard deviation of test accuracy over 10 trials for different estimators. (c) shows how the test accuracy changes over iterations for different methods in one trial.

classifier on top of X using data from two environments $\{\mathcal{D}_k\}_{k=1}^2$. In the first environment $\mathcal{D}_1$, $r_w = 95\%$ water birds appear on the water background and $r_l = 90\%$ land birds stay in land background. The spurious correlation numbers are $r_w = 75\%$ and $r_l = 70\%$ in $\mathcal{D}_2$. A good predictor should based on the core features related to the bird's appearance rather than the strong spurious correlation between the background and label. The trained model is evaluated in a test environment $|\mathcal{D}_3|$ where the spurious correlation reverses: $r_w = 2\%$ and $r_l = 2\%$. We repeat the experiment 10 times, where in each trial the training dataset and the test dataset are sampled from a larger dataset with sizes $n = |\mathcal{D}_1| = |\mathcal{D}_2| = 50k$ and $|\mathcal{D}_3| = 30k$. We compare our FAIR(-egulerized) estimator using $\mathcal{G} = \{\text{sig}(\beta^\top x)\}, \mathcal{F} = \{\beta^\top x\}$ and classification loss $\ell(y, v) = -\log(1 - v) - y \log\{v/(1 - v)\}$ (FAIR-GB) with invariant risk minimization (IRM) (Arjovsky et al., 2019) and group distributionally robust optimization (GroupDRO) (Sagawa et al., 2020). We also consider running Lasso on different environments for reference, including (1) using all the data $\mathcal{D}_1 \cup \mathcal{D}_2$ (Pooled Lasso); (2) using data in $\mathcal{D}_2$ (Lasso on D2); (3) using another randomized controlled environment $\mathcal{D}_4$ with $r_w = r_l = 50\%$ and $|\mathcal{D}_4| = n$ (Oracle). All the models are linear, and the performance of (3) can be seen as the upper bound of the performance using linear models; see data collection and experiential configuration details in Appendix C.4.

The performances are reported in Fig. 7 (b). Fig. 7 (c) also depicts how test accuracy changes as iterations in one trial. FAIR-GB performs similar to Oracle and significantly outperforms Lasso on D2, while other methods (IRM, DRO) falls behind Lasso on D2. This indicates that these methods cannot go beyond interpolating the spurious associations in $\mathcal{D}_1$ and $\mathcal{D}_2$, while our method can nearly eliminate the spurious association using the relatively small perturbations in the two environments.

Acknowledgement

We thank Yiran Jia for helpful discussions on presenting a generic identification result on SCM using the unified graph including E, Yimu Zhang for the help with the numerical implementation in Section 5.4, and Xinwei Shen for suggestions of using Gumbel approximation in implementation.

Appendix

The appendix is organized as follows:

Appendix A contains the omitted discussions in the main text, including the applicable scenarios for the nonparametric invariance pursuit, some discussions and extensions on the method, and some discussions on the conditions in Section 2 and Section 3.

Appendix B contains the complete result that is sketched in Section 4.3.

Appendix C contains omitted discussions and results in experiments section.

A Omitted Discussions and Results

A.1 Applicable Scenarios for Nonparametric Invariance Pursuit

This section is devoted to providing a self-contained introduction to the motivation behind the nonparametric invariance pursuit using statements akin to previous literature (Peters et al., 2016; Rojas-Carulla et al., 2018; Fan et al., 2023).

Causal Discovery. If we can expect $\mathcal{E}$ to be heterogeneous enough, recovering $S^\star$ in nonparametric invariance pursuit coincides with discovering the direct cause of Y when the multi-environment data come from SCM with intervention on X setting.

Proposition 3. *Under the model (3.2), if we further assume that $\mathbb{E}[|Y^{(e)}|^2] < \infty$ for any $e \in \mathcal{E}$, then (1.1) holds with $S^\star = \mathtt{pa}(d+1)$.*

The SCM (3.2) and Proposition 3 extend the framework described in Peters et al. (2016) (specifically Section 4.1 and Proposition 1). This model accommodates nonlinear structural assignments. Critically, the residuals $\varepsilon^{(e)} = Y^{(e)} - \mathbb{E}[Y^{(e)}|X_{S^\star}^{(e)}]$, do not need to be independent of $X_{S^\star}^{(e)}$ or remain invariant across various environments as represented by $\varepsilon^{(e)} \sim \mu_\varepsilon$. Such flexibility broadens the scope for various applications, including binary classification. According to Proposition 3, when restricted to model (3.2), a specific instantiation of our generic statistical model (1.1), identifying the true important variable set $S^\star$ is tantamount to pinpointing the direct cause of the target variable Y. Concurrently, unveiling the invariant association $m^\star$ aligns with uncovering the causal mechanism between Y and its direct causes.

Transfer Learning. Consider we collect data $\{(X_i^{(e)}, Y_i^{(e)})\}_{e\in\mathcal{E}, i\in[n]}$ from $|\mathcal{E}|$ distinct sources and aim to develop a model that produces decent predictions on the data $\{X_i^{(t)}\}_{i\in[n_t]}$ in an unseen environment t. A significant portion of transfer learning algorithms fundamentally relies on the covariate shift assumption, represented as

$$\mathbb{E}[Y^{(t)}|X^{(t)}] \equiv \mathbb{E}[Y^{(e)}|X^{(e)}] \qquad \forall e \in \mathcal{E}.$$

However, as illustrated in Fan et al. (2023); Rojas-Carulla et al. (2018), it is hard for this to be true given collecting so many variables. Therefore, a more realistic assumption is that information from true important variables is transferable, articulated as $\mathbb{E}[Y^{(t)}|X_{S^\star}^{(t)}] = \mathbb{E}[Y^{(e)}|X_{S^\star}^{(e)}]$. The subsequent proposition suggests that though $m^\star$ might not be the optimal predictor in the unseen environment t, it does minimize the worst-case L_2 risk, and the associated excess risk can be decomposed as follows.

We suppose both the distribution $\mu^{(e)}$ we observed in $\mathcal{E}$ and the future distributions ν come from the following distribution family.

$$\mathcal{U}_{S^\star,m^\star,\sigma^2} = \Big\{\mu : \mathbb{E}_\mu[Y^2] < \infty, \mathbb{E}_\mu[Y|X_{S^\star}] = m^\star(X_{S^\star}), \mathbb{E}_\mu[\mathrm{Var}_\mu(Y|X_{S^\star})] \vee \max_{1\le j\le d} \mathbb{E}_\mu[X_j^2] \le \sigma^2\Big\},$$

Proposition 4. *Let $\nu \in \mathcal{U}_{S^\star,m^\star,\sigma^2}$ be arbitrary. Define $\mathsf{R}_{\mathtt{oos}}(m;\nu_x) = \sup_{\mu\in\mathcal{U}_{S^\star,m^\star,\sigma^2},\mu_x\sim\nu_x} \mathbb{E}_{(X,Y)\sim\mu}[|Y - m(X)|^2]$ and $\Theta^{(t)} = L_2(\nu_x)$. We have*

$$\forall m \in \Theta^{(t)} \qquad \mathsf{R}_{\mathtt{oos}}(m;\nu_x) - \mathsf{R}_{\mathtt{oos}}(m^\star;\nu_x) = \|m - m^\star\|_{L_2(\nu_x)}^2 + 2\sigma\|m - \widetilde{m}\|_{L_2(\nu_x)},$$

where $\widetilde{m}(x) = \mathbb{E}_{X\sim\nu_x}[m(X)|X_{S^\star} = x_{S^\star}]$. The term $2\sigma\|m - \widetilde{m}\|_{L_2(\nu_x)}$ is zero when $m \in \Theta_{S^\star}^{(t)}$.

Given the framework described above, our proposed method solving problem in Section 1.1 can be integrated with the re-weighting technique (Gretton et al., 2009), a strategy addressing discrepancies within the marginal distribution of X, to yield reliable predictions in the previously unobserved environment t.

A.2 Discussion on the Methods

We provide a discussion in a question-and-response manner.

[Q] You are doing "focused regularizer" that are of combinatorial nature in computation, can it be removed?
Answer: The short answer is No. The regularizer will be the same as running least squares if we do not enforce the discriminator using the same variables that the predictor uses. This is also the main computational difficulty in our framework and why we use randomness relaxation and Gumbel approximation in implementation. Indeed, even for linear invariance pursuit, there are certain fundamental computational limits in this such that no polynomial-time algorithm can attain consistent estimation in pursuing invariance without relying on additional structures other than invariance.

[Q] The method has a similar form to IRM, what's the major difference?
Answer: The main difference is we should at least let $\Theta_f \supseteq \Theta_g$, such a constraint leverage the idea of over-identification and make identification possible even when $|\mathcal{E}| = 2$ provided enough heterogeneity. Suppose our regularizer, which can be seen as a "correct" method to pursue condition expectation invariance, is to make $u^{(1)} = u^{(2)}$ for two s-dimensional parameter vectors $u^{(1)}, u^{(2)} \in \mathbb{R}^s$, what IRM does is to let $\sum_{i=1}^s u_i^{(1)} = \sum_{i=1}^s u_i^{(2)}$. It is hard to say the latter constraint will make sense and can obtain a similar effect as the former.

[Q] Could your proposed framework be extended to the representation-level invariance like IRM?
Answer: The short answer is Yes given its algorithmic nature. But identification with two or constant-level environments is impossible now: a linear-in-dimension number of environments is required even for linear representation learning. For example, one can find some linear representation $\Phi : \mathbb{R}^d \to \mathbb{R}^r$ such that

$$\mathbb{E}[Y^{(e)}|\Phi X^{(e)}] \equiv m^\star(\Phi X^{(e)})$$

However, $|\mathcal{E}| \geq r$ is the necessary condition for identification even when the heterogeneity is enough and r is pre-known to us. We conjecture that any finite number of environments $|\mathcal{E}| < \infty$ may be impossible for identification if Φ lies in some nonparametric function class.

A.3 Extensions to General Environment Variable and Loss Function

In the main text, we propose an estimation framework leveraging conditional expectation invariance with respect to discrete environment variables. It is worth noticing that our adversarial estimation framework is indeed more versatile than this: one can easily extend it to other conditional point prediction invariance with respect to more general environment covariates. We briefly discuss the direct extension here and leave a rigorous treatment as future work. In the following discussions, suppose we observe data $\{(X_i, Y_i, E_i)\}_{i=1}^n$ drawn i.i.d. from some distribution μ_0, where $X \in \mathbb{R}^d$ is the covariate we used for prediction, $Y \in \mathbb{R}$ is the target response, $E \in \mathbb{R}^q$ is the environment covariate we wish our prediction should be invariant with respect to.

Let $\ell(u, y) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be the user-defined risk whose population-level minimizer may not necessarily be conditional expectation but satisfying certain regularity conditions. Let $\ell_u(u, y) = \partial \ell(u, y)/\partial u$ be the partial sub-gradient with respect to the prediction. Suppose the following general invariance structure with respect to ℓ and environment covariate holds, that there exists $S^\star \subseteq [d]$ and a function $g^\star$ that only depends $x_{S^\star}$ such that

$$\mathbb{E}\left[\ell_u(g^\star(X_{S^\star}), Y)|X_{S^\star}, E\right] \equiv 0. \tag{A.1}$$

It coincides with the main problem of study when E is discrete and ℓ satisfies (4.4), but also allows for other loss and continuous environment label. Other losses include but not limited to Huber loss for robust regression, or L_1 loss for median regression.

We consider the following optimization minimax objective containing a min-max game between a predictor $g : \mathbb{R}^d \to \mathbb{R}$ and a discriminator $f : \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}$:

$$\min_{g\in\mathcal{G}} \max_{f\in\mathcal{F}_{S_g}} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(g(X_i), Y_i)}_{\widehat{\mathsf{R}}(g)} + \gamma \underbrace{\frac{1}{n}\sum_{i=1}^{n} \left[\ell_u(g(X_i), Y_i) f(X_i, E_i) - 0.5\{f(X_i, E_i)\}^2\right]}_{\widehat{\mathsf{J}}(g,f)}, \tag{A.2}$$

where γ is the hyper-parameter to be determined, and $\mathcal{F}_{S_g} = \{f(x, e) \in \mathcal{F}, f(x, e) = w(x_{S_g}, e) \text{ for some } w\}$. Similar to the calculation in Section 1.2, one can expect that minimizing the population counterpart of the focused adversarial invariance regularizer $\max_{f\in\mathcal{F}_{S_g}} \widehat{\mathsf{J}}(g, f)$ shares a similar nature of imposing (A.1). One can derive non-asymptotic identification and estimation error results akin to Theorem 4 and Theorem 5 provided strong convexity and certain Lipschitz property of the loss $\ell(u, y)$. We leave this for future studies.

A.4 Discussion on Relaxing Nonparametric Invariance Pursuit Identification Condition

Given our FAIR criterion search for the most predictive variable set whose conditional expectations remain across different environments, that is, when $\gamma \to \infty$ and $n = \infty$, our population-level objective is equivalent to the following program,

$$\min_{g\in\Theta} \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \mathbb{E}\left[|Y^{(e)} - g(X^{(e)})|^2\right] \quad s.t. \;\; g(x) \equiv m^{(e,S_g)}(x) \;\; \forall e \in \mathcal{E}.$$

We say a set S is an invariant set if $m^{(e,S)} \equiv \bar{m}^{(S)}$ for any $e \in \mathcal{E}$. Therefore, one can slightly relax the identification condition as: $S^\star$ is the most predictive invariant set, that is,

$$\forall S \subseteq [d], \qquad \text{if } \; m^{(e,S)} \equiv \bar{m}^{(S)}, \qquad \text{then either } \|\bar{m}^{(S)}\|_2 < \|\bar{m}^{(S^\star)}\|_2 \text{ or } \bar{m}^{(S)} = m^\star. \tag{A.3}$$

The above condition is definitely weaker than Condition 2 because Condition 2 essentially requires the set $S^\star$ is the maximum invariant set,

$$\forall S \subseteq [d], \qquad \text{if } \; m^{(e,S)} \equiv \bar{m}^{(S)}, \qquad \text{then } \bar{m}^{(S\cup S^\star)} = \bar{m}^{(S^\star)}. \tag{A.4}$$

Here (A.4) just rewrites Condition 2 in a manner similar to (A.3).

It is easy to derive results similar to Theorem 1 under (A.3) rather than Condition 2. We can construct cases where (A.3) holds but Condition 2 does not. Examples include Example 1 below with $s^{(1)}s^{(2)} = 1$ under which both $\{1\}$ and $\{2\}$ are invariant set but the set $\{1, 2\}$ is not. In this case, Condition 2 no longer holds. However, our algorithm can still consistently estimate $m^\star$ provided (A.3) holds, that the variable X_1 has better prediction power. In the main text, we still adopt Condition 2 instead of (A.3). The main reasons are as follows. All our discussions are under the SCM with interventions setting.

Firstly, as further shown in Section 3, the cases where Condition 2 fails to hold are degenerate cases. When the interventions are nondegenerate, there always exists a maximum invariant set $S^\star$, i.e., Condition 2 holds. This means, (A.3) is somewhat "marginally" weaker than Condition 2.

