Causal Link Discovery with Unequal Edge Error Tolerance

Joni Shaska
University of Southern California
shaska@usc.edu

Urbashi Mitra
University of Southern California
ubli@usc.edu

Abstract—This paper proposes a novel framework for causal discovery with asymmetric error control, called *Neyman-Pearson causal discovery*. Despite the importance of applications where different types of edge errors may have different importance, current state-of-the-art causal discovery algorithms do not differentiate between the types of edge errors, nor provide any finite-sample guarantees on the edge errors. Hence, this framework seeks to minimize one type of error while keeping the other below a user-specified tolerance level. Using techniques from information theory, fundamental performance limits are found, characterized by the Rényi divergence, for Neyman-Pearson causal discovery. Furthermore, a causal discovery algorithm is introduced for the case of linear additive Gaussian noise models, called $\epsilon - CUT$, that provides finite-sample guarantees on the false positive rate, while staying competitive with state-of-the-art methods.

Index Terms—causal inference, finite-sample results, error control

I. INTRODUCTION

Understanding the underlying causes of phenomena affected by multiple variables can often be done via the representation of causal graphs [1]. These graphs are often assumed to be directed acyclic graphs. Applying causal graph discovery can have utility in disciplines as diverse as topology inference in wireless networks [2], gene networks in biology (e.g. [3]), the impact of medications, and optimizing the effect of advertising [4]. As a result, there is a wide range of research regarding causal graph discovery, such as constraint-based methods like the PC and fast-PC algorithms [5], [6], score-based methods [7]–[12], and discovery methods for time-series data and state-spaces [13]–[16]. However, despite modern applications in which one type of error is less favorable, most algorithms, with only a few notable exceptions [11], [14], do not distinguish between different types of edge errors. Furthermore, most recoverability results in the literature are asymptotic in the number of samples and do not provide any guarantees on error rates in a finite sample setting.

Hence, the purpose of the paper is two-fold. First, we introduce *Neyman-Pearson causal discovery*, which seeks to minimize one type of edge error while keeping the other below a fixed user-specified threshold. Second, we develop

an algorithm for the case of linear additive Gaussian noise models called $\epsilon - CUT$, which stands for *Causal discovery with Unequal edge error Tolerance*, which keeps the false positive rate below a user-specified tolerance level for any number of samples.

Although work on confidence regarding causal estimates has recently started to emerge, much of this work either focuses on confidence regarding causal effects (rather than direct edges) between nodes or focuses on time-series data. For instance, [17] derives performance limits for causal effect estimation on individuals, rather than populations and [18] constructs confidence intervals for causal effects in the linear additive Gaussian setting. For time-series data, [19] derives converse bounds for the achievable false positive and false negative rates, and [14] derives consistency conditions based on a false connection score, similar to our false positives rate.

In contrast, our work deals specifically with the error rates for individual edges between nodes. In particular, we formulate an optimization problem that seeks to minimize the false negative rate of the declared edges, while keeping the false positive rate below a user-specified threshold. We derive fundamental performance limits for our framework and show that our proposed algorithm, $\epsilon - CUT$, satisfies the false-positive constraint for any number of samples.

To provide finite-sample guarantees on the false positive rate, we consider the setting of linear structure equation models (LSEMs) with additive Gaussian noise, which has been extensively studied in the literature [10], [18], [20], [21]. Prior algorithms in this setting, such as AReCI [22] and ReCIT [23], rely on *the principle of independent mechanisms* [6], which assumes that the causes and mechanisms producing the effect (i.e., the specific functional form) are independent. Then, [22] and [23] provide independence tests by comparing estimation residuals. However, these algorithms only provide asymptotic guarantees on recoverability, which do not hold in our finite-sample setting. Our algorithm also considers the difference between residuals but differs in two important ways. The first key observation is the variance (or energy) of a child node is always higher than that of its parent, and the second is that the residuals follow a chi-squared distribution, enabling us to derive straightforward threshold tests that facilitate finite sample analysis and yield performance guarantees.

Our algorithm also differs from traditional methods typically using a sparsity constraint or penalty. For example, [10] uses the ℓ_0 penalty for linear additive Gaussian models, [15]

enforces sparsity through a Bayesian prior based on Gaussian processes in a time-series setting, and [14] uses a combination of ℓ_1 and ℓ_2 regularizers for inference in multivariate autoregression models. Continuous optimization formulations, such as NOTEARS [7], DAGMA [8], and DAGMA-DCE [9], will typically invoke sparsity through either the ℓ_1 or ℓ_2 regularizers. Hence, the methods often require a hyper-parameter that must be selected, and other than a few notable exceptions such as [11], there is typically no understanding of how different hyper-parameter values will affect the error rate *a priori*. In contrast, our work requires no tuning of hyper-parameters. Once the user has specified a tolerance level, all parameters in our algorithm are defined, along with *a priori* guarantees on the false positive rate for any number of given samples.

Our contributions are as follows:

1) We pose edge detection in causal discovery as a collection of Neyman-Pearson problems. An aggregated false negative rate is minimized subject to a pre-specified tolerance on the aggregated false positive rate. Aggregation is over all edges in the causal graph.
2) We derive the optimal detector for our framework and use it to derive upper and lower bounds on the achievable false negative rate for a given tolerance on the false positive rate (the optimal detector and bounds were given in [24] but without proof). Hence, we derive fundamental performance limits for our framework characterized by the Rényi divergence.
3) We derive an algorithm for the setting of linear additive Gaussian noise models that seeks to minimize the false negative rate while keeping the false positive rate below a pre-specified tolerance level ϵ. We call this method ϵ-CUT.
4) We show ϵ-CUT's false positive rate will always satisfy the user-specified tolerance (the proof is absent in [25]). This result requires no asymptotic assumptions on the convergence of distributions or consistency conditions and hence, is a finite-sample result. We investigate the performance of ϵ-CUT compared to other causal discovery methods through a series of numerical experiments.

The rest of this paper is organized as follows. Section II introduces the setup and problem formulation, Section III presents the optimal detector and fundamental performance limits for Neyman-Pearson causal discovery. Section IV introduces $\epsilon - CUT$. Section V briefly summarized state-of-the-art methods that are used as benchmarks for the numerical studies. Section VI presents numerical results, and Section VII concludes the paper. Much of our notation mimics that used in [19], [24] and [26].

II. PROBLEM FORMULATION

Consider a *directed acyclic graph* (DAG) $\mathcal{G}$ with edge set $\mathcal{E}$, vertex set $\mathcal{V}$ with $|\mathcal{V}| = d$, and corresponding weighted adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$. We assume a prior on the adjacency matrix, denoted by $\pi_{\boldsymbol{A}}$. We have a series of n
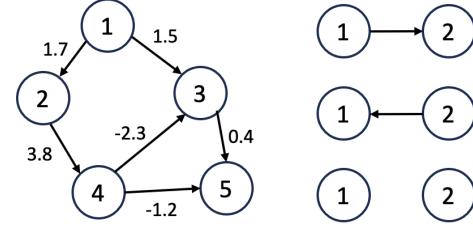

Fig. 1: (L) An example, five-node, directed, acyclic graph with edge weights. (R) Two-node graphs corresponding to the matrices $\boldsymbol{A}_0$, $\boldsymbol{A}_1$, and $\boldsymbol{A}_2$, respectively.

observations $\{X_k\}_{k=1}^n$ generated by the underlying distribution $P_{\boldsymbol{A}}$ conditioned on the realization of $\boldsymbol{A}$. Given a vertex $i \in \mathcal{V}$, let $X_k(i)$ be the k-th measurement $k = 1, .., n$ of vertex i. Moreover, let $\mathcal{Z}(i)$ denote the parent set of vertex i. Assuming $\mathcal{Z}(i)$ is non-empty, let i_j be the j-th parent of vertex i, $j = 1, ..., |\mathcal{Z}(i)|$. Define the vector

$$Z_k(i) = [X_k(i_1), X_k(i_2), ..., X_k(i_{|\mathcal{Z}(i)|})]^\top, \qquad (1)$$

i.e., the vector comprised of the k-th measurements of the parents of i. Moreover, we assume a linear structural equation model (LSEM), i.e.,

$$X_k = \boldsymbol{A}X_k + W_k, \qquad (2)$$

where the exogenous input terms $\{W_k\}_{k=1}^n$ are zero-mean indenpendent Gaussian random variables with equal covariance σ^2 (which we assume is known). This assumption is commonly made in the literature [10], [11], [18], [19], and it is also worth underscoring that if one removes either the equivariance or the independence assumption, recoverability of $\boldsymbol{A}$ is no longer guaranteed even in the asymptotic case [10]. Given the underlying adjacency matrix $\boldsymbol{A}$, the goal is to recover the *support* $\boldsymbol{\chi}$ of $\boldsymbol{A}$, where $\chi_{i,j} = 1 \iff \boldsymbol{A}_{i,j} \neq 0$ or $\boldsymbol{A}_{j,i} \neq 0$.

