DARING: Differentiable Causal Discovery with Residual Independence

Yue He
Tsinghua University
heyue18@mails.tsinghua.edu.cn

Peng Cui
Tsinghua University
cuip@tsinghua.edu.cn

Zheyan Shen
Tsinghua University
shenzy17@mails.tsinghua.edu.cn

Renzhe Xu
Tsinghua University
xrz199721@gmail.com

Furui Liu
Noah's Ark Lab, Huawei
liufurui2@huawei.com

Yong Jiang
Tsinghua University
jiangy@sz.tsinghua.edu.cn

ABSTRACT

Discovering causal structure among a set of variables is a crucial task in various scientific and industrial scenarios. Given finite i.i.d. samples from a joint distribution, causal discovery is a challenging combinatorial problem in nature. The recent development in functional causal models, especially the NOTEARS provides a differentiable optimization framework for causal discovery. They formulate the structure learning problem as a task of maximum likelihood estimation over observational data (i.e., variable reconstruction) with specified structural constraints such as acyclicity and sparsity. Despite its success in terms of scalability, we find that optimizing the objectives of these differentiable methods is not always consistent with the correctness of learned causal graph especially when the variables carry heterogeneous noises (i.e., different noise types and noise variances) in real data from wild environments. In this paper, we provide the justification that their proneness to erroneous structures is mainly caused by the over-reconstruction problem, i.e., the noises of variables are absorbed into the variable reconstruction process, leading to the dependency among variable reconstruction residuals, and thus raise structure identifiability problems according to FCM theories. To remedy this, we propose a novel differentiable method DARING by imposing explicit residual independence constraint in an adversarial way. Extensive experimental results on both simulation and real data show that our proposed method is insensitive to the heterogeneity of external noise, and thus can significantly improve the causal discovery performances.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**;

KEYWORDS

DARING, Causal Discovery, Mutual Independence, Adversarial Learning

KDD '21, August 14–18, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8332-5/21/08…$15.00
https://doi.org/10.1145/3447548.3467439

ACM Reference Format:
Yue He, Peng Cui, Zheyan Shen, Renzhe Xu, Furui Liu, and Yong Jiang. 2021. DARING: Differentiable Causal Discovery with Residual Independence. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3447548.3467439

1 INTRODUCTION

Causal discovery is a fundamental problem in machine learning, aiming to understand the causal mechanism of data generation. The learned causal graph, in form of a direited acyclic graph (DAG), plays an important role in many areas such as biology [18], healthcare [31] and economics [16], with itself contained in the study of algorithmic interpretability [15], stability [11], and fairness [10, 28].

Interventional experiments by randomized controlled trial is a golden rule of causal discovery, but they are often costly and even impossible in practice. A more realistic and attractive setting is to learn from observational data, where conditional independence criteria is the standard for assessment. Constraint-based algorithms [26, 27, 32] directly conduct independence tests to detect causal skeleton and determine the edge orientation on pruned search space with elaborately designed strategy. Some score-based algorithms [2, 9] adopt score functions that are consistent with the conditional independence statistics to increase the fitness of the target graph with finite data. However, these methods can only find the Markov equivalence class [8] under wild faithfulness assumption.

By virtue of additional hypothesis on structural equation and data distribution, functional causal models (FCMs) could identify the true causal structure from equivalence class, typically including linear Structural Equation Models (SEMs) [25], Additive Noise Models (ANMs) [23] and Post-nonlinear Causal Models (PNLs) [33]. However, the exhaustive and heuristic search for DAG structures in these methods lead to combinatorial explosion issue with the increasing scale of nodes, as well as the local optima issue. Recently, NOTEARS [34] formulate the DAG constraint as a continuous optimization term that can be solved by gradient descent methods, successfully bringing causal discovery problem into existing effective learning frameworks. Along with this line, further studies start to pay attention to the form of structural equation [13], convergence [17] , optimization technique [36] and applications [12].

It should be noted that, by maximizing likelihood of observational data, FCMs (including NOTEARS) rely on strong assumptions to find the true causal structure [21]. To name a few, the variable noises should be identifiable with uniform variances, and there is

no model misspecification problem. But in realistic and wild settings, variables are generated with heterogeneous noises of mixed types or mixed variances. Reconstructing variables with best efforts will erroneously absorb noises into the reconstruction process, leading to the dependencies among reconstruction residules which violates the FCM theories. This over-reconstruction problem may cause spurious and missing edges in the resulted structure. We justify this argument with illustrative examples as shown in Table 1. We find that all of the three basic causal structures (chain, fork, and collider) can easily be erroneously identified by NOTEARS due to over-reconstruction problem. In other words, traditional reconstruction-based objective functions are not sufficient to guarantee the correctness of discovered causal graphs.

In this paper, we propose a novel differentiable method Differentiable Adversarial Residual INdependence for causal Graphs (DARING) by imposing explicit residual independence constraint. More specifically, we design the residual independence constraint with a theoretically derived measure on mutual-independence, combine it with the reconstruction process, and formulate the problem as a min-max joint optimization problem. We solve it in an adversarial process, where the discriminator is constantly trained to seek out the maximum correlation among reconstruction residuals; and the generator is trained to learn the graph structure and structural equations to minimize the correlation among reconstruction residuals. The newly introduced residual independence constraint does not weaken the virtue of traditional differentiable causal discovery methods in scalability. We prove that the combinatorial problem of mutual independence test for d variables can be solved by a d neural networks in a supervised learning task. To verify the superiority of our proposed method, we conduct extensive experiments to compare the performance of ours and baselines in multiple settings, including mixed noise type, mixed noise variance, and real-world dataset.

In summary, our contributions are highlighted as :

- We identify the over-reconstruction problem of traditional FCM models, and provide the justification that heterogeneous noises can easily make them fail.
- we propose a novel differentiable method DARING by imposing explicit residual independence constraint in a adversarial way.
- Extensive experiments demonstrate that our proposed method can significantly improve the robustness of causal discovery in wild settings.

The rest of the paper is organized as follows. Section 2 reviews the literature of related fields. Section 3 introduces our new FCM architecture **DARING** in detail. Section 4 gives the settings and results of the experiment to show the availability of our model. Finally, we conclude this work at the end of the paper.

2 RELATED WORKS

In this section, we review the works of some related fields with this work, including causal structure learning and adversarial learning.

Generally speaking, there are two types of algorithms for causal discovery (with I.I.D. data). Constraint-based algorithms learns the equivalence class of causal graphs according to conditional independence criteria under faithfulness assumption. If no unobserved confounder exists, Peter-Clark (PC) [26] implements the independence test for each pair of connected variables conditioned on subsets of their neighbor nodes and decides the edge direction by some designed rules which meet the necessary requirements of causal graph. In the presence of unobserved confounders, Fast Causal Inference algorithm (FCI) [27] also calls independence judgement like PC, but targets at a extended causal graph with bidirected edges indicating that there is at least one unmeasured confounder between the ends. Recently, ReCIT [32] measures the conditional independence of fork structure by independent regression residuals on the two child nodes for linear SEM.

