On the Role of Entropy-Based Loss for Learning Causal Structure With Continuous Optimization

Weilin Chen®, Jie Qiao®, Ruichu Cai®, Senior Member, IEEE, and Zhifeng Hao, Member, IEEE

Abstract—Causal discovery from observational data is an important but challenging task in many scientific fields. A recent line of work formulates the structure learning problem as a continuous constrained optimization task using an algebraic characterization of directed acyclic graphs (DAGs) and the least-square loss function. Though the least-square loss function is well justified under the standard Gaussian noise assumption, it is limited if the assumption does not hold. In this work, we theoretically show that the violation of the Gaussian noise assumption will hinder the causal direction identification, making the causal orientation fully determined by the causal strength as well as the variances of noises in the linear case and by the strong non-Gaussian noises in the nonlinear case. Consequently, we propose a more general entropy-based loss that is theoretically consistent with the likelihood score under any noise distribution. We run extensive empirical evaluations on both synthetic data and real-world data to validate the effectiveness of the proposed method and show that our method achieves the best in structure Hamming distance, false discovery rate (FDR), and true-positive rate (TPR) matrices.

Index Terms—Acyclicity constraint, causal discovery, entropy-based loss, least-square loss.

NOMENCLATURE

| Notation | Description |
|----------|-------------|
| $\mathcal{G}$ | Directed acyclic graph (DAG). |
| $\tilde{X}$ | Random variables vector $\tilde{X} = \{X_1, X_2, \ldots, X_d\}$. |
| $X_i$ | $i$th variable (element) of $\tilde{X}$. |
| $\tilde{N}$ | Random variables vector $\tilde{N} = \{N_1, N_2, \ldots, N_d\}$. |
| $N_i$ | Noise variable of $X_i$. |
| $X_{pa(i)}$ | Parent variable of $X_i$. |
| $f_i$ | Mapping from $X_{pa(i)}$ to $X_i$. |
| $X$ | Design matrix, $X \in \mathbb{R}^{m \times d}$. |
| $Y$ | Cause variable in pairwise $X, Y$. |
| $Y$ | Effect variable in pairwise $X, Y$. |
| $N_X$ | Noise of cause variable. |
| $N_Y$ | Noise of effect variable. |
| $f$ | Nonlinear mapping from $X$ to $Y$ in causal direction. |
| $g$ | Nonlinear mapping from $Y$ to $X$ in anti-causal direction. |
| $\hat{N}_X$ | Residual of $X$ denoted in anti-causal direction, $\hat{N}_X := X - g(Y)$ or $\hat{N}_X := X - \hat{a}Y$. |
| $\hat{N}_Y$ | Residual of $Y$ denoted in anti-causal direction, $\hat{N}_Y := Y$. |
| $p$ | Probability density function. |
| $q$ | Probability density function of standard Gaussian distribution. |
| $I(X, Y)$ | Mutual information between variable $X$ and $Y$. |
| $H(X)$ | Entropy of variable $X$. |
| $\sigma_X^2$ | Variance of variable $X$. |

I. INTRODUCTION

LEARNING causal structure from observational data has become an important topic in many scientific fields, such as economics [1], biology [2], neuroinformatics [3], and social science [4]. Due to the expensive cost or the ethic of randomized experiments, the task of causal discovery from purely observational data has drawn much attention.

Many approaches have been proposed for learning causal structure. Traditionally, by utilizing the conditional independence property among observed variables, the constraint-based approaches have been proposed, e.g., PC algorithm [5], but only identify the underlying DAG up to Markov equivalence class [6]. Alternatively, by introducing a certain class of structure causal model (SCM), and further assuming the causal mechanism that the noise and the hypothetical cause are independent, the functional-based causal models have been proposed, e.g., the linear non-gaussian acyclic model (LiNGAM) [7], the additive noise model (ANM) [8], and the post-nonlinear (PNL) causal model [9]. However, due to the intractable search space superexponential in the number of graph nodes, learning DAGs using functional-based causal models is challenging.

Recently, Zheng et al. [10] propose a method named NOTEARS: Non-combinatorial Optimization via Trace
Exponential and Augmented lagRangian for Structure learning, which formulates the causal discovery problem as a continuous optimization problem using least-square loss with a DAG constraint under the ANM assumption. Such a technique has been extensively developed and applied to learning linear or nonlinear causal structures. Yu et al. [11] introduce a variational autoencoder framework for modeling the generative process of a causal structure equipped with evidence lower bound (ELBO) with a Gaussian prior of noise which is implemented by least-square loss. Ng et al. [12] and Zheng et al. [13] extend a linear causal model into a nonlinear causal model using neural networks but still rely on the least-square loss for reconstruction.

However, our analysis shows that using least-square loss as the score function will hinder the causal direction identification. In a linear system, such loss will be highly influenced by the causal strengths and the noise variances leading to a bias estimation compared with the likelihood score. In a more general nonlinear system, the causal direction will be incorrectly identified if noise distribution has a strong non-Gaussianity property.

In this article, we show that the entropy-based loss is consistent with the likelihood score under the ANM using any noise distribution, and thus we advocate using the entropy-based loss instead of least-square loss. Overall, our contributions are as follows.

1) In Section IV, we study the limitations of least-square loss and provide theoretical bounds that least-square loss will fail to identify the causal direction in a linear and nonlinear system, respectively.

2) In Section V, we build a connection between the entropy-based loss and the log-likelihood score and further provide the theory for its validity.

3) In Section VI, we run extensive experiments over linear and nonlinear systems by using both synthetic data and real-world data and show that our method performs more effectively and stably.

II. RELATED WORKS

A. Related to Least-Square Loss

Blöbaum et al. [14] address the problem of inferring the causal relation between two variables using least-square loss under the modularity property, i.e., the independence between the function and the distribution of cause. Peters and Bühlmann [15] prove the full identifiability of the linear Gaussian structural equation in the case that all noise variables have the same variance, which can be recovered by the least-square loss in our analysis. Loh and Bühlmann [16] prove the identification of a linear Gaussian structural causal model with known noise variances.

B. Related to Causal Discovery

One of the categories of causal discovery approaches is to learn causal structures by iteratively conducting independence tests, such as IC [17], PC [5], M2LC [18], and M3B [19]. Another category of approaches is based on causal mechanism assumptions. Shimizu et al. [7] propose LiNGAM by assuming linearity and non-Gaussianity. Many variants improve the efficiency [20], [21], extend to time-series data [22], or relax the causal sufficiency assumption [23]. Hoyer et al. [8] propose ANM with additive noise assumption. Peters et al. [24] study the identifiability of discrete ANM, and Wei et al. [25] extend the orientation rule of ANM by Wasserstein distance. Zhang and Hyvärinen [9] and [26] propose PNL causal model with inner additive noise and give its identifiability condition. Bühlmann et al. [27] propose the causal additive model (CAM) by assuming the structural equations are additive. Recently, based on the acyclic constraint, the gradient-based approaches has drawn attention. Zheng et al. [10] propose a DAG constraint named NOTEARS for learning causal structure with continuous optimization. Ng et al. [28] apply soft sparsity and DAG constraints to learn linear DAGs based on a likelihood-based score function. Yu et al. [11] propose an alternative characterization of acyclicity and utilize a generative model to learn the nonlinear causal structure. Ng et al. [12] utilize a graph autoencoder framework to extend the linear SCM to nonlinear SCM. Ng et al. [29] extend to nonlinearity by putting the weighted matrix on the first layer of MLP. Lachapelle et al. [30] express weighted matrix by neural network paths to deal with the nonlinear case. Zheng et al. [13] propose a more general acyclicity constraint based on partial derivatives to support nonlinear models. Liang et al. [31] propose the Gradient-based local causal structures learning algorithm (GraN-LCS) based on the DAG constraint. Wren et al. [32] learn discrete DAG using such a constraint. Moreover, many methods combine NOTEARS with other domains. Zhu et al. [33] use reinforcement learning to recover the causal structure with the best scoring under the acyclicity constraint. Pamfil et al. [34] extend to recovering the causal structure on time-series data. Brouillard et al. [35] make use of interventional data for differentiable causal discovery. Wehenkel and Louppe [36] utilize the normalizing flow to learn the causal mechanism with the DAG constraint. Zeng et al. [37] extend to learning the causal structure on multidomain data. Faria et al. [38] learn causal structures under latent intervention based on DAG constraint. Ng and Zhang [39] propose a federated Bayesian network structure learning method using such a DAG constraint. Yang et al. [40] and Mao et al. [41] apply such a DAG constraint to learn the causal representation. Zhang et al. [42] apply DAG constraint to learn a forest-structured neural topic model to capture relationships between topics.

