The Annals of Statistics
2025, Vol. 53, No. 5, 2230–2257
https://doi.org/10.1214/25-AOS2541

CAUSALITY PURSUIT FROM HETEROGENEOUS ENVIRONMENTS VIA NEURAL ADVERSARIAL INVARIANCE LEARNING

BY YIHONG GU[1,a], CONG FANG[2,c], PETER BÜHLMANN[3,d] AND JIANQING FAN[1,b]

[1]*Department of Operations Research and Financial Engineering, Princeton University,* [a]*yihongg@princeton.edu,* [b]*jqfan@princeton.edu*

[2]*School of Intelligence Science and Technology, Peking University,* [c]*congfang@pku.edu.cn*

[3]*Seminar for Statistics, ETH Zürich,* [d]*buhlmann@stat.math.ethz.ch*

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. The proposed Focused Adversarial Invariant Regularization (FAIR) framework utilizes an innovative minimax optimization approach that drives 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 nonasymptotic 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 a stochastic gradient descent ascent algorithm. The procedures are demonstrated using simulated and real-data examples.

1. Introduction. A fundamental problem in statistics and machine learning is to use collected data to predict the response variable Y based on explanatory covariates $X \in \mathbb{R}^d$. 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 by Legendre (1805) and Gauss (1809). The problem of achieving sample-efficient estimation of m_0 has been extensively studied, and there are many methods that attempt to exploit a low-dimensional structure such as sparsity, low-rankness, or additivity, and develop corresponding optimal methods tailored to this assumed structure (Hastie, Tibshirani and Friedman (2009), Wainwright (2019), Fan et al. (2020)). However, these methods may 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 and Gu (2024)) with no supervision of function structure. This nature endows them with universal applicability across various tasks and data.

Received December 2024; revised May 2025.
MSC2020 subject classifications. Primary 62G05; secondary 62D20.
Key words and phrases. Adversarial estimation, causal discovery, conditional moment restriction, Gumbel approximation, invariance, neural networks.

Despite many celebrated efforts for the efficient estimation of m_0 or its variants like quantile functions, the ultimate goal of statistical learning is to predict on unseen data, elucidate the causal relationships among variables, and guide 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 we 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 attribution and prediction in an unseen environment. 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*", namely, the conditional expectation of the residual for the "potential" interested (causal) $m_\star$ is not zero given all the explanatory variables, that is, $\mathbb{E}[Y - m_\star(X)|X] \neq 0$. Such a problem will lead to a misalignment between m_0 and $m_\star$, that is, $m_0(X) - m_\star(X) \neq 0$. Hence, traditional regression techniques for estimating m_0 will result in an unsatisfactory solution.

Causal inference methods offer remarkable remedies to the curse of endogeneity. Based on the potential outcome (Rubin (1974)) or structural causal model (SCM) (Pearl (2009)), efficient estimation via various regression techniques (Chernozhukov et al. (2018), Athey, Tibshirani and Wager (2019)) is possible. However, all these methods rely on relatively strong assumptions that are often untestable from data. This, in turn, leads to a high risk of severe misspecification of models and assumptions.

This paper proposes an algorithmic remedy for the "curse of endogeneity" taking advantage of data from multiple sources and a high-level invariance principle. Motivated by causal discovery under the SCM framework (Peters, Bühlmann and Meinshausen (2016)), the invariance principle argues that causal relations remain constant across different environments from multiple sources. Leveraging this invariance principle, we propose an algorithmic framework that estimates the most predictive association, which we refer to as *data-driven causality*, that is invariant across diverse environments. Methodologically and in contrast to previous work, our framework is nonparametric and assumption-lean, making it scalable and robust to model misspecification. From a statistical viewpoint, our estimator requires a minimal number of environments and achieves optimal sample complexity. Furthermore, our approach identifies the causal structure in the setting of an SCM under minimal assumptions of heterogeneity across different environments.

1.1. *The canonical model under study.* Consider the following multi-environment regression problem. Let $\mathcal{E}$ be the set of environments. For each environment $e \in \mathcal{E}$, n i.i.d. data $\{(X_i^{(e)}, Y_i^{(e)})\}_{i=1}^n$ are drawn from $\mu^{(e)}$—the joint distribution of $(X^{(e)}, Y^{(e)})$ satisfying

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

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 important 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 formulation is just for expository simplicity; the extension to varying $n^{(e)}$ is straightforward. We refer to the above problem as *nonparametric invariance pursuit*.

Now we temporarily refrain from causal discussions and frame it as a pure statistical estimation problem. We will use a running example in Section 1.3 right after introducing our method to provide a causal interpretation of $S^\star$ and then 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, that is, $\varepsilon^{(e)} \sim F_\varepsilon$ and independent of $X^{(e)}_{S^\star}$, as typically required for causal discovery (Peters, Bühlmann and Meinshausen (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 $m^\star$, that is, $\mathbb{E}[Y^{(e)}|X^{(e)} = x] \neq m^\star(x_{S^\star})$. This mismatch is due to $\mathbb{E}[\varepsilon^{(e)}|X^{(e)}] \neq 0$. Such a "curse of endogeneity" problem is the main challenge we need to address. Including even one of the endogenously spurious variables, for example, X_2 background color in the above example, in the regression function will create an inconsistent estimation of $m^\star$. Thus, it is essential to design an algorithm to eliminate all endogenously 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

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

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 $S_g = \mathrm{supp}(g)$ used by the regression function g, where $\mathcal{F}_{S_g} = \{f \in \mathcal{F} : f(x) = h(x_{S_g}) \text{ for some } h : \mathbb{R}^{|S_g|} \to \mathbb{R}\}$ 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 $f^{(e)} \in \mathcal{F}_{S_g}$,

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

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

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

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

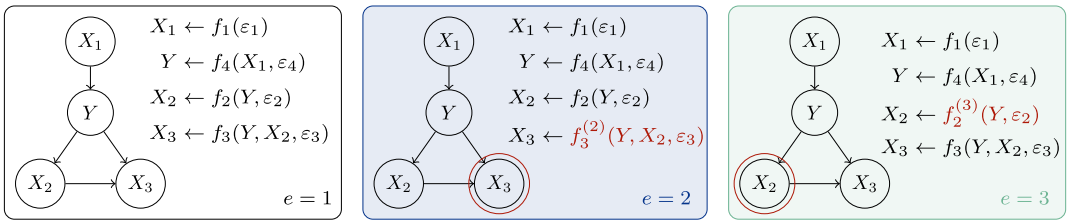


FIG. 1. *The running example when $d = 3$ and $|\mathcal{E}| = 3$. The arrow from node x to y indicates that x affects y directly. The data-generating process of $(X_1, \ldots, X_3, Y)$ in each environment is described by the set of assignments in each panel, and $\varepsilon_1, \ldots, \varepsilon_4$ are independent noises. Compared with the first environment $e = 1$, the assignment for X_3 perturbs in $e = 2$ and the assignment for X_2 perturbs in $e = 3$, which are marked by red.*

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

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

Choosing the multiplier $\lambda = 1/2$ gives rise to the objective function (1.2).

FAIR penalty screens out all the endogenously 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 $g = m^\star$ for large γ. After deleting endogenously spurious variables, we can apply the commonly-used statistical variable selection methods (Hastie, Tibshirani and Friedman (2009), Wainwright (2019), Fan et al. (2020)) to further eliminate exogenously spurious or weak causal variables such as the time and temperature at which the photos were taken.

1.3. *A running example and the roadmap.* Here we use an example to demonstrate the philosophy of our method, namely, to describe the target regression function of our method when data from different environments are observed. Leveraging this example, we also illustrate the key idea of the main theoretical results and offer an overview of this paper.

Let us use the running example with $d = 3$ in Figure 1 to illustrate the causal interpretation of $S^\star$ which our FAIR estimation pursues. The data-generating processes of (X, Y) in the environments are described by the SCMs shown in Figure 1: for example, the data-generating process of the first environment is described by the four assignments in the left panel, where $\{f_j\}_{j=1}^4$ are arbitrary nonparametric functions and $\varepsilon_1, \ldots, \varepsilon_4$ are some independent noises. Here, the presentation of the assignments and the cause-effect relationship is for illustration, our algorithm is blind to this knowledge.