The second reason is the lack of semantic meaning of $S^\star$ under this case. When the interventions are non-degenerate, $S^\star$ can be interpreted as "contemporary/pragmatic direct causes" that can be expressed as direct causes + unaffected children + parents of unaffected children in Proposition 4, such a variable set also has certain robust transfer learning properties as stated in Proposition 2. All the above semantic meanings are valid even if the interventions are insufficient. However, when the interventions are degenerate such that Condition 2 may not hold but (A.3) may hold, e.g., Example 1, all the two properties will no longer hold. If the true causal mechanism is $X_1 \to Y \to X_2$, then it is possible to construct data generating process such that $S^\star$ can be either $\{1\}$ or $\{2\}$ in (A.3).

A.5 Discussion on the Nondegenerate Intervention Condition

The conditions (a) and (b) in Condition 5 are imposed to eliminate some degenerate cases. To illustrate the intuitions why such two conditions are needed, and how such a condition will hold in general. We consider the following two examples.

Introduction of condition (a) From a high-level viewpoint, the introduction of condition (a) is to eliminate the cases where though there are shifts in condition distributions among different environments, it happens that there are no shifts in conditional expectations. This can be illustrated in the following example.

Example 1. *Consider the following canonical model also presented in Example 4.1 in Fan et al. (2023).*

$$\begin{aligned} X_1^{(e)} &\leftarrow \sqrt{0.5}U_1 \\ Y^{(e)} &\leftarrow X_1^{(e)} + \sqrt{0.5}U_3 \\ X_2^{(e)} &\leftarrow s^{(e)}Y^{(e)} + U_2 \end{aligned}$$

where U_1, U_2, U_3 are independent standard normal variables, and $\mathcal{E} = \{1, 2\}$. We let $e = 1$ be the observational environment and $e = 2$ be the interventional environment where the linear effect of Y on X_2 are intervened ($s^{(1)} \neq s^{(2)}$). We also focus on the regime where $s^{(1)} + s^{(2)} \neq 0$ such that running least squares will lead to a biased solution.

In the above model, we can see that

$$Y^{(e)}|X_2^{(e)} \sim \mathcal{N}\left(\frac{s^{(e)}}{(s^{(e)})^2+1}X_2^{(e)}, \frac{1}{(s^{(e)})^2+1}\right)$$

It is easy to check under the case of no-degenerated child ($s^{(1)} + s^{(2)} \neq 0$) and faithfulness on $\widetilde{M}$ ($s^{(1)} \neq s^{(2)}$). We have

$$Y^{(1)}|X_1^{(1)} \overset{d}{\neq} Y^{(2)}|X_2^{(2)},$$

or in other words, $Y \perp\!\!\!\perp E|X_2$. However, when $s^{(1)} = 1/s^{(2)} = s$, the following holds

$$\mathbb{E}[Y^{(1)}|X_2^{(1)} = x] = \frac{s^{(1)}}{(s^{(1)})^2+1}x = \frac{s}{s^2+1}x = \frac{s^{(2)}}{(s^{(2)})^2+1}x = \mathbb{E}[Y^{(2)}|X_2^{(2)} = x]$$

The introduction of Condition 5 (a) is to rule out the cases where $s^{(1)} = 1/s^{(2)} = s$. And it is easy to see when $s^{(1)}$ and $s^{(2)}$ are independently generated from some prior distribution that is absolute continuous with respect to Lebesgue measure on $\mathbb{R}$, i.e., $S^{(1)}, S^{(2)} \sim p_s$, then

$$\mathbb{P}\left[S^{(1)}S^{(2)} = 1\right] = 0.$$

Introduction of condition (b). The condition (b), that the faithfulness condition on $\widetilde{M}$, is to eliminate the cases where though the interventions are applied, it happens that such interventions do not make an impact on the variables intervened. The following example presents such an example.

Example 2. *Consider the case where $\mathcal{E} = \{1, 2\}$, and the data generating process is as follows*

$$\begin{aligned} Y^{(e)} &\leftarrow U_3 \\ X_1^{(e)} &\leftarrow Y^{(e)} + e + U_1 \\ X_2^{(e)} &\leftarrow 0.5Y^{(e)} - sX_1^{(e)} + e + U_2. \end{aligned}$$

where U_1, U_2, U_3 are independent standard normal variables, $s \neq 0.5$ is a fixed parameter. We let $e = 1$ be the observational environment and $e = 2$ be the interventional environment where shifts in mean are applied to the variables X_1 and X_2.

In the above case, we have $S^\star = \mathtt{pa}(3) = \emptyset$, and there exists a effective simultaneous intervention on (X_1, X_2). However, such an intervention will not affect X_2 if and only if $s = 1$ because its direct effect on X_2 and the indirect effect passing through X_1 get canceled provided $s = 1$. To be specific, $X_2^{(e)}$ can be written as

$$X_2^{(e)} = 0.5Y^{(e)} - s(Y^{(e)} + e + U_1) + e + U_2 = (0.5 - s)Y^{(e)} - sU_1 + U_2 + e(1 - s).$$

This implies that

$$Y \perp\!\!\!\perp E|X_2$$

provided $s = 1$, under which the faithfulness on $\widetilde{M}$ fails to hold because we have $Y \not\perp\!\!\!\perp_{\widetilde{G}} E|X_2$ since the path $Y \to X_2 \leftarrow E$ is not blocked by X_2. However, if the parameter s is also generated from some prior distribution that is absolute continuous with respect to Lebesgue measure on $\mathbb{R}$, i.e., $S \sim p_s$, then

$$\mathbb{P}[S = 1] = 0.$$

A.6 The Complete Statement of Proposition 2

Specifically, we construct a unified SCM $(X, Y, E) \sim \bar{M}(\bar{\mathcal{S}}, \nu)$ based on $M^{(0)}$ and new environment $M^{(t)}$ as follows:

$$\begin{aligned} E &\leftarrow \text{Uniform}(\{0, t\}) \\ X_j &\leftarrow \begin{cases} \bar{f}_j(X_{\mathtt{pa}(j)}, U_j) := f_j^{(0)}(X_{\mathtt{pa}(j)}, U_j) & \forall j \in [d] \setminus I \\ \bar{f}_j(X_{\mathtt{pa}(j)}, E, U_j) := f_j^{(t)}(X_{\mathtt{pa}(j)}, U_j) & \forall j \in I \end{cases} \\ Y &\leftarrow \bar{f}_{d+1}(X_{\mathtt{pa}}(d+1), U_{d+1}) := f_{d+1}(X_{\mathtt{pa(d+1)}}, U_{d+1}). \end{aligned}$$

We suppose the following condition similar to Condition 5 holds in the constructed graph.

Condition 6. *The following holds for $\bar{M}$: (1) $\forall S \subseteq [d]$ containing Y's descendants, i.e., $d+1 \in \cup_{j \in S} \mathtt{at}(j)$, if $E \not\perp\!\!\!\perp_{\bar{M}} Y|X_S$, then $(\mu^{(0)} \wedge \mu^{(t)})(\{m^{(0,S)} \neq m^{(t,S)}\}) > 0$; (2) $\bar{M}$ is faithful, that is,*

$$\forall \text{ Disjoint } A, B, C \subseteq [d+2], \qquad Z_A \perp\!\!\!\perp Z_B|Z_C \quad \overset{(a)}{\Longrightarrow} \quad Z_A \perp\!\!\!\perp_{\bar{G}} Z_B|Z_C,$$

where $Z_A \perp\!\!\!\perp_{\widetilde{G}} Z_B|Z_C$ means the node set A and B and d-separated conditioned on C in the graph $\bar{G} = G(\bar{M})$.

We are ready to give a complete statement of Proposition 2.

Proposition 5 (Formal Statement of Proposition 2)**.** *Under the setting of Theorem 2, for a new environment t with SCM $M^{(t)} = \{\mathcal{S}^{(t)}, \nu\}$ satisfying $f_j^{(t)} \equiv f_j^{(0)}$ for any $j \in [d+1] \setminus I$, i.e., only X_I is intervened, we also have $\mathbb{E}[Y^{(t)}|X_{S_\star}^{(t)}] \equiv \mathbb{E}[Y^{(0)}|X_{S_\star}^{(0)}]$. Suppose further that Condition 6 holds for the constructed SCM $\bar{M}$. Then $S_\star$ is the unique largest set whose conditional expectation is transferable, i.e., for any $S \subseteq [d]$ such that $\mathbb{E}[Y^{(t)}|X_{S_\star \cup S}^{(t)}] \neq \mathbb{E}[Y^{(t)}|X_{S_\star}^{(t)}]$, one has $\mathbb{E}[Y^{(t)}|X_S^{(t)}] \neq \mathbb{E}[Y^{(0)}|X_S^{(0)}]$.*

B Generic Results and Its Applications

B.1 Main Result for the General FAIR Least Squares Estimator

This section is designed to offer a unified main result characterizing when the FAIR least squares estimator can identify the target regression function together with a non-asymptotic L_2 error bound for general $(\mathcal{G}, \mathcal{F})$. We first introduce some standard regularity conditions.

Condition 7 (Data Generating Process)**.** *We collect data from $|\mathcal{E}| \in \mathbb{N}^+$ environments. For each environment $e \in \mathcal{E}$, we observe $(X_1^{(e)}, Y_1^{(e)}), \ldots, (X_n^{(e)}, Y_n^{(e)}) \overset{i.i.d.}{\sim} \mu^{(e)}$.*

Condition 8 (Sub-Gaussian Response)**.** *For any $e \in \mathcal{E}$ and $t \geq 0$, $\mathbb{P}\left[|Y^{(e)}| \geq t\right] \leq C_y e^{-t^2/(2\sigma_y^2)}$, where $\sigma_y > 0$ and $C_y > 0$ are some constants independent of e and t.*

To impose statistical complexity on the function classes we used, we introduce the definition of *localized population Rademacher complexity*, described as follows.

Definition 5 (Localized Population Rademacher Complexity). *For a given radius $\delta > 0$, function class $\mathcal{H}$, and distribution ν, define*

$$R_{n,\nu}(\delta; \mathcal{H}) = \mathbb{E}_{X,\varepsilon}\left[\sup_{h\in\mathcal{H}, \|h\|_{L_2(\nu)}\le\delta}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i h(X_i)\right|\right],$$

where $X_1, \ldots, X_n$ are i.i.d. samples from distribution ν, and $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher variables taking values in $\{-1, +1\}$ with equal probability which are also independent of $(X_1, \ldots, X_n)$.

Condition 9 (Function Class). *Suppose the following holds for the function class $\mathcal{G}$ and $\mathcal{F}$ we use:*

(1). It is uniformly bounded by $B \ge 1$, i.e., $\sup_{h\in\mathcal{G}\cup\mathcal{F}} \|h\|_\infty \le B$.
(2). $0 \in \mathcal{F}$ and the statistical complexity of the function classes $\mathcal{G} + \mathcal{F} := \{g + f : g \in \mathcal{G}, f \in \mathcal{F}_{S_g}\}$ is upper-bounded by δ_n. In particular, there exists some quantity $1/n \le \delta_n < 1$ such that

$$R_{n,\mu^{(e)}}(\delta; \partial\mathcal{G}) \le B\delta_n\delta \qquad \text{and} \qquad R_{n,\mu^{(e)}}(\delta; \partial(\mathcal{G}+\mathcal{F})) \le 2B\delta_n\delta$$

for any $e \in \mathcal{E}$ and $\delta \in [\delta_n, 2B]$, where $\partial\mathcal{H} = \{h - h' : h, h' \in \mathcal{H}\}$.

Note that when $-\mathcal{G} = \mathcal{G}$, $R_{n,\mu^{(e)}}(\delta; \partial\mathcal{G}) = R_{n,\mu^{(e)}}(\delta; \mathcal{G})$. The above three assumptions Condition 7, 8, 9 are standard in the theoretical analysis of regression. Recall the definition of $m^{(e,S)}$ and $\bar{m}^{(S)}$ in Section 2.1, now we introduce the specific assumption in our multi-environment regression setting.

Condition 10 (Invariance and Identification). *For any S, let $\overline{\mathcal{G}_S} \supseteq \mathcal{G}_S$, $\overline{\mathcal{F}_S} \supseteq \mathcal{F}_S$ be closed subspaces of Θ_S satisfying $\overline{\mathcal{G}_S} \subseteq \overline{\mathcal{F}_S}$. In this case, we can define $\Pi_{\mathcal{A}}(h) = \operatorname{argmin}_{a\in\mathcal{A}} \|a - h\|_2$ and $\Pi^{(e)}_{\mathcal{A}}(h) = \operatorname{argmin}_{a\in\mathcal{A}} \|a - h\|_{2,e}$ when $\mathcal{A} \in \{\overline{\mathcal{F}_S}, \overline{\mathcal{G}_S}\}$ and $h \in \Theta_S$. Suppose the following holds:*

1. *(Invariance) There exists some index set $S^\star \subseteq [d]$ such that*

$$\forall e \in \mathcal{E} \qquad \Pi^{(e)}_{\overline{\mathcal{F}_{S^\star}}}(m^{(e,S^\star)}) = \Pi_{\overline{\mathcal{G}_{S^\star}}}(\bar{m}^{(S^\star)}) := g^\star$$

2. *(Heterogeneity) For each $S \subseteq [d]$, if $\mathsf{b}_{\mathcal{G}}(S) > 0$, then $\bar{\mathsf{d}}_{\mathcal{G},\mathcal{F}}(S) > 0$, where*

$$\mathsf{b}_{\mathcal{G}}(S) = \|\Pi_{\overline{\mathcal{G}_{S\cup S^\star}}}(\bar{m}^{(S\cup S^\star)}) - g^\star\|_2^2 \quad \text{and} \quad \bar{\mathsf{d}}_{\mathcal{G},\mathcal{F}}(S) = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \|\Pi^{(e)}_{\overline{\mathcal{F}_S}}(m^{(e,S)}) - \Pi_{\overline{\mathcal{G}_S}}(\bar{m}^{(S)})\|_{2,e}^2. \quad \text{(B.1)}$$

3. *(Nondegenerate Covariate) For any $S \subseteq [d]$ such that $S^\star \setminus S \neq \emptyset$, we have $\inf_{g\in\overline{\mathcal{G}_S}} \|g - g^\star\|_2^2 \ge s_{\min}$ for some constant $s_{\min} > 0$.*

The first condition "invariance" specifies the target regression function $g^\star$ of interests and states the invariance structure imposed for our theoretical analysis. It relaxes the general conditional expectation invariance (1.1) when $\overline{\mathcal{F}_S} \subsetneq \Theta_S$. Two leading examples are (1) the fully nonparametric class $\overline{\mathcal{G}_S} = \overline{\mathcal{F}_S} = \Theta_S$, and (2) linear class $\overline{\mathcal{G}_S} = \overline{\mathcal{F}_S} = \{f(x) = \beta_S^\top x_S : \beta_S \in \mathbb{R}^{|S|}\}$. In the first example, we are interested in estimating the invariant conditional expectation $g^\star = m^\star$, and the invariance condition requires the conditional expectation invariance (1.1), that

$$\forall e \in \mathcal{E} \qquad m^{(e,S^\star)}(x) = m^\star(x_{S^\star}).$$

In the second example, when the covariance matrices $\mathbb{E}[X^{(e)}(X^{(e)})^\top]$ across all the environments are all positive definite, we are interested in estimating the invariant linear predictor $g^\star(x) = x^\top\beta^\star$, and such the "invariance" condition only requires that

$$\forall e \in \mathcal{E} \qquad \beta^{(e,S^\star)} \equiv \beta^\star \qquad \text{where} \qquad \beta^{(e,S^\star)} = \operatorname*{argmin}_{\beta\in\mathbb{R}^d, \beta_{(S^\star)^c}=0} \mathbb{E}[|Y^{(e)} - \beta^\top X^{(e)}|^2],$$

that is, the best linear predictors constrained on $S^\star$ among all the environment are the same. In this case, the conditional expectations $m^{(e,S^\star)}(x)$ can be nonlinear or different.