A. Error Metrics

We define the following error rates [1] which were first introduced in [26] and further studied in [19], [24],

$$\textit{false positive rate} : \epsilon^+ = \frac{\mathbb{E} \sum_{i,j} \mathbb{1}\{\hat{\boldsymbol{\chi}}_{i,j} = 1, \chi_{i,j} = 0\}}{\mathbb{E} \sum_{i,j} \mathbb{1}\{\chi_{i,j} = 0\}}, \quad (3)$$

$$\textit{false negative rate} : \epsilon^- = \frac{\mathbb{E} \sum_{i,j} \mathbb{1}\{\hat{\boldsymbol{\chi}}_{i,j} = 0, \chi_{i,j} \neq 0\}}{\mathbb{E} \sum_{i,j} \mathbb{1}\{\chi_{i,j} \neq 0\}}, \quad (4)$$

where the expectations are taken with respect to detector $\hat{\chi}$ as well as the prior distribution on the matrix $\boldsymbol{A}$, $\pi_{\boldsymbol{A}}$. We wish to find the detector $\hat{\chi}$ that solves the following optimization problem.

$$\inf_{\hat{\chi}} \epsilon^- \quad \text{s.t. } \epsilon^+ \leq \epsilon, \qquad (5)$$

where $0 < \epsilon < 1$. We underscore that the detector $\hat{\chi}$ captures the detection of **all** edges in the causal graph, and that ϵ^+

---

[1]It is important to highlight that these are in general *not the probabilities of error*, and rather an averaging over errors of individual edges. To see this note that in equation (12) we can write out the rates in terms of individual edge error probabilities, but that the individual edge error events *are not disjoint*, since we could simultaneously make errors on different edges.

2

and ϵ^- are *rates, not probabilities*; we shall provide a strategy for the detection of each individual edge, but our performance analysis will be over all edges.

B. Definitions

Throughout this work, we will consider a binary hypothesis test for each individual edge. In particular, for an observed data set $\{X_k\}_1^n$, define the hypothesis testing problem

$$H_0 : \boldsymbol{\chi}_{i,j} = 0, \qquad \{X_k\}_1^n \sim P_{i,j} \tag{6}$$

$$H_1 : \boldsymbol{\chi}_{i,j} = 1, \qquad \{X_k\}_1^n \sim Q_{i,j} \tag{7}$$

where $P_{i,j}$ denotes the density of $\{X_k\}_1^n$ conditioned on $\boldsymbol{\chi}_{i,j} = 0$ and $Q_{i,j}$ denotes the density of $\{X_k\}_1^n$ conditioned on $\boldsymbol{\chi}_{i,j} = 1$. Thus, our problem is actually mixed in nature, since we assume a prior distribution $\pi_{\boldsymbol{A}}$ on the graph matrix $\boldsymbol{A}$; however we wish to control the aggregated error rate on the aggregated detector $\hat{\chi}$ as in a classical Neyman-Pearson-type detection problem.

Definition 1. *Assuming $\boldsymbol{A} \sim \pi_{\boldsymbol{A}}$, we define the following probabilities and weights:*

$$P_{i,j}^+ = \mathbb{P}(\hat{\boldsymbol{\chi}}_{i,j} = 1 | \boldsymbol{\chi}_{i,j} = 0), \tag{8}$$

$$Q_{i,j}^- = \mathbb{P}(\hat{\boldsymbol{\chi}}_{i,j} = 0 | \boldsymbol{\chi}_{i,j} = 1). \tag{9}$$

$$w_{i,j}^+ = \frac{\mathbb{P}(\boldsymbol{\chi}_{i,j} = 0)}{\sum_{k,l} \mathbb{P}(\boldsymbol{\chi}_{k,l} = 0)}, \tag{10}$$

$$w_{i,j}^- = \frac{\mathbb{P}(\boldsymbol{\chi}_{i,j} \neq 0)}{\sum_{k,l} \mathbb{P}(\boldsymbol{\chi}_{k,l} \neq 0)}. \tag{11}$$

$\square$

With Definition 1, we can substitute (8) - (11) into (3) and (4) and straightforwardly write out our error rates as,

$$\epsilon^+ = \sum_{i,j} w_{i,j}^+ P_{i,j}^+, \qquad \epsilon^- = \sum_{i,j} w_{i,j}^- Q_{i,j}^-. \tag{12}$$

An interesting observation is that for any edge (i, j)

$$\mathbb{P}(\hat{\boldsymbol{\chi}}_{i,j} = 1 \cap \boldsymbol{\chi}_{i,j} = 0)$$
$$\text{(Probability of false positive on edge } (i, j)) \tag{13}$$

$$\leq \mathbb{P}\left(\bigcup_{i \neq j} \hat{\boldsymbol{\chi}}_{i,j} = 1 \cap \boldsymbol{\chi}_{i,j} = 0 \right) \tag{14}$$
$$\text{(Probability of false positive on any edge)}$$

$$\leq \sum_{i \neq j} \mathbb{P}(\hat{\boldsymbol{\chi}}_{i,j} = 1 | \boldsymbol{\chi}_{i,j} = 0)\mathbb{P}(\boldsymbol{\chi}_{i,j} = 0) \tag{15}$$

$$= \left(\sum_{i \neq j} \mathbb{P}(\boldsymbol{\chi}_{i,j} = 0) \right) \sum_{i \neq j} \mathbb{P}(\hat{\boldsymbol{\chi}}_{i,j} = 1 | \boldsymbol{\chi}_{i,j} = 0)$$

$$\frac{\mathbb{P}(\boldsymbol{\chi}_{i,j} = 0)}{\left(\sum_{i \neq j} \mathbb{P}(\boldsymbol{\chi}_{i,j} = 0) \right)} \tag{16}$$

$$= \left(\sum_{i \neq j} \mathbb{P}(\boldsymbol{\chi}_{i,j} = 0) \right) \sum_{i,j} w_{i,j}^+ P_{i,j}^+ = \epsilon^+ \sum_{i \neq j} \mathbb{P}(\boldsymbol{\chi}_{i,j} = 0). \tag{17}$$
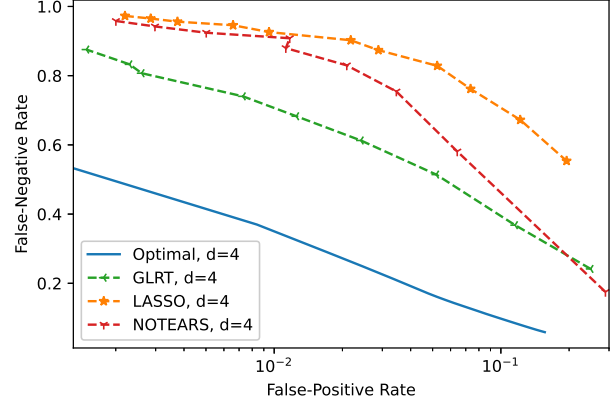

Fig. 2: Performance comparison showing the large gap between the optimal detector and several state-of-the-art methods on $n = 10$ observations with $\sigma^2 = 1$.

Since $\sum_{i \neq j} \mathbb{P}(A_{i,j} = 0)$ is a constant (for a fixed graph prior), we see that our false positive rate is an upper bound (up to a constant) for both the false positive probability on a given edge and the false positive probability on any edge. Hence, controlling the false positive rate also controls the false positive probability. Moreover, the false positive rate is often more computationally tractable than the probability of a false positive on any edge, so one may think of the false positive rate as a computationally tractable relaxation of the false positive probability.

III. FUNDAMENTAL LIMITS AND OPTIMAL DETECTOR

Herein We derive the optimal detector for Problem (5) which consists of a series of likelihood ratio tests between the distributions $P_{i,j}$ and $Q_{i,j}$ for each pair of vertices i, j. We derive upper and lower performance bounds for the optimal detector and thus the Neyman-Pearson causal discovery problem of (5); both bounds will described by the Rényi divergence.

Theorem 1. *Assume $\boldsymbol{A} \sim \pi_{\boldsymbol{A}}$. Let U be a uniform random variable on $[0, 1]$. Then, the optimal detector $\hat{\boldsymbol{\chi}}^*$ for Problem 5 under the prior $\pi_{\boldsymbol{A}}$ is given as follows, where the per edge detector is given by,*

$$\hat{\boldsymbol{\chi}}_{i,j}^* = \begin{cases} 0, & \frac{dP_{i,j}}{dQ_{i,j}} > \frac{w_{i,j}^-}{w_{i,j}^+}\gamma \\ 0, & \frac{dP_{i,j}}{dQ_{i,j}} = \frac{w_{i,j}^-}{w_{i,j}^+}\gamma \text{ and } U \leq \eta \\ 1, & \frac{dP_{i,j}}{dQ_{i,j}} < \frac{w_{i,j}^-}{w_{i,j}^+}\gamma \\ 1, & \frac{dP_{i,j}}{dQ_{i,j}} = \frac{w_{i,j}^-}{w_{i,j}^+}\gamma \text{ and } U > \eta \end{cases} \tag{18}$$

where γ and $\eta \in [0, 1]$ are chosen so that $\epsilon^+ = \epsilon$.