However, limited sample size easily results in failure of statistics tests due to instability issue. The daunting cost of checking every candidate sub-structure is intolerable. To overcome these drawback, some score-based algorithms are purposed to alternatively employ a score function to measure the correctness of conditional independence of target graph with finite data. Taking a score function called BDeu, Greedy Equivalence Search (GES) [2] starts from an empty graph and adopt greedy strategy to add and cut edges successively until the score could not be improved. BiweiH develops a new class of generalized score functions by exploiting a particular regression problem in Reproducing Kernel Hilbert Space (RKHS) [9] to capture the dependence between random variables in a nonparametric way.

Compared to the above methods that only find the Markov equivalence class, other class of score-based method, FCMs can identify true causal graph from the same equivalence class with additional hypothesis. For example, PNL [33] proves it's definitely identifiable for two-variable setting except 5 special cases by testing if the disturbance is independent of direct causes for each variable. However, traditional approaches search the DAG structure for multiple-variable in a combinatorial manner, e.g. topological ordering of causality diagram into lower triangular matrix (LiNGAM [25]), which actually daunt the learning process of FCMs. By converting acyclicity constraint into a continuous program, NOTEARS [34] can directly apply a standard numerical solver for constrained optimization, such as augmented Lagrangian method, to achieve a global approximate solution. Further, DAG-GNN [30] propose a variant of gradient optimized constraint formulation that is more suitable for implementation and solve generalized linear SEM in autoencoder architecture; NOTEARS-MLP [35] and Gran-DAG [13] extend the framework of NOTEARS to deal with nonlinear functions using neural network or orthogonal basis expansion on each variable separately and adapt the acyclicity constraint at the level of neural network paths; RL-BIC [36] introduce Reinforcement Learning (RL) to search for the DAG with the best scoring while generating graph adjacency matrices that are used to compute rewards; GOLEM [17] apply a likelihood-based objective with soft sparsity and DAG constraints instead of constrained optimization, making linear SEM problem much easier to solve. For all these methods, maximum likelihood estimation of observational data is the guarantee of causal discovery, that is to say causal graph should conform to data generation mechanism. But the heterogeneity of external noise item in structural equation would lead to the inconsistence of perfect reconstruction and the correctness of causal graph, even without model misspecification. It goes worse for agnostic data distribution in wild settings. To this end, we propose to improve

Table 1: Three cases that traditional differentiable FCMs would find wrong causal graphs while the residual independence regularizer can help to identify the true graphs. We give examples on three basic causal structures (chain, fork, and collider from top to bottom). The graphs in green lines denote the ground truth. The red ones are the wrong structures learned by traditional differentiable FCMs (owing to the minimal reconstruction loss). However, true graphs have minimal reconstruction losses among graphs that satisfy the residual independence regularizer. Detailed data generating processes are as follows. Chain example: $A = \epsilon_A (\sim \mathcal{N}(0,1))$, $B = A + \epsilon_B (\sim \mathcal{N}(0,4))$, $C = B/5 + \epsilon_C (\sim \mathcal{N}(0,1))$. **Fork example:** $B = \epsilon_B (\sim \mathcal{U}(-2,2))$, $A = B/2 + \epsilon_A (\sim \mathcal{U}(-1,1))$, $C = B/2 + \epsilon_C (\sim \mathcal{U}(-1,1))$. **Collider example:** $A = \epsilon_A (\sim \mathcal{N}(0,1))$, $C = \epsilon_C (\sim \mathcal{N}(0,1))$, $B = A/3 + C/3 + \epsilon_B (\sim \mathcal{N}(0,1/9))$.

Predicted Graph (Chain)												
Reconstruction Loss	6.00	6.97	6.17	6.17	6.96	6.33	6.16	6.80	6.33	7.00	5.65	7.00
Residuals Mutually Independent?	✓			✓				✓				
Predicted Graph (Fork)												
Reconstruction Loss	1.67	2.17	1.83	1.67	2.17	1.83	1.83	2.00	1.83	2.33	1.78	2.33
Residuals Mutually Independent?								✓				
Predicted Graph (Collider)												
Reconstruction Loss	1.89	2.00	2.22	1.89	2.00	2.22	2.22	1.67	2.22	1.83	2.11	1.83
Residuals Mutually Independent?											✓	

the robustness of FCMs of continuous optimization with the help of mutually independent residuals of fitting generation functions.

Another line is the thought of adversarial learning [7]. Taking Generative Adversarial Nets (GAN) as an example, the core concept of GAN is to play a zero-sum game where a discriminator learns to distinguish real distribution from synthetic data distribution, and a generator attempt to generate data as realistic as possible so that it cannot be identified by the discriminator. As a result, both two players can achieve a much promising solution after upgrading each other alternately. Proposed by Goodfellow et al. [7], adversarial learning has been widely applied in various fields, including Computer Vision [14], Natural Language Processing [29], Graph Neural Networks [4] and so on. In this work, we leverage adversarial thought to capture the mutual independence between multi-dimensional variables.

3 ALGORITHM

3.1 Problem Definition

Structural Causal Model (SCM) defined on a set of random variables $X = \{X_i\}_{i=1}^d$ consist of a causal directed acyclic graph $G = (V(G), E(G))$ and the structural equations, defining the generative process of X. The joint distribution $P(X)$ encoded in SCM is Markov with respect to G, that is $P(X) = \prod_{i=1}^d P\left(X_i | X_{pa(i)}\right)$, where $X_{pa(i)}$ denotes the parent set of X_i in G. In this paper, we assume the structural equations satisfying Additive Noise Models[23] (as in Equation 1), where F_i is the mapping function that denotes the generative process on X_i and ϵ_i is the external noise of X_i.

$$X_i = F_i\left(X_{pa(i)}\right) + \epsilon_i. \tag{1}$$

Problem 1. *Given i.i.d. samples* $\mathbf{X} = \left\{\mathbf{x}^{(k)}\right\}_{k=1}^n$ *from the joint distribution* $P(X)$*, our goal is to infer the unknown causal graph* G *from* $\mathbf{X}$*, assuming the data generative mechanism follows ANMs.*

3.2 Preliminary

Here, we first recall the continuous optimization task of causal diagram learning after NOTEARS. Although advanced variants have been proposed to improve performance from different perspectives, its general formulation could be written as a minimization problem of an objective composed of data reconstruction, graph sparsity and acyclicity constraint (Equation 2).