The second condition "heterogeneity" is for identification and is fundamental to derive the population-level strong convexity with respect to $g^\star$. The two quantities in (B.1) are general forms of the bias mean

and the bias variance, respectively. We refer to $\mathsf{b}_{\mathcal{G}}(S)$ as the bias mean because $\mathsf{b}_{\mathcal{G}}(S)$ is the precise bias of the estimator that regress Y on X_S when $S^\star \subseteq S$ using all the data. This can be formally presented in the following proposition, which asserts that in the absence of our proposed regularizer, a vanilla least squares estimator will not consistently estimate $g^\star$, and the discrepancy $\|\widehat{g} - g^\star\|_2^2$ is approximately equal to $\mathsf{b}(S)$ when n is large.

Proposition 6 (Inconsistency of Least Squares Estimator)**.** *Let S be an index set such that $S^\star \subseteq S \subseteq [d]$. Assume Condition 7, 8, 9–10 hold, and $\mathsf{b}_{\mathcal{G}}(S) > 0$. Suppose further that $U\delta_{n,\log n} + \inf_{g\in\mathcal{G}_S} \|g - \Pi_{\overline{\mathcal{G}_S}}(\bar{m}^{(S)})\|_2 = o(1)$, where U and $\delta_{n,t}$ are two constants defined in Theorem 4 below. Then the estimator $\widehat{g}_{\mathtt{R}}$ that minimizes (4.6) in $\mathcal{G}_S$ satisfies, for large enough n,*

$$0.99 \leq \frac{\|\widehat{g}_{\mathtt{R}} - g^\star\|_2^2}{\mathsf{b}_{\mathcal{G}}(S)} \leq 1.01$$

with probability at least $1 - \{C_y(\sigma_y + 1) + 1\}n^{-100}$.

On the other hand, our proposed FAIR estimator will not converge to the biased solution under the condition "heterogeneity". The condition "heterogeneity" is an abstraction of the "identification" condition in previous subsections, for example, Condition 2 for FAIR-NN.

The last condition "nondegenerate covariate" ensures that the target regression function $g^\star$ cannot be exactly fitted by any function g whose dependent variable set S_g does not cover $S^\star$. It reduces to be "non-collinearity" when $\mathcal{G}$ is linear.

In practice, we may only get access to the approximate solution. In our theoretical analysis, we focus on the performance of the approximate solution $(\widehat{g}, \widehat{f^{\mathcal{E}}})$ satisfying

$$\sup_{f^{\mathcal{E}}\in\{\mathcal{F}_{S_{\widehat{g}}}\}^{|\mathcal{E}|}} \widehat{\mathsf{Q}}_\gamma(\widehat{g}, f^{\mathcal{E}}) - (\gamma+1)\delta_{\mathtt{opt}}^2 \leq \widehat{\mathsf{Q}}_\gamma(\widehat{g}, \widehat{f^{\mathcal{E}}}) \leq \inf_{g\in\mathcal{G}} \sup_{f^{\mathcal{E}}\in\{\mathcal{F}_{S_g}\}^{|\mathcal{E}|}} \widehat{\mathsf{Q}}_\gamma(g, f^{\mathcal{E}}) + (1+\gamma)\delta_{\mathtt{opt}}^2 \tag{B.2}$$

with some optimization error $\delta_{\mathtt{opt}}^2 > 0$, here γ in $(1+\gamma)$ is the same as that in $\widehat{\mathsf{Q}}_\gamma$. Now we are ready to state the main result regarding the statistical rate of convergence of our estimator $\widehat{g}$ to $g^\star$, that is,

$$\|\widehat{g} - g^\star\|_2 = \left\{\int (\widehat{g} - g^\star)^2 \bar{\mu}_x(dx)\right\}^{1/2}.$$

Theorem 4 (Main Result for the FAIR Estimator with ℓ_2 Loss)**.** *Assume Conditions 7–10 hold. Define the critical threshold*

$$\gamma^\star := \sup_{S\subseteq[d]:\mathsf{b}_{\mathcal{G}}(S)>0} \frac{\mathsf{b}_{\mathcal{G}}(S)}{\bar{\mathsf{d}}_{\mathcal{G},\mathcal{F}}(S)}.$$

There exists some universal constant C such that, for any $\gamma \geq 8\gamma^\star$, the following holds:
(1) General L_2 error rate. Let $t > 0$ be arbitrary. Define general approximation errors with respect to the function class $\mathcal{G}$ and $\mathcal{F}$ as

$$\delta_{\mathtt{a},\mathcal{G}} = \inf_{g\in\mathcal{G}_{S^\star}} \|g - g^\star\|_2 \quad \text{and} \quad \delta_{\mathtt{a},\mathcal{F},\mathcal{G}}(S) = \sqrt{\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \sup_{g\in\mathcal{G}:S_g=S} \inf_{f\in\mathcal{F}_{S_g}} \|\Pi_{\overline{\mathcal{F}_S}}^{(e)}(m^{(e,S)}) - g - f\|_{2,e}^2},$$

and the stochastic error as $\delta_{n,t} = \delta_n + \{(\log(nB|\mathcal{E}|)+t+1)/n\}^{1/2}$, where δ_n is the quantity in Condition 9. Let $U = B(B + \sigma_y\sqrt{\log(n|\mathcal{E}|)})$, then

$$\|\widehat{g} - g^\star\|_2 \leq C(1+\gamma)\left(U\delta_{n,t} + \delta_{\mathtt{a},\mathcal{G}} + \delta_{\mathtt{a},\mathcal{F},\mathcal{G}}(S_{\widehat{g}}) + \delta_{\mathtt{a},\mathcal{F},\mathcal{G}}(S^\star) + \delta_{\mathtt{opt}}\right). \tag{B.3}$$

with probability at least $\mathfrak{p} = 1 - 6e^{-t} - 2C_y(\sigma_y+1)n^{-100}$.
(2) Faster L_2 error rate. Moreover, if

$$\delta_{\mathtt{opt}}^2 + \sup_{S\subseteq[d]} \delta_{\mathtt{a},\mathcal{F},\mathcal{G}}^2(S) + \delta_{\mathtt{a},\mathcal{G}}^2 + UB\delta_{n,t} \leq \left\{1 \wedge \frac{s_{\min}}{\gamma+1} \wedge \left(\frac{\gamma}{\gamma+1} \inf_{S:\bar{\mathsf{d}}_{\mathcal{G},\mathcal{F}}(S)>0} \bar{\mathsf{d}}_{\mathcal{G},\mathcal{F}}(S)\right)\right\}/C \tag{B.4}$$

then the following holds, with probability at least $\mathfrak{p}$,

$$\|\widehat{g} - g^\star\|_2 \leq C\left(U\delta_{n,t} + \delta_{\mathtt{a},\mathcal{G}} + \delta_{\mathtt{a},\mathcal{F},\mathcal{G}}^\star + \delta_{\mathtt{opt}}\right), \tag{B.5}$$

where $\delta_{\mathtt{a},\mathcal{F},\mathcal{G}}^\star = \{\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \sup_{g\in\mathcal{G}} \inf_{f\in\mathcal{F}_{S_g}} \|g^\star - g - f\|_{2,e}^2\}^{1/2}$.

Theorem 4 generalizes Theorem 4.4 in Fan et al. (2023) to a broad spectrum of $(\mathcal{G}, \mathcal{F})$ configurations. After specifying the function class $(\mathcal{G}, \mathcal{F})$, one can further derive the corresponding identification condition by calculating $(\mathsf{b}_{\mathcal{G}}(S), \bar{\mathsf{d}}_{\mathcal{G},\mathcal{F}}(S))$ and establish a high probability bound on the L_2 error by substituting approximation errors $(\delta_{\mathsf{a},\mathcal{G}}, \delta_{\mathsf{a},\mathcal{F},\mathcal{G}}(S), \delta^{\star}_{\mathsf{a},\mathcal{F},\mathcal{G}})$ and stochastic error δ_n for the function class $(\mathcal{G}, \mathcal{F})$. In particular, when $\mathcal{G}$ and $\mathcal{F}$ are restricted to the linear function class, they not only match but also significantly improve the result in Fan et al. (2023); see Appendix B.6. All the results in Table 2 are direct corollaries of our abstract result Theorem 4.

It is required that γ should be greater than a constant-level critical threshold $8\gamma^{\star}$ for consistent estimation of $g^{\star}$. Theorem 4 further establishes a crude instant-dependent and oracle-type error bound (B.3) that holds for arbitrary $n \geq 2$ and scales linearly with γ. Furthermore, when the stochastic error and approximation errors all go to 0 as n increases and n is large enough such that (B.4) holds, we have (B.5), which improves the L_2 error bound (B.3) in two aspects – the error bound is no longer dependent on either γ or other $m^{(e,S)}$ with $S \neq S^{\star}$. The quantities in the RHS of (B.4) can be interpreted as the smaller of (1) the signal of true important variables and (2) the signal of heterogeneity. When one of these signals is weak, one can expect to demand more data to differentiate whether it is signal or noise.

One important ingredient in the FAIR estimator is the choice of regularization hyper-parameter γ that promotes the invariance. Theorem 4 offers some insights on choosing γ. Firstly, $\gamma \geq C\gamma^{\star}$ is required such that it will correctly identify $g^{\star}$ from a population-level perspective. Second, it will influence the L_2 error rate when n is not large enough such that (B.4) does not hold. Furthermore, the final L_2 error rate (B.5) when n is large enough is independent of γ. This indicates that the estimator's performance is somewhat not very sensitive to the choice of hyper-parameter γ. In this case, one can adopt a slightly conservative large γ to meet the population condition $\gamma \geq C\gamma^{\star}$.

B.2 Extension to the General Risk Loss under the Nonparametric Setting

Condition 11 (Risk Loss). *Define* $\mathcal{V} = [\inf_{g\in\mathcal{G}\cup\{g^{\star}\}} \sup_L\{g(X) \geq L, \bar{\mu}_x\text{-}a.s.\}, \sup_{g\in\mathcal{G}\cup\{g^{\star}\}} \inf_U\{g(X) \leq U, \bar{\mu}_x\text{-}a.s.\}]$ *be the value that* $g(X)$ *takes, and* $\mathcal{Y} = [\sup_l\{Y \geq l, \bar{\mu}_x\text{-}a.s.\}, \inf_u\{Y \leq u, \bar{\mu}_x\text{-}a.s.\}]$ *be the value that* Y *takes. The loss* $\ell(\cdot,\cdot)$ *satisfies*

(1) $\ell(y, v) < \infty$ *for any* $y \in \mathcal{Y}$ *and* $v \in \mathcal{V}$ *and twice continuously differentiable in* $\mathcal{Y} \times \mathcal{V}$. $\frac{\partial \ell(y,v)}{\partial v} = (v - y)\psi(v)$ *for some continuously differentiable* $\psi(v) : \mathbb{R} \to \mathbb{R}$.

(2) There exists some universal constant $\zeta \geq 1$ *such that*

$$|\psi(v)| \leq \zeta \qquad \text{and} \qquad \zeta^{-1} \leq \frac{\partial^2 \ell}{\partial v^2}(Y, v) \leq \zeta \qquad \forall v \in \mathcal{V} \text{ and } \bar{\mu}\text{-}a.s. .$$

The assumptions on risk loss in Condition 11 is standard: (1) ensures that ℓ is well-defined on optimal solutions and linear combination of them, (2) requires that the population-level global minima is conditional mean, (3) guarantees that the loss function is strongly convex and smooth in the domain, and satisfies $|\ell(y, v) - \ell(y, v')| \leq \zeta|y - \widetilde{v}||v - v'|$ for some universal constant ζ, which slightly relaxes the Lipschitz condition in Farrell et al. (2021) and Foster & Syrgkanis (2019).

We now state the invariance and identification condition when the general risk loss is adopted.

Condition 12 (Invariance and Identification for General Risk Loss). *Suppose the following holds*

1. *(Invariance) There exists some index set* $S^{\star} \subseteq [d]$ *such that*

$$\forall e \in \mathcal{E} \qquad m^{(e,S^{\star})} = \bar{m}^{(S^{\star})} =: m^{\star}$$

2. *(Heterogeneity) For each* $S \subseteq [d]$, *if* $\mathsf{b}(S) > 0$, *then* $\bar{\mathsf{d}}(S) > 0$, *where*

$$\mathsf{b}(S) := \|\bar{m}^{(S\cup S^{\star})} - m^{\star}\|_2^2, \qquad \bar{\mathsf{d}}(S) := \frac{1}{|\mathcal{E}|}\sum_{e=1}^{m} \|m^{(e,S)} - \bar{m}^{(S)}\|_{2,e}^2. \tag{B.6}$$

3. *(Nondegenerate Covariate) For any* $S \subseteq [d]$ *such that* $S^{\star} \setminus S \neq \emptyset$, *we have* $\inf_{g\in\Theta_S} \|g - m^{\star}\|_2^2 \geq s_{\min}$ *for some constant* $s_{\min} > 0$.

We are now ready to state the main result in this case.