Proof. The proof is given in Section Appendix D $\qquad \square$

Similar to the optimal detector in Neyman-Pearson hypothesis testing, the random variable U controls our randomization

to achieve the false negative rate exactly [2]. The proof of Theorem 1 is similar to that of the classical Neyman-Pearson lemma (Proposition II.D.1 in [27]) and is given in Appendix D. In Figure 2, we compare the performance of the optimal detector with various state-of-the-art methods on $n = 10$ observations with $\sigma^2 = 1$. A description of each method is provide in Section V. For the graph prior $\pi_{\boldsymbol{A}}$, let $\mathcal{D}$ denote the set of all matrices with entries equal to either 0 or 1 corresponding to a directed acyclic graph with d vertices. For Figure 2, we uniformly, at random, select a matrix from $\mathcal{D}$. The uniform graph prior and optimal detector can be computed for small graph sizes, but become expensive to compute for larger graphs. Thus, in the sequel, Section V, we provide an alternative algorithm with finite sample performance guarantees whose complexity has a more favorable scaling with graph size. In addition, a lower complexity graph prior as an approximation to the uniform prior is also considered. Perhaps surprisingly, Figure 2 shows that the optimal detector strongly outperforms other methods, especially in the low false-positive regime. This gap suggests the need to develop improved finite-sample methods for low false-positive tolerance levels.

An interesting observation is that the threshold γ and the randomization η *do not depend on the edge being considered.* The effect of the edge in question is captured by the edge weights in the factor $\frac{w^-_{i,j}}{w^+_{i,j}}$ (which is completely determined by the prior $\pi_{\boldsymbol{A}}$). This feature simplifies design and analysis. For example, the next proposition follows relatively easily due to the form of $\hat{\chi}^*$.

Proposition 1. *Assume* $\boldsymbol{A} \sim \pi_{\boldsymbol{A}}$. *For the detector* $\hat{\chi}^*$ *given in Proposition 1, we have, for any* $\lambda \in [0, 1]$

$$\epsilon^- \leq \sum_{i,j} (w^-_{i,j})^{1-\lambda}(w^+_{i,j})^{\lambda}\frac{1}{\gamma^{\lambda}}e^{-(1-\lambda)D_{\lambda}(P_{i,j}||Q_{i,j})}, \quad (19)$$

where $D_{\lambda}(P_{i,j}||Q_{i,j})$ *is the the Rényi divergence of order* λ *between the series of distributions* $P_{i,j}$ *and* $Q_{i,j}$ *which is defined in Appendix A.*

Proof. The proof is given in Appendix B □

The next theorem is a converse bound for any detector $\hat{\chi}$.

Proposition 2. *Assume* $\boldsymbol{A} \sim \pi_{\boldsymbol{A}}$. *For any detector* $\hat{\chi}$ *that satisfies* $\epsilon^+ \leq \epsilon$, *we have that for any* $\lambda \in [0, 1]$.

$$\begin{aligned}\epsilon^- \geq \frac{1}{2}\sum_{i,j} w^-_{i,j}\exp\big\{ -(1-\lambda)D_{\lambda}(P_{i,j}||Q_{i,j}) \\ -\lambda D'_{\lambda}(P_{i,j}||Q_{i,j}) - \lambda\sqrt{2D''_{\lambda}(P_{i,j}||Q_{i,j})}\big\} \\ -\epsilon \max_{i,j}\Big\{ \frac{w^-_{i,j}}{w^+_{i,j}}\exp\big\{ -D'_{\lambda}(P_{i,j}||Q_{i,j}) \\ +(1-2\lambda)\sqrt{2D''_{\lambda}(P_{i,j}||Q_{i,j})}\big\}\Big\},\end{aligned} \quad (20)$$

where $D'_{\lambda}(P_{i,j}||Q_{i,j})$ and $D''_{\lambda}(P_{i,j}||Q_{i,j})$ are defined in Appendix A.

Proof. The proof is given in Appendix C □

Propositions 1 and 2 show that the Rényi divergence controls the false negative rate for a given tolerance ϵ. There are some important notes about Theorem 1 and Propositions 1 and 2.

1) The detector specified in Theorem 1 is optimal, but is prior, $\pi_{\boldsymbol{A}}$, dependent as the computations of $P_{i,j}$ and $Q_{i,j}$ do depend on $\pi_{\boldsymbol{A}}$.

2) Theorem 1 is general. The proof of Theorem 1 makes no assumptions on the specific model class, and hence holds for any general class of models, such as differentiable systems with additive noise, linear systems with non-Gaussian noise, *etc*. This is *not* to say that the detector given by Theorem 1 is necessarily *easy to compute* for any system model, as some models are more computationally tractable (such as the linear Gaussian case) than others. The theorem does provide insight into how to control error rates for the general causal discovery problem (and hence any general class of model).

3) It is shown in [24] that the proposed bounds are tighter for small n and small false positive tolerance than previously considered bounds for the same metrics, such as those presented in [19].

4) As expected, the Rényi divergence provides tighter bounds in the finite sample regime than classical Chernoff bounds and further provides an explicit dependence on ϵ. As such, for appropriately selected λ, the bound given in (20) reduces to an expression akin to classical bounds given in [29].

Unfortunately, although the form of the optimal detector $\hat{\chi}^*$ is simple to express, in practice, it is often computationally intractable, even for small graphs and relatively simple priors.

To see this, consider a system with two vertices, see Figure 1. We receive observation $\{X_k\}^n_1$ with $X_k = [X_k(1), X_k(2)]^{\top}$, $k = 1, 2, ..., n$. As stated before, we assume an LSEM,

$$\begin{bmatrix}X_k(1) \\ X_k(2)\end{bmatrix} = \boldsymbol{A}\begin{bmatrix}X_k(1) \\ X_k(2)\end{bmatrix} + \begin{bmatrix}W_k(1) \\ W_k(2)\end{bmatrix}, \quad (21)$$

where the $[W_k(1), W_k(2)]^{\top}$ vectors are *i.i.d.* Gaussian vectors with zero mean and covariance matrix $\sigma^2 I$. Since we restrict ourselves to directed acyclic graphs, the adjacency matrix $\boldsymbol{A}$ can only take one of three possible forms, which we denote as follows,

$$\boldsymbol{A}_0 = \begin{bmatrix}0 & 0 \\ 0 & 0\end{bmatrix}, \qquad \boldsymbol{A}_1(a) = \begin{bmatrix}0 & 0 \\ a & 0\end{bmatrix}, \qquad \boldsymbol{A}_2(a) = \begin{bmatrix}0 & a \\ 0 & 0\end{bmatrix}, \quad (22)$$

where $a \in \mathbb{R} \setminus \{0\}$. Assume that the prior $\pi_{\boldsymbol{A}}$ selects from the structures of $\boldsymbol{A}_0$, $\boldsymbol{A}_1(a)$, and $\boldsymbol{A}_2(a)$ uniformly at random, and that $a = 1 + \zeta$ where ζ is a standard exponential random variable. In order to implement $\hat{\chi}^*$ we must first compute the conditional distributions $P_{1,2}$ and $Q_{1,2}$.

---

[2]If the priors have different support for each hypothesis or the observations are discrete valued, randomization may be necessary per classical Neyman-Pearson hypothesis testing, *e.g.* [27], [28].

To compute $P_{1,2}$, observe that if $\boldsymbol{\chi}_{1,2} = 0$, then $\boldsymbol{A}_0$ must be the true graph structure, and so $X_k(1)$ and $X_k(2)$ are simply *i.i.d* Gaussian random variables with variance σ^2. To compute $Q_{1,2}$, notice that if $\boldsymbol{\chi}_{1,2} = 1$, the true graph may correspond to either $\boldsymbol{A}_1$ or $\boldsymbol{A}_2$. In either case, $X_{1,k}$ and $X_{2,k}$ are zero-mean jointly Gaussian random variables with covariance matrix $\sigma^2(I - \boldsymbol{A}_1(a))^{-1}(I - \boldsymbol{A}_1(a))^{-\top}$ under $\boldsymbol{A}_1(a)$, and $\sigma^2(I - \boldsymbol{A}_2(a))^{-1}(I - \boldsymbol{A}_2(a))^{-\top}$ under $\boldsymbol{A}_2(a)$. Then, we have that

$$Q_{1,2}(X_k) =$$
$$\frac{1}{2}\int_1^\infty \sum_{i=1}^2 \frac{e^{-\frac{1}{2\sigma^2}X_k^\top(I-\boldsymbol{A}_i(a))^\top(I-\boldsymbol{A}_i(a))X_k}}{2\pi|\sigma^2(I-\boldsymbol{A}_i(a))^{-1}(I-\boldsymbol{A}_i(a))^{-\top}|}e^{-(a-1)}da. \tag{23}$$

Moreover, observe that this is only for one observation, and that the joint distribution becomes more complex since the observations are *not independent a priori* and are only independent conditioned on $\boldsymbol{A}$. That is, for the joint distribution, the terms in the sum of (23) become products of Gaussian distributions. From this example, it is not difficult to see how computing the distributions $P_{i,j}$ and $Q_{i,j}$ can become computationally infeasible as the number of vertices increases, or as the graph prior $\pi_{\boldsymbol{A}}$ becomes more complex.