When only data from the first environment is observed, the standard least squares will use all the variables to predict Y if $\{f_j, \varepsilon_j\}_{j=1}^4$ are "*nondegenerate*". This is because besides X_1 (direct cause of Y), both X_2 and X_3 can help predict the noise ε_4, excluding some "degenerate" cases that rarely happen; for example, when $f_2(y, \varepsilon_2) = h(y) + \varepsilon_2$ and $f_3(y, x_2, \varepsilon_3) = h(y) - x_2 + \varepsilon_3 = -\varepsilon_2 + \varepsilon_3$, only X_1 and X_2 contribute to prediction; see a formal definition of the "*nondegenerate*" cases in Section 3. The FAIR estimation will pursue the same $S^\star = \{1, 2, 3\}$ as the invariance constraint trivially holds when $|\mathcal{E}| = 1$.

Things will be different when data from the second environment $e = 2$ becomes accessible. Here, the change of assignment of X_3 (Figure 1 middle panel), or the intervention on

X_3, will make the conditional moment invariance no longer hold for any variable set containing X_3 under nondegenerate cases if such an intervention is also "*nondegenerate*". Therefore, the FAIR method will pursue the *maximum invariant set* $S^\star = \{1, 2\}$. We use the name "maximum invariant set" because it is the most predictive set that preserves the conditional moment invariance constraint. The set $S^\star = \{1, 2\}$ includes both the direct cause of Y (X_1) and the effect of Y (X_2), and this is the best we can get from the currently available data. In this case, all the sets $\{1, 2\}, \{1\}, \{2\}, \varnothing$ preserve the invariant structure. The rule we follow is to pick one from the four candidates based on the following question: given available data from $\mathcal{E} = \{1, 2\}$, what is the best prediction model from a pragmatic perspective? A model including X_3 is not robust—because we have observed the perturbations of association (i.e., the "noninvariance") when X_3 is included in the prediction model, the adversarial effect of X_3 can make the prediction very bad in an unknown future environment. Here, making predictions on $\{X_1, X_2\}$ may be the best choice. This is because if we hold the belief that in the future, the interventions are made within X_3, then the association between $X_{S^\star} = X_{\{1,2\}}$ and Y will be maintained and is the most predictive one among all the maintaining associations. Therefore, the maximum invariant set FAIR estimation pursues can be interpreted as either contemporary direct causes (the candidate direct causes that haven't been falsified) or pragmatic direct causes (for pragmatic considerations in future predictions).

Finally, when we observe additional data from environment $e = 3$ (Figure 1 right panel), the maximum invariant set $S^\star$ will match the exact direct causes X_1 under this model. As a comparison, when $\mathcal{E} = \{1, 2, 3\}$ are observed, the standard least squares, group distributional robust optimization procedure (Meinshausen and Bühlmann (2015), Duchi and Namkoong (2021), Sagawa et al. (2020), Agarwal and Zhang (2022)), and IRM (Arjovsky et al. (2019)) will produce prediction models using all the variables; and the previous hypothesis-test based procedure from nonlinear ICP (Heinze-Deml, Peters and Meinshausen (2018)) will result in the null prediction because $\varnothing$ is also a set maintaining invariant structure.

We also remark that it is possible to recover the direct causes X_1 nontrivially when only one environment is observed. But it is at the cost of imposing additional structural assumptions, for example, assuming f_4 is linear (Fan and Liao (2014)). This can be implemented in our framework by restricting $\mathcal{G}$ within a linear class and choosing a nonparametric $\mathcal{F}$.

Roadmap. The theoretical claims in the paper will extend the above intuitions to arbitrary multivariate cases in a rigorous manner. Section 2 and Section 3 focus on the method and theoretical results for the nonparametric invariance pursuit (1.1). Section 2 considers the pure estimation problem *nonparametric invariance pursuit* itself. Theorem 2.1 shows that under the existence of the maximum invariant set (which is testable), realizing the prediction and testing function class $\mathcal{G}$ and $\mathcal{F}$ by neural networks can allow us to estimate the regression function $m^\star$ induced by the maximum invariant set efficiently in several aspects. Section 3 offers a causal interpretation of the maximum invariant set under extra structural assumptions in SCMs: Theorem 3.1 shows that there always exists a maximum invariant set under nondegenerate cases, it can be represented as pragmatic direct causes in general (Proposition 3.3) and will match the direct causes under sufficient interventions (Proposition 3.2).

The above nonparametric invariance pursuit, as a special instance, 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 the nonasymptotic results are derived as corollaries, along with other applications in Section 4.3. This includes the case that is identifiable using only one environment. 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. A robust prediction of water birds and land birds, similar to the thought experiment, is deferred to Appendix C.3 (Gu et al. (2025)).

1.4. *New contributions*. We propose a unified, algorithmic, and sample-efficient methodological framework that can discover the invariant regression function, that is, 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 and Langer (2021).

We also establish the *first* general causal interpretation of the $S^\star$ in the canonical model (1.1) under SCM with interventions on X setting. Specifically, we demonstrate that under *arbitrary nondegenerate interventions*, there always exists a set—explicitly characterized by the cause-effect relationships and the intervened variables and referred to as "pragmatic direct causes"—that satisfies the aforementioned general identification condition. As a corollary, we establish the general sufficient and necessary condition under which the direct causes of Y can be recovered, while previous recovery results (Peters, Bühlmann and Meinshausen (2016)) only consider special cases. Moreover, even when the interventions are insufficient to identify direct causes, such an $S^\star$ also has implications in robust transfer learning. These results are, to the best of our knowledge, the *first* in the literature and are of independent interest.

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 (Jang, Gu and Poole (2017), Maddison, Mnih and Teh (2017)) 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, Bühlmann and Meinshausen (2016)) and asymmetry principle (Janzing, Chaves and Schölkopf (2016)) for observational data into a unified framework.

1.5. *Related works and comparisons*. Starting from the pioneering work of Peters, Bühlmann and Meinshausen (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, Bühlmann and Meinshausen (2016), Heinze-Deml, Peters and Meinshausen (2018), Pfister, Bühlmann and Peters (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 $\widehat{S} = \varnothing$ due to the inherent worst-case construction in the algorithm. Additionally, the introduction of hypothesis tests also hinders their seamless integration by machine learning algorithms, limiting their scalability. On the other hand, some optimization-based methods (Ghassami et al. (2017), Rothenhäusler, Bühlmann and Meinshausen (2019), Rothenhäusler et al. (2021)) focus on linear $m^\star$ and tackle the problem under additional structures such as linear SCMs with additive interventions (Rothenhäusler, Bühlmann and Meinshausen (2019)). This limitation curtails its applicability to a broader nonparametric setting. Some optimization-based methods (Pfister et al. (2021), Yin, Wang and Blei (2024)) 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. (2024) 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 nonasymptotic guarantees in terms of both $|\mathcal{E}|$ and n. Additionally, it is the first to provide a general sufficient and near-necessary conditions for interventions that enable the exact recovery of direct causes within the SCM framework.

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, Ravikumar and Risteski (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, Glymour and Scheines (2000), Richardson (1996), Chickering (2002), Hyttinen et al. (2013), Hyttinen, Eberhardt and Järvisalo (2014)). However, most of them only attain identification up to Markov equivalent class (Geiger and Pearl (1990)). To overcome the issue, existing methods can be roughly divided into two categories—one based on the invariance principle and the other based on the asymmetry principle. The invariance-based approaches (Peters, Bühlmann and Meinshausen (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 and 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 and Wager (2021)), and so on. We adopted adversarial

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 trick in the implementation jointly relax the combinatorial nature in (1.3) and make a variant of gradient descent continue to work numerically.

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 constants that may depend on the other defined constants.

In the context of the multi-environment setup, 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, that is, $\mathcal{H}_S = \{h \in \mathcal{H}, h(x) \equiv u(x_S)$ 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

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

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 1.1 (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 $\mathcal{H}_{\text{nn}}(d, L, N, B) = \{\widetilde{g}(x) = \text{Tc}_B(g(x)) : g(x)$ in (1.4)$\}$, where $\text{Tc}_B : \mathbb{R} \to \mathbb{R}$ is the truncation operator defined as $\text{Tc}_B(z) = \min\{|z|, B\} \cdot \text{sign}(z)$.