Our work is different from the works above. Many mentioned works [10], [11], [12], [28], [29], [30], [34], [39] are based on the least-square loss to orient causal direction. However, our analysis shows that it has no identifiability for causal discovery (see Section IV), and we propose a better entropy-based loss that identifies causal direction correctly (see Section V).

III. PRELIMINARY

A. Problem Formulation

We tackle the problems of multivariate causal discovery, aiming to learn a DAG \( \mathcal{G} \) that represents a joint distribution \( p(\mathbf{X}) \) over a set of random variables \( \mathbf{X} = \{X_1, X_2, \ldots, X_d\} \). A directed arrow from \( X_i \) to \( X_j \) represents the cause-effect relation between two variables, where we say \( X_i \) is the parent of \( X_j \). Let \( X_{\text{par}(i)} \) denote all direct parents of \( X_i \) with respect to a DAG \( \mathcal{G} \). The dependence relation between \( X_i \) and \( X_{\text{par}(i)} \) can be represented by a conditional distribution \( p(X_i | X_{\text{par}(i)}) \).

Given a set of observational data \( \{x_i^{(0)}, x_i^{(1)}, \ldots, x_i^{(0,m)}\}_{i=1}^d \), which is sampled from a joint distribution \( p(\mathbf{X}) \), we assume that the
joint distribution is Markov with respect to a ground truth DAG $G$ and can be factorized as $p(\mathbf{X}) = \prod_{i=1}^{d} p(X_i|X_{pa(i)})$. Then, the corresponding design matrix is $\mathbf{X} \in \mathbb{R}^{m \times d}$. In this article, we focus on the task of learning the underlying DAG given observational data. The mathematical notations and the corresponding descriptions are summarized in Table I.

B. Additive Noise Model

Causal discovery problems can be formalized by a SCM [43]. Given a set of random variables $\tilde{X} = \{X_1, X_2, \ldots, X_d\}$ and the corresponding noises $\mathcal{N} = \{N_1, N_2, \ldots, N_d\}$, SCM between $X_i$ and its direct parents $X_{pa(i)}$ with respect to a DAG $G$ is defined as $X_i = f_i(X_{pa(i)}, N_i)$, where $f_i$ could be a linear or nonlinear function. Following the previous works [10], we assume the causal sufficiency and consider the ANM:

$$X_i = f_i(X_{pa(i)}) + N_i$$

where $N_i$ is the exogenous additive noise and $N_i \not\perp X_{pa(i)}$. The identifiability of ANM depends on the asymmetry independence property, i.e., independence between the causes and the noise but not vice versa. Such an asymmetry greatly motivates our entropy-based loss for causal orientation in Section V.

C. NOTEARS

Zheng et al. [10] proposed a continuous optimization for learning causal structure, using least-square loss with an acyclicity constraint. In particular, the directed graph in a linear SCM can be encoded by a weighted adjacency matrix $W \in \mathbb{R}^{d \times d}$, i.e., $\mathcal{X} = \mathcal{W} \cdot \mathcal{X} + \mathcal{N}$. NOTEARS shows that $W$ represents a DAG if and only if $\text{tr}(e^{W \odot W}) - d = 0$ holds, where $\odot$ denotes the Hadamard product. Then, NOTEARS formulates the causal structure learning problem as the following continuous optimization problem:

$$\begin{align*}
\arg\min_{W \in \mathbb{R}^{d \times d}} & \frac{1}{2m} \| \mathbf{X} - \mathbf{XW} \|_F^2 + \lambda \| W \|_1 \\
\text{s.t.} & \text{tr}(e^{W \odot W}) - d = 0
\end{align*}$$

where $(1/2m)\| \mathbf{X} - \mathbf{XW} \|_F^2$ is the least-square loss and is equal, up to constant, to the log-likelihood score of a linear Gaussian DAG with equal noise variances, and $\| W \|_1$ denotes the $\ell_1$ penalty term on the causal structure.

D. Differential Entropy Estimator

To estimate the entropy in continuous optimization, we use the estimator proposed by Hyvärinen [44]. In detail, let $H(\mathbf{X}) = -\int p(\mathbf{X}) \log p(\mathbf{X}) d\mathbf{X}$ denote the entropy of variable $\mathbf{X}$. It gives a way to estimate entropy as follows:

$$H(\mathbf{X}) \approx H(\nu) - k_1 (E\{G_1(\nu)\})^2 + k_2 (E\{G_2(\nu)\} - E\{G_2(\nu)\})^2$$

where $k_1 = 36/(8 \sqrt{3} - 9)$, $k_2 = 24/(16 \sqrt{3} - 27)$, $G_1(\mathbf{X}) = X \exp(-X^2/2)$, $G_2(\mathbf{X}) = \exp(-X^2/2)$, $G_2(\nu) = \sqrt{1/2\pi}$, and $H(\nu) = (1/2)(1 + \log(2\pi))$. For more details, see [44].

IV. LIMITATIONS OF LEAST-SQUARE LOSS

This section discusses why the least-square loss-based methods fail to discover the underlying causal structure, and we theoretically show the conditions that they fail in both linear and nonlinear cases.

Without loss of generality, we analyze the causal pair of two variables and assume that

$$\begin{align*}
X &= N_X \\
Y &= \begin{cases} 
\alpha X + N_Y, \quad \text{Linear} \\
f(X) + N_Y, \quad \text{Nonlinear}
\end{cases}
\end{align*}$$

where the variance $\text{Var}(N_X) = \sigma_{N_X}^2$, $\text{Var}(N_Y) = \sigma_{N_Y}^2$, $N_X \not\perp N_Y$, and $f$ is a nonlinear function.

Without loss of generality, we denote the ANM in the anticausal direction as

$$\begin{align*}
Y &= \hat{N}_Y \\
X &= \begin{cases} 
\beta Y + \hat{N}_X, \quad \text{Linear} \\
g(Y) + \hat{N}_X, \quad \text{Nonlinear}
\end{cases}
\end{align*}$$

where the noises are not independent $\hat{N}_X \not\perp \hat{N}_Y$. Then, using the least-square loss, one may identify the causal direction by comparing the difference of least-square loss between $\text{LS}_{X \to Y}$ and $\text{LS}_{Y \to X}$ on the causal direction and anticausal direction, receptively, and taking the smaller one as the causal direction.

However, one of the main issues of using least-square loss for causal discovery is that the score in each direction can be easily manipulated by the causal strength or the variance of noises. As shown in Fig. 1, we run linear regression using the least-square loss 100 times in each experiment and show the difference of least-square loss between the causal direction and the anticausal direction with different causal strengths and variance of noises in linear Gaussian, Uniform, and Gumbel data, respectively. We can see that the least-square loss, even fails to identify the causal direction under the non-Gaussian distribution that is notably identifiable [7]. Thus, the least-square loss should not be a proper criterion for identifying the causal direction and a new criterion is urgently needed for methods like NOTEARS. To achieve this, we first show why the least-square loss is not a proper criterion in Sections IV-A and IV-B and propose a corresponding remedy in Section V.

A. Linear Case

In this section, we show that using the least-square loss, the identifiability of linear ANM can be fully determined by its causal strength and the noise variance.

To be clear, let $X \to Y$ denote the causal direction from cause $X$ to effect $Y$, and let $Y \to X$ denote the anticausal direction. Let $\sigma^2$ denote the variance of variable $X$. Let $\beta_{Y/X}$ be the linear regression coefficient obtained from regressing $Y$ on $X$, and let $\sigma_{Y/X}^2 = \text{Var}(Y - \beta_{Y/X} \cdot X)$ be the variance of the residual of regressing $Y$ on $X$. The least-square loss in causal direction $\text{LS}_{X \to Y}$ and in the anticausal direction $\text{LS}_{Y \to X}$ can be represented by

$$\begin{align*}
\text{LS}_{X \to Y} &= \mathbb{E}\left[ (X - 0)^2 \right] + \mathbb{E}\left[ (Y - \beta_{Y/X} \cdot X)^2 \right] \\
\text{LS}_{Y \to X} &= \mathbb{E}\left[ (X - \beta_{Y/X} \cdot Y)^2 \right] + \mathbb{E}\left[(Y - 0)^2 \right].
\end{align*}$$