IV. ϵ-CUT

Because of the shortcomings mentioned above, we derive an alternative algorithm, the ϵ-*rate Causal discovery with Unequal edge error Tolerance* ($\epsilon-CUT$) algorithm. Recall that detecting the absence or presence of an edge is equivalent to a binary hypothesis test (see (6) and (7)) for which we provide an alternative algorithm. This algorithm relies on the fact that children have higher variances than their parents, which we highlight through an illustrative example below. In addition, we provide time complexity analysis for $\epsilon - CUT$.

Via an example, we provide the intuition for our new detector. Assume we have the same two-vertex given by equation (21), with the understanding that a may be any non-zero real number (which may be either random or deterministic). If $\boldsymbol{A} = \boldsymbol{A}_0$, we have that for all k,

$$\mathbb{E}[X_k(1)^2] = \mathbb{E}[W_k(1)^2] = \sigma^2, \tag{24}$$
$$\mathbb{E}[X_k(2)^2] = \mathbb{E}[W_k(2)^2] = \sigma^2. \tag{25}$$

Alternatively, if $\boldsymbol{A} = \boldsymbol{A}_1$, we still have $\mathbb{E}[X_k(1)^2] = \sigma^2$, but

$$\mathbb{E}[X_k(2)^2] = \mathbb{E}[(aX_k(1) + W_k(2))^2] = (a^2+1)\sigma^2. \tag{26}$$

Similarly, if $\boldsymbol{A} = \boldsymbol{A}_2$, then we have

$$\mathbb{E}[X_k(1)^2] = (a^2+1)\sigma^2, \qquad \mathbb{E}[X_k(2)^2] = \sigma^2. \tag{27}$$

Hence, we can study the following equivalent hypothesis testing problem,

$$H_0 : \mathbb{E}[X_k(1)^2] = \mathbb{E}[X_k(2)^2], \tag{28}$$
$$H_1 : \mathbb{E}[X_k(1)^2] \neq \mathbb{E}[X_k(2)^2], \tag{29}$$

since, in the linear Gaussian setting, any solution to the above problem is also a solution to (6)-(7).

Our hypothesis test is given as follows. If H_0 is true, then the empirical estimates of the variances should be roughly equal with high probability. That is, for a properly specified threshold τ, the event $|\hat{\sigma}_1^2 - \hat{\sigma}_2^2| \le \tau$, where

$$\hat{\sigma}_1^2 = \sum_{k=1}^n X_k(1)^2, \quad \hat{\sigma}_2^2 = \sum_{k=1}^n X_k(2)^2, \tag{30}$$

should occur with high probability. To extend the intuition to the case of more than two vertices, notice that if we condition on the parents of vertex i, $\hat{\sigma}_i^2$ is now given by

$$\hat{\sigma}_i^2 = \sum_{k=1}^n \left(X_k(i) - \alpha_i^\top Z_k(i)\right)^2, \tag{31}$$

where α_i is the vector of non-zero edge weights in the ith row of $\boldsymbol{A}$. Recall that tests based on the residual sum of squares has been previously considered [22], [23]. However, these works test *independence* between residuals and only provide asymptotic guarantees. In contrast, our algorithm compares the difference between residuals to detect energy differences. Moreover, we can select the threshold τ to satisfy the false positive constraint in a *finite-sample setting*. The following lemma, whose proof is in Appendix F, enables the setting of the needed threshold to achieve our finite-sample constraints.

Lemma 1. *For any n and $i \in \mathcal{V}$ let $\mathcal{Z}(i)$ and $Z_k(i)$ be defined as in the top of Section II. Define*

$$\hat{\sigma}_i^{*2} = \sum_{k=1}^n \left(X_k(i) - \hat{\alpha}_i^\top Z_k(i)\right)^2, \tag{32}$$

where $\hat{\alpha}_i$ is the vector of coefficients resulting from performing least squares for vertex i on $\{Z_k(i)\}_{k=1}^n$. Then, conditioned on $\{Z_k(i)\}_{k=1}^n$, $\hat{\sigma}_i^{*2}/\sigma^2$ follows a chi-squared distribution with $n - p$ degrees of freedom, where $p = |\mathcal{Z}(i)|$.

The complete algorithm for $\epsilon - CUT$ is given in Algorithm 1. With Lemma 1, we can prove the following result.

Theorem 2. *Assume we have a LSEM, and that $\boldsymbol{A} \sim \pi_{\boldsymbol{A}}$. Then, for any number of samples n and any given false positive tolerance ϵ, the false positive rate of ϵ-CUT, denoted by ϵ_1^+, satisfies $\epsilon_1^+ \le \epsilon$.*

Some important notes regarding ϵ-CUT and Theorem 2.
1) Theorem 2 is a finite-sample result. This differs from much of the current literature which focuses on asymptotic recoverability guarantees of causal discovery [1], [10], [18], [30], [31]. Finite-sample results have appeared in the literature, with a notable result appearing in [11] (Theorem 3). However, the result in [11] deals with connected components instead of individual edges, which is our main objective.
2) The exact algorithm for $\epsilon - CUT$, outlined in Algorithm 1, does not depend on the choice of prior $\pi_{\boldsymbol{A}}$. Hence, $\epsilon - CUT$ may be used in either a Bayesian or non-Bayesian setting.
3) The procedure outlined in Algorithm 1 requires finding a different threshold for each value of p and q given in

5

Algorithm 1: ϵ-CUT

Specify user false alarm tolerance $0 < \epsilon \le 1$ and the variance of the noise terms σ^2. Let $2^{\mathcal{V}}$ denote the power set of $\mathcal{V}$ (the vertex set). Let $\mathcal{B}$ denote the set of all unique pairs of edges, i.e., the set of all pairs (i, j), such that $i \ne j$, $i, j \in \mathcal{V}$. The cumulative distribution function (cdf) of a chi-squared distribution with l degrees of freedom is denoted by $F_l(x)$.

For each $(i, j) \in \mathcal{B}$:

1) For each $\hat{\mathcal{Z}}(i) \in 2^{\mathcal{V}}$ and $\hat{\mathcal{Z}}(j) \in 2^{\mathcal{V}}$ with $j \notin \hat{\mathcal{Z}}(i)$ and $i \notin \hat{\mathcal{Z}}(j)$ (i.e., the potential parent sets of nodes i and j):

   a) Perform linear least squares regression for nodes i and j on $\hat{\mathcal{Z}}(i)$ and $\hat{\mathcal{Z}}(j)$, respectively, obtaining the vectors of coefficients $\hat{\alpha}$ and $\hat{\beta}$, i.e.,

$$\hat{\alpha} = (\hat{\boldsymbol{Z}}(i)\hat{\boldsymbol{Z}}(i)^\top)^{-1}\hat{\boldsymbol{Z}}(i)^\top \boldsymbol{X}(i)$$

   where $\hat{\boldsymbol{Z}}(i) = [\hat{Z}_1(i), \hat{Z}_2(i), ..., \hat{Z}_n(i)]^\top$, $\hat{Z}_k(i) = [X_k(i_1), X_k(i_2), ..., X_k(i_{|\hat{\mathcal{Z}}(i)|})]^\top$ for $i_l \in \hat{\mathcal{Z}}(i)$, $l = 1, 2, ..., |\hat{\mathcal{Z}}(i)|$, and $\boldsymbol{X}(i) = [X_1(i), X_2(i), ..., X_n(i)]^\top$. Similarly for $\hat{\beta}$.

   b) Compute

$$\hat{\sigma}_i^2 = \sum_{k=1}^{n} \left(X_k(i) - \hat{\alpha}^\top \hat{Z}_k(i) \right)^2, \quad \hat{\sigma}_j^2 = \sum_{k=1}^{n} \left(X_k(j) - \hat{\beta}^\top \hat{Z}_k(j) \right)^2,$$

   i.e., the residual sum of squares for vertices i and j.

   c) Define $\tau'_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)} = (\tau_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)} - |q - p|\sigma^2)/(2\sigma^2)$ and compute $\tau_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)}$ such that

$$2 - F_{n-p}(\tau'_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)} + n - p) + F_{n-p}(-\tau'_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)} + n - p)$$
$$- F_{n-q}(\tau'_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)} + n - q) + F_{n-q}(-\tau'_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)} + n - q) = \epsilon,$$

   where $p = |\hat{\mathcal{Z}}(i)|$ and $q = |\hat{\mathcal{Z}}(j)|$.

   d) If $|\hat{\sigma}_i^2 - \hat{\sigma}_j^2| \le \tau_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)}$, declare $\hat{\chi}_{i,j} = 0$.

2) If all potential parent sets $\hat{\mathcal{Z}}(i) \in 2^{\mathcal{V}}$ and $\hat{\mathcal{Z}}(j) \in 2^{\mathcal{V}}$ have been tested without declaring $\hat{\chi}_{i,j} = 0$, declare $\hat{\chi}_{i,j} = 1$.