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 invariance pursuit.

The main messages of this section are twofold. From a theoretical perspective, it shows that sample-efficient estimation (in both n and $|\mathcal{E}|$) in the general nonparametric invariance

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-dimensional structures.

2.1. *Setup*. 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 2.1 (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* $t \ge 0$, $\mathbb{P}[|Y^{(e)}| \ge t] \le C_0 e^{-t^2/(2C_0)}$.

(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*: $\forall S \subseteq [d]$ *with* $S^\star \setminus S \neq \varnothing$, $\inf_{m\in\Theta_S} \|m - m^\star\|_2^2 \ge s_{\min} > 0$.

Condition 2.1(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. This condition is not necessary for deriving the L_2 error rate, but it is necessary for the variable selection. The target (invariant) regression function in nonparametric invariance pursuit is $m^\star$.

2.2. *Proposed FAIR-NN least squares estimator*. Given 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

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

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

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

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 function class $\mathcal{F}$ as

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

for neural network architecture hyper-parameters N, L and truncation parameter $B = C_0$. Here B can be larger than C_0 but should satisfy $B = O(1)$. One can also adopt a larger width, depth, and truncation parameter for $\mathcal{F}$. Our choice 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. *Nonasymptotic result for FAIR-NN.*

CONDITION 2.2 (Identification for Nonparametric Invariance 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 2.1 (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 the conditional expectation $m^{(e,S)}$ across environments. In other words, $S^\star$ is the maximum set preserving the invariant associations. This condition is minimal. If it is violated, it would imply that $\exists \widetilde{S} \subseteq [d]$ with $\widetilde{S} \setminus S^\star \neq \varnothing$ such that

$$\forall e \in \mathcal{E}, \quad \mathbb{E}\big[Y^{(e)} | X_{\widetilde{S}}^{(e)}\big] \equiv g\big(X_{\widetilde{S}}^{(e)}\big) \quad \mu^{(e)}\text{-a.s. for some } g : \mathbb{R}^{|S|} \to \mathbb{R},$$

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, Bühlmann and Meinshausen (2016)). These approaches require the shifts of conditional distributions for all the S with $\bar{m}^{(S)} \neq m^\star$ for identifying $S^\star$.

The following theorem provides an oracle-type inequality for the FAIR-NN least squares estimator in a structure-agnostic manner. It shows that under Condition 2.2, one can expect consistent estimation and further establish nonasymptotic upper bounds on the L_2 error between the estimator (2.1) and the invariant regression function $m^\star$. In addition, the theorem quantifies the amount of penalty needed, namely $\gamma_{\mathrm{NN}}^\star$, which is of constant order and is related to the signal-to-noise ratio of the problem.

THEOREM 2.1 (Oracle-type Inequality for FAIR-NN Least Squares Estimator). *Assume Conditions 2.1 and 2.2 hold. Then $\gamma_{\mathrm{NN}}^\star = \sup_{S \subseteq [d]: \mathsf{b}_{\mathrm{NN}}(S) > 0} (\mathsf{b}_{\mathrm{NN}}(S) / \bar{\mathsf{d}}_{\mathrm{NN}}(S)) < \infty$, where*

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

Consider the estimator that solves (2.1) using $\gamma \geq 8\gamma_{\mathrm{NN}}^\star$ 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 $n \geq 3$,

$$\frac{\|\widehat{g} - m^\star\|_2}{\widetilde{C}} \leq \max_{e \in \mathcal{E}} \inf_{h \in \mathcal{G}_{S^\star}} \big\|m^\star - h\big\|_{2,e} + \frac{NL \log^{3/2} n}{\sqrt{n}} + 1_{\{\delta_{\mathrm{NN},1} > \mathsf{s}\}} \cdot (\gamma \delta_{\mathrm{NN},1}) \tag{2.5}$$

occurs with probability at least $1-\widetilde{C}n^{-100}$. *Here* $\delta_{\mathrm{NN},1}=\max_{e\in\mathcal{E},S\subseteq[d]}\inf_{h\in\mathcal{G}_S}\|m^{(e,S)}-h\|_{2,e}+\frac{NL\log^{3/2}n}{\sqrt{n}}$ *and* $\mathsf{s}=\widetilde{C}^{-1}[1\wedge s_{\min}\wedge\{\gamma\inf_{S:\bar{\mathsf{d}}_{\mathrm{NN}}(S)>0}\bar{\mathsf{d}}_{\mathrm{NN}}(S)\}]/(1+\gamma)$, *where* $s_{\min}$ *is defined in Condition* 2.1(*e*). *Moreover, under the above event, if* $\delta_{\mathrm{NN},1}\le\mathsf{s}$, *then the variable selection property holds, for* $\widehat{S}=S_{\widehat{g}}$,

$$S^\star\subseteq\widehat{S}\quad\text{and}\quad\forall e\in\mathcal{E},\quad m^{(e,\widehat{S})}=m^\star. \tag{2.6}$$

REMARK 2.2 (Interpretation of $\mathsf{b}_{\mathrm{NN}}(S)$ and $\bar{\mathsf{d}}_{\mathrm{NN}}(S)$). We refer to $\mathsf{b}_{\mathrm{NN}}(S)$ as bias mean since it exactly characterizes the bias of the least squares estimator in the presence of endogenously spurious variables like the background color in the thought experiment. In particular, letting $\widehat{g}_{\mathrm{LSE}(S)}$ be the least squares estimator that regresses Y on X_S using all the data, namely, the FAIR-NN estimator with $\gamma=0$, Proposition B.1 implies

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

We refer to $\bar{\mathsf{d}}_{\mathrm{NN}}(S)$ 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 $\bar{\mathsf{d}}_{\mathrm{NN}}(S)$ can be viewed as the variance of the bias concerning the uniform distribution on $\mathcal{E}$ since $\bar{\mathsf{d}}_{\mathrm{NN}}(S)=\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\|(m^{(e,S)}-m^\star)-(\bar{m}^{(S)}-m^\star)\|_{2,e}^2$. We have $\bar{\mathsf{d}}_{\mathrm{NN}}(S^\star)=0$ by the invariance structure in Condition 2.1(b).

REMARK 2.3 (Identification). Theorem 2.1 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 2.2 holds, then one can consistently estimate $m^\star$ provided γ is larger than some threshold $8\gamma^\star_{\mathrm{NN}}$ that is independent of n.

Here $\delta_{\mathrm{NN},1}$ can be interpreted as the sum of the worst-case approximation error of neural networks to all the conditional moments $\{m^{(e,S)}\}_{e\in\mathcal{E},S\subseteq[d]}$ and the stochastic error. One can expect $\delta_{\mathrm{NN},1}=o(1)$ if $NL\log^{1.5}n=o(\sqrt{n})$ and all the conditional moments are Lipschitz functions. Moreover, s can be explained as the minimum of the signal of true important variables in $S^\star$ and the signal of heterogeneity. Given s is of constant order and $\delta_{\mathrm{NN},1}=o(1)$, the error bound (2.5) shows that as $\delta_{\mathrm{NN},1}\le\mathsf{s}$, that is, if n is large enough, the L_2 error is composed of the approximation error of neural networks to $m^\star$ and the stochastic error. In this case, all endogenously spurious variables can be surely screened (Fan and Lv (2008)), that is, (2.6), and $m^\star$ can be estimated as well as if the invariant set of variables $S^\star$ is known. At the same time, given our results are nonasymptotic, for a given (not large enough) n, we may not be able to eliminate all endogenously spurious variables as in (2.6). The error rate in this case will be $(\gamma+1)\delta_{\mathrm{NN},1}$ as in (2.5) given that our method may select wrong variables. The error bounds and $\delta_{\mathrm{NN},1}$ will be presented explicitly as (2.7) in Corollary 2.2 when we impose assumptions on the function class.

2.4. *Adapting to the low-dimensional structures algorithmically*. In this section, we present the convergence rate of the FAIR-NN when $m^\star$ lies within the hierarchical composition model (Bauer and Kohler (2019)). This is the function class that neural networks can efficiently estimate (Schmidt-Hieber (2020), Kohler and Langer (2021), Fan and Gu (2024)) with little guidance regarding the forms of functions. We show that FAIR-NN can obtain the same result as standard regression blind to both the knowledge of $S^\star$ and function structure. This example demonstrates our framework's ability to fully leverage the sample efficiency of the adopted machine learning model while also providing a concrete instance that realizes several quantities defined in the structure-agnostic setting of Theorem 2.1.