$$\min \mathcal{L}^{(0)} = \mathcal{L}_{\text{rec}}(G, \mathbf{X}, \theta) + \alpha \mathcal{L}_{\text{DAG}}(G) + \beta \mathcal{L}_{\text{sparse}}(G). \tag{2}$$

$\mathcal{L}_{\text{rec}}(G, \mathbf{X}, \theta)$ refers to the ability of graph G and learnable model f with parameter set $\theta = (\theta_1, ..., \theta_d)$ to recover the data generative process of X, where θ_i is the parameter of f to approximate function F_i for variable X_i. For linear SEM, θ is a weighted matrix W. If considering nonlinear correlation, we can take a predefined nonlinear formulation or a neural network to fit agnostic F_i. A multilayer perception (MLP) [5] with parameter θ_i can be represented as,

$$\text{MLP}\left(X, \theta_i = \left(W^{(1)}, \ldots, W^{(l_i)}\right)\right) = \delta\left(W^{(l_i)} \delta\left(\cdots \delta\left(W^{(1)} X\right)\right)\right), \tag{3}$$

where $\delta : \mathbb{R} \to \mathbb{R}$ is an activation function and l_i is the number of layers for this MLP. To measure the distance between X and $f(X)$, $\mathcal{L}_{\text{rec}}(G, \mathbf{X}, \theta)$ could be regression-based metrics such as Euclidean distance (i.e. $||X - f(X)||_2^2$), or likelihood-based objectives for specific model.

Lemma 3.1 (Zheng et al. [34]). *A matrix $G \in \mathbb{R}^{d\times d}$ is a DAG if and only if*

$$\operatorname{tr}\left(e^{G\circ G}\right) - d = 0. \tag{4}$$

Lemma 3.1 uses the matrix exponential of Hadamard product of G to count the paths of arbitrary length between all the two nodes in graph. To ease the numerical difficulty of computing $\operatorname{tr}\left(e^{G\circ G}\right)$, Yu et al. [30] gives an alternative constraint that is more convenient.

$$\mathcal{L}_{\mathrm{DAG}}(G) = \operatorname{tr}\left[(I + \varphi(G\circ G)^d\right] - d, \quad \varphi > 0. \tag{5}$$

$\mathcal{L}_{\mathrm{sparse}}(G)$ denotes the number of edge in graph (i.e., $|G|_0$), and always optimized by $|G|_1$ in practice. $\mathcal{L}_{\mathrm{DAG}}(G)$ and $\mathcal{L}_{\mathrm{sparse}}(G)$ actually consider the structure characteristic of causal graph.

3.3 Motivation

However, it is not always consistent with the correctness of causal discovery if minimizing $\mathcal{L}^{(0)}$ only, especially in wild settings. The predefined reconstruction objectives, e.g., regression-based metrics, easily lead to misspecification problems due to the agnostic external noise types. Even if there is no such problem on both structural equations and noise types, i.e., for general linear Gaussian models, traditional methods will still fail in the situation of heterogeneous noise variance. Generally speaking, we can always decompose a complicated causal graph into three kinds of basic structure: chain, fork and, collider. In Table 1, we have shown that the truth graph would be mistakenly learned to another false structure for any basic structure, owing to the heterogeneity of external noise. Taking the chain as an example, the graph in green lines denotes the ground truth G_{true}, but the red one is the false structure G_{false} learned by traditional objective function $\mathcal{L}^{(0)}$. Because all the DAGs in Table 1 has two edges, they have the same value of $\mathcal{L}_{\mathrm{sparse}} = 2$ and $\mathcal{L}_{\mathrm{DAG}} = 0$. But $\mathcal{L}_{\mathrm{rec}}(G_{\mathrm{true}}) \le \mathcal{L}_{\mathrm{rec}}(G_{\mathrm{false}})$ under given generative mechanism, causing the anti-causal issue in this example. The similar situation also happens to fork and collider. Intuitively, the reason is that it prefers to use all related variables to regress the variable with external noise of larger variance to reduce the overall reconstruction loss, under the same constraint of acyclicity and sparsity. Over-reconstruction problem can easily confuse FCMs to mistakenly consider totally false structure (collider example). Traditional differentiable FCM methods are sensitive to the heterogeneity of noises, which is almost everywhere in realistic scenes.

Compared to FCMs, constraint-based methods are more robust for agnostic data distribution because they directly conduct independence test in data. However, testing independence between variables conditioned on parent node set is combinatorial and cannot be adapted into continuous optimization frameworks.

Actually, the mutual independence between external noises of variables is the foundation to ensure the identifiability of causal structure [6]. For additive noise models, the reconstruction residual $R_i = X_i - f_i\left(X_{pa(i)}\right)$ is the noise term of X_i. However, we notice that the residuals in learned false structure of Table 1 are not strictly independent. That indicates the violated independence condition in Equation 1 leads to the failure of differential FCMs. But driving the independent residuals as well, we can hence reduce the solution space and identify the true graph. Therefore, we propose to measure the mutual independence of residuals and introduce it into the traditional objective function 2 to capture the correctness of learned graphs better.

3.4 Proposed Model

In this section, we will introduce the details of our model, which defines a mutual independence statistic and employ it as a regularization of differentiable FCMs to ensure the independent residuals in an adversarial way. For clarity, we first present the technology of mutual independence measure for multi-dimensional variables with neural networks. Afterward, we describe the approach of applying it to discovering causal diagrams.

3.4.1 Mutual Independence.

Lemma 3.2 (Daudin [3]). *X and Y are independent if and only if for all functions $h \in L_X^2$, $g \in L_Y^2$,*

$$\operatorname{Cov}[h(X), g(Y)] = 0, \tag{6}$$

where

$$\begin{aligned} L_X^2 &= \left\{h(X) \mid \mathbb{E}\left[h(X)^2\right] < \infty\right\}, \\ L_Y^2 &= \left\{g(Y) \mid \mathbb{E}\left[g(Y)^2\right] < \infty\right\}, \end{aligned} \tag{7}$$

are square summable functions on X and Y.

Lemma 3.2 tells us the variables X and Y are completely independent, if they are always linearly independent after mapping by all the square summable functions. Further, we can extend it to the mutual independence of multi-dimensional variables.

Theorem 3.3. *Let $R = \{R_i\}_{i=1}^d$ be a set of random variables and $R_{-i} = \{R_1, ..., R_{i-1}, R_{i+1}, ..., R_d\}$. All variables of R are mutually independent if and only if $\forall h_i \in L^2_{R_{-i}}, \forall g_i \in L^2_{R_i}$, $i \in \{1, \ldots, d\}$,*

$$\operatorname{Cov}[h_i(R_{-i}), g_i(R_i)]| = 0. \tag{8}$$

Similar with Equation 7, $L^2_{R_{-i}}$ and $L^2_{R_i}$ are the spaces of square summable functions on R_{-i} and R_i respectively.