Theorem 5 (Main Result for the FAIR Estimator with General Risk Loss). *Assume Condition 7,8,9, and Condition 11–12 hold. Define the critical threshold*

$$\gamma^\star := \sup_{S\subseteq[d]:\mathsf{b}(S)>0} \frac{\mathsf{b}(S)}{\bar{\mathsf{d}}(S)}.$$

There exists some universal constant C such that, for any $\gamma \geq 8\zeta^2\gamma^\star$, the following holds:
(1) General L_2 error rate. Let $t > 0$ be arbitrary. Define general approximation errors with respect to the function class $\mathcal{G}$ and $\mathcal{F}$ as

$$\delta_{\mathsf{a},\mathcal{G}} = \inf_{g\in\mathcal{G}_{S^\star}} \|g - m^\star\|_2 \quad \text{and} \quad \delta_{\mathsf{a},\mathcal{F},\mathcal{G}}(S) = \sqrt{\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \sup_{g\in\mathcal{G}:S_g=S} \inf_{f\in\mathcal{F}_{S_g}} \|m^{(e,S)} - g - f\|_{2,e}^2},$$

and the stochastic error as $\delta_{n,t} = \delta_n + \{(\log(nB|\mathcal{E}|)+t+1)/n\}^{1/2}$, where δ_n is the quantity in Condition 9. Let $U = B(B + \sigma_y\sqrt{\log(n|\mathcal{E}|)})$, then

$$\|\widehat{g} - m^\star\|_2 \vee \|\widehat{g} - m^\star\|_n \leq C(\zeta+\gamma)\zeta\left(U\delta_{n,t} + \delta_{\mathsf{a},\mathcal{G}} + \delta_{\mathsf{a},\mathcal{F},\mathcal{G}}(S_{\widehat{g}}) + \delta_{\mathsf{a},\mathcal{F},\mathcal{G}}(S^\star) + \delta_{\mathtt{opt}}\right). \tag{B.7}$$

with probability at least $\mathfrak{p} = 1 - 6e^{-t} - 2C_y(\sigma_y+1)n^{-100}$.
(2) Faster L_2 error rate. Moreover, if

$$\delta_{\mathtt{opt}}^2 + \sup_{S\subseteq[d]} \delta_{\mathsf{a},\mathcal{F},\mathcal{G}}^2(S) + \delta_{\mathsf{a},\mathcal{G}}^2 + UB\delta_{n,t} \leq \left\{1 \wedge \frac{s_{\min}}{(\gamma+\zeta)\zeta} \wedge \left(\frac{\gamma}{\gamma+\zeta} \inf_{S:\bar{\mathsf{d}}_{\mathcal{G},\mathcal{F}}(S)>0} \bar{\mathsf{d}}_{\mathcal{G},\mathcal{F}}(S)\right)\right\}/C \tag{B.8}$$

then the following holds, with probability at least $\mathfrak{p}$,

$$\|\widehat{g} - m^\star\|_2 \vee \|\widehat{g} - m^\star\|_n \leq C\zeta^2\left(U\delta_{n,t} + \delta_{\mathsf{a},\mathcal{G}} + \delta_{\mathsf{a},\mathcal{F},\mathcal{G}}^\star + \delta_{\mathtt{opt}}\right), \tag{B.9}$$

where $\delta_{\mathsf{a},\mathcal{F},\mathcal{G}}^\star = \{\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \sup_{g\in\mathcal{G}} \inf_{f\in\mathcal{F}_{S_g}} \|m^\star - g - f\|_{2,e}^2\}^{1/2}$.

B.3 Key Ideas and Proof Sketch of Theorem 4

We first introduce some additional notations. Let

$$\mathsf{A}^{(e)}(g, f^{(e)}) = \mathbb{E}\left[\{Y^{(e)} - g(X^{(e)})\}f^{(e)}(X^{(e)}) - \frac{1}{2}\{f^{(e)}(X^{(e)})\}^2\right]$$
$$\widehat{\mathsf{A}}^{(e)}(g, f^{(e)}) = \frac{1}{n}\sum_{i=1}^n \{Y_i^{(e)} - g(X_i^{(e)})\}f^{(e)}(X_i^{(e)}) - \frac{1}{2}\{f^{(e)}(X_i^{(e)})\}^2.$$

Define the population-level pooled risk and FAIR estimator loss as

$$\mathsf{R}(g) = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \mathbb{E}\left[\frac{1}{2}|Y^{(e)} - g(X^{(e)})|^2\right] \qquad \text{and} \qquad \mathsf{Q}_\gamma(g, f^{\mathcal{E}}) = \mathsf{R}(g) + \gamma\mathsf{J}(g, f^{\mathcal{E}})$$

We will use the following theorem establishing approximate strong convexity with respect to $g^\star$.

Theorem 6. *Assume Condition 10 hold, $\ell(y,v) = \frac{1}{2}(y-v)^2$. Let $\delta \in (0,1)$ be arbitrary. Then the following holds, for any $\gamma \geq 4\delta^{-1}\gamma^\star$,*

$$\begin{aligned}\mathsf{Q}_\gamma(g, f^{\mathcal{E}}) - \mathsf{Q}_\gamma(\widetilde{g}, \widetilde{f}^{\mathcal{E}}) \geq& \frac{1-\delta}{2}\|g - \widetilde{g}\|_2^2 + \frac{\gamma}{4}\bar{\mathsf{d}}_{\mathcal{G},\mathcal{F}}(S) + \frac{\gamma}{2}\|g - \Pi_{\overline{\mathcal{G}_S}}(\bar{m}^{(S)})\|_2^2 \\ &- \frac{\gamma}{2|\mathcal{E}|}\sum_{e\in\mathcal{E}} \|f^{(e)} - \{\Pi_{\overline{\mathcal{F}_S}}^{(e)}(m^{(e,S)}) - g\}\|_{2,e}^2 - (\delta^{-1} + \gamma/2)\|\widetilde{g} - g^\star\|_2^2\end{aligned}$$

for any $g \in \mathcal{G}$, $\widetilde{g} \in \mathcal{G}_{S^\star}$ and $S_{\widetilde{g}} = S^\star$, $f^{\mathcal{E}} \in \{\overline{\mathcal{F}_{S_g}}\}^{|\mathcal{E}|}$, and $\widetilde{f}^{\mathcal{E}} \in \{\overline{\mathcal{F}_{S^\star}}\}^{|\mathcal{E}|}$.

Recall our definition of

$$\delta_{n,t} = \delta_n + \sqrt{\frac{t + \log(nB|\mathcal{E}|) + 1}{n}} \qquad \text{and} \qquad U = B(B + \sigma\sqrt{\log(n|\mathcal{E}|)})$$

The first proposition establishes instance-dependent error bounds on

$$\Delta_{\mathsf{R}}(g, \widetilde{g}) := \{\widehat{\mathsf{R}}(g) - \widehat{\mathsf{R}}(\widetilde{g})\} - \{\mathsf{R}(g) - \mathsf{R}(\widetilde{g})\},$$

and is standard in nonparametric regression literature.

Proposition 7 (Instance-dependent error bounds for pooled risk)**.** *Suppose Condition 7,8, 9 hold. There exists some universal constant C such that for any $\eta > 0$ and $t > 0$, the following event*

$$\forall g, \widetilde{g} \in \mathcal{G}, \quad |\Delta_{\mathsf{R}}(g, \widetilde{g})| \le CU \left\{ \delta_{n,t}^2 + \delta_{n,t} \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \|g - \widetilde{g}\|_{2,e} \right\}$$

occurs with probability at least $1 - 3e^{-t} - C_y(\sigma_y + 1)n^{-100}$.

The analysis of the focused adversarial invariance regularizer is more involved. The next proposition establishes the instance-dependent error bound for the regularizer. We define

$$\Delta_{\mathsf{A}}^{(e)}(g, \widetilde{g}, f^{(e)}, \widetilde{f}^{(e)}) = \mathsf{A}^{(e)}(f, g^{(e)}) - \mathsf{A}^{(e)}(\widetilde{f}, \widetilde{g}^{(e)}) - \left\{ \widehat{\mathsf{A}}^{(e)}(f, g^{(e)}) - \widehat{\mathsf{A}}^{(e)}(\widetilde{f}, \widetilde{g}^{(e)}) \right\}$$

and

$$\mathcal{M}(\mathcal{G}, \mathcal{F}) = \left\{ (g, \widetilde{g}, f, \widetilde{f}) : g, \widetilde{g} \in \mathcal{G} \text{ and } f \in \mathcal{F}_{S_g}, \widetilde{f} \in \mathcal{F}_{S_{\widetilde{g}}} \right\}.$$

Proposition 8 (Instance-dependent error bounds for regularizer)**.** *Suppose Condition 7, 8, 9 hold. There exists some universal constant C such that for any $t > 0$, the following event*

$$\forall e \in \mathcal{E},\ \forall (g, \widetilde{g}, f^{(e)}, \widetilde{f}^{(e)}) \in \mathcal{M}(\mathcal{G}, \mathcal{F}),$$
$$|\Delta_{\mathsf{A}}^{(e)}(g, \widetilde{g}, f^{(e)}, \widetilde{f}^{(e)})| \le CU \left(\delta_{n,t} \left(\|\widetilde{g} - g\|_{2,e} + \|\widetilde{g} + \widetilde{f}^{(e)} - g - f^{(e)}\|_{2,e} \right) + \delta_{n,t}^2 \right)$$

occurs with probability at least $1 - 3e^{-t} - C_y(\sigma_y + 1)n^{-100}$.

We first utilize Proposition 8 in a way that g and $\widetilde{g}$ are the same. In this case, the optimization problem of max-$\mathcal{F}$ in one single environment $e \in \mathcal{E}$ for fixed $g \in \mathcal{G}$ is similar to least squares regression that fits the target regression function

$$\Pi_{\overline{\mathcal{F}_S}}^{(e)}(m^{(e,S)}) - g.$$

Thus one can establish high probability error bounds on the $\|\cdot\|_{2,e}$ norm between the empirical loss maximizer $\widehat{f}_g^{(e)}$ and the above target function in terms of statistical error $\delta_{n,t}$ and approximation error rate $\delta_{\mathsf{a},\mathcal{F},\mathcal{G}}(e, S_g)$, defined as

$$\delta_{\mathsf{a},\mathcal{F},\mathcal{G}}(e, S) := \sup_{g \in \mathcal{G}: S_g = S} \inf_{f \in \mathcal{F}_S} \|\Pi_{\overline{\mathcal{F}_S}}^{(e)}(m^{(e,S)}) - g - f\|_{2,e}$$

We formally present the above intuition in the following instance-dependent error bound in Proposition 9 in a way that the optimization gap term is maintained in the error bound.

Proposition 9 (Instance-dependent characterization of approximately optimal discriminator)**.** *Let $0 < \eta < 1/2$ be arbitrary, under the event defined in Proposition 8, the following holds,*

$$\forall e \in \mathcal{E}, \forall g \in \mathcal{G}, \forall f^{(e)} \in \mathcal{F}_{S_g},$$
$$\|\Pi_{\overline{\mathcal{F}_S}}(m^{(e,S)}) - g - f^{(e)}\|_{2,e}^2 \le \frac{2\eta^{-1} + 2 - 4\eta}{1 - 2\eta} \delta_{\mathsf{a},\mathcal{F},\mathcal{G}}^2(e, S_g) + \frac{2\eta^{-1} + 4}{1 - 2\eta} C^2 U^2 \delta_{n,t}^2$$
$$+ \frac{4}{1 - 2\eta} \left\{ \sup_{\breve{f} \in \mathcal{F}_{S_g}} \widehat{\mathsf{A}}^{(e)}(g, \breve{f}) - \widehat{\mathsf{A}}^{(e)}(g, f^{(e)}) \right\}$$

where C is the universal constant defined in Proposition 8. Averaging over all the $e \in \mathcal{E}$, we obtain

$$\forall g \in \mathcal{G},\ \forall f^{\mathcal{E}} \in \{\mathcal{F}_{S_g}\}^{|\mathcal{E}|},$$
$$\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \|\Pi_{\overline{\mathcal{F}_S}}(m^{(e,S)}) - g - f^{(e)}\|_{2,e}^2 \le \frac{2\eta^{-1}+2-4\eta}{1-2\eta}\delta^2_{\mathsf{a},\mathcal{F},\mathcal{G}}(S_g) + \frac{2\eta^{-1}+4}{1-2\eta}C^2U^2\delta^2_{n,t}$$
$$+\gamma^{-1}\frac{4}{1-2\eta}\left\{\sup_{\breve{f}^{\mathcal{E}}\in\{\mathcal{F}_{S_g}\}^{|\mathcal{E}|}} \widehat{\mathsf{Q}}_\gamma(g,\breve{f}) - \widehat{\mathsf{Q}}_\gamma(g,f^{\mathcal{E}})\right\}$$

Now we are ready to prove Theorem 4.
For the proof of (2) faster L_2 rate, we will divide the proof into two main steps as follows.

1. In the first step, we establish a variable selection property claim that when the Eq. (B.4) holds, and the events defined in Proposition 7 and 8 occurs, then $\widehat{S}$ satisfies
$$\forall e \in \mathcal{E} \qquad \Pi^{(e)}_{\overline{\mathcal{F}_{\widehat{S}}}}(m^{(e,\widehat{S})}) = g^\star$$
using proof by contradiction that any g such that such that the above constrain is violated in S_g, will not be the approximate solution of the minimax optimization $\inf_g \sup_{f^{\mathcal{E}}} \widehat{\mathsf{Q}}_\gamma(g, f^{\mathcal{E}})$. This can be summarized as the following Proposition 10.

2. In the second step, we proceed conditioned on the above claim and derive a sharp L_2 error bound. To derive a sharp error bound, we combine (1) the approximate strong convexity with respect to $g^\star$, i.e., Theorem 6, (2) the instance-dependent error bound for J and R, i.e., Proposition 7 and 8, and (3) the key fact that, if the claim in step 1 holds, then
$$\begin{aligned}\|\widetilde{g} + \widetilde{f}^{(e)}_{\widetilde{g}} - g - f^{(e)}_g\|_{2,e} &\le \|\widetilde{g} + \widetilde{f}^{(e)}_{\widetilde{g}} - g^\star + \Pi^{(e)}_{\overline{\mathcal{F}_{S_g}}}(m^{(e,S_g)}) - g - f^{(e)}_g\|_{2,e}\\ &\lesssim \|\widetilde{g} + \widetilde{f}^{(e)}_{\widetilde{g}} - g^\star\|_{2,e} + \|g^\star - g - f^{(e)}_g\|_{2,e}\\ &\lesssim \delta_{n,t} + \delta^\star_{\mathsf{a},\mathcal{F},\mathcal{G}}.\end{aligned}$$

The proof of (1) is similar to the second step in the proof of (2), but now we no longer have $g^\star = \Pi^{(e)}_{\overline{\mathcal{F}_{S_g}}}(m^{(e,S_g)})$. The key challenge here is to establish an upper bound on $\|g^\star - \Pi^{(e)}_{\overline{\mathcal{F}_{S_g}}}(m^{(e,S_g)})\|_{2,e}$ without imposing other population-level condition like Condition 7 in an early version of Fan et al. (2023). Instead, we will use the following instance-dependent bound, that

$$\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \|g^\star - \Pi^{(e)}_{\overline{\mathcal{F}_{S_g}}}(m^{(e,S_g)})\|_{2,e}^2 \le C\left((1+\gamma^\star)\bar{\mathsf{d}}_{\mathcal{G},\mathcal{F}}(S_g) + \|g-g^\star\|_2^2\right)$$

Such a bound is a population-level instance-dependent bound in that both the R.H.S. and L.H.S. are dependent on the function g.

Proposition 10. *Under the event defined in Proposition 8 and 7, we have the event*

$$\mathcal{A}_+ := \left\{\forall e \in \mathcal{E} \qquad \Pi^{(e)}_{\overline{\mathcal{F}_{\widehat{S}}}}(m^{(e,\widehat{S})}) = g^\star \qquad \textit{for}\ \ \widehat{S} = S_{\widehat{g}}\right\} \tag{B.10}$$

occurs if the condition (B.4) with some large universal constant C holds.