DEFINITION 2.1 ((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 nonnegative 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.

DEFINITION 2.2 (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 and 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 $l \geq 1$,

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

Following Kohler and 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 (Fan and Gu (2024)) 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 2.3 (Function Complexity). *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*) $\max\{C_0, d, l, C_h, \sup_{(\beta,t)\in\mathcal{O}}(\beta \vee t), \sup_{(\beta,t)\in\mathcal{O}^\star}(\beta \vee t)\} \leq C_1$ *for some constant* $C_1 > 1$.
(*d*) *The neural network architecture hyper-parameters diverge*: $(\log n)/(N \wedge L) = o(1)$.

COROLLARY 2.2 (Convergence Rate for FAIR-NN). *Under the setting of Theorem 2.1, assume further that Condition 2.3 holds. Then, for any* $n \geq 3$, *with probability at least* $1 - \widetilde{C}n^{-100}$, *the following holds*

$$\frac{\|\widehat{g} - m^\star\|_2}{\widetilde{C}\log^{1.5\vee 4\alpha^\star}(n)} \leq (NL)^{-2\alpha^\star} + \frac{NL}{\sqrt{n}} + 1_{\{n<n_0\}\gamma}\underbrace{\left[(NL)^{-2\alpha_0} + \frac{NL}{\sqrt{n}}\right]}_{\delta_{\mathtt{NN},1}}, \tag{2.7}$$

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. *Under the optimal choice of network architecture hyper-parameters* N, L *satisfying* $LN \asymp n^{\frac{1}{2(2\alpha^\star+1)}}$, *the R.H.S. of* (2.7) *is* $n^{-\alpha^\star/(2\alpha^\star+1)} + 1_{\{n<n_0\}}\gamma n^{-\alpha_0/(2\alpha^\star+1)}$

From Corollary 2.2, 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. First, similar to using neural networks in nonparametric regression (Schmidt-Hieber (2020)), 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 depend 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, that is, their adaptivity capability in the regularization part of FAIR.

REMARK 2.4 (Error Guarantees for All n). The error bound (2.7) is applicable for any $n \geq 3$, even when it selects the wrong variables. This is the benefit brought by our proposed regularized least squares, which cannot be easily attained by alternative two-state procedures, such as first running a variable selection procedure similar to ICP and then refitting the model. Furthermore, the error bound will not inflate if the invariant signal $s_{\min}$ and the heterogeneity signal $\inf_{S\subseteq[d]:\bar{\mathsf{d}}_{\mathrm{NN}}(S)>0} \bar{\mathsf{d}}_{\mathrm{NN}}(S)$ is small. Though the error bound scales linearly with γ, the estimator 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 $\mathsf{b}_{\mathrm{NN}}(S)$ to heterogeneity $\bar{\mathsf{d}}_{\mathrm{NN}}(S)$ is controlled.

REMARK 2.5 (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 $\gamma \geq 8\gamma^\star_{\mathrm{NN}}$ 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, if there exists a "maximum invariant set" $S^\star$ satisfying

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

simultaneously, then both $S^\star$ and the induced $m^\star$ can be estimated well as if standard regression by the FAIR-NN estimator. It is natural to ask

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

We offer a general answer to the question under the SCM with arbitrary interventions (on X) setting. The short answer is: Yes, 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 (Pearl, Glymour and Jewell (2016)). See Figure 2 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 3.1 (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 $\{f_1, \ldots, f_p\} = \mathcal{S}$:

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

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, Bühlmann and Meinshausen (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

$$\begin{aligned} X_j^{(e)} &\leftarrow f_j^{(e)}\big(Z^{(e)}_{\mathtt{pa}(j)}, U_j\big), \quad j = 1, \ldots, d \\ Y^{(e)} &\leftarrow f_{d+1}\big(X^{(e)}_{\mathtt{pa}(d+1)}, U_{d+1}\big). \end{aligned} \tag{3.2}$$

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 Figure 2(a). Here we restrict to the case without hidden confounders; see the statement under the presence of hidden confounders in Appendix A.6.

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 setting X_j to be v_j, or a soft intervention that slightly perturbs the association, for example, replacing $X_j \leftarrow 2X_k + U_j$ 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

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

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 Figure 2(b).

We summarize the above construction as a condition.

CONDITION 3.1 (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 being 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. 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 $\mathtt{at}(k) = \mathtt{pa}(k) \cup \bigcup_{j\in\mathtt{pa}(k)} \mathtt{at}(j)$ in the topological order of G. The following condition rules out some degenerate cases.

CONDITION 3.2 (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, $\forall$ Disjoint $A, B, C \subseteq [d+2]$, if $Z_A \perp\!\!\!\perp Z_B | Z_C$, then $Z_A \perp\!\!\!\perp_{\widetilde{G}} Z_B | Z_C$. Here $Z_A \perp\!\!\!\perp_{\widetilde{G}} Z_B | Z_C$ means the node set A and B are d-separated by C in the graph $\widetilde{G}$; see Definition 2.4.1 in Pearl, Glymour and Jewell (2016) for a formal definition of d-separation.*

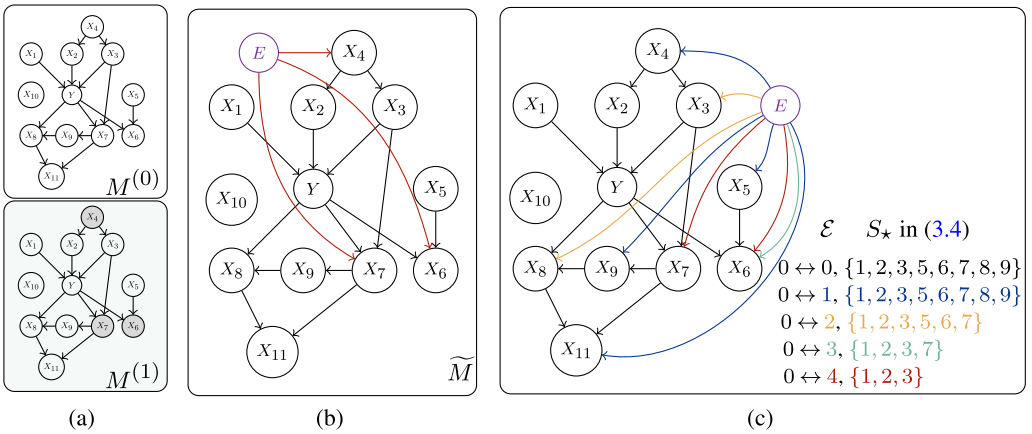


FIG. 2. *(a) is an illustration of the two-environment model, the SCMs in the two environments share the same associated graph: $M^{(0)}$ is an observational environment, and $M^{(1)}$ is an intervention environment where some unknown intervention is applied to (X_4, X_6, X_7), where $M^{(0)}$ and $M^{(1)}$ 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 3.1 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.*

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 3.2 such that the dependence on E in the conditional expectation of Y given X_S with any $S \subseteq [d]$ can be represented by the graph $\widetilde{G}$ itself. The imposed Condition 3.2 rules out the possibility of some degenerate cases; see the justifications for Condition 3.2 and some degenerate examples in Appendix A.4. It should be noted that our general results in Theorem 3.1 and Proposition 3.2 apply to arbitrary forms of interventions under Condition 3.2, which is a mild condition as the violation of faithfulness in Condition 3.2 occurs with probability zero under some suitable measure on the model (Spirtes, Glymour and Scheines (2000)).

THEOREM 3.1 (Existence of Maximum Invariant Set). *Under Condition 3.1, 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 $A(I) = \{j : j \in \mathtt{ch}(d+1), j \notin I, \mathtt{at}(j) \cap \mathtt{ch}(d+1) \cap I = \varnothing\}$, we have $m^{(e,S_\star)} \equiv \bar{m}^{(S_\star)} := m_\star$. Suppose further Condition 3.2 holds, then Condition 2.2 holds with $(S^\star, m^\star) = (S_\star, m_\star)$.