Algorithm 1 c). While this is not computationally difficult it is possible to use only a single threshold in Algorithm 1 while maintaining the guarantee that the false positive rate is less than ϵ. To see this, observe that for two chi-squared random variables with p and q degrees of freedom, respectively, where $p \le q$ we have that for any $x \in \mathbb{R}$, $F_p(x) \ge F_q(x)$, where F_p is the cdf of a chi-squared random variable with p degrees of freedom. This tells us that the term in Algorithm 1 c) is upper-bounded by

$$2\big[1 - F_n(\tau' + n) + F_{n-d+2}(-\tau' + n - d + 2)\big]. \quad (33)$$

where $\tau' = (\tau - (d - 2)\sigma^2)/(2\sigma^2)$. Hence, using only a single threshold while maintaining the false positive constraint is possible.

4) As mentioned before, different solutions to the hypothesis-testing problem between H_0 and H_1, lead to different causal discovery algorithms. Another way to circumvent the computational challenges associated with the optimal detector $\hat{\chi}^*$ is to use a generalized likelihood ratio test (GLRT). This is done in [18] to obtain confidence intervals on the causal effect between two vertices, rather than directly declaring the presence or absence on an edge. Unfortunately, the main results in [18] are asymptotic. Hence, ϵ-CUT circumvents the

computational issues associated with computing $\hat{\chi}^*$ while providing finite-sample results.

5) Unlike several popular algorithms such as BIC [10], LASSO [11], and NOTEARS [7], ϵ-CUT requires no hyper-parameter tuning. That is, in the algorithms mentioned, a sparsity constraint is added through a regularization term (ℓ_0 in [10] and ℓ_1 in [7], [11]). The constant λ controls the sparsity of the resulting graph. Unfortunately, there is no current way to determine *a priori* how the constant λ affects the error rates, and if these rates satisfy the constraint in (5). Hence, one needs to experiment with different regularization constants. In contrast ϵ-CUT, once ϵ is given, everything in ϵ-CUT is completely specified.

Proof Sketch of Theorem 2. Since $\epsilon^+ = \sum_{i,j} w_{i,j}^+ P_{i,j}^+$, it suffices to show that for all i, j, $P_{i,j}^+ \le \epsilon$. Note that

$$P_{i,j}^+ = \int_{\boldsymbol{A}} \int_{\boldsymbol{X}^{\backslash i,j}} \mathbb{P}_{\boldsymbol{A}}(\hat{\chi}_{i,j} = 1 | \boldsymbol{X}^{\backslash i,j}) \mathbb{P}_{\boldsymbol{A}}(\boldsymbol{X}^{\backslash i,j})$$
$$\frac{\mathbb{1}\{\boldsymbol{A}_{i,j} = 0 \cap \boldsymbol{A}_{j,i} = 0\}\pi_{\boldsymbol{A}}}{\mathbb{P}(\chi_{i,j} = 0)}, \quad (34)$$

where $\boldsymbol{X}^{\backslash i,j}$ denotes the set of all measurements except those from the ith and jth vertex. That is, $\boldsymbol{X}^{\backslash i,j} = \{X_k^{\backslash i,j}\}_{k=1}^{n}$ where $X_k^{\backslash i,j} = [X_k(1), ..., X_k(i-1), X_k(i+1), ..., X_k(j-1), X_k(j+1), ..., X_k(d)]^\top$. Then, it suffices to show that for any $\boldsymbol{A}$ and $\boldsymbol{X}^{\backslash i,j}$ we have $\mathbb{P}_{\boldsymbol{A}}(\hat{\chi}_{i,j} = 1 | \boldsymbol{X}^{\backslash i,j}) \le \epsilon$. This

is done by first noticing that for $\hat{\sigma}_i^2$ and $\hat{\sigma}_j^2$ as defined in Algorithm 1,

$$\mathbb{P}_{\boldsymbol{A}}(\hat{\boldsymbol{\chi}}_{i,j} = 1 | \boldsymbol{X}^{\setminus i,j}) \tag{35}$$

$$= \mathbb{P}_{\boldsymbol{A}}\left(\bigcap_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)} |\hat{\sigma}_i^2 - \hat{\sigma}_j^2| > \tau_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)} | \boldsymbol{X}^{\setminus i,j}\right) \tag{36}$$

$$\overset{(a)}{\leq} \mathbb{P}_{\boldsymbol{A}}\left(|\hat{\sigma}_i^{*2} - \hat{\sigma}_j^{*2}| > \tau_{\mathcal{Z}(i),\mathcal{Z}(j)} | \boldsymbol{X}^{\setminus i,j}\right), \tag{37}$$

where (a) holds since the event $|\hat{\sigma}_i^2 - \hat{\sigma}_j^2| > \tau_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)}$ must hold for *all potential parent sets* $\hat{\mathcal{Z}}(i), \hat{\mathcal{Z}}(j)$, including the true parent sets $\mathcal{Z}(i), \mathcal{Z}(j)$, and so

$$\left\{\bigcap_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)} |\hat{\sigma}_i^2 - \hat{\sigma}_j^2| > \tau_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)}\right\} \tag{38}$$
$$\subset \{|\hat{\sigma}_i^{*2} - \hat{\sigma}_j^{*2}| > \tau_{\mathcal{Z}(i),\mathcal{Z}(j)}\}.$$

Using Lemma 1 together with a series of inequalities and algebraic manipulations, (37) is upper bounded by the expression given in Algorithm 1 c), which completes the proof. The full proof is given in Appendix E. $\square$

Notice that in the proof of Theorem 2 it is shown that for any $\boldsymbol{A}$ (with both $\boldsymbol{A}_{i,j} = 0$ and $\boldsymbol{A}_{j,i} = 0$ and $\boldsymbol{X}^{\setminus i,j}$ we have $\mathbb{P}_{\boldsymbol{A}}(\hat{\boldsymbol{\chi}}_{i,j} = 1 | \boldsymbol{X}^{\setminus i,j}) \leq \epsilon$. This then implies that

$$\mathbb{P}_{\boldsymbol{A}}(\hat{\boldsymbol{\chi}}_{i,j} = 1) = \int_{\boldsymbol{X}^{\setminus i,j}} \mathbb{P}_{\boldsymbol{A}}(\hat{\boldsymbol{\chi}}_{i,j} = 1 | \boldsymbol{X}^{\setminus i,j})\mathbb{P}_{\boldsymbol{A}}(\boldsymbol{X}^{\setminus i,j}) \tag{39}$$

$$\leq \epsilon \int_{\boldsymbol{X}^{\setminus i,j}} \mathbb{P}_{\boldsymbol{A}}(\boldsymbol{X}^{\setminus i,j}) = \epsilon. \tag{40}$$

Hence, if the true underlying graph does not have an edge between i and j, the probability of a false positive does not exceed ϵ. Observe that this is a non-Bayesian result, and so $\epsilon - CUT$ may be used in a Bayesian or non-Bayesian setting while preserving finite-sample results.

A. Complexity Analysis

For each pair of vertices i and j, we must conduct a threshold test between the residual sum of squares for potentially every pair of vertex subsets that do not contain i or j. Since there are 4^{d-2} possible pairs of these subsets for each pair of vertices, we must compute at most $\frac{d(d-1)}{2}4^{d-2}$ linear regressions. The computational complexity for linear regression is $\mathcal{O}(d^2(n+d))$ (the details are in the supplemental file located at https://github.com/shaskajo/epsilonCUT). For each pair of subsets, we must also perform a threshold test on the difference between the residuals, which has complexity $\mathcal{O}(1)$. Hence, the overall complexity of ϵ-CUT is $\mathcal{O}(d^3(d-1)4^{d-2}(n+d))$.

V. COMPARISON ALGORITHMS

In Section VI we consider numerical examples to compare ϵ-CUT to other popular algorithms such as NOTEARS [7], DAGMA [8], LASSO neighborhood selection [11], and the generalized likelihood ratio test [18]. Hence, we briefly summarize all prior algorithms used in the sequel. We also briefly discuss the computational complexity of LASSO and the generalized likelihood ratio test. We focus on these algorithms since they possess some form of finite-sample guarantee of recovering the underlying graph. Moreover, continuous optimization formulations, such as NOTEARS and DAGMA, suffer from a highly non-convex optimization space, and hence cannot guarantee convergence to global optima for finite samples, as well as making computational complexity analysis rather difficult and beyond the scope of this paper. However, we provide runtime results in Section VI for all algorithms used. Because of this, we also briefly mentioned how each method is implemented.

A. LASSO

For a given observed node $X_k(i)$, let $Y_k^i = [X_k(1), ..., X_k(i-1), X_k(i+1), ..., X_k(d)]^\top$ for $k = 1, 2, ..., n$. Then, assuming a linear model for the observations, $\hat{X}_k(i) = A_i^\top Y_{i,k}$, where A_i is a vector of coefficients, LASSO [11] minimizes the following sparsity penalized loss over the coefficients of A_i,

$$\frac{1}{2n}\sum_{k=1}^n \|X_k(i) - A_i^\top Y_k^i\|_2^2 + \lambda\|A_i\|_1, \tag{41}$$

for each node i, where λ is a regularizer term.