B.4 Applications of Theorem 4 and Connection to the Predecessors

We present some examples here, sorted by the potential approximation capability of the function class $(\mathcal{G}, \mathcal{F})$.

Example 3 (Linear $\mathcal{G}$, Linear $\mathcal{F}$)**.** *The simplest case is that $\mathcal{G}$ and $\mathcal{F}$ are all linear function classes, that*

$$\mathcal{G} = \mathcal{F} = \{h(x) = \beta^\top x : \beta \in \mathbb{R}^d\} := \mathcal{H}_{\mathtt{lin}}(d).$$

	$\mathcal{G}$	$\mathcal{F}$	Category	Short Name	Result
Example 3	$\mathcal{H}_{\texttt{lin}}(d)$	$\mathcal{H}_{\texttt{lin}}(d)$	$\mathcal{G} \asymp \mathcal{F}$	FAIR-Linear	Theorem 8
Example 7	$\mathcal{H}_{\texttt{nn}}(d, L_g, N_g, B_g)$	$\mathcal{H}_{\texttt{nn}}(d, L_f, N_f, B_f)$	$\mathcal{G} \asymp \mathcal{F}$	FAIR-NN	Theorem 1
Example 4	$\mathcal{H}_{\texttt{lin}}(d)$	$\mathcal{H}_{\texttt{alin}}(d, \phi)$	$\mathcal{G} \ll \mathcal{F}$	FAIR-AugLinear	Theorem 9
Example 5	$\mathcal{H}_{\texttt{lin}}(d)$	$\mathcal{H}_{\texttt{nn}}(d, L_f, N_f, B_f)$	$\mathcal{G} \ll \mathcal{F}$	FAIR-NNLinear	Theorem 10
Example 6	$\mathcal{H}_{\texttt{ann}}(d, L_g, N_g, B_g)$	$\mathcal{H}_{\texttt{nn}}(d, L_f, N_f, B_f)$	$\mathcal{G} \ll \mathcal{F}$	FAIR-ANN	Theorem 7

Table 2: A Glimpse of Estimators

The objective takes on a form that closely resembles the EILLS objective proposed in Fan et al. (2023). To see this, the EILLS objective is expressed as $\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\widehat{\mathbb{E}}[|Y^{(e)} - g(X^{(e)})|^2] + \frac{\gamma}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\|\widehat{r}_g^{(e)}\|_2^2$ where $\widehat{r}_g^{(e)} = \widehat{\mathbb{E}}[\{Y^{(e)} - g(X^{(e)})\}X_{S_g}^{(e)}]$. If we take the supremum over all the $f^{(e)} \in \mathcal{F}_{S_g}$ with $e \in \mathcal{E}$, the objective in (4.5) transforms into

$$\sup_{f^{\mathcal{E}}\in\{\mathcal{F}_{S_g}\}^{|\mathcal{E}|}} \widehat{\mathsf{Q}}_\gamma(g, f^{\mathcal{E}}) = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\widehat{\mathbb{E}}[|Y^{(e)} - g(X^{(e)})|^2] + \frac{\gamma}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}(\widehat{r}_g^{(e)})^\top\{\widehat{\mathbb{E}}[X_S^{(e)}(X_S^{(e)})^\top]\}^{-1}(\widehat{r}_g^{(e)}).$$

It slightly stabilizes the EILLS objective in that the regularizer has a matched moment index compared with the pooled least squares loss; see a detailed explanation and theoretical justification in Appendix B.6.

Example 4 (Linear $\mathcal{G}$, Augmented Linear $\mathcal{F}$)**.** *Consider the case where $\mathcal{F}$ is potentially larger than $\mathcal{G}$, that is, $\mathcal{G} = \mathcal{H}_{\texttt{lin}}(d)$ and $\mathcal{F} = \{f(x) = \beta^\top x + \beta_\phi^\top\bar{\phi}(x) : \beta, \beta_\phi \in \mathbb{R}^d\} := \mathcal{H}_{\texttt{alin}}(d, \phi)$, where $\bar{\phi}(x) = (\phi(x_1), \ldots, \phi(x_d))$ applies a transformation function $\phi : \mathbb{R} \to \mathbb{R}$ to each entry of the vector x.*

The proposed estimator utilizes both the heterogeneity among different environments and the strong prior knowledge that the true regression function admits linear form. It bridges the EILLS estimator in Fan et al. (2023) and the Focused GMM estimator in Fan & Liao (2014) when the instrumental variables are $[X_S, \bar{\phi}(X_S)]$ and reduces to an improved version of the latter when $|\mathcal{E}| = 1$.

Example 5 (Linear $\mathcal{G}$, Neural Network $\mathcal{F}$)**.** *We consider a more algorithmic version of Example 4 that uses neural networks to automatically learn the transformation function, that is, $\mathcal{G} = \mathcal{H}_{\texttt{lin}}(d)$ and $\mathcal{F} = \mathcal{H}_{\texttt{nn}}(d, L_f, N_f, B_f)$ with neural network architecture hyper-parameters of (L_f, N_f, B_f).*

The above three estimators focus on linear $\mathcal{G}$, the simplest structural function class. We now consider a more complicated structural function class when we know the invariant association admits additive form.

Example 6 (Additive Neural Network $\mathcal{G}$, Neural Network $\mathcal{F}$)**.** *We let $\mathcal{G} = \mathcal{H}_{\texttt{ann}}(d, L_g, N_g, B_g) := \{g(x) = \mathrm{Tc}_{B_g}(\sum_{j=1}^d g_j(x_j)) : g_j \in \mathcal{H}_{\texttt{nn}}(1, L_g, N_g, \infty)\}$ and $\mathcal{F} = \mathcal{H}_{\texttt{nn}}(d, L_f, N_f, B_f)$. Here (L_g, N_g, B_g) and (L_f, N_f, B_f) are all neural network architecture hyper-parameters.*

Finally, we present the most algorithmic estimator, the *FAIR-NN* estimator, in which both $\mathcal{G}$ and $\mathcal{F}$ are realized by fully-connected neural networks with no additional imposed structures.

Example 7 (Neural Network $\mathcal{G}$, Neural Network $\mathcal{F}$)**.** *We let $\mathcal{G} = \mathcal{H}_{\texttt{nn}}(d, L_g, N_g, B_g)$ and $\mathcal{F} = \mathcal{H}_{\texttt{nn}}(d, L_f, N_f, B_f)$ with neural network architecture hyper-parameters (L_g, N_g, B_g) and (L_f, N_f, B_f).*

Our framework requires $\mathcal{G} \subseteq \mathcal{F}$. We can divide the above estimators into two main categories that (1) $\mathcal{G}$ has roughly the same representation power as $\mathcal{F}$, denoted as $\mathcal{G} \asymp \mathcal{F}$, and (2) $\mathcal{F}$ has at least as good representation power as $\mathcal{G}$, denoted as $\mathcal{G} \ll \mathcal{F}$. For the former, our framework uses only heterogeneity among different environments to identify the invariant association. For the latter, our framework utilizes both the heterogeneity and strong prior structural assumption that the invariant association cannot be significantly better approximated by $\mathcal{F}$ than by $\mathcal{G}$ to jointly identify the invariant association. We summarize the proposed estimators above and divide them into these two categories in Table 2.

B.5 FAIR-ANN: Bridging Invariance and Additional Structural Knowledge

We next consider the estimator that utilizes both heterogeneity and the strong structural assumption that the invariant association $m^\star$ admits additive form to identify $m^\star$, which can be summarized as the following assumption.

Condition 13 (Invariance and Nondegenerate Covariate for FAIR-ANN). *There exists some set $S^\star$ and $m^\star : \mathbb{R}^{|S^\star|} \to \mathbb{R}$ such that $m^{(e,S^\star)}(x) \equiv m^\star(x_{S^\star}) = \sum_{j\in S^\star} m_j^\star(x_j)$ for any $e \in \mathcal{E}$. Moreover, for any $S \subseteq [d]$ with $S^\star \setminus S \neq \emptyset$, $\inf_{m\in\Theta_S} \|m - m^\star\|_2^2 \geq s_{\min} > 0$.*

Condition 14 (Boundedness in Nonparametric Regression). *There exists some constants b_x and b_m such that (1) $X \in [-b_x, b_x]^d$ $\bar{\mu}$-a.s. and (2) $\|m^{(e,S)}\|_\infty \leq b_m$ for any $S \subseteq [d]$ and $e \in \mathcal{E}$.*

Condition 15. *There exists some constant C_a such that*

$$\left\|\sum_{j=1}^d m_j(x_j)\right\|_2^2 \geq C_a^{-1} \sum_{j=1}^d \|m_j(x_j)\|_2^2 \qquad \forall (m_1,\ldots,m_d) \in \prod_{j=1}^d \Theta_{\{j\}} \text{ with } \int m_j(x_j)\bar{\mu}_x(dx) \equiv 0.$$

The above condition is referred to as the nonparametric version of the restricted strong convexity condition, which is widely used in the theoretical analysis for nonparametric high-dimension additive models (Van de Geer, 2008; Raskutti et al., 2012; Yuan & Zhou, 2016). This condition is imposed to let $\prod_{j\in S} \Theta_{\{j\}}$ be a closed subspace of Θ_S, where we can define

$$A_S(h) = \operatorname*{argmin}_{u\in\prod_{j\in S}\Theta_{\{j\}}} \|h - u\|_2,$$

which finds a unique additive function dependent on x_S that fits h best in $\|\cdot\|_2$ norm.

Condition 16 (Identification for FAIR-ANN). *For any $S \subseteq [d]$ such that $\bar{\mu}(\{m^\star \neq A_{S\cup S^\star}(\bar{m}^{(S\cup S^\star)})\}) > 0$, either of the two holds: (1) there exists some $e, e' \in \mathcal{E}$ such that $(\mu^{(e)} \wedge \mu^{(e')})(\{m^{(e,S)} \neq m^{(e',S)}\}) > 0$, or (2) $\bar{\mu}(\{\bar{m}^{(S)} \neq A_S(\bar{m}^{(S)})\}) > 0$.*

With network hyper-parameter N, L, we realize the $\mathcal{G}$ and $\mathcal{F}$ as

$$\mathcal{G} = \mathcal{H}_{\texttt{ann}}(d, L, N, b_m) \qquad \text{and} \qquad \mathcal{F} = \mathcal{H}_{\texttt{nn}}(d, L+2, 2dN, 2b_m). \tag{B.11}$$

Similarly to the choice of for FAIR-NN (2.3), the choice of $\mathcal{F}$ is to ensure $\mathcal{G} - \mathcal{G} \subseteq \mathcal{F}$.

Theorem 7 (Optimal Rate for FAIR-ANN Least Squares Estimator). *Assume Condition 7,8, and 13–16 hold. Assume further that all the conditional moments $\{m^{(e,S)}\}_{e\in\mathcal{E}, S\subseteq[d]}$ are (β', C')-smooth for some $\beta' > 0$ and $C' > 0$, and $\delta_{\texttt{opt}} = o(1)$. Consider the FAIR-ANN estimator that solves (B.2) with $\ell(y, v) = \frac{1}{2}(y - v)^2$ using $\gamma \geq 8\gamma^\star_{\texttt{AN}}$ with*

$$\gamma^\star_{\texttt{AN}} := \sup_{S\subseteq[d]:\bar{\mu}(\{m^\star \neq A_{S\cup S^\star}(\bar{m}^{(S\cup S^\star)})\})>0} \frac{\|m^\star - A_{S\cup S^\star}(\bar{m}^{(S\cup S^\star)}))\|_2^2}{\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \|m^{(e,S)} - A_S(\bar{m}^{(S)}))\|_{2,e}^2}, \tag{B.12}$$

and function class (B.11) with L, N satisfying $LN \asymp \{n(\log n)^{8\beta^\star-3}\}^{\frac{1}{2(2\beta^\star+1)}}$ and $(\log n)/(N \wedge L) = o(1)$. Then, we have (1) $\gamma^\star_{\texttt{AN}} \leq \gamma^\star_{\texttt{NN}}$, and (2) for n large enough, the following event occurs with probability at least $1 - \widetilde{C}n^{-100}$

$$\sup_{\substack{m^\star = \sum_{j\in S^\star} m_j^\star(x_j) \text{ with } m_j^\star \in \mathcal{H}_{\texttt{HS}}(1,\beta^\star,C^\star) \\ \|m^\star\|_\infty \leq b_m}} \|\widehat{g} - m^\star\|_2 \leq \widetilde{C}\left\{\delta_{\texttt{opt}} + \left(\frac{\log^7 n}{n}\right)^{-\frac{\beta^\star}{2\beta^\star+1}}\right\}, \tag{B.13}$$

where $\widetilde{C}$ is a constant that depends on $(C_1, d, \beta^\star, C^\star, \sigma_y, C_y, b_x, b_m)$ but independent of $\gamma, \delta_{\texttt{opt}}$ and n.

The choice of N, L, and the convergence rate align with FAIR-NN with $\alpha^\star = \beta^\star$. Given the strong structural prior knowledge that the true regression function is additive, FAIR-ANN requires weaker identification condition Condition 16 and also smaller critical threshold of γ. In particular, Condition 16 requires that for any S such that regressing Y on $X_{S\cup S^\star}$ via additive models yields biased estimation, there should be either (1) a shift in conditional moments $m^{(e,S)}$ across different environments, or (2) one of the conditional moments $m^{(e,S)}$ is non-additive. This characteristic is called the "*double identifiable*" property since meeting either of these conditions can consistently estimate $m^\star$. Notably, the critical threshold $\gamma^\star_{\texttt{AN}}$ can be smaller than that of the FAIR-NN estimator. A small γ can be adopted if either the signal of violating the additive structure or the signal of heterogeneity is strong.

B.6 Theoretical Analysis for Linear $\mathcal{G}$

In this section, we apply our result in Theorem 4 to the cases where the target regression function $g^\star$ is linear. As such, we use linear function class $\mathcal{H}_{\texttt{lin}}(d)$ as our predictor function class $\mathcal{G}$. Our theorem suggests that enhancing the potential approximation ability of the discriminator function class $\mathcal{F}$ will result in (1) a stronger condition on invariance, and (2) a weaker identification condition and a reduced choice of critical threshold $\gamma^\star$.

B.6.1 Linear $\mathcal{F}$

We first consider the case where we use linear discriminator function class $\mathcal{F} = \mathcal{H}_{\texttt{lin}}(d)$. We introduce some notations used in linear regression and state some standard regularity conditions used in linear regression and are also imposed in Fan et al. (2023).

Condition 17. *Suppose the following holds:*

(1) The data satisfies Condition 7 with $|\mathcal{E}| \le n^{C_1}$ for some constant C_1.

(2) The covariance matrix $\Sigma^{(e)} = \mathbb{E}[X^{(e)}(X^{(e)})^\top] \in \mathbb{R}^{d\times d}$ in each environment satisfies $\lambda(\Sigma^{(e)}) \ge \kappa_L$ for some constant $\kappa_L > 0$.