By analyzing the square-least loss in different directions, we can conclude the following theorem.
The least-square loss will fail to identify the correct causal direction if the causal strength $\alpha$, the noise variances $\sigma^2_{N_X}$ and $\sigma^2_{N_Y}$ satisfy the following inequality:

$$\alpha^2 < 1 - \frac{\sigma^2_{N_Y}}{\sigma^2_{N_X}}. \tag{6}$$

**Proof:** For the causal direction, we have

$$\sigma^2_{X|\phi} = \sigma^2_{N_X}, \quad \beta_{Y|X} = \frac{\alpha \sigma^2_{X}}{\sigma^2_{X}} = \alpha, \quad \sigma^2_{Y|X} = \sigma^2_{N_Y}. \tag{7}$$

For the anticausal direction, we have

$$\sigma^2_{Y|\phi} = \alpha^2 \sigma^2_{N_X} + \sigma^2_{N_Y}, \quad \beta_{X|Y} = \frac{\alpha \sigma^2_{N_X}}{\sigma^2_{Y|\phi}} = \frac{\alpha \sigma^2_{X}}{\alpha^2 \sigma^2_{N_X} + \sigma^2_{N_Y}}$$

$$\sigma^2_{X|Y} = \text{Var}(X - \beta_{X|Y} Y) = \frac{\sigma^2_{X} \sigma^2_{N_Y}}{\alpha^2 \sigma^2_{N_X} + \sigma^2_{N_Y}}. \tag{8}$$

Thus, using (7), the least-square loss in the direction $X \rightarrow Y$ is given as follows:

$$\text{LS}_{X \rightarrow Y} = \mathbb{E}[(X - 0)^2] + \mathbb{E}[(Y - \beta_{Y|X} \cdot X)^2]$$

$$= \sigma^2_{N_X} + \sigma^2_{N_Y} \tag{9}$$

where $\beta_{Y|X}$ is the coefficient obtained from linear regression. Intuitively, the least-square loss is to calculate the sum of the residual’s variance, and therefore the loss in the causal direction is equal to the sum of noise variance. Similarly, using (8), we obtain

$$\text{LS}_{Y \rightarrow X} = \mathbb{E}[(X - \beta_{X|Y} Y)^2] + \mathbb{E}[(Y - 0)^2]$$

$$= \frac{\sigma^2_{N_X} \sigma^2_{N_Y}}{\alpha^2 \sigma^2_{N_X} + \sigma^2_{N_Y}} + \alpha^2 \sigma^2_{N_X} + \sigma^2_{N_Y}. \tag{10}$$

We can see that different from the least-square loss in the causal direction, the least-square loss in the anticausal direction has much more complicated terms, which is closely related to the causal strength and the noise variance. Thus, if we consider the condition that the loss in the causal direction $X \rightarrow Y$, is larger than the loss in the anticausal direction $Y \rightarrow X$, i.e., $\text{LS}_{X \rightarrow Y} > \text{LS}_{Y \rightarrow X}$, using (9) and (10), we have the following inequality:

$$\sigma^2_{N_X} + \sigma^2_{N_Y} > \frac{\sigma^2_{N_X} \sigma^2_{N_Y}}{\alpha^2 \sigma^2_{N_X} + \sigma^2_{N_Y}} + \alpha^2 \sigma^2_{N_X} + \sigma^2_{N_Y}$$

A similar result has been shown in [45] from the view of conditional variances, of which the main contribution is to prove the identifiability of ANM with unknown heterogeneous error variances.

**Theorem 1** indicates that, under the linear ANM assumption, the least-square loss fails to identify the correct causal direction unless the inequality (6) does not hold. It is also interesting to see that if all noises follow the standard Gaussian distribution, we have $\sigma^2_{N_X} = \sigma^2_{N_Y}$, then the causal direction can still be correctly identified because the inequality $\alpha^2 < 0$ must not hold. However, it is unrealistic that all noises follow the same standard Gaussian distribution in the real world. And a simple standardization that is usually used for preprocessing will invalidate the assumption.

Moreover, **Theorem 1** also explains the phenomenon in Fig. 1. For example, based on (9) and (10), given a fix $\sigma^2_{N_X} = 4$ and $\sigma^2_{N_Y} = 1$, then we have $\text{LS}_{X \rightarrow Y} = 5$ and $\text{LS}_{Y \rightarrow X} = (4/(4 \alpha^2 + 1)) + 4 \alpha^2 + 1$. As a result, the $\text{LS}_{Y \rightarrow X}$ will be even smaller than $\text{LS}_{X \rightarrow Y}$ if we have $|\alpha| < (\sqrt{3}/2) \approx 0.866$, and reaches its minimum $\text{LS}_{Y \rightarrow X} = 4$ when $|\alpha| = 0.5$. As a result, the causal direction will be misdirected as the anticausal direction has a smaller least-square loss.

Therefore, we conclude that the least-square loss is not a suitable score for learning linear causal structures.

**B. Nonlinear Case**

In this section, we show that using the least-square loss, the identifiability of ANM is closely related to the measured mutual information between the cause and regression residual in the nonlinear case.

Specifically, we study SCM between $X \rightarrow Y$ with the nonlinear additive form $Y = f(X) + N_Y$. To show the limitation of least-square loss in the nonlinear case, we aim to show that using the least-square loss can be viewed as minimizing the mutual information between the noise and cause variables but with bias such that the independence measure will be not reliable, leading to identifying the wrong direction, which is the main theoretical result of this section, i.e., **Theorem 2**. The proof of **Theorem 2** depends on Lemma 1, 2, and 3. To clearly show how these theories connect, we provide the proof sketch in Fig. 2.
Proof Sketch: First, we show that maximizing log-likelihood is equivalent to minimizing the mutual information between the cause and the noise variables (Lemma 1). Second, we further show that maximizing the likelihood under the standard Gaussian noise assumption is equivalent to minimizing the least-square loss (Lemma 2). Third, by combing Lemma 1 and Lemma 2, we conclude that using least-square loss, the mutual information will be biased (Lemma 3). Finally, by considering the bias in mutual information, we provide the condition that least-square loss fails to identify the causal direction (Theorem 2).

Lemma 1: Given the samples \( \{x^{(i)}, y^{(i)}\}_{i=1}^{m} \) and causal model \( Y = f(X; \theta) + N_Y \) with any parameters \( \theta \) and as \( m \to \infty \), the average log-likelihood \( I_{X \to Y}(\theta) \) and the mutual information between \( X \) and \( N_Y \) are related in the following way:

\[
I(X, N_Y; \theta) = \mathbb{E}_{x, y \sim p(x, y)}[\log p(X = x, Y = y)] - \lim_{m \to \infty} l_{X \to Y}(\theta) \tag{11}
\]

where

\[
l_{X \to Y}(\theta) = \frac{1}{m} \sum_{i=1}^{m} \log p(X = x^{(i)})
+ \frac{1}{m} \sum_{i=1}^{m} \log p(N_Y = y^{(i)} - f(x^{(i)}); \theta). \tag{12}
\]

The proof is in Appendix B. Lemma 1 indicates that in the pairwise ANM, maximizing the log-likelihood [see (12)] is equivalent to minimizing the mutual information [see (11)]. Next, we further build a connection between the log-likelihood and the least-square loss in the following lemma.

Lemma 2: For the ANM \( Y = f(X) + N_Y \), as \( m \to \infty \), maximizing the average log-likelihood \( I_{X \to Y}(\theta) \) with the standard Gaussian noise assumption

\[
l_{X \to Y}(\theta) = \frac{1}{m} \sum_{i=1}^{m} \log \left( \frac{1}{\sqrt{2\pi}} \exp \left( \frac{-(x^{(i)} - 0)^2}{2} \right) \right)
+ \frac{1}{m} \sum_{i=1}^{m} \log \left( \frac{1}{\sqrt{2\pi}} \exp \left( \frac{-(y^{(i)} - f(x^{(i)}; \theta))^2}{2} \right) \right) \tag{13}
\]

is equivalent to minimizing the least-square loss \( LS_{X \to Y} \)

\[
LS_{X \to Y} = \mathbb{E}[(X - 0)^2] + \mathbb{E}[(Y - f(X))^2]. \tag{14}
\]

The proof is in Appendix C. Lemma 2 indicates that under the standard Gaussian noise assumption, minimizing the log-likelihood [see (13)] is equivalent to minimizing the least-square loss [see (14)]. It implies that choosing the least-square loss as the objective function will introduce the standard Gaussian distribution in log-likelihood, causing a distribution mismatch when the underlying noise distribution is not the standard Gaussian. That is, the least-square loss will introduce bias to the mutual information according to the connection between log-likelihood and mutual information in Lemma 1. To be precise, we have the following lemma.