Theorem 3.1 exactly characterizes what $S^\star$ is in our nonparametric invariance 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 2.2 simultaneously. Second, 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. As shown in the definition of $A(I)$, the "unaffected" children include the children of Y unaffected by both direct interventions in I (itself is not included in I) and indirect interventions (it does not have

an ancestor that is both Y's child and suffer from intervention). Theorem 3.1 states explicitly that the pursued set of invariant variables $S^\star$ is the union of parents of Y, unaffected children of Y, and parents of these unaffected children. The size of that set $S^\star$ will keep decreasing when I enlarges. It will finally match the direct causes of Y when I includes "root children set" $I^\star$ as stated in Proposition 3.2 below; see an illustration in Figure 2(c).

PROPOSITION 3.2 (Direct Cause Recovery). *(Sufficiency) Under Condition 3.1, define* $I^\star = \{j : j \in \mathtt{ch}(d+1), \mathtt{at}(j) \cap \mathtt{ch}(d+1) = \varnothing\}$. *If Condition 3.2 holds and* $I \supseteq I^\star$, *then Condition 2.2 holds with* $S^\star = \mathtt{pa}(d+1)$.

(Necessity) Moreover, if $\bar{m}^{(S^\star \cup S)} \neq m^\star$ *for any* S *with* $\mathtt{ch}(d+1) \cap S \neq \varnothing$, *that is, Y does not have degenerate children, then Condition 2.2 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, nondegenerate 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 Figure 2. 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 $I^\star \nsubseteq I$ 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, $S^\star = S_\star$ in this setup can still be interpreted as the "contemporary direct causes" or "pragmatic direct causes" of Y based on the observed environments. We refer to it as "pragmatic direct causes" from the perspective of future prediction. The direct causes of Y have implications in robust transfer learning because the conditional moment of Y given direct causes is the most predictive one among all the transferable associations under the worst case where all the covariates are arbitrarily strongly intervened. The "pragmatic direct causes" can be understood similarly if future interventions are made within the intervened variables X_I. Particularly, if the future interventions are made within the set I, then $S^\star$ can be regarded as the direct causes from a pragmatic perspective since the conditional expectation of Y given $X_{S^\star}$ will remain invariant in a new environment t. Moreover, it depicts the most predictive one among all the associations in the observational environment $e = 0$ that remains in the environment t.

PROPOSITION 3.3 (Robust Transfer Learning). *Under Condition 3.1, for a new environment* t *with SCM* $M^{(t)} = \{\mathcal{S}^{(t)}, \nu\}$ *satisfying* $f_j^{(t)} \equiv f_j^{(0)}$ *for any* $j \in [d+1] \setminus I$, *that is, 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 3.2 holds and* $M^{(t)}$ *satisfies a condition akin to Condition 3.2* (*see Appendix A.5*), *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 losses 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 nonasymptotic 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 and testing functions, respectively. Our goal is to estimate the *invariant regression function* $g^\star \in \Theta_g$ satisfying the invariance structure

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

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 causality 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, that is, estimating $g^\star(x) = (\beta^\star)^\top x = (\beta_{S^\star}^\star)^\top x_{S^\star}$ with $\beta^\star \in \mathbb{R}^d$ satisfying $\operatorname{supp}(\beta^\star) = S^\star$ in the multi-environment linear regression (Fan et al. (2024)) with linear invariance structure

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

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, for all $e \in \mathcal{E}$, $j \in S^\star$,

$$\mathbb{E}\big[(Y^{(e)} - (\beta_{S^\star}^\star)^\top X_{S^\star}^{(e)})X_j^{(e)}\big] = \mathbb{E}\big[(Y^{(e)} - (\beta_{S^\star}^\star)^\top X_{S^\star}^{(e)})\phi(X_j^{(e)})\big] = 0. \tag{4.3}$$

It coincides with the problem considered by Fan and 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 $\mathbb{E}[Y^{(e)}|X_{S^\star}^{(e)}] = (\beta_{S^\star}^\star)^\top X_{S^\star}^{(e)}$, which is much weaker than the assumption $\mathbb{E}[Y^{(e)}|X^{(e)}] = (\beta_{S^\star}^\star)^\top X_{S^\star}^{(e)}$ 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 endogenously spurious variables, making traditional variable selection methods applicable.

REMARK 4.1. We point out here that there are two kinds of spurious variables. One is endogenously spurious variables such as $X_2 =$ background color, and the other is exogenously 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 (Fan and Lv (2008)) the endogenously spurious variables while keeping all the important variables as in (2.6). Exogenously spurious variables can be reduced by using commonly used variable selection methods such as Lasso, SCAD, and best subsets. See Appendix A.7 for how to attain variable selection consistency.

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 key idea is to exploit both the heterogeneity among different environments and the above invariance structure (4.1) to pinpoint the invariant regression function $g^\star$.

Algorithm 1 FAIR Estimation

1: **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. Determine risk loss $\ell(\cdot,\cdot)$.
2: Choose predictor function class $\mathcal{G}$.
3: Choose testing function class $\mathcal{F} = \partial\mathcal{G}$, unless with prior knowledge that the target function $\notin \Theta_f \setminus \partial\Theta_g$.
4: Choose invariance hyper-parameter γ.
5: Solve the minimax program in (4.5).

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, that is, the "invariance principle", and the mentioned prior structure knowledge, that is, 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

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

which is slightly more general than the quasi-likelihood in the generalized linear model (Nelder and 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 $\{\{(X_i^{(e)}, Y_i^{(e)})\}_{i=1}^n\}_{e\in\mathcal{E}}$ 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:

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

γ is the hyper-parameter to be determined, and $\widehat{\mathsf{J}}(g, f^{\mathcal{E}})$ is defined the same as (2.2). We summarize the framework proposed in Algorithm 1.

The difference compared with the standard empirical risk minimizer is outlined in *red*: the choice of testing function can be the default $\mathcal{F} = \partial\mathcal{G}$ in the absence of additional priors. Though one additional hyper-parameter γ is introduced, our theorem and empirical studies show it has no effect when n is large. So we recommend picking a large enough γ like $\gamma = 36$ for the causal discovery task and can use either one additional validation set or leave-one-out cross-validation to optimize the prediction error; see the idea of data-driven determination of

γ in Appendix D.7 of Fan et al. (2024). Our Section 5.1 proposes an efficient implementation of Step 5 if running least squares on $\mathcal{G}$ can be solved by gradient descent, which is quite mild.

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 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 nonasymptotic results in Section 2 can be extended to the general FAIR estimation framework, formally stated in Theorem B.1, 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 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 $\bar{g}^{(S)}(x) = \operatorname{argmin}_{g \in [\Theta_g]_S} \|g - \bar{m}^{(S)}\|_2$ and $f^{(e,S)}(x) = \operatorname{argmin}_{f \in [\Theta_f]_S} \|f - m^{(e,S)}\|_{2,e}$. Then 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 2.2, 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 2.2 is

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

It requires that whenever incorporating more variables in S leads to better prediction performance, the set S will not satisfy the invariance structure (4.1). Condition 2.2 instantiates (4.8) by letting $\bar{\mathsf{d}}(S) = \bar{\mathsf{d}}_{\mathrm{NN}}(S)$ and $\mathsf{b}(S) = \mathsf{b}_{\mathrm{NN}}(S)$ with $(\mathsf{b}_{\mathrm{NN}}(S), \bar{\mathsf{d}}_{\mathrm{NN}}(S))$ defined in (2.4).