(3) Define the pooled covariance matrix $\Sigma := |\mathcal{E}|^{-1}\sum_{e\in\mathcal{E}}\Sigma^{(e)}$. There exists some positive constant C_x, σ_x such that

$$\forall e \in \mathcal{E},\ \forall v \in \mathbb{R}^d \text{ with } \|v\|_2 = 1,\ \forall t \in [0,\infty), \qquad \mathbb{P}\left(|v^\top(\Sigma)^{-1/2}X^{(e)}| \ge t\right) \le C_x e^{-t^2/(2\sigma_x^2)}$$

(4) Condition 8 holds.

Under Condition 17 that the covariance matrices are all positive definite, we can define

$$\beta^{(e,S)} = \operatorname*{argmin}_{\beta\in\mathbb{R}^d:\beta_{S^c}=0} \mathbb{E}[|Y^{(e)} - \beta^\top X^{(e)}|^2]$$

We can state the invariance and identification condition in this case.

Condition 18 (Invariance in Linear $\mathcal{G}$ and Linear $\mathcal{F}$)**.** *There exists some $S^\star \subseteq [d]$ and $\beta^\star \in \mathbb{R}^d$ with $\beta^\star_{(S^\star)^c} = 0$ and $\min_{j\in S^\star}|\beta^\star_j| = \beta_{\min} > 0$ such that*

$$\forall e \in \mathcal{E} \qquad \beta^{(e,S)} = \beta^\star. \tag{B.14}$$

Let $\varepsilon^{(e)} = Y^{(e)} - (\beta^\star)^\top X^{(e)}$, the above invariance equality (B.14) is equivalent to that $X_{S^\star}$ are exogenous across all the environments, that is,

$$\forall e \in \mathcal{E} \qquad \mathbb{E}[\varepsilon^{(e)}X^{(e)}_{S^\star}] = 0$$

Condition 19 (Identification for Linear $\mathcal{G}$ and Linear $\mathcal{F}$)**.** *For any $S \subseteq [d]$ with $\sum_{e\in\mathcal{E}}\mathbb{E}[X^{(e)}_S\varepsilon^{(e)}] \ne 0$, there exists $e, e' \in \mathcal{E}$ such that $\beta^{(e,S)} \ne \beta^{(e',S)}$.*

We are ready to state the result using truncated linear function class with bounded L_2 norm, that is,

$$\mathcal{H}_{\texttt{lin}}(d, B_1, B_2) = \left\{f(x) = \mathrm{Tc}_{B_2}(\beta^\top x) : \beta \in \mathbb{R}^d, \|\Sigma^{1/2}\beta\|_2 \le B_1\right\}.$$

Theorem 8 (Linear $\mathcal{G}$ and Linear $\mathcal{F}$)**.** *Suppose Condition 17–19 hold, and we choose*

$$\mathcal{G} = \mathcal{H}_{\texttt{lin}}(d, C_2, C_2\sqrt{\log n}) \qquad \text{and} \qquad \mathcal{F} = \mathcal{H}_{\texttt{lin}}(d, 2C_2, 2C_2\sqrt{\log n})$$

with some constant $C_2 \geq 2(\sigma_x \vee 1)\max_{e\in\mathcal{E},S\subseteq[d]} \|\Sigma^{1/2}\beta^{(e,S)}\|_2$. *Then, there exists some constant* $\widetilde{C}$ *that only depends on* $(C_1, C_2, \sigma_x, C_x, \sigma_y, C_y)$ *such that the FAIR least squares estimator using the above function class and hyper-parameter* γ *satisfying* $\gamma \geq 8\gamma^\star_{\mathtt{LL}} = 8\sup_{S:\mathsf{b}_{\mathtt{LL}}(S)>0} \mathsf{b}_{\mathtt{LL}}(S)/\bar{\mathsf{d}}_{\mathtt{LL}}(S)$, *where*

$$\begin{aligned}
\mathsf{b}_{\mathtt{LL}}(S) &= \left\| \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \mathbb{E}[X^{(e)}_{S\cup S^\star}\varepsilon^{(e)}] \right\|^2_{(\bar{\Sigma}_{S\cup S^\star})^{-1}} \leq (\kappa_L)^{-1}\left\| \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \mathbb{E}[X^{(e)}_{S}\varepsilon^{(e)}] \right\|^2_2, \\
\bar{\mathsf{d}}_{\mathtt{LL}}(S) &= \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \|\beta^{(e,S)}_S - \beta^{(S)}_\dagger\|^2_{\Sigma^{(e)}_S} \geq \kappa_L \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \|\beta^{(e,S)} - \bar{\beta}^{(S)}\|^2_2
\end{aligned} \tag{B.15}$$

with $\beta^{(S)}_\dagger = (\bar{\Sigma}_S)^{-1}\{\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \mathbb{E}[X^{(e)}_S Y^{(e)}]\}$ *and* $\bar{\beta}^{(S)} = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \beta^{(e,S)}$, *satisfies, with probability at least* $1 - \widetilde{C}n^{-100}$,

$$\forall n \geq 3 \qquad \|\Sigma^{1/2}(\beta_{\widehat{g}} - \beta^\star)\|_2 \leq \widetilde{C}(1+\gamma)\sqrt{\frac{d\log^5(n)}{n}}, \tag{B.16}$$

for $\widehat{g}(x) = \mathrm{Tc}_B(\beta^\top_{\widehat{g}} x)$. *Moreover, if* $d = o((1+\gamma^2)n/(\log^6 n))$, *then for large enough* n, *we further have*

$$\|\Sigma^{1/2}(\beta_{\widehat{g}} - \beta^\star)\|_2 \leq \widetilde{C}\sqrt{\frac{d\log^5(n)}{n}} \tag{B.17}$$

Remark 8. *We present the results using truncated function classes, and there exist poly-*$\log n$ *factors in the non-asymptotic* L_2 *error bounds. These are for technical convenience such that we can directly apply our result Theorem 4 which focuses on uniformly bounded function classes. Indeed, one can use a finer analysis and obtain the* ℓ_2 *error bound*

$$\sqrt{\frac{d+\log n}{n}}$$

using unbounded linear function class.

The obtained results in Theorem 8 align with (up to $\log(n)$ factors) and offer significant enhancements over Theorem 2 & 3 from Fan et al. (2023). Firstly, the "invariance" condition gets relaxed, we only assume that the noise $\varepsilon^{(e)}$ and the true important variables $X^{(e)}_{S^\star}$ are uncorrelated rather than conditional independent across different environments. Meanwhile, the identification condition Condition 19 exactly matches that in Fan et al. (2023) (refer to Condition 5 therein), and the choice of critical threshold $\gamma^\star$ gets reduced as indicated by the inequality in (B.15) and given that $\kappa_L = O(1)$. Such an improvement can be attributed to the term $-\frac{1}{2}\{f^{(e)}\}^2$ in our minimax regularization that stabilizes the objective. To see this, consider β with $\mathrm{supp}(\beta) = S^\star$, the population-level EILLS objective can be written as

$$(\beta-\beta^\star)^\top\Sigma(\beta-\beta^\star) + \gamma\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}(\beta-\beta^\star)^\top_{S^\star}(\Sigma^{(e)}_{S^\star})^2(\beta-\beta^\star)_{S^\star},$$

where a square of the covariance matrix appears in the regularizer. This does not match what it is in the empirical risk part and will make the objective less stable. Meanwhile, the population-level FAIR objective with sup-f in this case is

$$(1+\gamma)(\beta-\beta^\star)^\top\Sigma(\beta-\beta^\star),$$

which the problem of mismatched covariance matrix order disappears.

We've also refined the non-asymptotic L_2 error bounds. On the one hand, we can derive the error bound without further imposing stronger population-level conditions (Condition 7 required by Theorem 3 in Fan et al. (2023)). On the other, the faster ℓ_2 error bound for sufficiently large n remains independent of the hyper-parameter γ we choose. These refinements result from our tighter characterization of the instance-dependent error bounds compared to the ones in Fan et al. (2023); see the discussion on technical novelties in Appendix B.3.

B.6.2 Augmented Linear $\mathcal{F}$

Here we consider the case where the discriminator function class $\mathcal{F}$ is potentially larger than the predictor function class $\mathcal{G}$. We introduce the following notations. We let $[x, y]$ be the concatenation of two vectors $x \in \mathbb{R}^{d_1}$ and $y \in \mathbb{R}^{d_2}$ as a d_1+d_2 dimensional vector. For each $S \subseteq [d]$, we define $\widetilde{X}_S^{(e)} = [X_S^{(e)}, \bar{\phi}(X_S^{(e)})] \in \mathbb{R}^{2|S|}$, $\widetilde{\Sigma}_S^{(e)} = \mathbb{E}[\widetilde{X}_S^{(e)}(\widetilde{X}_S^{(e)})^\top] \in \mathbb{R}^{(2|S|)\times(2|S|)}$ and let $\widetilde{X}^{(e)} = \widetilde{X}_{[d]}^{(e)}$ and $\widetilde{\Sigma}^{(e)} = \widetilde{\Sigma}_{[d]}^{(e)}$. We impose additional regularity conditions due to the incorporation of basis function ϕ.

Condition 20. *There exists some constant $\widetilde{\kappa}_L > 0$ such that $\lambda_{\min}(\widetilde{\Sigma}^{(e)}) \geq \widetilde{\kappa}_L$ for any $e \in \mathcal{E}$. Moreover, define $\widetilde{\Sigma} := |\mathcal{E}|^{-1}\sum_{e\in\mathcal{E}} \widetilde{\Sigma}^{(e)}$. There exists some positive constant $C_{\widetilde{x}}, \sigma_{\widetilde{x}}$ such that*

$$\forall e \in \mathcal{E},\ \forall v \in \mathbb{R}^{2d} \text{ with } \|v\|_2 = 1,\ \forall t \in [0,\infty), \qquad \mathbb{P}\left(|v^\top(\widetilde{\Sigma})^{-1/2}\widetilde{X}^{(e)}| \geq t\right) \leq C_{\widetilde{x}} e^{-t^2/(2\sigma_{\widetilde{x}}^2)}$$

Under Condition 20 such that the covariance matrix for $\widetilde{X}$ are positive definite, we can define

$$\widetilde{\beta}^{(e,S)} = [\breve{\beta}, \breve{\beta}^\phi] \quad \text{with} \quad (\breve{\beta}, \breve{\beta}^\phi) = \underset{(\beta,\beta^\phi)\in(\mathbb{R}^d)^2, \beta_{S^c}=\beta^\phi_{S^c}=0}{\operatorname{argmin}} \mathbb{E}[|Y^{(e)} - \beta^\top X^{(e)} - (\beta^\phi)^\top \bar{\phi}(X^{(e)})|^2],$$

and $\widetilde{\beta}_S^{(e,S)} = [\breve{\beta}_S, \breve{\beta}^\phi_S]$ be a $2|S|$-dimensional vector. The invariance and identification conditions in this case are as follows.

Condition 21 (Invariance in Linear $\mathcal{G}$ and Augmented Linear $\mathcal{F}$)**.** *There exists some $S^\star \subseteq [d]$ and $\beta^\star \in \mathbb{R}^d$ with $\beta^\star_{(S^\star)^c} = 0$ and $\min_{j\in S^\star}|\beta^\star_j| = \beta_{\min} > 0$ such that*

$$\forall e \in \mathcal{E} \qquad \widetilde{\beta}^{(e,S)} = [\beta^\star, 0]. \tag{B.18}$$

Let $\varepsilon^{(e)} = Y^{(e)} - (\beta^\star)^\top X^{(e)}$ be the noise, the above invariance equality (B.14) is equivalent to that both $X_{S^\star}$ and $\bar{\phi}(X_{S^\star})$ are uncorrelated with noise across all the environments, that is,

$$\forall e \in \mathcal{E} \qquad \mathbb{E}[\varepsilon^{(e)} X_{S^\star}^{(e)}] = \mathbb{E}[\varepsilon^{(e)}\bar{\phi}(X_{S^\star}^{(e)})] = 0$$

Condition 22 (Identification for Linear $\mathcal{G}$ and Augmented Linear $\mathcal{F}$)**.** *For any $S \subseteq [d]$ with $\sum_{e\in\mathcal{E}} \mathbb{E}[X_S^{(e)}\varepsilon^{(e)}] \neq 0$, either (1) there exists some $e \in \mathcal{E}$ such that $\widetilde{\beta}^{(e,S)} \neq [\beta^{(e,S)}, 0]$, or (2) there exists $e, e' \in \mathcal{E}$ such that $\beta^{(e,S)} \neq \beta^{(e',S)}$.*

For technical convenience, we also used truncated function class the discriminator class, defined as $\mathcal{H}_{\texttt{alin}}(d,\phi,B) = \{\widetilde{f} = \mathrm{Tc}_B(f) : f \in \mathcal{H}_{\texttt{alin}}\}$.

Theorem 9 (Linear $\mathcal{G}$ and Augmented Linear $\mathcal{F}$)**.** *Suppose Condition 17, 20–22 hold, and we choose*

$$\mathcal{G} = \mathcal{H}_{\texttt{lin}}(d, C_2, C_2\sqrt{\log n}) \qquad \text{and} \qquad \mathcal{F} = \mathcal{H}_{\texttt{alin}}(d, \phi, 2C_2\sqrt{\log n})$$

with some constant $C_2 \geq 2(\sigma_{\widetilde{x}} \vee 1)\max_{e\in\mathcal{E}, S\subseteq[d]} \|\widetilde{\Sigma}^{1/2}\widetilde{\beta}^{(e,S)}\|_2$. Then, there exists some constant $\widetilde{C}$ that only depends on $(C_1, C_2, \sigma_{\widetilde{x}}, C_{\widetilde{x}}, \sigma_y, C_y)$ such that the FAIR least squares estimator using the above function classes and hyper-parameter γ satisfying $\gamma \geq 8\gamma^\star_{\texttt{LA}} = 8\sup_{S:\mathsf{b}_{\texttt{LL}}(S)>0} \mathsf{b}_{\texttt{LL}}(S)/\bar{\mathsf{d}}_{\texttt{LA}}(S)$, where

$$\bar{\mathsf{d}}_{\texttt{LA}}(S) = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \|\widetilde{\beta}_S^{(e,S)} - [\beta_\dagger^{(S)}, 0]\|^2_{\widetilde{\Sigma}_S^{(e)}} \geq \bar{\mathsf{d}}_{\texttt{LL}}(S) \qquad \text{with } \beta_\dagger^{(S)} \text{ defined in Theorem 8}, \tag{B.19}$$

satisfies the L_2 error bound (B.16) with probability at least $1 - \widetilde{C}n^{-100}$. Moreover, if $d = o((1+\gamma^2)n/(\log^6 n))$, for large enough n, the error bound (B.17) also holds with probability at least $1-\widetilde{C}n^{-100}$.

We can see that the proposed estimator utilizes both the heterogeneity among different environments and strong prior knowledge that the true regression function admits linear form to help the identification. It bridges the EILLS estimator in Fan et al. (2023) and the Focused GMM (FGMM) estimator in Fan & Liao (2014) when the instrumental variables are $[X_S, \bar{\phi}(X_S)]$ and hence has some advantages over the individual ones. We illustrate this as follows.