We provide a proof of Theorem 3.3 in Appendix.

Theorem 3.3 helps to reduce the combinatorial problem of mutual independence test for $\{R_i\}_{i=1}^d$ to d sub-problems. Afterward, we can define a statistic $\mathrm{M}(R)$ to measure the mutual independence of a variable set R as,

$$\mathrm{M}(R) = \sum_{i=1}^{d} \sup_{h_i \in L^2_{R_{-i}},\, g_i \in L^2_{R_i}} \left\| \frac{\operatorname{Cov}[h_i(R_{-i}), g_i(R_i)]}{\sqrt{\operatorname{Var}[h_i(R_{-i}))]} \cdot \sqrt{\operatorname{Var}[g_i(R_i)]}} \right\|. \tag{9}$$

where h_i, g_i are not constant mappings.

The value of $\mathrm{M}(R) \in [0, d]$ denotes the correlation strength of variable set R. We consider to use multilayer perception (MLP) with parameters ϕ_i to approximate $h_i \in L^2_{R_{-i}}$, $g_i \in L^2_{R_i}$ and learn them to reach the supremum of $\mathrm{M}(R)$. However, with limited data size, if both h_i, g_i are modeled by different MLPs respectively, the information from observational data of R will be quite weak due to huge function space. As a result, we predefine $g_i(R_i) = R_i$ for single variable R_i. Then we can design an objective function, which maximizes $\mathcal{L}_{\mathrm{M}}(R, \phi)$ by optimizing all the h_i in d supervised task, i.e.,

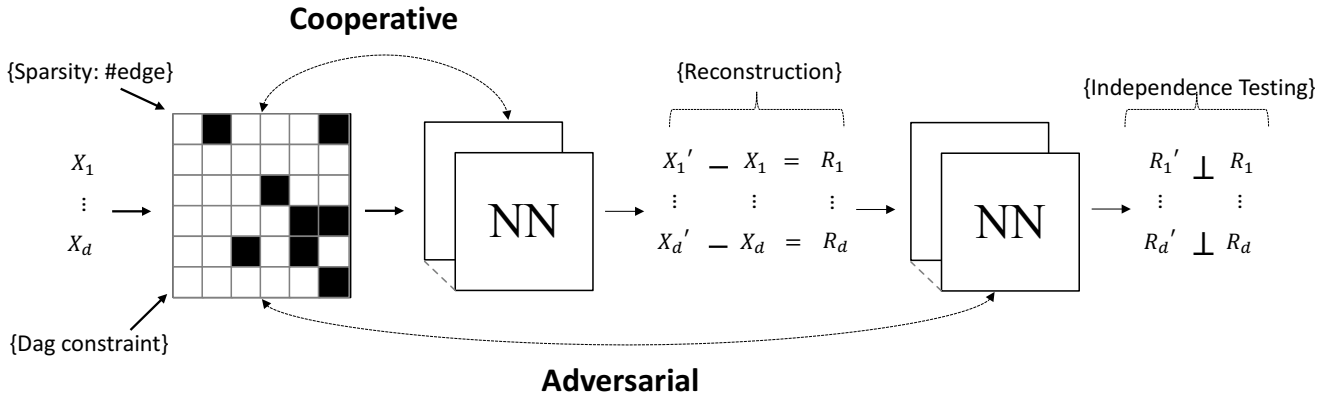

Figure 1: The framework of our proposed DARING: the learnable graph structure cooperates with a parameterized model (e.g. neural network) to recover the generative mechanism of observational data; another neural network, adversarial with learned graph, is used to conduct mutual independence test of all the reconstruction residuals and ensure independent residuals. Further additional constraints, including DAG and Sparsity, enforce the learned graph having causal structure characteristic. The overall framework is continuously optimized.

$$\max_{\phi} \mathcal{L}_{\mathrm{M}}(R, \phi) = \sum_{i=1}^{d} \left\| \frac{\mathrm{Cov}[\mathrm{MLP}(R_{-i}, \phi_i), R_i]}{\sqrt{\mathrm{Var}[\mathrm{MLP}(R_{-i}, \phi_i)]} \cdot \sqrt{\mathrm{Var}[R_i]}} \right\|_2^2. \quad (10)$$

For each variable R_i, ϕ_i is learned as the mapping function of R_{-i} with maximal correlation with R_i. To better adapt $\mathrm{M}(R)$ to continuous optimization framework, we replace 1-norm (in Equation 9) with 2-norm (in Equation 10).

3.4.2 Causal Discovery with Adversarial Learning. As explained in Section Motivation, we have shown that traditional FCMs easily fail in finding the true causal graph due to pursuit of perfect reconstruction. However, the independent residuals can help to shrink down the solution space of traditional differentiable FCMs.

(1) Independence of residuals $R_i = X_i - f_i(X_{pa(i)})$ is the necessary condition (Equation 1) for correctly identifying the true causal graph under ANMs' assumption;
(2) The false structure learned as best solution by traditional methods would get residuals with strong correlations.
(3) The truth graph, missed by traditional methods, would be identified with independence measure (examples in Table 1).

The point (1), (2) and (3) indicates a independence regularization can always benefit the causal discovery performance. Therefore, we acknowledge a graph describing causal mechanism if it can satisfy the following conditions.

(1) Based on its diagram, a learnable model $f(\theta)$ can recover the generative process of observational data.
(2) All the residuals $(X - f(X, \theta))$ are mutually independent.
(3) It is a sparse DAG structure.

Continuous optimization helps to promote using effective learning technologies to ensure condition (1) and (3) abundantly. Furthermore, we hope to satisfy condition (2) with the advantage of $\mathrm{M}(R)$ defined before. But it's not straightforward to combine $\mathrm{M}(R)$ with traditional methods, because $(X - f(X, \theta))$ keeps changing along with the updating of G and θ.