THEOREM 4.1 (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* $\min_{S: S^\star \setminus S \neq \varnothing} \inf_{g \in [\Theta_g]_S} \|g - g^\star\|_2$, *and* C_1 *is a universal constant independent of the two quantities.*

The complete and rigorous statement is deferred to Theorem B.1 in Appendix B.1, with more loss function ℓ in Theorem B.2. 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 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

TABLE 1

Applications of Theorem 4.1. 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

Θ_g	Θ_f	$\mathcal{G}$	$\mathcal{F}$	Priors	$\lvert\mathcal{E}\rvert = 1$ Ident	Result
Linear	Linear	Linear	Linear	None	Impossible	Thm B.5
Linear	Linear w/ ϕ	Linear	Linear w/ ϕ	Nearly Linear	Possible	Thm B.6
Linear	Θ	Linear	NN	Linear	Possible	Thm B.7
Additive	Θ	Additive NN	NN	Additive	Impossible	Thm B.4
Θ	Θ	NN	NN	None	Impossible	Thm 2.1

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 nonasymptotic 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" that enforces $f^{(e)} \in \mathcal{F}_{S_g}$. Here we consider a re-parameterization trick that disentangles the function g and 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/absence of variables. Then we can write the objective (4.5) as

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

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:

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

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

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

for which its gradient can be taken. Let $A_\tau(U,w)=(V_\tau(U_1,w_1),\dots,V_\tau(U_d,w_d))^\top\in\mathbb{R}^d$ with $\{U_j\}_{j=1}^d$ being i.i.d. uniform random variables. One can thus approximate (5.1) by

$$(\widehat{\theta},\widehat{w})\in \operatorname*{argmin}_{\theta\in\mathbb{R}^{N_g},w\in\mathbb{R}^d}\sup_{\forall e\in\mathcal{E},\phi^{(e)}\in\mathbb{R}^{N_f}} \mathbb{E}_U\big[\widehat{\mathsf{L}}\big(A_\tau(U,w),\theta,\{\phi^{(e)}\}_{e\in\mathcal{E}}\big)\big], \tag{5.3}$$

with $\widehat{\mathsf{L}}(A,\theta,\{\phi^{(e)}\}_{e\in\mathcal{E}})]=\widehat{\mathsf{R}}(g(A\odot\cdot;\theta))+\gamma\widehat{\mathsf{J}}(g(A\odot\cdot),f^{\mathcal{E}}(A\odot\cdot;\{\phi^{(e)}\}_{e\in\mathcal{E}}))$, where parametrizations of $g\in\mathcal{G}$ and $f^e\in\mathcal{F}$ are used. Since $\operatorname{logit}(U_j)\overset{d}{=}U_{j,1}-U_{j,2}$ 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 Gumbel approximation.

One can use similar implementation tricks widely used in stochastic gradient descent with Gumbel approximation that gradually anneals the Gumbel approximation hyperparameter τ; see the pseudo-code in Appendix C.1. We include the simulation for linear models and applications of causal discovery in the main text and defer the simulation for FAIR-NN estimator to Appendix C.2 and robust prediction of water/land birds to Appendix C.3.

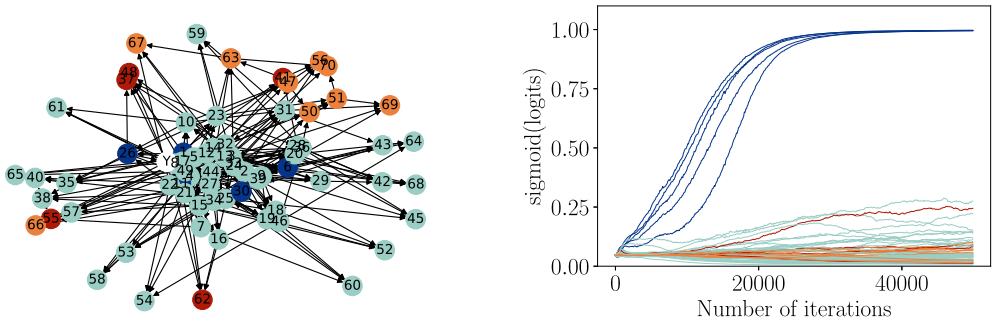
5.2. *Simulations for FAIR-linear estimator.* In this section, we present the simulation result for the FAIR-Linear estimator implemented by the Gumbel approximation trick and gradient descent ascent algorithm.

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 $\{Z_1,\dots,Z_{i-1}\}$, 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 $(Z_1,\dots,Z_{35},Z_{36},Z_{37},\dots,Z_{71})=(X_1,\dots,X_{35},Y,X_{36},\dots,X_{70})$. 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

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

where $(\varepsilon_1,\dots,\varepsilon_{71})$ are independent standard normal distributed. 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 $m^\star(x)=\sum_{k\in\mathtt{pa}(36)}C_{36,k}^{(e)}x_k$ in pursuit. In this case, we let $S^\star=\mathtt{pa}(36)$ 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, that is, $C_{36,36}^{(0)}\neq C_{36,36}^{(1)}$. Now, the model only has conditional expectation invariance rather than the full conditional distribution invariance. Figure 3(a) visualizes the induced graph in one trial. The complex cause-effect relationships in high-dimensional variables make it very challenging to estimate $\beta^\star$.

Implementation. For the FAIR-Linear estimator, we realize $\mathcal{G}$ and $\mathcal{F}$ by linear function classes, that is, $\mathcal{G}=\{g(x)=\beta_g^\top x:\beta_g\in\mathbb{R}^d\}$ and $\mathcal{F}=\{f(x)=\beta_f^\top x:\beta_f\in\mathbb{R}^d\}$, and run



(a) (b)

FIG. 3. *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*.

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 trick that outputs $\beta_g \odot \mathrm{sig}(w)$.
(3) FAIR-RF: it selects the variables x_j with $\mathrm{sig}(w_j) > 0.9$ of the fitted model in (2), that is, $\widehat{S} = \{j : \mathrm{sig}(w_j) > 0.9\}$, 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.
(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. It is unbiased yet has a larger variance compared with the Oracle one.

Figure 3(b) visualizes how the Gumbel gate values for different covariates $\mathrm{sig}(w)$ 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.

Results. The results are shown in Figure 4(a). We can see that the square of the ℓ_2 estimation error $\|\widehat{\beta} - \beta^*\|_2^2$ for the pooled least squares estimator (×) 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 (♦) decays as n grows (≈ 0.01 when $n = 1k$) and lies in between that for least squares on X_{G^c} (Semi-Oracle ▼) and least squares on X_{S^*} (Oracle ▲). This is expected to happen since the FAIR-Linear estimator is not designed to screen out all the exogenously spurious variables: They can be further regularized using the commonly used variable selection techniques; see Remark 4.1. We also observe that the training dynamics of adversarial estimation are highly nonstable: 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 ($n = 5000$).

We also compare our FAIR-Linear estimator with the cousin estimator EILLS (►) in Fan et al. (2024) and other invariance learning estimators (dotted lines), including invariant causal prediction (Peters, Bühlmann and Meinshausen (2016)) (ICP▼), invariant risk minimization (Arjovsky et al. (2019)) (IRM+), anchor regression (Rothenhäusler et al. (2021)) (Anchor●) in a similar but smaller dimension setting with $d = 15$, under which ICP and EILLS can be

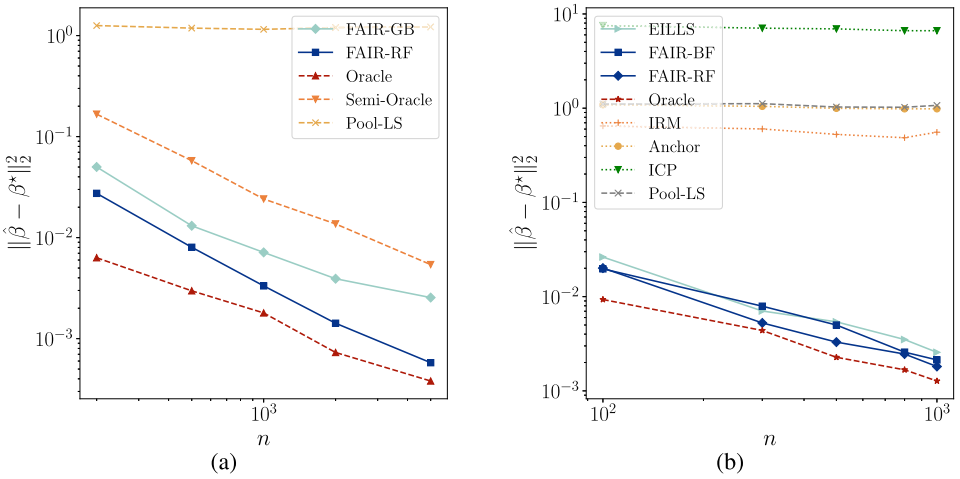


FIG. 4. *The simulation results for linear models with* (*a*) $d = 70$ *and* (*b*) $d = 15$. *Both figures depict how the median estimation errors* (*based on* 50 *replications*, *shown in log scale*) *for different estimators* (*marked with different shapes and colors*) *change when* n *varies in* (*a*) {200, 500, 1000, 2000, 5000} *and* (*b*) {100, 200, 500, 800, 1000}, *respectively.*

computed within affordable time. For the FAIR-Linear estimator, we report the performance of the FAIR-RF (♦) and the one with brute force search (FAIR-BF■). The results are shown in Figure 4(b): we can see that the FAIR family estimators (▶■♦ 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.4.1.