Lemma 3: For the ANM, as \( m \to \infty \), minimizing the least-square loss is equivalent to minimizing the mutual information under the standard Gaussian noise assumption with the following form:

\[
I_q(X, N_Y; \theta) = \mathbb{E}_{x, y \sim p(x, y)}[\log p(X = x, Y = y) - \log q(X = x)] - \log q(N_Y = y - f(x); \theta) \tag{15}
\]

where \( q \) is the density function of standard Gaussian distribution.

The proof is in Appendix D. Lemma 3 shows that the least-square loss is equivalence to the biased mutual information given in (15) which explains exactly why the least-square loss will fail to identify the correct causal direction. The reason is that based on the independence property of ANM, i.e., \( X \perp N_Y \) but \( Y \not\perp N_Y \), using the standard mutual information is able to capture such independence property and to identify the causal direction by testing whether the noise and cause are independent in the causal direction but not independent in the anticausal direction, i.e., \( I(Y; N_X) > I(X, N_Y) = 0 \). However, if the least-square loss is used and the underlying noise distribution is not standard Gaussian, the mutual information is biased as stated in Lemma 3 such that the independence property \( I_q(Y, N_X) > I_q(X, N_Y) \) does not necessarily hold. That is, in the following theorem, we show that when the condition \( I_q(Y, N_X) > I_q(X, N_Y) \) occurs, the least-square loss will fail to identify the causal direction.

Theorem 2: Let \( X \to Y \) be the causal direction following the data generation mechanism \( Y = f(X) + N_Y \) and we assume \( m \to \infty \). Using least-square loss, the causal direction is nonidentifiable if the following inequality holds:

\[
- \int p(X) \log q(X) dX - \int p(N_Y) \log q(N_Y) dN_Y
+ \int p(Y) \log q(Y) dY - \int p(\hat{N}_X) \log q(\hat{N}_X) d\hat{N}_X \tag{16}
\]

where \( q \) is the density function of standard Gaussian distribution.

The proof is in Appendix E. Theorem 2 gives the certain condition of no-identifiability of the least-square loss. The key reason why the least-square cannot correctly identify causal direction is the distribution mismatch between the underlying distribution \( p \) and standard Gaussian distribution \( q \). As shown in inequality 16, when the divergence between the standard Gaussian distribution \( q \) and the underlying distribution \( p(X) \) or \( p(N_Y) \) is large, the left-hand side will be larger than the right-hand side, making the causal direction misdirected using the least-square loss.

For example, let \( Y = \text{sigmoid}(X) + N_Y \) where \( X \sim \text{Uniform}(-C, C) \) and \( N_Y \sim \text{Uniform}(-1, 1) \). In this case, we can simply increase \( C \) such that \( - \int p(X) \log q(X) dX \) will tend to infinity while \( Y \) and \( \hat{N}_X \) are bounded due to the sigmoid function, such that the right-hand side of (16) is also
bounded, and hence the inequality must hold. In this case, the least-square loss will identify incorrectly the causal direction.

Therefore, it is necessary to use the correct distribution setting \( q = p \) to obtain the correct result, which inspires a way that uses the entropy-based loss instead.

V. Structure Learning Using Entropy-Based Loss

As discussed in Section IV, using least-square loss, which is equivalent to assuming the distribution of all noise terms is standard Gaussian distribution, will result in incorrect causal identification. The reason for these errors is the mismatch between the underlying noise distribution and the assumed standard Gaussian distribution.

Instead of using the square operator in least-square loss, we can replace it with the entropy as follows:

\[
\min_{N_i} \sum_{i=1}^{d} H(N_i) + \lambda \| W \|_1 \quad \text{s.t.} \quad \text{tr}(e^{W \cdot W}) - d = 0
\]  

(17)

where \( N_i \) denotes the independent noise and \( H(N_i) \) denotes the entropy of noise \( N_i \).

The intuition that we use entropy-based loss is shown in Fig. 3. Intuitively, the sum of entropy can be rewritten as follows \( H(N_X) + H(N_Y) = H(N_X, N_Y) + I(N_X, N_Y) \). As a result, since all residuals in the causal direction are independent of each other, we should have the lower sum of entropy under which \( I(N_X, N_Y) = 0 \). While the residuals in the anticausal direction depend on each other, we should have the larger sum of entropy as the mutual information \( I(N_X, N_Y) > 0 \).

To explain why we use entropy-based loss theoretically, we first show that the entropy-based loss is consistent with the log-likelihood score under the ANM (see Theorem 3). Based on Theorem 3, by utilizing identifiability results in terms of likelihod (see Lemma 4), it is easy to show that the entropy-based loss is also identifiable (see Corollary 1).

**Theorem 3:** In the ANM, the entropy-based score has a consistency with the log-likelihood score when the sample sizes \( m \to \infty \), i.e.,

\[
\lim_{m \to \infty} \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{d} \log p(\hat{x}_i^{(j)} \mid x_{pa(i)}) = -\sum_{i=1}^{d} H(N_i).
\]  

(18)

The proof is in Appendix F. Theorem 3 indicates that, under the ANM, the entropy-based loss is consistent with the log-likelihood score, while Peters et al. [46] had proven that for the ANM, the log-likelihood is able to distinguish the causal direction under a mild assumption, which is illustrated in the following lemma.

**Lemma 4 (Theorem 1 in [8]):** For the ANM \( Y = f(X) + N_Y \), there is a forward model of the form \( p(X, Y) = p(Y - f(X)) p(X) \). If there is a backward model of the same form \( p(X, Y) = p(X - g(Y)) p(Y) \), then for all \( X, Y \) with the three-time differentiable \( f \) and \( v''(Y - f(X)) f''(X) \neq 0 \), the following equality holds:

\[
\xi'' = \xi'' \left( -\frac{v'' f''}{v''} + \frac{f''}{f'} \right) - 2v'' f'' v' + \nu f'' + \nu v'' + \frac{1}{v'} \left( \frac{v''}{v'} \right)^2 (19)
\]

where \( v := \log p(N_Y) \), \( \xi := \log p(X) \).

Lemma 4 describes a necessary condition of the identifiability of ANM, i.e., if there exists a backward ANM with the independence property between cause and noise, the condition (19) will hold. In other words, if the condition (19) does not hold, there must not exist a backward ANM such that the cause and noise are independent and we can identify the causal direction using the log-likelihood score. Fortunately, such a condition only holds in restricted cases, e.g., the bivariate linear Gaussian noises model [46] and thus ANM can be identified in most cases.

Given backward ANM does not exist, i.e., (19) does not hold, we generalize the result in Lemma 4 to the following corollary by incorporating the result of Theorem 3, showing that the entropy-based loss is able to identify the causal direction.

**Corollary 1:** For each pair of ANM \( Y = f(X) + N_Y \), if \( v''(Y - f(X)) f''(X) \neq 0 \) and the condition in (19) does not hold, then using the entropy-based loss each pair of ANM is identifiable and the following inequality holds:

\[
H(X) + H(N_Y) < H(Y) + H(N_X).
\]  

(20)

The proof is in Appendix G. Corollary 1 shows the identifiability of entropy-based loss under ANM assumption. It indicates that the entropy of residuals is lower in the causal direction than in the anticausal direction, making the causal direction can be identified by entropy. The example in Fig. 3 also shows such an asymmetry of the entropy of residuals between the causal direction and the anticausal direction. The regression residuals in the causal direction [blue points in Fig. 3(a)] are independent, while the regression residuals in the anticausal direction [yellow points in Fig. 3(b)] are not independent. This independence property is characterized by the entropy of residuals, which is consistent with the inequality 20 in Corollary 1.

Moreover, for the task of causal discovery, Corollary 1 illustrates the validity of the entropy-based loss under the ANM, i.e., (17), which is based on the consistency with the likelihood score, indicating that by comparing the entropy of residuals between causal and anticausal directions, we can successfully identify the causal direction.

VI. EXPERIMENTS

To investigate the effectiveness of the entropy-based loss, we compare it with baseline algorithms on both synthetic and real-world data. The synthetic data are generated from linear pairwise SCM, linear non-Gaussian SCM, and nonlinear additive noise SCM, respectively. The following algorithms are taken as the baseline: NOTEARS [10], LiNGAM [7], Gradient-based optimization of dag-regularized likelihood for learning linear dag models with nonequal noise.
variances (GOLEM-NV) [28], gradient-based DAG learning method using the graph neural network (DAG-GNN) [11], nonlinearity extension of NOTEAR using a multilayer perceptron (NOTEARS-MLP) [13], and GraN-DAG [30]. We reuse the parameter settings for those baseline algorithms in their original papers and codes.