1. When there are multiple environments $|\mathcal{E}| > 1$, the identification condition Condition 22 is weaker to both the EILLS and FGMM estimators. In particular, a consistent estimate $\beta^\star$ is attainable if incorporating variables x_j with $\sum_{e\in\mathcal{E}} \mathbb{E}[X_j^{(e)}\varepsilon^{(e)}] \neq 0$ will result in either (1) a shift in the best linear predictor across environments or (2) the fitted residuals is strongly correlated with some nonlinear basis. We refer to this property as "*double identifiable*" property, given satisfying either condition can lead to the consistent estimation of the true parameter. Furthermore, the critical threshold $\gamma^\star$ can be smaller than that of the EILLS estimator according to the inequality $\bar{\mathsf{d}}_{\mathtt{LA}}(S) \geq \bar{\mathsf{d}}_{\mathtt{LL}}(S)$. This implies that the estimation is sample efficient, which allows for a small γ, if either the signal of nonlinear basis or the signal of heterogeneity is strong.
2. If there is only one environment $|\mathcal{E}| = 1$, it reduces to an estimator similar to the FGMM estimator. Consistent estimation remains feasible in this case but completely impossible for EILLS estimator. Moreover, the identification condition, in this case, resembles and relaxes that in Fan & Liao (2014).

At the same time, it should be noted that the above advantages over the EILLS estimator (linear $\mathcal{F}$) are at the cost of imposing stronger invariance condition Condition 21, which assures that the noise should not only be uncorrelated with $X_j^{(e)}$ but also be uncorrelated with $\phi(X_j^{(e)})$ for any $j \in S^\star$ and $e \in \mathcal{E}$.

B.6.3 Neural Network $\mathcal{F}$

We impose some regularity conditions on the regression function.

Condition 23. *There exists some constant (C_m, σ_m) such that $m^{(e,S)}$ is C_m Lipschitz and $|m^{(e,S)}(0)| \leq C_m$ for any $e \in \mathcal{E}$ and $S \subseteq [d]$ and*

$$\mathbb{P}(|m^{(e,S)}(X_S^{(e)})| \geq t) \leq C_m e^{-t^2/(2\sigma_m^2)} \qquad \forall t \in [0,\infty)$$

In this case, we consider the strongest invariance condition together with the weakest identification when the predictor function class $\mathcal{G}$ is linear.

Condition 24 (Invariance in Linear $\mathcal{G}$ and Neural Network $\mathcal{F}$)**.** *There exists some $S^\star \subseteq [d]$ and $\beta^\star \in \mathbb{R}^d$ with $\beta^\star_{(S^\star)^c} = 0$ and $\min_{j\in S^\star} |\beta_j^\star| = \beta_{\min} > 0$ such that*

$$\forall e \in \mathcal{E} \qquad \mathbb{E}[Y^{(e)}|X_{S^\star}^{(e)}] \equiv (\beta^\star)^\top X^{(e)} \tag{B.20}$$

Condition 25 (Identification for Linear $\mathcal{G}$ and Neural Network $\mathcal{F}$)**.** *For any $S \subseteq [d]$ with $\sum_{e\in\mathcal{E}} \mathbb{E}[X_S^{(e)}\varepsilon^{(e)}] \neq 0$, either (1) there exists some $e \in \mathcal{E}$ such that $\mu^{(e)}(\{m^{(e,S)} \neq X^\top \beta^{(e,S)}\}) > 0$, or (2) there exists $e, e' \in \mathcal{E}$ such that $\beta^{(e,S)} \neq \beta^{(e',S)}$.*

Theorem 10 (Linear $\mathcal{G}$ and Neural Network $\mathcal{F}$)**.** *Suppose Condition 17, 23–25 hold, and we choose the function classes $\mathcal{G} = \mathcal{H}_{\mathtt{lin}}(d, C_2, C_2\sqrt{\log n})$ and $\mathcal{H}_{\mathtt{nn}}(d, \log^d n, \log^d n, C_2\sqrt{\log n})$ with some constant $C_2 \geq (1 \vee \sigma_x \vee \sigma_m) \max_{e\in\mathcal{E}, S\subseteq[d]} \|\Sigma^{1/2}\beta^\star\|_2$. Then, there exists some constant $\widetilde{C}$ that only depends on $(C_1, C_2, d, \sigma_m, C_m, \sigma_y, C_y, \sigma_x, C_x)$ such that the FAIR estimator using the above function classes and hyper-parameter γ satisfying $\gamma \geq 8\gamma^\star_{\mathtt{LN}} = 8\sup_{S:\mathsf{b}_{\mathtt{LL}}(S)>0} \mathsf{b}_{\mathtt{LL}}(S)/\bar{\mathsf{d}}_{\mathtt{LN}}(S)$, where*

$$\bar{\mathsf{d}}_{\mathtt{LN}}(S) = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \|m^{(e,S)} - (\beta_\dagger^{(S)})^\top x_S\|_{2,e} \geq \mathsf{d}_{\mathtt{LA}}(S), \tag{B.21}$$

satisfies, for large enough n,

$$\|\beta_{\widehat{g}} - \beta^\star\|_2 \leq \widetilde{C}(\log^{d+3} n)n^{-1/2}$$

with probability at least $1 - \widetilde{C}n^{-100}$.

The estimator can be viewed as an advanced version of the one using $\mathcal{F} = \mathcal{H}_{\mathtt{alin}}(d, \phi)$. It leverages neural networks to search for appropriate basis function ϕ with strong signals. With the proper choice of the neural network hyper-parameters, the estimator still maintains a parametric optimal rate (up to logarithmic factors). Additionally, it requires a weaker identification condition as described by Condition 25 and reduced critical threshold $\gamma^\star$ according to the inequality $\bar{\mathsf{d}}_{\mathtt{LN}}(S) \geq \bar{\mathsf{d}}_{\mathtt{LA}}(S)$ in Theorem 10.

C Omitted Parts in Experiments

C.1 Pseudo-code of the Gradient Descent Ascent Algorithm

See Algorithm 1.

C.2 Detailed Simulation Configuration

C.2.1 Linear Model with $d = 15$

Data Generating Process. The data-generating process is similar to that described in Section 5.2.1. We also let $|\mathcal{E}| = 2$, and use the same procedure to generate parent-children relationship and structural assignment except that (1) we use $d = 15$ and let the variable Z_8 be Y; and (2) we enforce that Y has at least 3 parents and 3 children (3) the structural assignment for variable Y is

$$Y^{(8)} = Z^{(8)} \leftarrow \sum_{k \in \mathtt{pa}(8)} C_{8,k} Z_k^{(e)} + C_{8,8}\varepsilon_8,$$

that is we let the variance noise to be the same for the two environments. This is because we will include ICP in our simulation comparisons, which requires conditional distribution invariance.

Implementation. We use the same configurations in the implementation of FAIR-GB and FAIR-RF. We also use fixed $\gamma = 36$ for all the FAIR family estimators including EILLS. It is worth noticing that ICP, anchor regression, and IRM introduce an additional hyper-parameter, we pick it in an oracle way for them: that is, we enumerate all the candidate hyper-parameters and select the one that minimizes the L_2 estimation error. We report the performance for $n \in \{100, 200, 500, 800, 1000\}$.

Discussion of Results. For anchor regression and IRM, their performance and the corresponding relationships w.r.t. Pool-LS are similar to the 12 variable illustrations in Fan et al. (2023). The anchor regression is almost the same as Pool-LS because it is essentially the same as standard least squares when the environments are discrete: indeed, in $|\mathcal{E}| = 2$, it just runs least squares with a difference intercept for the interventional environment $e = 1$. The IRM is better than vanilla least squares by slightly decreasing the bias, while the performance improvement is negligible compared with the bias it has.

For ICP, the performance is even worse than pooled least squares because it collapses to conservative solutions like 0. Note that we apply interventions to all the variables in environment $e = 1$, under which it is possible for ICP to identify $\beta^\star$ and $S^\star$ when $n = \infty$. The large estimation error it depicts is due to its inefficiency in estimation.

We can also see that the performance of FAIR-BF and FAIR-RF are similar, demonstrating the effectiveness of our proposed gradient descent ascent algorithm with Gumbel approximation. The performance of FAIR-GF and FAIR-RF is slightly better than EILLS. This is because the FAIR estimator is essentially doing the most efficient pooled least squares when it selects the correct variable.

C.2.2 Nonlinear Model

Data Generating Process. For the structural assignment, we let $\varepsilon_i^{(e)} = \varepsilon_i$ for $i \leq 5$ and $\varepsilon_i^{(e)} = C_{i,i}^{(e)}\varepsilon_i$ where $(\varepsilon_1, \ldots, \varepsilon_{26})$ are independent Uniform($[-1.5, 1.5]$) random variables to let the covariates to be uniformly bounded and $C_{i,i}^{(e)}$ are scalars that are randomly generated in each trial. ε_0 is standard normal distributed that is independent of $(\varepsilon_1, \ldots, \varepsilon_{26})$.

For the assignments for the children of Y, we let $f_{i,0}^{(e)}(u) = C_{i,0}^{(e)} \tanh(u)$, where $C_{i,0}^{(e)}$ are scalars that is randomly sampled from Uniform($[-1.5, 1.5]$) for $e = 0$ and Uniform($[-5, 5]$) for $e = 1$, the noise level $C_{i,i}^{(e)}$ is a scalar generated from Uniform($[1, 1.5]$). For the assignments for other variables X_i with $i \geq 10$, we let $f_{i,j}^{(e)}(u) = C_{i,j}^{(e)} h_{i,j}^{(e)}(u)$ where $h_{i,j}^{(e)}$ are randomly picked from the function set $\{\tanh(x), \sin(x), \cos(x)\}$, the noise level $C_{i,i}^{(e)}$ is a scalar generated from Uniform($[2, 3]$). For $m_1^\star$, it is $\sum_{k=1}^{5} f_{0,j}(x)$ with $f_{0,j}(x)$ randomly picked from $\{\tanh(x), \sin(x), \max(0, x), x\}$.

Algorithm 1 FAIR Gradient Descent Ascent Training

SGD Hyper-parameters: iteration steps T, batch size m, predictor/discriminator iter steps T_g/T_f.
FAIR Hyper-parameters: invariance regularization γ.
Annealing Hyper-parameters: Initial τ_0 and final τ_T.
Models: predictor $g(x;\theta)$, discriminators $\{f^{(e)}(x;\phi^{(e)})\}_{e\in\mathcal{E}}$, gate w.
Input: data $\{\mathcal{D}^{(e)}\}_{e\in\mathcal{E}}$ with $\mathcal{D}^{(e)} = \{(x_i^{(e)}, y_i^{(e)})\}_{i=1}^n$ from $|\mathcal{E}|$ environments, loss function $\ell(\cdot,\cdot)$.
Output: Parameters of the prediction model: w and θ

Initialize $\theta, \{\phi^{(e)}\}_{e\in\mathcal{E}}$ with random weights
Set $w = 0$

for $t \in \{1,\dots,T\}$ **do**
  Set $\tau_t = \tau_0 \times (\tau_T/\tau_0)^{t/T}$
  for $t_f \in \{1,\dots,T_f\}$ **do** ▷ Discriminator Ascent
    Sample $\{u_j\}_{j=1}^d = \{(u_{j,1}, u_{j,2})\}_{j=1}^d$ from Gumbel$(0,1)$.
    Calculate $a = (a_1,\dots,a_d)$ with $a_j = V_{\tau_t}(w_j, u_j)$, where $V(\cdot)$ is defined in (5.2).
    for $e \in \mathcal{E}$ **do** ▷ Update $f^{(e)}$
      Sample minibatch of m examples $\{(x^{(e,i)}, y^{(e,i)})\}_{i=1}^m$ from $\mathcal{D}^{(e)}$.
      Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\phi^{(e)}} \frac{\gamma}{m} \sum_{i=1}^{m} \left[\{y^{(e,i)} - g(x^{(e,i)})\} f_{\phi^{(e)}}(x^{(e,i)}) - \frac{1}{2}\{f_{\phi^{(e)}}(\phi^{(e)})\}^2 \right]$$

      where

$$g(x) = g(a(w)\odot x;\theta) \qquad \text{and} \qquad f_{\phi^{(e)}}(x) = f(a(w)\odot x;\phi^{(e)})$$

    end for
  end for
  for $t_g \in \{1,\dots,T_g\}$ **do** ▷ Predictor Descent
    Sample $\{u_j\}_{j=1}^d = \{(u_{j,1}, u_{j,2})\}_{j=1}^d$ from Gumbel$(0,1)$.
    Calculate $a = (a_1,\dots,a_d)$ with $a_j = V_{\tau_t}(w_j, u_j)$, where $V(\cdot)$ is defined in (5.2).
    for $e \in \mathcal{E}$ **do** ▷ Enumerate Environments
      Sample minibatch of m examples $\{(x^{(e,i)}, y^{(e,i)})\}_{i=1}^m$ from $\mathcal{D}^{(e)}$.
      Calculate loss as function of θ and w, that is

$$L^{(e)}(\theta, w) = \frac{\gamma}{m} \sum_{i=1}^{m} \left[\{y^{(e,i)} - g_{w,\theta}(x^{(e,i)})\} f_w(x^{(e,i)}) - \frac{1}{2}\{f_w(x^{(e,i)})\}^2 \right] + \frac{1}{m} \sum_{i=1}^{m} \left[\ell\left(y^{(e,i)}, g_{w,\theta}(x^{(e,i)})\right) \right]$$

      where

$$g_{w,\theta}(x) = g(a(w)\odot x;\theta) \qquad \text{and} \qquad f_w(x) = f(a(w)\odot x;\phi^{(e)})$$

    end for
    Update the predictor weights w, θ by descending its stochastic gradient:

$$\nabla_{(\theta,w)} \sum_{e\in\mathcal{E}} L^{(e)}(\theta, w)$$

  end for
end for

Implementation. For the FAIR-NN implementation using Gumbel approximation, we also run gradient descent ascent using the Adam optimizer using a learning rate of 1e-3, batch size 64. The number of iterations is $70k$ for $m_1^\star$ and $80k$ for $m_2^\star$. In each iteration, one gradient descent update of the neural network parameters in g and the Gumbel logits parameter w is conducted followed by three gradient ascent updates of the neural network parameters in $f^{(0)}$ and $f^{(1)}$. We also use fixed $\gamma = 36$. The implementation details for the estimators are:

(1) Pool-LS: it simply runs least squares on the full covariate X using all the data.

(2) FAIR-GB: Our FAIR-NN estimator with Gumbel approximation, its prediction on the test dataset is evaluated by averaging the predictions over 100 Gumbel samples.

(3) FAIR-RF: it first selects the variables x_j in the fitted model in (2) with $\text{sig}(w_j) > t$, i.e., $\widehat{S} = \{j : \text{sig}(w_j) > t\}$, and runs least squares again on $X_{\widehat{S}}$ using all the data. Here we let $t = 0.6$ for $n \leq 2000$ and $t = 0.9$ for $n > 2000$.