Algorithm 1 Causal Discovery with DARING

Input: $\mathbf{X} = \left\{\mathbf{x}^{(k)}\right\}_{k=1}^{n}$ i.i.d. sampled from $P(X)$ and threshold Δ
Output: Causal graph G
Initial G, parameters of causality fitting model θ $(\theta_1, ..., \theta_d)$ and parameters of independence test model ϕ $(\phi_1, ..., \phi_d)$
Pretrain G and θ to minimize $\mathcal{L}^{(0)}$ for τ_0 steps
while not arriving maximal iteration or triggering termination conditions **do**
  for $t = 1$ to τ_1 **do**
    Fix G, θ and calculate $\mathcal{L}_{\mathrm{M}}(R, \phi)$ in Equation 10
    Update ϕ to maximize $\mathcal{L}_{\mathrm{M}}(R, \phi)$
  end for
  for $t = 1$ to τ_2 **do**
    Fix ϕ and calculate total $\mathcal{L}$ in Equation 11
    Update G, θ to minimize $\mathcal{L}$
  end for
end while
Prune the edges less than Δ of G
return: G

To deal with this issue, we formulate the independent residual condition/constraint as a min-max problem and introduce adversarial learning to solve it. In continuous framework, alternatively updating two models for finite steps against each other, it's said that both models can reach the global optimal solution in ideal settings. Hence, we adopt adversarial learning to design an architecture as in Figure 1 to jointly optimize causal graph G, fitting model $f(\theta)$ and mutual independence measure model $h(\phi)$.

We first define a new objective function $\mathcal{L}$ as in Equation 11 to ensure independent residuals while reconstructing data in adversarial process, under DAG and sparsity constraint. The training

details of our **DARING** are as follows: if $R^{(t)} = X - f\left(X, \theta^{(t)}\right)$ is the residuals after t epoch; at $(t+1)$ epoch, we fix $G^{(t)}$ and $\theta^{(t)}$ and update $\phi^{(t+1)}$ for τ_1 steps to maximize $\mathcal{L}_{\mathrm{M}}\left(R^{(t)}, \phi\right)$ that measures the mutual correlation of residuals; then fixing $\phi^{(t+1)}$, we update $G^{(t+1)}$ and $\theta^{(t+1)}$ for τ_2 steps to minimize $\mathcal{L}$ in turn. Although $\mathcal{L}_{\mathrm{M}}\left(R^{(t)}, \phi\right)$ possibly increases in its training steps, its value would keep decline on the whole in learning process owing to independent residuals on the correct graph.

$$\min_{G,\theta} \max_{\phi} \mathcal{L}(\mathbf{X}, G, \theta) = \mathcal{L}_{\mathrm{rec}}(G, \mathbf{X}, \theta) + \alpha \mathcal{L}_{\mathrm{DAG}}(G) + \beta \mathcal{L}_{\mathrm{sparse}}(G) + \gamma \mathcal{L}_{\mathrm{M}}(X - f(X, \theta), \phi). \tag{11}$$

For better convergence, we can pretrain G and θ according to $\mathcal{L}^{(0)}$ for a few epochs at first. Then optimizing G, θ and ϕ until arriving maximal iteration or triggering termination conditions, e.g., DAG requirement has been met, training process ends at a Nash equilibrium solution ideally. Finally, a causal graph is outputted after post-processing on G, e.g., the edge is cut off if its weight value less than a threshold Δ.

4 EXPERIMENTS

In this section, we carry out extensive experiments to verify the effectiveness of our proposed method **DARING** on learning causal structures. Our experimental settings cover linear and nonlinear cases, structure identifiable and non-identifiable cases. All the experiments conducted in this paper are implemented in Python and Pytorch [20] framework.

4.0.1 Baselines. The proposed method for residual independence regularization is not confined to a specific form of differentiable FCM model. Without lose of generality, we take NOTEARS as backbone, and compare with the models based on the same backbone. We choose NOTEARS and GOLEM as baselines for linear models, NOTEARS-MLP for nonlinear models, and add independence regularization term into them respectively as the implementations of our method. NOTEARS takes a matrix W as causal graph and the weight on it $W_{i,j}$ denotes the causal effect of variable X_i to X_j. W is learnt by regression-based objective using augmented Lagrangian method. GOLEM uses likelihood-based objective, together with a soft constraint term $-\log|\det(I - W)|$, to learn W for better optimization. NOTEARS-MLP approximates generative mechanism of X_j by MLP with parameters $\theta_j = \left(A_j^{(1)}, ..., A_j^{(l)}\right)$ and takes $W(\theta)_{i,j} = \left\| \left(A_j^{(1)}\right)_{[:,k]} \right\|_2$ as causal graph. We adopt the same post-processing strategy for all the methods, cutting off the edges with values less than 0.3.

4.0.2 Metrics. We evaluate the estimated DAG structure using the following common metrics:

- Precision: proportion of correctly detected edges to the total detected edges;
- Recall: proportion of correctly detected edges to the total edges in true graph;
- Structural Hamming Distance (SHD): the number of missing, falsely detected or reversed edges;
- Structural Interventional Distance (SID) [22]: the number of pair (X_i, X_j) such that the interventional distribution $p(X_j|\mathrm{do}(X_i = x))$ would be miscalculated if we choose the parent adjustment set in estimated graph.

4.1 Synthetic Data of Linear Models

4.1.1 Simulation. We examine the structure learning performance of NOTEARS, GOLEM and proposed method (NOTEARS + DARING, GOLEM + DARING) in three kinds of wild settings for linear data: mixed noise type, identifiable mixed noise variance and non-identifiable mixed noise variance. For each setting, we varied the number of nodes (10, 20, 40) and sampled 10 datasets of 1000 examples as follows: first, a ground truth DAG G is randomly sampled following Erdos-Renyi (ER) or Scale-Free (SF) scheme with different edge density (degree = 2 or 4); then the data is generated according to linear SEM $X = BX + \epsilon$, where each value of matrix B is uniformly sampled in $(-2, -0.5) \bigcup (0.5, 2)$ and the external noise ϵ comes from a specific scheme. The detailed configurations in each settings are as follows,

- Mixed noise type setting. We randomly select one kind of noise from *Gaussian* distribution $\mathcal{N}(0, \sigma^2)$, $Uniform$ distribution $U(-\varsigma, \varsigma)$, *Exponential* distribution $\mathcal{E}(\lambda)$ and *Gumbel* distribution $\mathcal{G}(0, \kappa)$ for each node and keep $\sigma^2 = \varsigma = \lambda = \kappa = 1$. We adopt SF graph with edge density to be 4 in this setting.
- Identifiable mixed noise variance setting.We randomly select $\varsigma = 1, 2, 3, 4$ for each node and sample noise from $U(-\varsigma, \varsigma)$. SF graph with edge density to be 2 is adopted.
- Non-Identifiable mixed noise variance setting. We randomly select $\sigma = 1, 2, 3, 4$ for each node and sample noise from $\mathcal{N}(0, \sigma^2)$. This setting is a well-known non-Identifiable case because all the external noise is sampled from *Gaussian* distribution. ER graph with edge density to be 2 is adopted.