Our implementation, denoted by ours (linear) and ours-MLP (nonlinear), is based on the codes of NOTEARS [10] and NOTEARS-MLP [13], respectively. Following the previous works, the numerical optimize algorithms limited memory broyden–fletcher–goldfarb–Shanno algorithm (L-BFGS) [47] is used for optimization, and we will prune the edges after training with a small threshold $\omega = 0.3$ to rule out cycle-inducing edges.

In the linear pairwise SCM, the data are generated according to the following SCM: $X = N_X, Y = \alpha X + N_Y$ with causal strength $\alpha \in [0.5, 1.0]$ and the noise $N_X, N_Y \sim \sigma_X \ast \text{Gaussian}(0, 1)$ or $N_X, N_Y \sim \sigma_X \ast \text{Uniform}(-\sqrt{3}, \sqrt{3})$ or $N_X, N_Y \sim \sigma_X \ast \text{Gumbel}(0, (\sqrt{6}/\pi))$ with standard deviation $\sigma_X \in [0.6, 1.6]$.

In linear non-Gaussian case, the data are generated according to the following linear SCM: $X_i = \Sigma_{j \in \text{pa}(i)} \beta_{i,j} X_j + N_i$ with random causal strength $\beta_{i,j} \sim \text{Uniform}(-0.4, -0.8) \cup \text{Uniform}(0.4, 0.8)$ and the noise $N_i \sim \sigma_N \ast \text{Uniform}(-\sqrt{3}, \sqrt{3})$ or $N_i \sim \sigma_N \ast \text{Gumbel}(0, (\sqrt{6}/\pi))$ with random standard deviation $\sigma_N \sim \text{Uniform}(0.5, 1.0)$.

In the nonlinear case, the data are generated according to the following nonlinear SCM: $X_i = \text{tanh}(\Sigma_{j \in \text{pa}(i)} \beta_{i,j} X_j) + \cos(\Sigma_{j \in \text{pa}(i)} \beta_{2,i,j} X_j) + \sin(\Sigma_{j \in \text{pa}(i)} \beta_{3,i,j} X_j) + N_i$ with random causal strength $\beta_{k,i,j} \sim \text{Uniform}(0.5, 2.0)$ for $k = 1, 2, 3$ and the noise $N_i \sim \sigma_N \ast \text{Uniform}(-\sqrt{3}, \sqrt{3})$ with random standard deviation $\sigma_N \sim \text{Uniform}(0.5, 1.0)$.

Following previous works, structural Hamming distance (SHD), false discovery rate (FDR), and true-positive rate (TPR) are recorded as the evaluation metrics for all algorithms.

A. Pairwise Synthetic Data

In this section, we extend the experiments in Fig. 1, and design a series of controlled experiments with respect to the causal strength, the variance of cause, and the variance of effect. The number of variables is fixed at 2. At each experiment, we will control one of the parameters while fixing others. The range of the above controlled parameters are as follows: $\alpha = [0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$, $\sigma_X = [0.6, 0.8, 1.0, 1.2, 1.4, 1.6]$, and $\sigma_Y = [0.6, 0.8, 1.0, 1.2, 1.4, 1.6]$ where the default parameters are $\beta = 0.5, \sigma_X = 2$, and $\sigma_Y = 1$. All experiments will run 30 times.

As shown in Fig. 4, the least-square loss-based methods are sensitive to causal strength and variance of noise, which verifies Theorem 1 indicating that they fail when $\alpha^2 < 1 - (\sigma_N^2 / \sigma_X^2)$. NOTEARS and GOLEM-NV, based on least-square loss, only correctly orient causal direction when $\alpha^2 > 1 - (\sigma_N^2 / \sigma_X^2)$ regardless of noise’s distribution. DAG-GNN returns an empty adjacency matrix in many cases, and the reason might be that the autoencoder framework is overly complex and underfitting such a simple SCM. Based on the non-Gaussian property, our method and LiNGAM perform well in the Uniform and Gumbel cases and perform poorly in the Gaussian cases, which also verifies the Corollary 1. The experiments show that the entropy-based loss can correctly orient the causal direction with the help of non-Gaussian property, while the least-square-based loss orients causal direction by the inequality $\alpha^2 > 1 - (\sigma_N^2 / \sigma_X^2)$, which does not always hold in reality.

B. Synthetic Structures

In this section, we design a series of controlled experiments with respect to the sample sizes and the number of variables on the synthetic random causal structures. At each experiment, we will control one of the parameters while fixing others. The range of the above controlled parameters are as follows: the number of samples $N \in [200, 400, 600, 800, 1000]$ and the number of variables $V = [5, 10, 15, 20, 25, 50, 100]$ with in-degree $k = 2$ where the default setting is marked as bold. All experiments will run at least ten times.

1) Linear Case: In this part, we perform linear experiments on the synthetic linear non-Gaussian data. Here, we test our method on Uniform and the Gumbel distribution, respectively. Each algorithm runs 30 times for all experiments.

As shown in Fig. 5, our method outperforms all baseline methods on the non-Gaussian distributions experiments including Uniform and Gumbel distributions. The lower SHD means the causal structure recovered by our method is closer to the ground truth. The lower FDR and higher TPR mean we always recover more real edges without less misorientation. Moreover, by using the entropy-based method, we have the least deviation error, indicating the stability of our method. It is also interesting to see the significant difference between ours and NOTEARS, which verifies that the algorithm could learn the anticausal direction using least-square loss. In detail, in Fig. 5(a), as the sample size increases, all methods have better performances, and we can see that when the sample size is small, we still outperform the other gradient-based methods, which shows the robustness of our method. In Fig. 5(b) and (c), in general, the performance decreases as the number of variables increases, but compared with the baseline methods, our method decreases more slowly. Comparing Fig. 5(a) with (b) and (c), we find that LiNGAM
performs better in low-dimension cases, but its performance decreases rapidly with the increasing of the number of variables, which indicates that gradient-based methods are more robust to deal with high dimension cases. In addition, we can see that under different distributions, our method has similar performance within experimental errors, which verifies the results in Theorem 1.

2) Nonlinear Case: In this part, we test our entropy-based loss in nonlinear data. To further verify Theorem 3 and Corollary 1, we generate the nonlinear data using Uniform distribution with different variances. That is, the higher the variance, the more non-Gaussianity. In the controlled experiments, we further test our method on different noise variance $\{1, 2, 3, 4, 5\}$. Each algorithm will run at least ten times.

The results are given in Fig. 6. Overall, our method generally outperforms the baseline methods showing the robustness of our method in variable-varying and variance-varying cases. In detail, in Fig. 6(a), we notice that compared with the linear data, all methods require more samples to acquire a decent performance, but our method still outperforms the baseline methods. For our method, 600 is a decent choice for the sample sizes. In Fig. 6(b), with the growing number of variables, the performance of NOTEARS-MLP decreases rapidly while our method remains stable. The reason is that as the number of variables grows, the edges increase simultaneously, and the incorrectly identified edges will also increase. In the high-dimensional case, our method stays the lowest SHD, the lowest FDR, and the highest TPR, indicating that our method is less sensitive to the dimensional increase. In Fig. 6(c), we can see that our method is not sensitive to the noise variances while the performance of other methods decreases rapidly as the variances grow. The reason is that the variance controls the noise non-Gaussianity making it far from the standard Gaussian. Therefore, the entropy-based loss is stable and identifiable under any distribution while other methods will lose its identifiability.