The complexity of LASSO neighborhood selection for a single vertex is $\mathcal{O}(nd\min\{n, d\})$ [11]. Hence, the total complexity for LASSO is $\mathcal{O}(nd^2\min\{n, d\})$. We implement LASSO using the Python implementation provided in the sklearn library.

B. Generalized Likelihood Ratio Test (GLRT)

We consider a generalized likelihood ratio test (GLRT) for our hypothesis testing problem. [18]. In [18], the GLRT is used to construct confidence intervals on the causal effect rather than detect the presence or absence of an edge. We modify the algorithm in [18] to address our problem. If we let $\mathcal{M}$ denote the set of all DAGS and $\mathcal{M}_{i,j}^0$ the set of all DAGS with no edge between vertices i and j. Then, for given DAG $\mathcal{G}$ with weighted adjacency matrix $\boldsymbol{A}$, the likelihood function is given as

$$\ell(\mathcal{G}) = -\frac{np}{2}\log(2\pi\sigma^2) - \frac{n}{2\sigma^2}\text{Trace}\big((I-\boldsymbol{A})^\top(I-\boldsymbol{A})\hat{\Sigma}\big), \tag{42}$$

where $\hat{\Sigma}$ is the empirical covariance matrix. Then, define the test statistic

$$\Lambda_{i,j} = 2\big(\sup_{\mathcal{G}\in\mathcal{M}} \ell(\mathcal{G}) - \sup_{\mathcal{G}\in\mathcal{M}_{i,j}^0} \ell(\mathcal{G})\big), \tag{43}$$

which is compared against a threshold τ. If $\Lambda_{i,j} > \tau$ then an edge is declared, otherwise declare $\hat{\boldsymbol{\chi}}_{i,j} = 0$. The threshold τ is chosen so that $\Lambda_{i,j} > \tau$ occurs with probability ϵ when $\boldsymbol{\chi}_{i,j} = 0$ *asymptotically*. We note that reducing the computational complexity of the inherent exhaustive DAG search for ML/GLRT schemes remains an open problem.

To obtain the computational complexity of the GLRT, we use the procedure outlined in Section 5 of [18], with slight modifications to fit our problem (since our algorithm

7

Computational Complexities	
Algorithm	Complexity
LASSO	$\mathcal{O}(d^2 n \min\{n, d\})$
ϵ-CUT	$\mathcal{O}(d^3(d-1)4^{d-2}(n+d))$
GLRT	$\mathcal{O}((d!+1)d^3(d-1)^2(n+d))$

TABLE I: Computational complexities of LASSO, ϵ-CUT, and the GLRT.

outputs a potential edge rather than a confidence interval), and can be found in the following GitHub repository: https://github.com/shaskajo/epsilonCUT/tree/main. The details of our implementation along with the full analysis of the computational complexity are given in the supplemental material: https://github.com/shaskajo/epsilonCUT. The computational complexity of the GLRT is $\mathcal{O}((1+d!)d^3(d-1)^2(n+d))$.

C. Continuous Optimization Formulations

A popular framework used in causal discovery is mapping the combinatorial optimization problem into a continuous one. Typically, a loss function is minimized subject to a constraint that ensures the outputted matrix corresponds to a DAG. Throughout this paper, we assume the loss function is the penalized squared loss,

$$\frac{1}{2n}\sum_{k=1}^n \|X_k - \boldsymbol{A}X_k\|_2^2 + \lambda\|\boldsymbol{A}\|_1, \tag{44}$$

where λ is a regularizer term, $\|X_k - \boldsymbol{A}X_k\|_2$ is the l_2 norm of the vector $X_k - \boldsymbol{A}X_k$ and $\|\boldsymbol{A}\|_1$ is the l_1 norm of the matrix $\boldsymbol{A}$. Then, the following algorithms differ only in what algebraic constraint is used to ensure $\boldsymbol{A}$ corresponds to a DAG.

1) *NOTEARS*: It is shown in [7] that a weighted adjacency matrix $\boldsymbol{A}$ represents a directed acyclic graph if and only if

$$\text{Trace}(e^{\boldsymbol{A}\circ\boldsymbol{A}}) = d, \tag{45}$$

where $\circ$ denotes the Hadamard product (element-wise multiplication). The elements of the continuous-valued estimate $\hat{\boldsymbol{A}}$ are thresholded to determine the inactive edges, note that this threshold is another hyperparamter to be tuned. We use the implementation publicly provided in [7].

2) DAGMA [8] solves the continuous optimization problem subject to the constraint

$$-\log|s\boldsymbol{I} - \boldsymbol{A}\circ\boldsymbol{A}| + d\log s = 0, \tag{46}$$

where $s > 0$ is a hyperparameter. Similarly to NOTEARS, [8] shows that $\boldsymbol{A}$ is a DAG if and only if (46) holds for appropriately selected s. We use the implementation publicly provided in [8].

VI. NUMERICAL RESULTS

We consider a numerical example to compare ϵ-CUT to other popular algorithms that were summarized in Section V. We examine some interesting phenomena and show the potential gains of ϵ-CUT over the abovementioned algorithms.

A. Definition of Graph Prior

The number of DAGS grows super-exponentially in the number of vertices d [6], thus, generating all possible DAGS and uniform, random selection becomes computationally infeasible for increasing d. We use the following prior, $\pi_{\boldsymbol{A}}$ which is closely related to the priors used in [7], [19], and defined as follows:

1) For a given number of vertices d, let $\boldsymbol{D}$ denote a $d \times d$ strictly lower triangular matrix. Then, the lower diagonal entries (support) are zeroed out randomly and independently. Then, for each remaining edge, the weight is selected uniformly at random from the set $[-5, -.5]\cup[.5, 5]$.

2) Randomly relabel the vertices of $\boldsymbol{D}$. This operation is motivated by the fact that for any DAG, there exists a relabelling of the vertices such that the corresponding adjacency matrix is strictly lower triangular [6].

B. Results

Some numerical comparisons for a low number of samples ($n = 10$) are given in Figure 3. In Figure 3a we compare the performance between $\epsilon - CUT$ and the previous state-of-the-art methods for $d = 5$ vertices, and in Figure 3b for $d = 7$. In both cases, our algorithm is competitive with state-of-the-art methods (aside from the GLRT) in the low tolerance regime. In Figure 3a, our method outperforms state-of-the-art methods by nearly 20% in the low tolerance regime. While the GLRT significantly outperforms the other methods, it can be seen from Table I and Figure 3c that the GLRT suffers from high computational complexity even for moderately sized graphs (this complexity is the reason that the GLRT is omitted from Figure 3b, its complexity is too high). Moreover, in addition to $\epsilon - CUT$ offering competitive performance, we emphasize that, unlike LASSO, NOTEARS, and DAGMA, our method *does not require hyperparameter tuning*, in addition to possessing *finite-sample performance guarantees*.

There is another interesting point to be made about the numerical results. First, from the definition of $\pi_{\boldsymbol{A}}$ and the description of the algorithms used, the edge error probabilities $P_{i,j}^+$ and $Q_{i,j}^-$ are the same for all pairs (i,j). Hence, $\epsilon^+ = \sum_{i,j} w_{i,j}^+ P_{i,j}^+ = P_{1,2}^+ \sum_{i,j} w_{i,j}^+ = P_{1,2}^+$, and similarly for ϵ^-. Hence, the curves figures 3a and 3b are also for the *probability of error of individual edges with finite samples*. This is in contrast to most theoretical guarantees, which often deal with full graph recovery [7], [8], [10], or connected neighborhoods in the finite sample setting [11].