From the results in Figure 2, we can see that, across all the linear settings:

(1) Compared to NOTEARS, GOLEM has a higher precision value but less recall value, meaning that GOLEM prefers to learn a more sparse structure.
(2) For overall metric SHD and SID, GOLEM always outperforms NOTEARS, especially for *Gaussian* noise perfectly satisfying its assumption.
(3) No matter implemented based on vanilla NOTEARS or advanced GOLEM, our method consistently help the backbone to improve their performances under all metrics, implying our method can comprehensively mitigate the failure cases including missing, falsely detected or reversed edges.
(4) Our advantages become more prominent as the scale of the graph increases.

4.2 Synthetic Data of Nonlinear Models

4.2.1 Simulation. We examine the performance of NOTEARS-MLP and proposed method (NOTEARS-MLP + DARING) with the same backbone in four kinds of wild settings for nonlinear data: mixed noise type, identifiable mixed noise variance, identifiable large noise

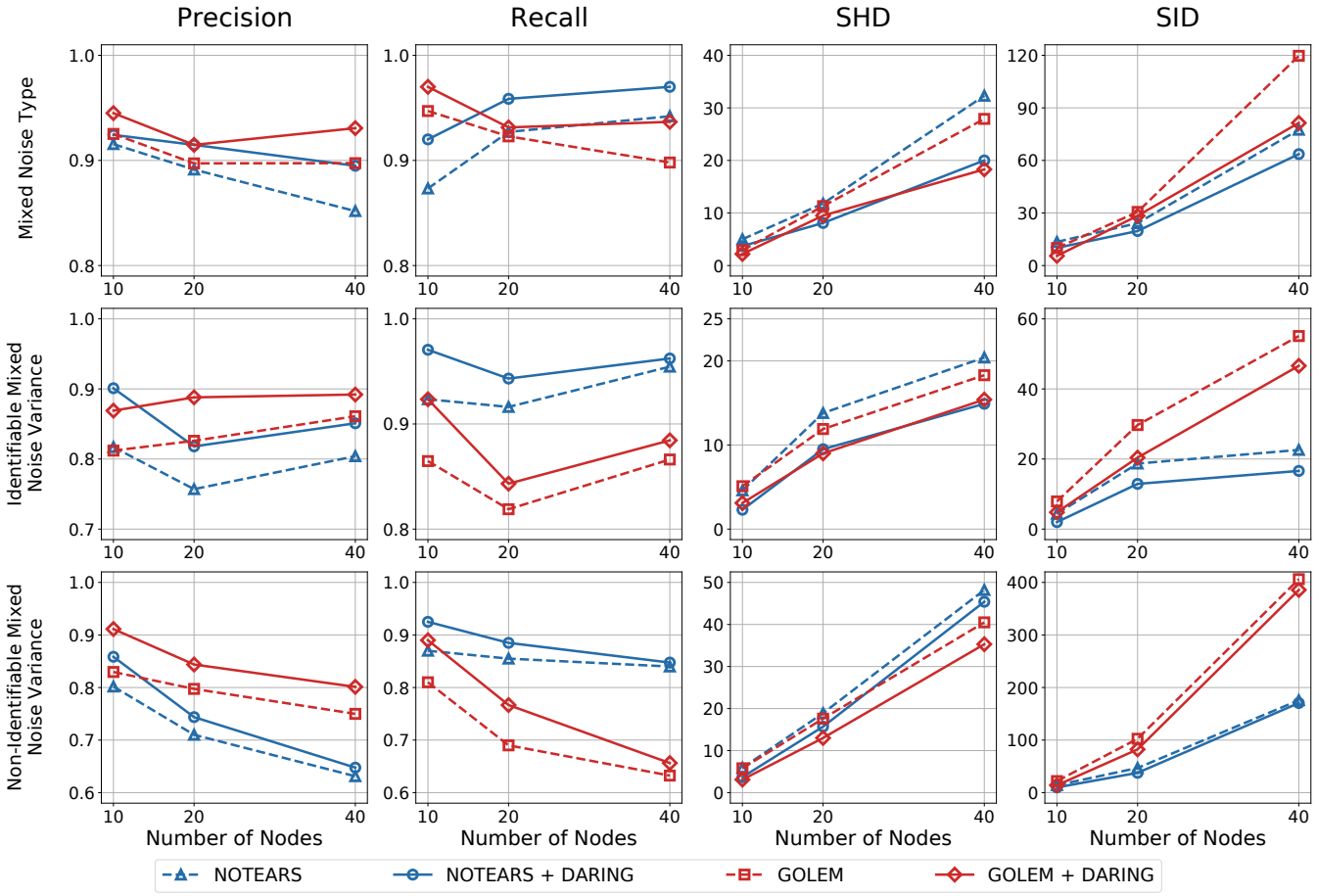

Figure 2: Empirical results on linear synthetic data. We show the performance of different methods on four metrics (precision, recall, SHD, and SID from left to right) in three wild settings with varied graph scales. For precision and recall, the higher, the better; for SHD and SID, the lower, the better. Our method (DARING) improves the backbones (NOTEARS or GOLEM) on all the metrics and achieves the best performance in almost every cases.

variance and non-identifiable small noise variance. For each setting, we varied the data size (400, 1000, 2000) and sampled 10 datasets of graph with 20 nodes as follows: first, a ground truth DAG G is randomly sampled following ER or SF scheme with different edge density (degree = 2, 3, or 4); then the data is generated according to ANM $X = f(X) + \epsilon$, where f is a two layers MLP with each value of parameter uniformly sampled in $(-2, -0.5) \bigcup (0.5, 2)$ and the external noise ϵ comes from a specific scheme. The detailed configurations in each settings are as follows,

- Mixed Noise Type Setting. We randomly select one kind of noise from $\mathcal{N}(0, \sigma^2)$, $U(-\varsigma, \varsigma)$, $\mathcal{E}(\lambda)$ and $\mathcal{G}(0, \kappa)$ for each node and keep $\sigma^2 = \varsigma = \lambda = \kappa = 1$. We adopt ER graph with edge density to be 3 in this setting.
- Identifiable Mixed Noise Variance Setting. We randomly select $\varsigma = 1, 2, 3, 4$ for each node and sample noise from $U(-\varsigma, \varsigma)$. ER graph with edge density to be 2 is adopted.
- Identifiable Large Noise Variance Setting. We randomly sample noise from $\mathcal{G}(0, 4)$ for each node. SF graph with edge density to be 3 is adopted.
- Non-identifiable Small Noise Variance Setting. We randomly sample noise from $\mathcal{N}(0, 1)$ for each node. ER graph with edge density to be 4 is adopted.