C. Real-World Data

Following previous works, we consider the real-world dataset for the discovery of a protein signaling network.
Fig. 6. Structure recovery in nonlinear data in terms of SHD, FDR, and TPR to the true graph. (a) Noises are uniform with variance $= 3.0$ and the number of variables and edges are 15 and 30. (b) Noises are Uniform with variance $= 3.0$ and the number of sample sizes is 600. (c) Noises are Uniform, the number of sample sizes is 600 and the number of variables and edges are 15 and 30. (a) Uniform noises with variables $= 15$. (b) Uniform noises with samples size $= 600$. (c) Uniform noises with variables $= 15$.

| Parameter                  | Symbol | Value | Applicable to                                                                 |
|----------------------------|--------|-------|--------------------------------------------------------------------------------|
| threshold on $W$           | $\omega$ | 0.3   | ours, ours-MLP, NOTEARS, NOTEARS-MLP, GOLEM-NV                                |
| threshold on $W$           | $\omega$ | 0.4   | GraN-DAG                                                                      |
| acyclicity penalty         | $h(W)$ | $\text{tr}(e^{W+W}) - d$ | ours, ours-MLP, NOTEARS, NOTEARS-MLP, GOLEM-NV, GraN-DAG                    |
| acyclicity penalty         | $h(W)$ | $\text{tr}((I + \frac{W}{\gamma})^2 - d)$ | DAG-GNN                                                                       |
| $h$ tolerance              | $\epsilon$ | $10^{-8}$ | ours, ours-MLP, NOTEARS, NOTEARS-MLP, GOLEM-NV, DAG-GNN, GraN-DAG            |
| $h$ progress rate          | $c$    | 0.25  | ours, ours-MLP, NOTEARS, NOTEARS-MLP, DAG-GNN, GraN-DAG                      |
| initial Lagrange multiplier| $\alpha_0$ | 0     | ours, ours-MLP, NOTEARS, NOTEARS-MLP, DAG-GNN, GraN-DAG                      |
| $\rho$ increase factor     | $\rho$ | 10    | ours, ours-MLP, NOTEARS, NOTEARS-MLP, DAG-GNN, GraN-DAG                      |
| $\rho$ maximum             | $\rho$ | $10^{16}$ | ours, ours-MLP, NOTEARS-MLP                                                  |
| optimize algorithms        |        |       | DAG-GNN                                                                       |
| optimize algorithms        |        |       | ours, ours-MLP, NOTEARS, NOTEARS-MLP                                          |
| optimize algorithms        |        |       | GOLEM-NV, DAG-GNN                                                            |
| optimize algorithms        |        |       | GraN-DAG                                                                      |
| seed                       |        | 123   | all                                                                           |

based on expression levels of proteins and phospholipids [48]. This dataset is commonly used for causal structures learning work, with experimental annotations accepted by the biological research community. There are 11 variables and 853 observational samples in the dataset, and the ground-truth structure is provided by Sachs et al. [48] containing 17 edges. On this benchmark dataset, compared with other methods, ours-MLP achieves the best SHD 11 with 13 estimated edges. It shows that in the real-world situation our method still works better.
D. Algorithm Parameter Settings

The parameter settings of baseline algorithms follow their original papers and codes. Importantly, we use the least-squares loss \((1/2m)\|X - XW\|_F^2 + \lambda \|W\|_1\) for NOTEARS [10] and NOTEARS-MLP [13] regardless of the noise type. We use the negative ELBO as the objective function under standard Gaussian noise assumption for DAG-GNN [11]. For GOLEM-NV [28] and GraN-DAG [30], we use the likelihood-based loss with (multivariate) Gaussian assumption. For our and our-MLP, we use entropy-based loss as the objective function, and the parameter settings for our method are shown in Table II. Additionally, the code of NOTEARS and NOTEARS-MLP is available at https://github.com/xunzheng/notears and the code of DAG-GNN is available at https://github.com/fishmoon1234/DAG-GNN. The code of GraN-DAG is available at https://github.com/kurowasan/GraN-DAG and the code of GOLEM-NV is available at https://github.com/ignavieng/golem. For LiNGAM, we use the causal discovery toolbox packages [49]. Our code is available at https://github.com/DMIRLAB-Group/CausalDiscoveryBasedOnEntropy.

VII. CONCLUSION

In this work, we have reexamined NOTEARS with the least-square loss for learning DAGs in causal discovery. Our analysis shows that NOTEARS with the least-square loss is disabled to discover underlying causal structure under some weak assumptions, and the entropy-based loss is a proper replacement of the least-square loss in NOTEARS framework. We further provide the theoretical justification for the proposed method, by showing the consistency with the likelihood. Our experimental results validate our theoretical analysis and the effectiveness of the proposed entropy-based method. A clear next step is to generalize the theory and algorithms to the more general causal mechanism, e.g., the PNL causal model.

APPENDIX A

PROOF OF PROPOSITION 1

We present the property of ANM in the following proposition showing that \(p(X, Y) = p(X, N_Y)\) in the causal direction and \(p(X, Y) = p(Y, N_X)\) in the anticausal direction.

**Proposition 1 (Noise Representation Property):** For a bivariate model of variables \(X_1\) and \(X_2\), if the noise \(N\) is additive, i.e., \(X_2 = f(X_1) + N\), then there exists the following distribution transformations:

\[
p(X_1, X_2) = p(X_1, N).
\]

If the noise \(N\) is independent of \(X_1\), we further have

\[
p(X_2|X_1) = p(N = X_2 - f(X_1)).
\]

**Proof:** For the ANM \(X_2 = f(X_1) + N\), to avoid ambiguity, we consider the transformation from \(p(X_1, X_2)\) to \(p(X_1', N)\), where we additionally denote \(X'_1 = X_1\) for understandability. Specifically, the transformation can be expressed as follows:

\[
\begin{align*}
X'_1 &= X_1 \\
N &= X_2 - f(X_1).
\end{align*}
\]

Then, the distribution transformations are given as follows:

\[
p(X_1, X_2) = p(X'_1, N) |\det(J)| = p(X'_1, N)
\]

where \(J\) denotes the Jacobian matrix of the transformation from \((X_1, X_2)^T\) to \((X'_1, N)^T\), and we have

\[
|\det(J)| = \det \begin{pmatrix}
\frac{\partial X'_1}{\partial N} & \frac{\partial X'_1}{\partial X_2} \\
\frac{\partial N}{\partial X_1} & \frac{\partial N}{\partial X_2}
\end{pmatrix}
\]

\[
= \det \begin{pmatrix}
\frac{\partial X_1}{\partial X_1} & \frac{\partial X_1}{\partial X_2} \\
\frac{\partial (X_2 - f(X_1))}{\partial X_1} & \frac{\partial (X_2 - f(X_1))}{\partial X_2}
\end{pmatrix}
\]

\[
= \det \begin{pmatrix}
1 & 0 \\
-\frac{\partial f(X_1)}{\partial X_1} & 1
\end{pmatrix}
\]

\[
= 1
\]

(A.5)

where the second equality is obtained by substituting (A.3) into the first equality.

Hence, substituting (A.5) into (A.4), we conclude

\[
p(X_1, X_2) = p(X'_1, N) = p(X_1, N).
\]

If the noise \(N\) is independent of \(X_1\), we further have

\[
p(X_1)p(X_2|X_1) = p(X_1)p(N)\tag{A.7}
\]

and by eliminating \(p(X_1)\) in both sides, we have

\[
p(X_2|X_1) = p(N)\tag{A.8}
\]

which finishes the proof.

**APPENDIX B

PROOF OF LEMMA 5

In the following lemma, we bridge a connection between the log-likelihood and mutual information for the ANM. The proof of Lemma 5 is based on the definition of mutual information.

**Lemma 5:** Given the samples \(x^{(i)}, y^{(i)}, i = 1, \ldots, m\) and causal model \(Y = f(X; \theta) + N_Y\) with any parameters \(\theta\) and as \(m \to \infty\), the average log-likelihood \(l_{X\to Y}(\theta)\) and the mutual information between \(X\) and \(N_Y\) are related in the following way:

\[
l(X, N_Y; \theta) = \mathbb{E}_{x,y \sim p(X,Y)} \left[ p(X = x, Y = y) \right] - \lim_{m \to \infty} l_{X\to Y}(\theta) \tag{B.1}
\]

where

\[
l_{X\to Y}(\theta) = \frac{1}{m} \sum_{i=1}^{m} \log p(x^{(i)}) + \frac{1}{m} \sum_{i=1}^{m} \log p(N_Y = y^{(i)} - f(x^{(i)}; \theta)).\tag{B.2}
\]