C. Runtime Comparisons

We compare the runtimes for the various algorithms in Figure 3c. Surprisingly, $\epsilon - CUT$ is competitive with continuous optimization formulations (DAGMA and NOTEARS) even for moderately sized graphs. Of course, as the number of vertices increases, $\epsilon - CUT$ begins to fall behind DAGMA and NOTEARS. This is to be expected since $\epsilon - CUT$ requires an exhaustive search procedure for each pair of vertices to satisfy the false-positive constraint, leading to increased runtimes. We
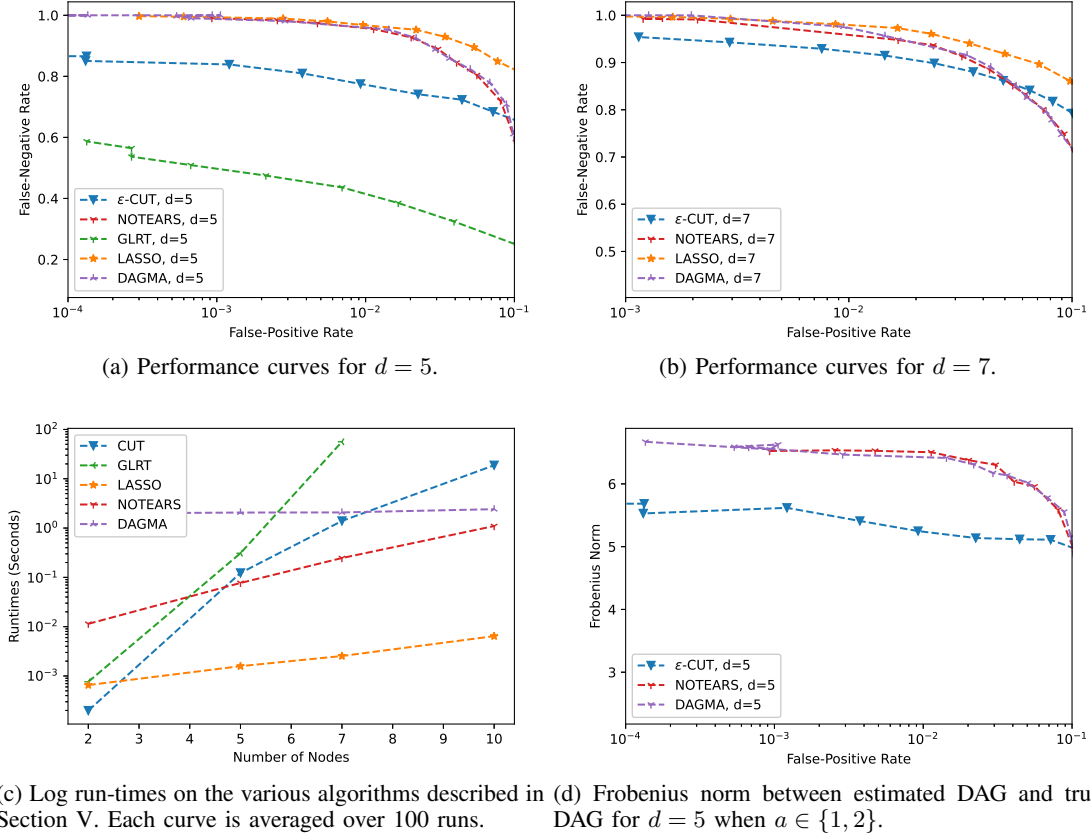
(a) Performance curves for $d = 5$.

(b) Performance curves for $d = 7$.

(c) Log run-times on the various algorithms described in Section V. Each curve is averaged over 100 runs.

(d) Frobenius norm between estimated DAG and true DAG for $d = 5$ when $a \in \{1, 2\}$.

Fig. 3: Numerical results of the performance of various detectors. For all the cases, $\sigma^2 = 1$ and $n = 10$ and the regularization constant λ is varied for LASSO, NOTEARS, and DAGMA to control the error rates. The regularizer values used to generate the plots are given in the supplemental file: https://github.com/shaskajo/epsilonCUT. Performance curves are averaged over 1500 iterations.

reiterate that $\epsilon - CUT$ does not require hyperparameter tuning to satisfy the false-positive constraint.

D. Recovering Directions From Support

Throughout this paper, we have focused only on the support recovery of a directed acyclic graph as this is the most challenging part of the graph identification problem. Note that the maximum likelihood estimate of the underlying DAG structure without any sparsity constraints can be found by considering only complete DAGs, i.e., DAGs with the maximum number of edges (Proposition 3.2 in [18]); thus one must consciously enforce sparsity. However, if one does not care to enforce sparsity during the maximum likelihood estimation, then the maximum likelihood estimate may be found efficiently [18], [21]. Thus, the maximum likelihood estimate can be used to estimate link direction and the associated weights.

We first use $\epsilon - CUT$ to estimate the support $\hat{\chi}$. Then, we then compute the *unrestricted maximum likelihood estimate*,

$$\min_{A \in \mathcal{M}} \frac{nd}{2} \log(2\pi\sigma^2) + \frac{1}{\sigma^2}\text{Trace}((\boldsymbol{I} - \boldsymbol{A})^\top(\boldsymbol{I} - \boldsymbol{A}\hat{\Sigma}) \quad (47)$$

where $\mathcal{M}$ is the set of all DAGs, and $\hat{\Sigma}$ is the empirical covariance matrix. Denoting the maximum likelihood estimate

as $\hat{\boldsymbol{A}}_{ML}$, the final estimate is $\hat{\boldsymbol{A}} = \hat{\boldsymbol{A}}_{ML} \circ \hat{\chi}$ which contains the edge weights and directions of the causal graph while inheriting the theoretical guarantees on the sparsity provided by $\epsilon - CUT$. In Fig. 3d we plot the Frobenius norm between the true matrix $\boldsymbol{A}$ and the estimated matrix $\hat{\boldsymbol{A}}$ obtained from the procedure described above. In addition, we compare the performance of this procedure with DAGMA and NOTEARS, both of which output a DAG (as opposed to only the support). Except in the regime for larger false alarm rates, ϵ-cut yields superior performance over NOTEARS and DAGMA for full DAG recovery. Thus, our method is competitive with both continuous optimization formulations, despite focusing only on support recovery.

VII. Conclusions

We have introduced a framework for *Neyman-Pearson causal discovery*. Often, one needs to control one type of edge error, so our framework seeks to minimize one error rate subject to the other being below a user-specified tolerance level. This approach allowed us to derive the optimal detector for this problem. We derived finite sample performance bounds for this optimal detector. Surprisingly, there is a strong performance gap between current methods and the

9

optimal detector. Unfortunately, the optimal detector becomes computationally infeasible even for modestly sized graphs, which motivated us to derive $\epsilon - CUT$, which scales more favorably with graph size. We showed that $\epsilon - CUT$ provides finite-sample guarantees, and hence is feasible in the Neyman-Pearson causal discovery framework. In addition, we showed that $\epsilon - CUT$ is competitive with state-of-the-art methods *without the need to tune regularizers or hyper-parameters while offering performance guarantees in the finite sample regime.* Finally, we considered the ability to recover both directions and edge weights, since $\epsilon - CUT$ only recovers support, and show that our method in tandem with unrestricted maximum likelihood estimations is better than state-of-the-art methods that output the full DAG for small false-alarm rate constraints.

APPENDIX

A. Definitions

Definition 2. *The <u>Rényi divergence of order λ</u> between two probability measures P and Q is given as*

$$D_\lambda(P||Q) \doteq \frac{1}{\lambda - 1} \log \int_\mathcal{X} \left(\frac{dP}{dQ}\right)^\lambda dQ, \tag{48}$$

where $\frac{dP}{dQ}$ is the Radon-Nikodym derivative of P with respect to Q.

Definition 3. *For any two probability measures P and Q, let*

$$D'_\lambda(P||Q) \doteq \int_\mathcal{X} F_\lambda(x; P, Q) \log \frac{dP}{dQ}, \tag{49}$$

$$D''_\lambda(P||Q) \doteq \int_\mathcal{X} F_\lambda(x; P, Q) \left(\log \frac{dP}{dQ}\right)^2 - \left(D'_\lambda(P||Q)\right)^2, \tag{50}$$

where $F_\lambda(x; P, Q) \doteq \frac{f(x)^\lambda g(x)^{1-\lambda}}{\int_\mathcal{X} f(x)^\lambda g(x)^{1-\lambda} d\mu}, \quad \text{for } \lambda \in [0, 1].$ $\tag{51}$

B. Proof of Proposition 1

Proof. for any i, j, we have

$$Q_{i,j}^- \le Q_{i,j}\left(\frac{dP_{i,j}}{dQ_{i,j}} \ge \frac{w_{i,j}^-}{w_{i,j}^+}\gamma\right) \tag{52}$$

$$= \int_\mathcal{X} \mathbb{1}\left\{\frac{dP_{i,j}}{dQ_{i,j}} \ge \frac{w_{i,j}^-}{w_{i,j}^+}\gamma\right\} dQ_{i,j} \tag{53}$$

$$\overset{(a)}{\le} \int_\mathcal{X} \left(\frac{w_{i,j}^+}{w_{i,j}^1}\frac{1}{\gamma}\frac{dP_{i,j}}{dQ_{i,j}}\right)^\lambda dQ_{i,j} \tag{54}$$

$$= \left(\frac{w_{i,j}^+}{w_{i,j}^+}\right)^\lambda \frac{1}{\gamma^\lambda} \int_\mathcal{X} \left(\frac{dP_{i,j}}{dQ_{i,j}}\right)^\lambda dQ_{i,j}, \tag{55}$$

where (a) holds since $\mathbb{1}\{a \ge \gamma\} \le \left(\frac{a}{\gamma}\right)^\lambda$ for any $a, \gamma, \lambda > 0$. Since, the above holds for any $\lambda > 0$. Multiplying by $w_{i,j}^-$ yields

$$w_{i,j}^- Q_{i,j}^- \le (w_{i,j}^-)^{1-\lambda} (w_{i,j}^+)^\lambda \frac{1}{\gamma^\lambda} e^{-(1-\lambda)D_\lambda(P_{i,j}||Q_{i,j})}. \tag{56}$$