The empirical results are reported in Figure 3. Causal discovery from nonlinear data produced by MLP is actually a quite challenging task, owing to the complicated structural equations. It becomes harder if facing with inadequate sample size or wild settings in our problem. As a result, the curves of all methods sharply drop for SHD and SID, and rise for precision and recall, with the increase of sample size. Mixed noise type or *Gumbel* noise with 400 samples are the most difficult. Even in these challenging cases, DARING can still bring substantial improvements in all metrics. The improvement margins are much larger for cases with small data size (e.g. 400 or 1000). Moreover, we find DARING can also work well in traditional mild setting, e.g. all *Gaussian* noise with small variance. A plausible

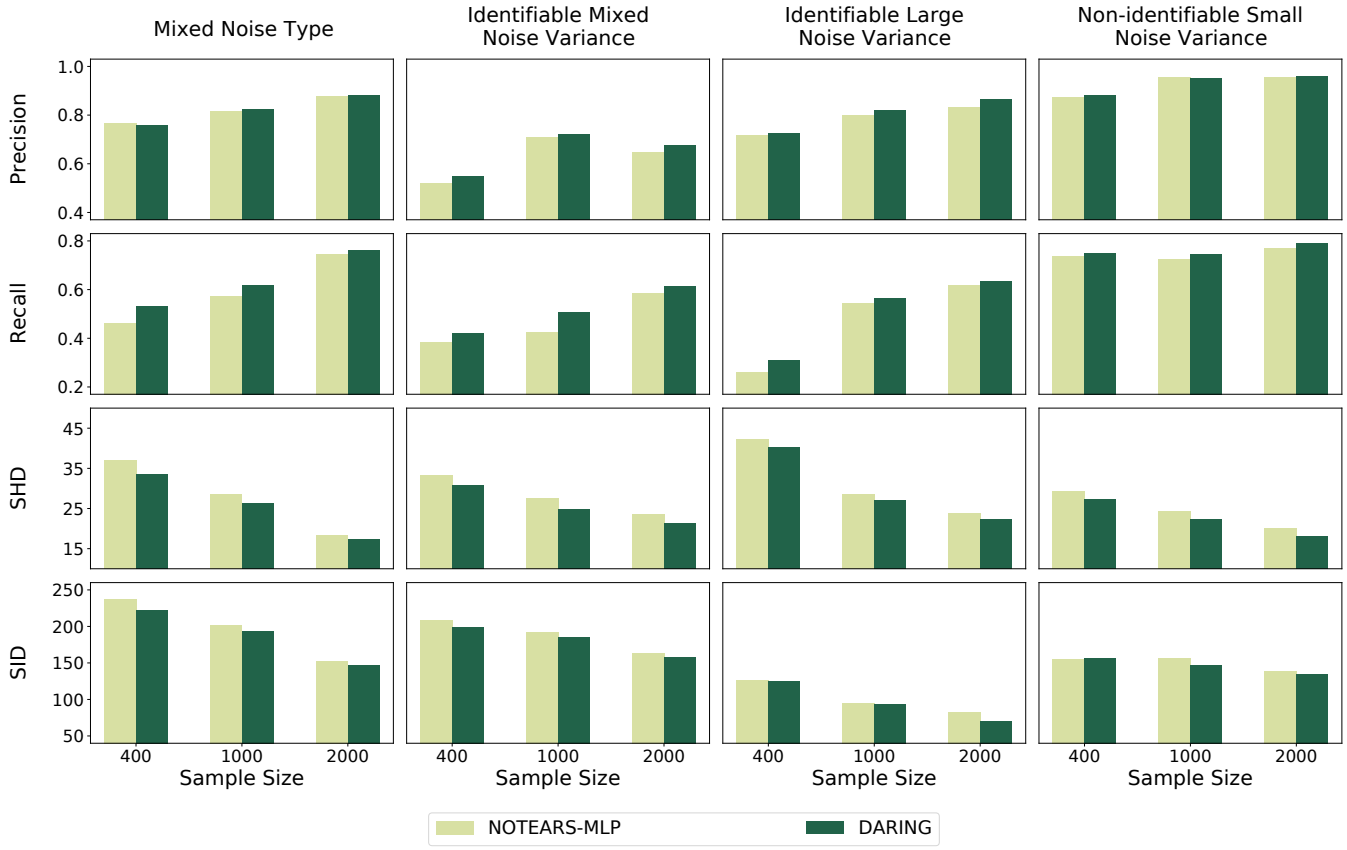

Figure 3: Empirical results on nonlinear synthetic data. We show the performance of NOTEARS-MLP and DARING on four metrics (precision, recall, SHD, and SID) in four wild settings with varied data sample sizes. For precision and recall, the higher, the better; for SHD and SID, the lower, the better. Our method (+DARING) outperforms NOTEARS-MLP in four settings.

Table 2: Empirical results on Sachs dataset.

Method	Total Edges	Correct Edges	SHD	SHD-C
RL-BIC	10	**7**	**11**	9
GraN-DAG	10	5	13	9
NOTEARS-MLP	11	6	**11**	6
DAG-GNN	15	6	16	12
GOLEM	11	6	14	12
NOTEARS	20	6	19	13
ICA-LiNGAM	8	4	14	11
CAM	10	6	12	9
DARING	15	**7**	**11**	**4**

reason is that DARING can alleviate the overfitting problem of MLP model to some extent by enforing independent residuals.

4.3 Real Data

To evaluate the preformance of DARING in real applicaitons, we consider a dataset that is to discover a protein signaling network on expression levels of different proteins and phospholipids in human cells[1] [24]. This dataset from biology community is a common benchmark of graphical models, containing both observational and interventional data. The true causal graph given by Sachs et al. [24] has 11 nodes and 17 edges. Here, we only consider the observational data with $n = 853$ samples, that is the same with Lachapelle et al. [13].

In this benchmark dataset, we compare with the recent continuous optimized FCM methods, containing NOTEARS, NOTEARS-MLP, DAG-GNN, GraN-DAG, RL-BIC and GOLEM, traditional FCM methods ICA-LiNGAM [25], and a combinational method CAM [1]. Because the true causal graph is so sparse that an empty graph can reach as low as 17 in SHD, we report the #total predicted edges, #correct edges, SHD and SHD-C (the SHD between corresponding CPDAG[2] [19]) in Table 2.

The poor performance of methods based on generalized linear SEM, including DAG-GNN, GOLEM, NOTEARS and ICA-LiNGAM, could be explained by their inability of modeling nonlinear mechanism in real data. Through decoupling the causal order search

[1] https://www.bnlearn.com/book-crc/code/sachs.data.txt.gz
[2] A Markov equivalence class can be characterized by a graphical object named a completed partially directed acyclic graph (CPDAG).

among variables from feature or edge selection, CAM shows competitive result. The nonlinear functional methods use the fitting capability of MLP to achieve the same level of score. RL-BIC estimates more correct edges with lower SHD, but NOTEARS-MLP leads in better detecting skeleton structure. Comparatively, DARING achieves the best performances in all the metrics. At the same time, we achieve state-of-the-art performance in detecting skeleton structures, indicating by SHD-C.