**Proof:** The average log-likelihood \(l_{X\to Y}(\theta)\) has the following form:

\[
l_{X\to Y}(\theta)
\]

\[
= \frac{1}{m} \sum_{i=1}^{m} \log p(X = x^{(i)}, Y = y^{(i)}; \theta)
\]

\[
= \frac{1}{m} \sum_{i=1}^{m} \log p(X = x^{(i)}, N_Y = y^{(i)} - f(x^{(i)}; \theta))
\]
### TABLE III

| Metric               | Control Param | Value | ours | NOTIEARS | GOLEM-NV | DAG-GNN | LINGAM |
|----------------------|---------------|-------|------|----------|----------|---------|--------|
|                      |               |       | mean | + std    | mean | + std | mean | + std | mean | + std |
| sample size          |               |       | 200  | 8.233 ± 5.219 | 14.2000 ± 4.9288 | 12.0000 ± 5.4160 | 29.7333 ± 8.1376 | 18.666 ± 7.9440 |
|                      |               |       | 400  | 7.0333 ± 7.3977 | 13.3000 ± 6.2779 | 10.1333 ± 6.3905 | 23.0000 ± 6.6812 | 2.7000 ± 4.5891 |
|                      |               |       | 600  | 6.0333 ± 7.935 | 12.9333 ± 8.5327 | 9.7333 ± 8.3527 | 20.3333 ± 8.9703 | 9.0000 ± 1.1357 |
|                      |               |       | 800  | 4.8666 ± 4.3568 | 12.5333 ± 5.2582 | 10.6333 ± 6.6557 | 20.6000 ± 6.0310 | 8.3533 ± 1.0027 |
|                      |               |       | 1000 | 3.8000 ± 4.6000 | 15.3666 ± 6.1614 | 13.1333 ± 6.3560 | 21.1333 ± 8.8615 | 7.6666 ± 0.7608 |
| SHD number of variable |           |       | 3    | 0.6333 ± 1.8458 | 2.8333 ± 4.8545 | 1.5333 ± 3.5997 | 7.7666 ± 1.0765 | 0.0000 ± 0.0000 |
|                      |               |       | 10   | 2.5666 ± 6.0800 | 8.6666 ± 3.8444 | 6.3666 ± 3.8428 | 13.4333 ± 3.9045 | 0.5666 ± 0.4674 |
|                      |               |       | 15   | 2.5666 ± 6.0800 | 8.6666 ± 3.8444 | 6.3666 ± 3.8428 | 13.4333 ± 3.9045 | 0.5666 ± 0.4674 |
|                      |               |       | 20   | 3.5666 ± 6.0800 | 8.6666 ± 3.8444 | 6.3666 ± 3.8428 | 13.4333 ± 3.9045 | 0.5666 ± 0.4674 |
|                      |               |       | 50   | 5.1000 ± 4.3778 | 18.4333 ± 7.3551 | 19.8333 ± 7.9200 | 33.6666 ± 4.3288 | 3.9333 ± 3.8981 |
|                      |               |       | 100  | 7.5666 ± 4.6695 | 20.9000 ± 7.0349 | 40.4667 ± 7.9103 | 62.6666 ± 4.9210 | 105.9000 ± 17.6943 |
|                      |               |       | 150  | 13.9000 ± 7.5024 | 35.4333 ± 10.7776 | 92.2000 ± 19.0723 | 135.7000 ± 34.8072 | 216.1666 ± 21.0223 |

The mutual information between $X$ and $N_Y$ can be represented by

$$
I(X, N_Y; \theta) = \text{E}_{x, y \sim p(X, Y)} \left[ \log \left( \frac{p(X = x, N_Y = y | f(x); \theta)}{p(X = x) p(N_Y = y | f(x); \theta)} \right) \right]
$$

where the second equality is based on Proposition 1 and the third equality is based on the independence between $X$ and $N_Y$. The mutual information between $X$ and $N_Y$ can be represented by

$$
= \frac{1}{m} \sum_{i=1}^{m} \log p(X = x^{(i)})
+ \frac{1}{m} \sum_{i=1}^{m} \log p(N_Y = y^{(i)} - f(x^{(i)}); \theta)
$$

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
\[ -\log p(X = x) - \log p(N_Y = y - f(x); \theta) = E_{X,Y \sim p(X,Y)}[\log p(X = x, Y = y)] - \lim_{m \to \infty} I_{X \to Y}(\theta) \]

where the third equality is based on Proposition 1 and the fourth equality is obtained by substituting (B.3).

**APPENDIX C**

**PROOF OF LEMMA 2**

In the following lemma, we bridge a connection between the log-likelihood and least-square loss for the ANM.

**Lemma 6:** For the ANM \( Y = f(X) + N_Y \), as \( m \to \infty \), maximizing the average log-likelihood \( I_{X \to Y}(\theta) \) with the standard Gaussian noise assumption

\[ I_{X \to Y}(\theta) = \frac{1}{m} \sum_{i=1}^{m} \log \left( \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{(x^{(i)} - 0)^2}{2} \right) \right) + \frac{1}{m} \sum_{i=1}^{m} \log \left( \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{(y^{(i)} - f(x^{(i)}; \theta))^2}{2} \right) \right) \]

is equivalent to minimizing the least-square loss \( L_{S_{X \to Y}} \)

\[ L_{S_{X \to Y}} = \mathbb{E}[ (X - \theta)^2 ] + \mathbb{E}[ (Y - f(X))^2 ] \]

**Proof:** Based on Lemma 5, as \( m \to \infty \), the average log-likelihood under the standard Gaussian distribution has
the following form:

$$\lim_{m \to \infty} I_{X \to Y} (\theta) = \lim_{m \to \infty} \left( \frac{1}{m} \sum_{i=1}^{m} \log q (X = x^{(i)}) + \frac{1}{m} \sum_{i=1}^{m} \log q (N_Y = y^{(i)} - f (x^{(i)}; \theta)) \right)$$

$$= E_{x,y \sim p(X,Y)} [\log q (X = x) + \log q (N_Y = y - f (X; \theta))]$$

$$= E_{x,y \sim p(X,Y)} \left[ \log \left( \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{(X - 0)^2}{2} \right) \right) \right. + \left. \log \left( \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{(Y - f(X))^2}{2} \right) \right) \right]$$

$$= -\frac{1}{2} (E_{x,y \sim p(x,y)} [(X - 0)^2 + (Y - f(X))^2] + \text{const})$$

$$= -\frac{1}{2} (L_{S_{x \to y}} + \text{const}) \quad \text{(C.3)}$$

where $q$ is the density of the standard Gaussian distribution, and the third equality is obtained by substituting the standard Gaussian distribution into $q$, and in the fourth equality, $\text{const} = \log 4\pi$.

**APPENDIX D**

**PROOF OF LEMMA 3**

In the following lemma, we bridge a connection between the least-square loss and mutual information for the ANM. Based on the following lemma, we can utilize the biased mutual information between the cause and the noise to analyze the least-square loss. The proof of Lemma 3 is based on Lemma 5 and Lemma 2.

**Lemma 7:** For the ANM, as $m \to \infty$, minimizing the least-square loss is equivalent to minimizing the mutual information under the standard Gaussian noise assumption with the following form:

$$I_q (X, N_Y; \theta) = E_{x,y \sim p(x,y)} [\log p (X = x, Y = y) - \log q (X = x) - \log q (N_Y = y - f (x); \theta)] \quad \text{(D.1)}$$

where $q$ is the density function of standard Gaussian distribution.

**Proof:** Let $q$ denote the density function of standard Gaussian distribution. By substituting $q$ into (B.1) in Lemma 5, the mutual information under the standard Gaussian noise assumption can be rewritten as

$$I_q (X, N_Y; \theta) = E_{x,y \sim p(x,y)} [\log p (X = x, Y = y) - \log q (X = x) - \log q (N_Y = y - f (x); \theta)]$$

$$= E_{x,y \sim p(x,y)} [\log p (X = x, Y = y)] + \frac{1}{2} (L_{S_{x \to y}} + \text{const})$$

where the second equality is based on (C.3) in Lemma 2. \hfill \square

**APPENDIX E**

**PROOF OF THEOREM 2**

In the following theorem, we establish a certain condition under which the least-square loss cannot identify the causal direction, indicating the least-square loss is not a suitable loss for the task of causal discovery. The proof of Theorem 2 is based on Lemma 3.

**Theorem 2:** Let $X \to Y$ be the causal direction following the data generation mechanism $Y = f (X) + N_Y$ and we assume $m \to \infty$. The causal direction is nonidentifiable using least-square loss if the following inequality holds:

$$- \int p(X) \log q (X) dX - \int p(N_Y) \log q (N_Y) dN_Y > \int p(\hat{N}_Y) \log q (\hat{N}_Y) d\hat{N}_Y \quad \text{(E.1)}$$

where $q$ is the density function of standard Gaussian distribution.