5.3. *Application I*: *Discovery in real physical systems*. We apply our method to perform causal discovery in the light tunnel datasets from Gamella, Peters and Bühlmann (2025). 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 are known such that we can get access to the ground-truth cause-effect relationship; see Figure 2(d) and Figure 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 and the knowledge of ground-truth cause-effect knowledge make it an excellent testbed for causal discovery algorithms.

Following the notations, 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 Figure 5(a). In this case, the variables $(R, G, B, \theta_1, \theta_2)$ are the direct causes, that is, $S^\star = \{1, 2, 3, 4, 5\}$, $\widetilde{V}_3$ are

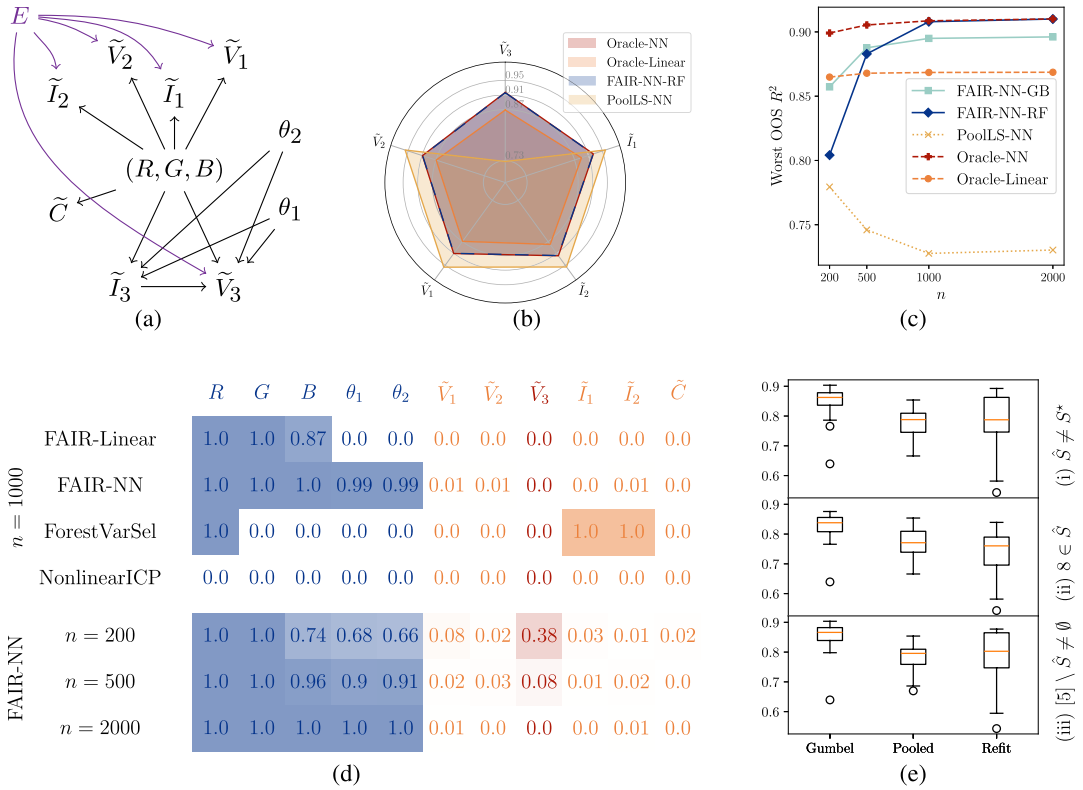


FIG. 5. *Discovery in Real Physical Systems*: (*a*) *the unified cause-effect relationship and interventions similar to Figure* 2(*b*). (*b*) *the average out-of-sample* R^2 *for different estimators using the spider chart*: *the axis annotated by placeholder variable* Z *corresponds to the test environment 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 descendant*. (*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* $(X_1, \ldots, X_5) = (R, G, B, \theta_1, \theta_2)$.

the spurious variables that will lead to biased estimation. The remaining variables are exogenous but have marginal predictive power, that is, $\mathrm{Var}[Y|X_j] > 0$ 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 exogenously/endogenously spurious variable Z is evaluated using the OOS R^2 in corresponding $\mathcal{D}_{2,Z}$ defined as

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

See the detailed data collection and experimental configuration in Appendix C.5.

The first four rows in Figure 5(d) report the variable selection result for several methods over 100 trials. The nonlinear ICP (Heinze-Deml, Peters and Meinshausen (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, Peters and Meinshausen (2018) cannot detect $(G, B, \theta_1, \theta_2)$ and falsely select $(\widetilde{I}_1, \widetilde{I}_2)$. The last three rows in Figure 5(d) illustrate how the variable selection rate for the FAIR-NN estimator changes when n grows.

Figure 5(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 $M \in \{\text{Linear, NN}\}$ 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 identically to that for Oracle-NN. This quantitative result illustrates its capability to correct nontrivial and strong biases without supervision and its efficiency in detecting nonlinear and weak signals.

Figure 5(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 Figure 5(d) is not because of truncating weak but nonnegligible spurious signals. Moreover, as shown in Figure 5(e), its performance significantly outperforms the least squares estimator using either the full covariate or the selected covariates when $n = 200$ and it selects the wrong variables. This further supports the theoretical claims and the advantages of adopting penalized least squares.

Acknowledgments. The authors would like to thank the anonymous referees, an Associate Editor, and the Editor for their constructive comments that improved the quality and accessibility of this paper. 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 for robust prediction application, and Xinwei Shen for drawing our attention to Gumbel approximation in the implementation.

Funding. J. Fan was supported by ONR grant N00014-25-1-2317 and NSF Grants DMS-2053832, DMS-2210833, and DMS-2412029.

Y. Gu was supported by the Charlotte Elizabeth Procter Honorific Fellowship from Princeton University.

C. Fang was supported by the NSF China No.s 92470117.

SUPPLEMENTARY MATERIAL

Supplement to "Causality pursuit from heterogeneous environments via neural adversarial invariance learning" (DOI: 10.1214/25-AOS2541SUPPA; .pdf). This supplementary material collects further theoretical results, discussions, experimental results, and all the technical proofs.

"FAIR-code.zip" (DOI: 10.1214/25-AOS2541SUPPB; .zip). It contains code and instructions to reproduce all the numerical results. We also offer a unified implementation such that the practitioner can easily customize FAIR estimators by specifying function classes. See also in https://github.com/wmyw96/FAIR.

REFERENCES

AGARWAL, A. and ZHANG, T. (2022). Minimax regret optimization for robust machine learning under distribution shift. In *Conference on Learning Theory* 2704–2729. PMLR.

ARJOVSKY, M., BOTTOU, L., GULRAJANI, I. and LOPEZ-PAZ, D. (2019). Invariant risk minimization. arXiv preprint. Available at arXiv:1907.02893.

ATHEY, S., TIBSHIRANI, J. and WAGER, S. (2019). Generalized random forests. *Ann. Statist.* **47** 1148–1178. MR3909963 https://doi.org/10.1214/18-AOS1709

BAUER, B. and KOHLER, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Ann. Statist.* **47** 2261–2285. MR3953451 https://doi.org/10.1214/18-AOS1747

BREIMAN, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statist. Sci.* **16** 199–231. MR1874152 https://doi.org/10.1214/ss/1009213726

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

CHERNOZHUKOV, V., NEWEY, W., SINGH, R. and SYRGKANIS, V. (2020). Adversarial estimation of Riesz representers. arXiv preprint. Available at arXiv:2101.00009.

CHICKERING, D. M. (2002). Optimal structure identification with greedy search. *J. Mach. Learn. Res.* **3** 507–554.

DIKKALA, N., LEWIS, G., MACKEY, L. and SYRGKANIS, V. (2020). Minimax estimation of conditional moment models. *Adv. Neural Inf. Process. Syst.* **33** 12248–12262.