5 CONCLUSION

In this paper, we target the problem of causal discovery using FCM based continuous optimization methods. We discuss the inconsistence issue of traditional objective function and the correctness of learnt graph in depth. To dispose of this disadvantage, we design a novel architecture called **DARING**. With access to the good convergence of adversarial learning in continuous framework, DARING enforces mutually independent residuals while fitting structural equations from passive data, which satisfies the external noise assumption in ANM formulation. Extensive experimental results of both simulation and real data prove that our method has more robustness and performs better in wild settings.

ACKNOWLEDGMENTS

This work was supported in part by National Key R&D Program of China (No. 2018AAA0102004, No. 2020AAA0106300), National Natural Science Foundation of China (No. U1936219, 61521002, 61772304), Beijing Academy of Artificial Intelligence (BAAI), and a grant from the Institute for Guo Qiang, Tsinghua University.

REFERENCES

[1] Peter Bühlmann, Jonas Peters, Jan Ernest, et al. 2014. CAM: Causal additive models, high-dimensional order search and penalized regression. *Annals of statistics* 42, 6 (2014), 2526–2556.

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

[3] JJ Daudin. 1980. Partial association measures and an application to qualitative regression. *Biometrika* 67, 3 (1980), 581–590.

[4] Nicola De Cao and Thomas Kipf. 2018. MolGAN: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973* (2018).

[5] Matt W Gardner and SR Dorling. 1998. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment* 32, 14-15 (1998), 2627–2636.

[6] Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of causal discovery methods based on graphical models. *Frontiers in genetics* 10 (2019), 524.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, Vol. 27. 2672–2680.

[8] Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. 2020. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)* 53, 4 (2020), 1–37.

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

[10] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 656–666.

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

[12] Trent Kyono, Yao Zhang, and Mihaela van der Schaar. 2020. CASTLE: Regularization via Auxiliary Causal Graph Discovery. *arXiv preprint arXiv:2009.13180* (2020).

[13] Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. 2020. Gradient-Based Neural DAG Learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

[14] Yuqing Ma, Yue He, Fan Ding, Sheng Hu, Jun Li, and Xianglong Liu. 2018. Progressive Generative Hashing for Image Retrieval.. In *IJCAI*. 871–877.

[15] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592* (2019).

[16] Leland Gerson Neuberg. 2003. Causality: Models, Reasoning, and Inference.

[17] Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. 2020. On the Role of Sparsity and DAG Constraints for Learning Linear DAGs. *Advances in Neural Information Processing Systems* 33 (2020).

[18] Rainer Opgen-Rhein and Korbinian Strimmer. 2007. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC systems biology* 1, 1 (2007), 1–10.

[19] Simon Parsons. 2011. *Probabilistic Graphical Models: Principles and Techniques* by Daphne Koller and Nir Friedman, MIT Press, 1231 pp., $95.00, ISBN 0-262-01319-3. *Knowl. Eng. Rev.* 26, 2 (2011), 237–238. https://doi.org/10.1017/S0269888910000275

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems* 32 (2019), 8026–8037.

[21] Jonas Peters and Peter Bühlmann. 2014. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika* 101, 1 (2014), 219–228.

[22] Jonas Peters and Peter Bühlmann. 2015. Structural intervention distance for evaluating causal graphs. *Neural computation* 27, 3 (2015), 771–799.

[23] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. 2014. Causal discovery with continuous additive noise models. (2014).

[24] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308, 5721 (2005), 523–529.

[25] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. 2006. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* 7, 10 (2006).

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

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

[28] Renzhe Xu, Peng Cui, Kun Kuang, Bo Li, Linjun Zhou, Zheyan Shen, and Wei Cui. 2020. Algorithmic Decision Making with Conditional Fairness. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2125–2135.

[29] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.

[30] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. 2019. DAG-GNN: DAG structure learning with graph neural networks. In *International Conference on Machine Learning*. PMLR, 7154–7163.

[31] Bin Zhang, Chris Gaiteri, Liviu-Gabriel Bodea, Zhi Wang, Joshua McElwee, Alexei A Podtelezhnikov, Chunsheng Zhang, Tao Xie, Linh Tran, Radu Dobrin, et al. 2013. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* 153, 3 (2013), 707–720.

[32] Hao Zhang, Shuigeng Zhou, and Jihong Guan. 2018. Measuring conditional independence by independent residuals: Theoretical results and application in causal discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[33] K Zhang and A Hyvärinen. 2009. In *25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)*. AUAI Press, 647–655.

[34] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. 2018. DAGs with NO TEARS: Continuous Optimization for Structure Learning. *Advances in Neural Information Processing Systems* 31 (2018), 9472–9483.

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

[36] Shengyu Zhu, Ignavier Ng, and Zhitang Chen. 2019. Causal Discovery with Reinforcement Learning. In *International Conference on Learning Representations*.

Appendix

Proof of Theorem 3.3

Theorem. *Let $R = \{R_i\}_{i=1}^{d}$ be a set of random variables and $R_{-i} = \{R_1, ..., R_{i-1}, R_{i+1}, ..., R_d\}$. All variables of R are mutually independent if and only if $\forall h_i \in L^2_{R_{-i}}, \forall g_i \in L^2_{R_i}, i \in \{1, \ldots, d\}$,*

$$\mathrm{Cov}[h_i(R_{-i}), g_i(R_i)]| = 0. \tag{12}$$

Similar with Equation 7, $L^2_{R_{-i}}$ and $L^2_{R_i}$ are the spaces of square summable functions on R_{-i} and R_i respectively.

Proof. On the basis of Lemma 3.1, $\forall i$, given the condition

$$\mathrm{Cov}[h_i(R_{-i}) \cdot g_i(R_i)] = 0, \quad \forall h_i \in L^2_{R_{-i}}, g_i \in L^2_{R_i},$$

we have $R_i \perp R_{-i}$, i.e.,

$$P(R) = P(R_i) \cdot P(R_{-i}).$$

Integrate the above function over $R_1, ..., R_{i-1}$, we have

$$P(R_i, ..., R_d) = P(R_i) \cdot P(R_{i+1}, \ldots, R_d).$$

Hence,

$$\begin{aligned} P(R) &= P(R_1)P(R_2, R_3, \ldots, R_d) \\ P(R) &= P(R_1)P(R_2)P(R_3, \ldots, R_d) \\ &= \cdots \\ &= \prod_{i=1}^{d} P(R_i). \end{aligned}$$

As a result, R are mutually independent. □