(4) Oracle: it runs least squares on $X_{S^\star}$ using all the data.

For FAIR-GB, we report the estimated MSE for the model in the last iteration. For other estimators, we also run gradient descent using the Adam optimizer for 10k iterations. We report the estimated MSE for the model with early stopping regularization: that is, we report the estimated MSE of the model that has the smallest validation error, and the validation data is sampled independently and identically to the training data with sample size $n_{\text{valid}} = \lfloor 3n/7 \rfloor$.

C.3 Details of the Discovery in Real Physical System Application

Data Collection We directly use the dataset 'lt_interventions_standard_v1' released in Gamella et al. (2024).

For the training dataset, given fixed sample size n, the data in the first environment $e = 0$ is sampled from the experimental setting 'uniform_reference'. For the second environment $e = 1$, a mixture of interventions is applied. To be specific, a weak intervention on the variables $\widetilde{V}_3, \widetilde{V}_1, \widetilde{V}_2, \widetilde{I}_1, \widetilde{I}_2$ with probability $(1/3, 1/6, 1/6, 1/6, 1/6)$, respectively. This is equivalent to sample data from the experimental setting 't_vis_3_weak', 't_vis_1_weak', 't_vis_2_weak,', 't_ir_1_weak', 't_ir_2_weak' with weights $(1/3, 1/6, 1/6, 1/6, 1/6)$.

For the test data used for evaluation in Fig. 6 (b)–(c), we use the data from the experimental setting 't_vis_3_strong', 't_vis_1_strong', 't_vis_2_strong', 't_ir_1_strong', 't_ir_2_strong'. Since there is an out-of-support issue for the intervention, i.e.,

$$|\text{Mean}_{\mu_{X,i}}(X) - \text{Mean}_{\bar{\mu}_n}(X)| > 1.6 \cdot \text{Std}_{\bar{\mu}_n}(X)$$

where $\mu_{X,i}$ is the empirical distribution of X in the experimental setting where strong intervention is intervened on X, and $\bar{\mu}_n$ is the empirical distribution of X in the training dataset. Thus, we recenter the variable X in the corresponding test intervention environment such that it has the same empirical mean as that in the training dataset.

Explanation on the Equivalent Graph We regress $\widetilde{I}_3$ on $(R, G, B, \theta_1, \theta_2, \widetilde{V}_1, \widetilde{V}_2, \widetilde{V}_3, \widetilde{I}_1, \widetilde{I}_2)$. There are several hidden confounders, hence there should be an arrow from $\widetilde{V}_3$ to $\widetilde{I}_3$ and an arrow from $\widetilde{I}_3$ to $\widetilde{V}_3$ if $\widetilde{V}_3$ is not intervened given the existence of hidden confounders $(L_{3,1}, L_{3,2})$. Introducing the variable $\widetilde{V}_3$ in predicting $\widetilde{I}_3$ can increase the predictive power given it can provide additional information of $(L_{3,1}, L_{3,2})$. The (equivalent) arrow from $\widetilde{V}_3$ to $\widetilde{I}_3$ do disappear because of the intervention on $\widetilde{V}_3$ will make the association perturbs.

Experimental Setup For the FAIR-NN implementation using Gumbel approximation, we also run gradient descent ascent using the Adam optimizer using a learning rate of 1e-3, batch size 64. The number of iterations is $100k$. In each iteration, one gradient descent update of the neural network parameters in g and the Gumbel logits parameter w is conducted followed by three gradient ascent updates of the neural network parameters in $f^{(0)}$ and $f^{(1)}$. We also use fixed $\gamma = 36$. The neural network architectures for all the estimators are the same and are the same as in the simulation of FAIR-NN. The implementation details for all the estimators are:

(1) Pooled-NN: it simply runs least squares on the full covariate X using all the data.

(2) FAIR-NN-GB: Our FAIR-NN estimator with Gumbel approximation, its prediction on the test dataset is evaluated by averaging the predictions over 100 Gumbel samples.

(3) FAIR-NN-RF: it first selects the variables x_j in the fitted model in (2) with $\text{sig}(w_j) > 0.9$, i.e., $\widehat{S} = \{j : \text{sig}(w_j) > t\}$, and runs least squares again on $X_{\widehat{S}}$ using all the data.

(4) Oracle-NN: it runs least squares on $X_{S^\star}$ using all the data and neural networks.

(5) Oracle-Linear: it runs least squares on $X_{S^\star}$ using all the data and linear model.

The out-of-sample R^2 for all the estimators is reported based on the model selection using the validation set that is sampled from the same source as training data with sample size $n' = 0.6n$. Such a model selection is adopted to prevent the model from over-fitting.

C.4 Details of the Prediction Based on Extracted Features

We generate datasets by combining the bird images in the CUB dataset (Wah et al., 2011) and the background images in the Places dataset (Zhou et al., 2017) using specific probabilities, which is similar to the waterbird setting in Sagawa et al. (2020) except the spurious correlation ratio. In each environment, there are 50% water birds and 50% land birds. The probabilities of each environment are as follows:

(a) Environment-1. We place 95% of all water birds against a water background, with the remaining 5% against a land background. We place 90% of all land birds against a land background, with the remaining 10% against a water background. The dataset is denoted by $\mathcal{D}_1$, with 50k images.

(b) Environment-2. We place 75% of all waterbirds against a water background, with the remaining 25% against a land background. We place 70% of all landbirds against a land background, with the remaining 30% against a water background. The dataset is denoted by $\mathcal{D}_2$, with 50k images.

(c) Environment-3 (Test Environment). We only place 2% of all waterbirds against a water background, with the remaining 98% against a land background. We place 2% of all landbirds against a land background, with the remaining 98% against a water background. The dataset is denoted by $\mathcal{D}_3$, with 30k images.

(d) Environment-4 (Oracle Environment). We place 50% of all waterbirds against a water background, with the remaining 50% against a land background. We place 50% of all landbirds against a land background, with the remaining 50% against a water background. The dataset is denoted by $\mathcal{D}_4$, with 30k images.

Class Identification We apply the CUB dataset Wah et al. (2011), which contains images of birds, along with pixel-level segmentation masks for each bird. When generating the dataset, we classify each bird into waterbird if it belongs to the seabird or waterfowl categories (e.g., albatross, auklet, cormorant, frigatebird, fulmar, gull, jaeger, kittiwake, pelican, puffin, tern, gadwall, grebe, mallard, merganser, guillemot, or Pacific loon) and the land birds if it does not belong to the seabird or waterfowl categories.

Image Generation When picking bird images from the CUB dataset, we use the provided pixel-level segmentation masks to crop each bird from its original background. Then we decide which environment they should be placed in and select either a water background like ocean and lake or a land background like bamboo forest and broadleaf forest sourced from the Places dataset Zhou et al. (2017). We randomly select 70% of the images in the CUB dataset as a training set and the remaining 30% as a testing set and generate our dataset for training and testing based on the split CUB dataset.

Feature Extraction Based on the dataset, we use the Pytorch torchvision implementation of the ResNet50 model He et al. (2016) with the pre-trained weights to extract the feature of the images, obtaining a dataset of the feature vector of 2048 dimensions. Then we apply principal components analysis (PCA) to reduce the dimensions of the feature vector to 500 based on the whole training data $\mathcal{D}_1$ and $\mathcal{D}_2$. We apply the same dimensionality reduction transformation to data in other environments.

Experiment Setup We run FAIR-Linear with Gumbel approximation on the dataset. Following the standard setting, we apply the logistic loss and Adam optimizer using a learning rate of $1e-2$, weight decay of $1e-4$, and batch size 4096 for 10000 iterations. In each iteration, one gradient descent update of the neural network parameters in g and the Gumbel logits parameter ω is conducted based on 5 gradient ascent updates of the neural network parameters in f. We fix γ as 200. The implementation details for all the estimators are:

(1) Oracle: it runs logistic regression with ℓ_1 penalty and penalty weight $\alpha = 0.001$ on the oracle environment for 1000 iterations.

(2) Pooled Lasso: it runs logistic regression with ℓ_1 penalty and penalty weight $\alpha = 0.001$ on the Environment-1 and Environment-2 for 1000 iterations.

(3) Lasso on D2: it runs logistic regression with ℓ_1 penalty and penalty weight $\alpha = 0.001$ on the Environment-2 for 1000 iterations.

(4) FAIR-GB: Our FAIR-Linear estimator with Gumbel approximation trained on Environment-1 and Environment-2 for 10000 iterations.

(5) IRM: it runs Invariant Risk Minimization (IRM) trained on Environment-1 and Environment-2 with ℓ_2 regularizer weight 0.001 and penalty weight 100 for 10000 iterations.

(6) GroupDRO: it runs Group Distributionally Robust Optimization (Group-DRO) on Environment-1 and Environment-2 using ResNet50 and $\gamma = 0.1$ for 10000 iterations.

References

Agarwal, A. & Zhang, T. (2022). Minimax regret optimization for robust machine learning under distribution shift. In *Conference on Learning Theory* (pp. 2704–2729).: PMLR.

Anthony, M. & Bartlett, P. L. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press.

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

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2), 1148.

Bartlett, P. L., Harvey, N., Liaw, C., & Mehrabian, A. (2019). Nearly-tight vc-dimension and psuedodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63), 1–17.

Bauer, B. & Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4), 2261–2285.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3), 199–231.

Chernozhukov, V., Newey, W., Singh, R., & Syrgkanis, V. (2020). Adversarial estimation of riesz representers. *arXiv preprint arXiv:2101.00009*.

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

Dikkala, N., Lewis, G., Mackey, L., & Syrgkanis, V. (2020). Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, 33, 12248–12262.

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

Fan, J., Fang, C., Gu, Y., & Zhang, T. (2023). Environment invariant linear least squares. *arXiv preprint arXiv:2303.03092*.

Fan, J. & Gu, Y. (2024). Factor augmented sparse throughput deep relu neural networks for high dimensional regression. *Journal of American Statistical Association*, to appear.

Fan, J., Gu, Y., & Zhou, W.-X. (2022). How do noise tails impact on deep relu networks? *arXiv preprint arXiv:2203.10418*.

Fan, J., Li, R., Zhang, C.-H., & Zou, H. (2020). *Statistical foundations of data science*. Chapman and Hall/CRC.

Fan, J. & Liao, Y. (2014). Endogeneity in high dimensions. *Annals of statistics*, 42(3), 872.

Fan, J. & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5), 849–911.

Farrell, M. H., Liang, T., & Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1), 181–213.

Foster, D. J. & Syrgkanis, V. (2019). Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*.

Gamella, J. L., Peters, J., & Bühlmann, P. (2024). The causal chambers: Real physical systems as a testbed for ai methodology. *arXiv preprint arXiv:2404.11341*.

Gauss, C. F. (1809). *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*. Cambridge University Press; Reissue edition (May 19, 2011).

Geiger, D. & Pearl, J. (1990). On the logic of causal models. In *Machine Intelligence and Pattern Recognition*, volume 9 (pp. 3–14). Elsevier.

Ghassami, A., Salehkaleybar, S., Kiyavash, N., & Zhang, K. (2017). Learning causal structures using regression invariance. *Advances in Neural Information Processing Systems*, 30.

Glymour, M., Pearl, J., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., Schölkopf, B., et al. (2009). Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4), 5.

Györfi, L., Kohler, M., Krzyżak, A., & Walk, H. (2002). *A Distribution-free Theory of Nonparametric Regression*, volume 1. Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Heinze-Deml, C., Peters, J., & Meinshausen, N. (2018). Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2).

Hirshberg, D. A. & Wager, S. (2021). Augmented minimax linear estimation. *The Annals of Statistics*, 49(6), 3206–3227.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., & Schölkopf, B. (2008). Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21.

Hyttinen, A., Eberhardt, F., & Järvisalo, M. (2014). Constraint-based causal discovery: Conflict resolution with answer set programming. In *Conference on Uncertainty in Artificial Intelligence* (pp. 340–349).: AUAI Press.

Hyttinen, A., Hoyer, P. O., Eberhardt, F., & Järvisalo, M. (2013). Discovering cyclic causal models with latent variables: A general sat-based procedure. In *Uncertainty in Artificial Intelligence* (pp. 301).: Citeseer.

Janzing, D., Chaves, R., & Schölkopf, B. (2016). Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference. *New Journal of Physics*, 18(9), 093052.

Janzing, D., Mooij, J., Zhang, K., Lemeire, J., Zscheischler, J., Daniušis, P., Steudel, B., & Schölkopf, B. (2012). Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182, 1–31.

Kamath, P., Tangella, A., Sutherland, D., & Srebro, N. (2021). Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics* (pp. 4069–4077).: PMLR.

Kennedy, E. H., Balakrishnan, S., Robins, J. M., & Wasserman, L. (2024). Minimax rates for heterogeneous causal effect estimation. *The Annals of Statistics*, 52(2), 793–816.

Kohler, M. & Langer, S. (2021). On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4), 2231–2249.

Legendre, A.-M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes [New Methods for the Determination of the Orbits of Comets] (in French)*. Paris: F. Didot.

Liang, T. (2021). How well generative adversarial networks learn distributions. *Journal of Machine Learning Research*, 22(228), 1–41.

Lu, J., Shen, Z., Yang, H., & Zhang, S. (2021). Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5), 5465–5506.

Meinshausen, N. & Bühlmann, P. (2015). Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, 43(4), 1801–1830.

Nelder, J. A. & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3), 370–384.

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

Peters, J., Mooij, J. M., Janzing, D., & Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15, 2009–2053.

Pfister, N., Bühlmann, P., & Peters, J. (2019). Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527), 1264–1276.

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

Popper, K. (2005). *The logic of scientific discovery*. Routledge.

Raskutti, G., J Wainwright, M., & Yu, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of machine learning research*, 13(2).

Richardson, T. (1996). *Feedback models: Interpretation and discovery*. PhD thesis, Ph. D. thesis, Carnegie Mellon.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427), 846–866.

Rojas-Carulla, M., Schölkopf, B., Turner, R., & Peters, J. (2018). Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1), 1309–1342.

Rosenfeld, E., Ravikumar, P., & Risteski, A. (2021). The risks of invariant risk minimization. *International Conference on Learning Representations*.

Rothenhäusler, D., Bühlmann, P., & Meinshausen, N. (2019). Causal dantzig: fast inference in linear structural equation models with hidden variables under additive interventions. *The Annals of Statistics*, 47(3), 1688–1722.

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

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

Sagawa, S., Koh, P. W., Hashimoto, T. B., & Liang, P. (2020). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *International Conference on Learning Representations*.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with relu activation function (with discussion). *The Annals of Statistics*, 48(4), 1875–1921.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., & Jordan, M. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10).

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

Van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2), 614–645.

Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset.

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

Yin, M., Wang, Y., & Blei, D. M. (2021). Optimization-based causal estimation from heterogenous environments. *arXiv preprint arXiv:2109.11990*.

Yuan, M. & Zhou, D.-X. (2016). Minimax optimal rates of estimation in high dimensional additive models. *The Annals of Statistics*, 44(6), 2564–2593.

Zhang, K. & Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)* (pp. 647–655).: AUAI Press.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6), 1452–1464.