DUCHI, J. C. and NAMKOONG, H. (2021). Learning models with uniform performance via distributionally robust optimization. *Ann. Statist.* **49** 1378–1406. MR4298868 https://doi.org/10.1214/20-aos2004

FAN, J., FANG, C., GU, Y. and ZHANG, T. (2024). Environment invariant linear least squares. *Ann. Statist.* **52** 2268–2292. MR4829488 https://doi.org/10.1214/24-aos2435

FAN, J. and GU, Y. (2024). Factor augmented sparse throughput deep ReLU neural networks for high dimensional regression. *J. Amer. Statist. Assoc.* **119** 2680–2694. MR4833907 https://doi.org/10.1080/01621459.2023.2271605

FAN, J., LI, R., ZHANG, C.-H. and ZOU, H. (2020). *Statistical Foundations of Data Science*. CRC Press/CRC, Boca Raton.

FAN, J. and LIAO, Y. (2014). Endogeneity in high dimensions. *Ann. Statist.* **42** 872–917. MR3210990 https://doi.org/10.1214/13-AOS1202

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322 https://doi.org/10.1111/j.1467-9868.2008.00674.x

GAMELLA, J. L., PETERS, J. and BÜHLMANN, P. (2025). The causal chambers: Real physical systems as a testbed for AI methodology. *Nat. Mach. Intell.* **7** 107–118.

GAUSS, C. F. (1809). *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*. Cambridge University Press.

GEIGER, D. and PEARL, J. (1990). On the logic of causal models. In *Uncertainty in Artificial Intelligence*, 4. *Mach. Intelligence Pattern Recogn.* **9** 3–14. North-Holland, Amsterdam. MR1166825 https://doi.org/10.1016/B978-0-444-88650-7.50006-8

GHASSAMI, A., SALEHKALEYBAR, S., KIYAVASH, N. and ZHANG, K. (2017). Learning causal structures using regression invariance. *Adv. Neural Inf. Process. Syst.* **30**.

GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A. and BENGIO, Y. (2014). Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **27**.

GU, Y., FANG, C., BÜHLMANN, P. and FAN, J. (2025). Supplement to "Causality Pursuit from Heterogeneous Environments via Neural Adversarial Invariance Learning." https://doi.org/10.1214/25-AOS2541SUPPA, https://doi.org/10.1214/25-AOS2541SUPPB

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294 https://doi.org/10.1007/978-0-387-84858-7

HEINZE-DEML, C., PETERS, J. and MEINSHAUSEN, N. (2018). Invariant causal prediction for nonlinear models. *J. Causal Inference* **6** Art. No. 20170016. MR4335430 https://doi.org/10.1515/jci-2017-0016

HIRSHBERG, D. A. and WAGER, S. (2021). Augmented minimax linear estimation. *Ann. Statist.* **49** 3206–3227. MR4352528 https://doi.org/10.1214/21-aos2080

HOYER, P., JANZING, D., MOOIJ, J. M., PETERS, J. and SCHÖLKOPF, B. (2008). Nonlinear causal discovery with additive noise models. *Adv. Neural Inf. Process. Syst.* **21**.

HYTTINEN, A., EBERHARDT, F. and JÄRVISALO, M. (2014). Constraint-based causal discovery: Conflict resolution with answer set programming. In *Conference on Uncertainty in Artificial Intelligence* 340–349. AUAI Press.

HYTTINEN, A., HOYER, P. O., EBERHARDT, F. and JÄRVISALO, M. (2013). Discovering cyclic causal models with latent variables: A general SAT-based procedure. In *Uncertainty in Artificial Intelligence* 301. Citeseer.

JANG, E., GU, S. and POOLE, B. (2017). Categorical reparameterization with Gumbel-softmax. In *International Conference on Learning Representations*.

JANZING, D., CHAVES, R. and SCHÖLKOPF, B. (2016). Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference. *New J. Phys.* **18** 093052.

JANZING, D., MOOIJ, J., ZHANG, K., LEMEIRE, J., ZSCHEISCHLER, J., DANIUŠIS, P., STEUDEL, B. and SCHÖLKOPF, B. (2012). Information-geometric approach to inferring causal directions. *Artificial Intelligence* **182/183** 1–31. MR2909089 https://doi.org/10.1016/j.artint.2012.01.002

KAMATH, P., TANGELLA, A., SUTHERLAND, D. and SREBRO, N. (2021). Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics* 4069–4077. PMLR.

KOHLER, M. and LANGER, S. (2021). On the rate of convergence of fully connected deep neural network regression estimates. *Ann. Statist.* **49** 2231–2249. MR4319248 https://doi.org/10.1214/20-aos2034

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*). F. Didot, Paris.

LIANG, T. (2021). How well generative adversarial networks learn distributions. *J. Mach. Learn. Res.* **22** Paper No. 228. MR4329807

MADDISON, C. J., MNIH, A. and TEH, Y. W. (2017). The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*.

MEINSHAUSEN, N. and BÜHLMANN, P. (2015). Maximin effects in inhomogeneous large-scale data. *Ann. Statist.* **43** 1801–1830. MR3357879 https://doi.org/10.1214/15-AOS1325

NELDER, J. A. and WEDDERBURN, R. W. (1972). Generalized linear models. *J. Roy. Statist. Soc., Ser. A, Statist. Soc.* **135** 370–384.

PEARL, J. (2009). *Causality*: *Models*, *Reasoning*, *and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166 https://doi.org/10.1017/CBO9780511803161

PEARL, J., GLYMOUR, M. and JEWELL, N. P. (2016). *Causal Inference in Statistics*: *A Primer*. Wiley, Chichester. MR3497861

PETERS, J., BÜHLMANN, P. and MEINSHAUSEN, N. (2016). Causal inference by using invariant prediction: Identification and confidence intervals. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 947–1012. MR3557186 https://doi.org/10.1111/rssb.12167

PETERS, J., MOOIJ, J. M., JANZING, D. and SCHÖLKOPF, B. (2014). Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.* **15** 2009–2053. MR3231600

PFISTER, N., BÜHLMANN, P. and PETERS, J. (2019). Invariant causal prediction for sequential data. *J. Amer. Statist. Assoc.* **114** 1264–1276. MR4011778 https://doi.org/10.1080/01621459.2018.1491403

PFISTER, N., WILLIAMS, E. G., PETERS, J., AEBERSOLD, R. and BÜHLMANN, P. (2021). Stabilizing variable selection and regression. *Ann. Appl. Stat.* **15** 1220–1246. MR4317406 https://doi.org/10.1214/21-aoas1487

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

ROSENFELD, E., RAVIKUMAR, P. and RISTESKI, A. (2021). The risks of invariant risk minimization. *Int. Conf. Learn. Represent.*.

ROTHENHÄUSLER, D., BÜHLMANN, P. and MEINSHAUSEN, N. (2019). Causal Dantzig: Fast inference in linear structural equation models with hidden variables under additive interventions. *Ann. Statist.* **47** 1688–1722. MR3911127 https://doi.org/10.1214/18-AOS1732

ROTHENHÄUSLER, D., MEINSHAUSEN, N., BÜHLMANN, P. and PETERS, J. (2021). Anchor regression: Heterogeneous data meet causality. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 215–246. MR4250274 https://doi.org/10.1111/rssb.12398

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688.

SAGAWA, S., KOH, P. W., HASHIMOTO, T. B. and LIANG, P. (2020). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *Int. Conf. Learn. Represent.*.

SCHMIDT-HIEBER, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function (with discussion). *Ann. Statist.* **48** 1875–1921. MR4134774 https://doi.org/10.1214/19-AOS1875

SHIMIZU, S., HOYER, P. O., HYVÄRINEN, A. and KERMINEN, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.* **7** 2003–2030. MR2274431

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation, Prediction, and Search*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR1815675

WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics: A Non-asymptotic Viewpoint. Cambridge Series in Statistical and Probabilistic Mathematics* **48**. Cambridge Univ. Press, Cambridge. MR3967104 https://doi.org/10.1017/9781108627771

YIN, M., WANG, Y. and BLEI, D. M. (2024). Optimization-based causal estimation from heterogeneous environments. *J. Mach. Learn. Res.* **25** Paper No. 168. MR4777410

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