**Proof:** Without loss of generality, we assume the identifiability condition of the ANM holds. Then, we have $I (X, N_Y) = I (Y, N_Y)$. However, based on (D.1) of Lemma 3 as $m \to \infty$, using the least-square loss, the mutual information becomes

$$I_q (X, N_Y) = E_{x,y \sim p(x,y)} [\log p (X = x, Y = y) - \log q (X = x) - \log q (N_Y = y - f (x))]$$

$$= E_{x,y \sim p(x,y)} [\log p (X = x, Y = y) - \log q (\hat{N}_Y = x - g(y))]$$

$$= E_{x,y \sim p(x,y)} [\log q (\hat{N}_Y) + \log q (\hat{N}_X)]$$

where $q \sim N (0, I)$. In this case, the inequality $I_q (X, N_Y) < I_q (Y, N_X)$ does not necessarily hold. In fact, by solving the inequality $I_q (X, N_Y) > I_q (Y, N_X)$, we obtain

$$E_{p(X)} [- \log q (X)] + E_{p(X,Y)} [- \log q (N_Y)]$$

$$> E_{p(Y)} [- \log q (Y)] + E_{p(X,Y)} [- \log q (\hat{N}_X)] \quad \text{(E.2)}$$

Note that, the expectation of $N_Y$ can be rewritten as

$$E_{p(X,Y)} [- \log q (N_Y)] = E_{p(X,N_Y)} [- \log q (N_Y)] = E_{p(Y)} [- \log q (N_Y)]$$

in which the first equality holds due to Proposition 1 (see Appendix A), and the second equality is obtained by integrating $X$. Similarly, we have

$$E_{p(X,Y)} [- \log q (\hat{N}_X)] = E_{p(\hat{N}_X)} [- \log q (\hat{N}_X)] \quad \text{(E.4)}$$

By substituting (E.3) and (E.4) into inequality (E.2), we obtain inequality (E.1). \hfill \square

**APPENDIX F**

**PROOF OF THEOREM 3**

In the following theorem, we show that, for the ANM, maximizing the log-likelihood score is equivalent to maximizing the entropy of noise. It inspires our entropy loss-based method for causal discovery.

**Theorem 3:** In the ANM, the entropy-based score has a consistency with the log-likelihood score when the sample sizes $m \to \infty$, i.e.,

$$\lim_{m \to \infty} \frac{1}{m} \sum_{j=1}^{d} \sum_{i=1}^{d} \log p (x^{(j)}_i | x^{(j)}_{\text{par}(i)}) = - \sum_{i=1}^{d} H (N_i).$$

**Proof:** Based on Lemma 5, we generalize the log-likelihood from two variables case to the multivariate case.
using Proposition 1 as follows:

\[
\frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{d} \log p\left(x_i^{(j)} | x_{\text{pa}(i)}^{(j)}\right) = \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{d} \log p\left(N_i = x_{\text{pa}(i)}^{(j)} - f_i\left(x_i^{(j)}\right)\right).
\]

Then, as \(m \to \infty\), we have

\[
limit_{m \to \infty} \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{d} \log p\left(N_i = x_{\text{pa}(i)}^{(j)} - f_i\left(x_i^{(j)}\right)\right) = - \sum_{i=1}^{d} \mathbb{E}[- \log p(N_i)] = - \sum_{i=1}^{d} H(N_i).
\]

**APPENDIX G**

**PROOF OF COROLLARY 2**

In the following corollary, we show that, for the ANM, the entropy-based loss can identify the causal direction correctly. It promises the correctness of our method. The proof of Corollary 2 is based on the independence property of ANM, i.e., \(I(X, N_Y) < I(Y, N_X)\).

**Corollary 2:** For each pair of ANM \(Y = f(X) + N_Y\), if \(v^Y - f(X)\)'s \(f(X)\) does not hold, then using the entropy-based loss pair of ANM is identifiable and the following inequality holds:

\[
H(X) + H(N_Y) < H(Y) + H(N_X).
\]  

**Proof:** To prove Corollary 2, we will compare the mutual information between the hypothetical cause and the regression residual in causal and anticausal directions. Based on Lemma 4, if (19) does not hold, i.e., no backward ANM model exists, then we must have the following equation:

\[
p(X, Y) = p(N_Y)p(X) \quad p(X, Y) \neq p(N_X)p(Y)
\]

from the causal direction and anticausal direction, respectively. It means that the noise and the hypothetical cause in the anticausal direction are not independent of each other, i.e., \(N_Y \perp \!
\!
\perp X\) in the causal direction but \(N_X \perp \!
\!
\perp Y\) in the anticausal direction. Such a property can be represented by mutual information as follows:

\[
I(X, N_Y) = 0 \quad \text{and} \quad I(Y, N_X) > 0.
\]

Hence, \(I(X, N_Y) < I(Y, N_X)\), i.e.,

\[
\int p(X, N_Y) \log \frac{p(X, N_Y)}{p(X)p(N_Y)} dXdN_Y < \int p(Y, N_X) \log \frac{p(Y, N_X)}{p(Y)p(N_X)} dYdN_X.
\]

By rewriting the inequality above using entropy, we have

\[
H(X) + H(N_Y) + \mathbb{E} \log p(X, N_Y) < H(Y) + H(N_X) + \mathbb{E} \log p(Y, N_X).
\]  

By utilizing Proposition 1, we have

\[
p(X, Y) = p(X, N_Y) = p(Y, N_X).
\]

By substituting (G.3) into (G.2), we obtain

\[
H(X) + H(N_Y) + \mathbb{E} \log p(X, N_Y) < H(Y) + H(N_X) + \mathbb{E} \log p(Y, N_X)
\]

and by eliminating the same term of both sides, we obtain

\[
H(X) + H(N_Y) < H(Y) + H(N_X)
\]

which finishes the proof.

**APPENDIX H**

**ADDITIONAL EXPERIMENTAL DETAILS**

For a better presentation of the experimental results, we provide the precise values of Figs. 5 and 6 in Tables III and IV, respectively.

**REFERENCES**

[1] E. Ghysels, J. B. Hill, and K. Motegi, “Testing for Granger causality with mixed frequency data,” *J. Econometrics*, vol. 192, no. 1, pp. 207–230, May 2016.

[2] M. Grosse-Wentrup, D. Janzing, M. Siegel, and B. Schölkopf, “Identification of causal relations in neuroimaging data with latent confounders: An instrumental variable approach,” *NeuroImage*, vol. 125, pp. 825–833, Jan. 2016.

[3] J. Liu, J. Ji, G. Xun, and A. Zhang, “Inferring effective connectivity networks from fMRI time series with a temporal entropy-score,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5993–6006, Oct. 2022.

[4] R. Cai, Z. Zhang, Z. Hao, and M. Winslett, “Understanding social causalities behind human action sequences,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 8, pp. 1801–1813, Aug. 2017.

[5] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman, *Causation, Prediction, and Search*. Cambridge, MA, USA: MIT Press, 2000.

[6] S. A. Andersson, D. Madigan, and M. D. Perlman, “A characterization of Markov equivalence classes for acyclic digraphs,” *Ann. Statist.*, vol. 25, no. 2, pp. 505–541, Apr. 1997.

[7] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan, “A linear non-Gaussian acyclic model for causal discovery,” *J. Mach. Learn. Res.*, vol. 7, no. 10, pp. 2003–2030, 2006.

[8] P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, “Nonlinear causal discovery with additive noise models,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 21, 2008.

[9] K. Zhang and A. Hyvarinen, “On the identifiability of the post-nonlinear causal model,” in *Proc. 25th Conf. Uncertainty Artif. Intell.*, Jun. 2009, pp. 647–655.

[10] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, “Dags with no tears: Continuous optimization for structure learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 9472–9483.

[11] Y. Yu, J. Chen, T. Gao, and M. Yu, “DAG-GNN: DAG structure learning with graph neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7154–7163.

[12] I. Ng, S. Zhu, Z. Chen, and Z. Fang, “A graph autoencoder approach to causal structure learning,” 2019, arXiv:1911.07420.

[13] X. Zheng, C. Dan, B. Aragam, P. Ravikumar, and E. Xing, “Learning sparse nonparametric dags,” in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 3414–3425.

[14] P. Blöbaum, D. Janzing, T. Washio, S. Shimizu, and B. Schölkopf, “Analysis of cause-effect inference by comparing regression errors,” *PeerJ Comput. Sci.*, vol. 5, e169, Jan. 2019.

[15] J. Peters and P. Bühlmann, “Identifiability of Gaussian structural equation models with equal error variances,” *Biometrika*, vol. 101, no. 1, pp. 219–228, Mar. 2014.

[16] P.-L. Loh and P. Bühlmann, “High-dimensional learning of linear causal networks via inverse covariance estimation,” *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3065–3105, Jan. 2014.

[17] J. Pearl and T. S. Verma, “A theory of inferred causation,” *Stud. Logica*, vol. 134, pp. 789–811, Jan. 1995.

[18] K. Yu, M. Cai, X. Wu, L. Liu, and J. Li, “Multilabel feature selection: A local causal structure learning approach,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 6, pp. 3044–3057, Jun. 2023.

[19] K. Yu, L. Liu, J. Li, and H. Chen, “Mining Markov blankets without causal sufficiency,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6333–6347, Dec. 2018.

[20] S. Shimizu et al., “DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model,” *J. Mach. Learn. Res.*, vol. 12, pp. 1225–1248, Apr. 2011.