Summing over i, j completes the proof. □

C. Proof of Proposition 2

Proof. It is relatively easy to show that for any pair (i, j),

$$Q_{i,j}(X_1,.., X_n) = \exp\{-(1-\lambda)D_\lambda(P_{i,j}||Q_{i,j}) \tag{57}$$

$$-\lambda \log \frac{dP_{i,j}}{dQ_{i,j}}\} F_\lambda(X_1, ..., X_n; P_{i,j}, Q_{i,j}), \tag{58}$$

$$P_{i,j}(X_1,.., X_n) = \exp\{-(1-\lambda)D_\lambda(P_{i,j}||Q_{i,j}) \tag{59}$$

$$+ (1-\lambda) \log \frac{dP_{i,j}}{dQ_{i,j}}\} F_\lambda(X_1, ..., X_n; P_{i,j}, Q_{i,j}). \tag{60}$$

Then, consider the following sets

$$\mathcal{X}_{i,j,\lambda} = \left\{ X_1, ..., X_n : |\log \frac{dP_{i,j}}{dQ_{i,j}} - D'_\lambda(P_{i,j}||Q_{i,j})| \le \sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})} \right\}, \tag{61}$$

$$\mathcal{X}_{i,j}^0 = \left\{ X_1, ..., X_n : \hat{\chi}_{i,j} = 0 \right\}, \tag{62}$$

$$\mathcal{X}_{i,j}^1 = \left\{ X_1, ..., X_n : \hat{\chi}_{i,j} = 1 \right\}. \tag{63}$$

Hence, for any $X_1, ..., X_n \in \mathcal{X}_{i,j,\lambda}$, we have that

$$Q_{i,j}(X_1,.., X_n) \ge$$
$$\exp\left\{ -(1-\lambda)D_\lambda(P_{i,j}||Q_{i,j}) - \lambda D'_\lambda(P_{i,j}||Q_{i,j}) - \lambda\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})} \right\} F_\lambda(X_1, ..., X_n; P_{i,j}, Q_{i,j}), \tag{64}$$

$$P_{i,j}(X_1,.., X_n) \ge$$
$$\exp\left\{ -(1-\lambda)D_\lambda(P_{i,j}||Q_{i,j}) + (1-\lambda)D'_\lambda(P_{i,j}||Q_{i,j}) - (1-\lambda)\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})} \right\} F_\lambda(X_1, ..., X_n; P_{i,j}, Q_{i,j}). \tag{65}$$

Notice that $D'_\lambda(P_{i,j}||Q_{i,j})$ is actually the mean of $\log \frac{dP_{i,j}}{dQ_{i,j}}$ with respect to the distribution F_λ, and $D''_\lambda(P_{i,j}||Q_{i,j})$ is its variance. Then, $\mathcal{X}_{i,j,\lambda}$ is the event that the log-likelihood ratio is within $\sqrt{2}$ standard deviations of its mean, and so from Chebychev's inequality,

$$\int_{\mathcal{X}_{i,j,\lambda}} F_\lambda(X_1, ..., X_n; P_{i,j}, Q_{i,j}) \ge \frac{1}{2}. \tag{66}$$

Then, from the union bound, we have that

$$\int_{\mathcal{X}_{i,j,\lambda} \cap \mathcal{X}_{i,j}^0} F_\lambda(X_1, ..., X_n; P_{i,j}, Q_{i,j}) + \int_{\mathcal{X}_{i,j,\lambda} \cap \mathcal{X}_{i,j}^1} F_\lambda(X_1, ..., X_n; P_{i,j}, Q_{i,j}) \ge \frac{1}{2}. \tag{67}$$

From (65), we have that

$$\int_{\mathcal{X}_{i,j,\lambda} \cap \mathcal{X}_{i,j}^1} F_\lambda(X_1, ..., X_n; P_{i,j}, Q_{i,j}) \tag{68}$$

$$\leq \exp\left\{(1-\lambda)D_\lambda(P_{i,j}||Q_{i,j}) - (1-\lambda)D'_\lambda(P_{i,j}||Q_{i,j}) \right.$$
$$\left. + (1-\lambda)\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}\right\} \int_{\mathcal{X}_{i,j,\lambda} \cap \mathcal{X}^1_{i,j}} P_{i,j}(X_1,...,X_n) \tag{69}$$

$$\leq \exp\left\{(1-\lambda)D_\lambda(P_{i,j}||Q_{i,j}) - (1-\lambda)D'_\lambda(P_{i,j}||Q_{i,j}) \right.$$
$$\left. + (1-\lambda)\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}\right\} P^+_{i,j} \tag{70}$$

and so

$$\int_{\mathcal{X}_{i,j,\lambda} \cap \mathcal{X}^0_{i,j}} F_\lambda(X_1,...,X_n; P_{i,j}, Q_{i,j}) \geq \frac{1}{2} - $$
$$\exp\left\{(1-\lambda)D_\lambda(P_{i,j}||Q_{i,j}) - (1-\lambda)D'_\lambda(P_{i,j}||Q_{i,j})\right. \tag{71}$$
$$\left. + (1-\lambda)\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}\right\} P^+_{i,j}.$$

Then, (64) gives

$$Q^-_{i,j} \geq \exp\left\{ - (1-\lambda)D_\lambda(P_{i,j}||Q_{i,j}) \right.$$
$$\left. - \lambda D'_\lambda(P_{i,j}||Q_{i,j}) - \lambda\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})} \exp \right\}$$
$$\int_{\mathcal{X}_{i,j,\lambda} \cap \mathcal{X}^0_{i,j}} F_\lambda(X_1,...,X_n; P_{i,j}, Q_{i,j}), \tag{72}$$

together with (71) we have

$$Q^-_{i,j} \geq \frac{1}{2}\exp\left\{ - (1-\lambda)D_\lambda(P_{i,j}||Q_{i,j}) - \lambda D'_\lambda(P_{i,j}||Q_{i,j}) \right.$$
$$\left. - \lambda\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}\right\} - \exp\left\{ - D'_\lambda(P_{i,j}||Q_{i,j}) \right.$$
$$\left. + (1-2\lambda)\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}\right\} P^+_{i,j}. \tag{73}$$

Multiplying both sides by $w^-_{i,j}$ and summing over all (i,j) pairs gives us

$$\sum_{i,j} w^-_{i,j} Q^-_{i,j} \geq \frac{1}{2}\sum_{i,j} w^-_{i,j} \exp\left\{ - (1-\lambda)D_\lambda(P_{i,j}||Q_{i,j}) \right.$$
$$\left. - \lambda D'_\lambda(P_{i,j}||Q_{i,j}) - \lambda\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}\right\} - \sum_{i,j} w^+_{i,j}\frac{w^-_{i,j}}{w^+_{i,j}}$$
$$\exp\left\{ - D'_\lambda(P_{i,j}||Q_{i,j}) + (1-2\lambda)\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}\right\} P^+_{i,j}, \tag{74}$$

which yields

$$\sum_{i,j} w^-_{i,j} Q^-_{i,j} \geq \frac{1}{2}\sum_{i,j} w^-_{i,j} \exp\left\{ - (1-\lambda)D_\lambda(P_{i,j}||Q_{i,j}) \right.$$
$$\left. - \lambda D'_\lambda(P_{i,j}||Q_{i,j}) - \lambda\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}\right\} - \max_{i,j}\left\{\frac{w^-_{i,j}}{w^+_{i,j}} \right.$$
$$\left. \exp\left\{ - D'_\lambda(P_{i,j}||Q_{i,j}) + (1-2)\lambda\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}\right\}\right\}$$
$$\sum_{i,j} w^+_{i,j} P^+_{i,j}. \tag{75}$$

Noticing that $\sum_{i,j} w^+_{i,j} P^+_{i,j} = \epsilon^+ \leq \epsilon$ completes the proof. $\quad\square$

D. Proof of Theorem 1.

We first present the proof of Theorem 1, which closely resembles that of the Neyman-Pearson Lemma (Proposition II.D.1 in [27]). Nonetheless, we include it here for completeness.

Proof. We proceed first by showing existence. Let $\gamma \geq 0$ be the largest γ such that

$$\sum_{i,j} w^+_{i,j} P_{i,j}\left(\frac{dP_{i,j}}{dQ_{i,j}} < \frac{w^-_{i,j}}{w^+_{i,j}}\gamma\right) \leq \epsilon. \tag{76}$$

Then, if the inequality is strict, select η to be

$$\eta = \frac{\epsilon - \sum_{i,j} w^+_{i,j} P_{i,j}\left(\frac{dP_{i,j}}{dQ_{i,j}} < \frac{w^-_{i,j}}{w^+_{i,j}}\gamma\right)}{\sum_{i,j} w^+_{i,j} P_{i,j}\left(\frac{dP_{i,j}}{dQ_{i,j}} = \frac{w^-_{i,j}}{w^+_{i,j}}\gamma\right)} \tag{77}$$

otherwise choose η arbitrarily. Hence, we have that

$$\epsilon^+ = \sum_{i,j} w^+_{i,j} P^+_{i,j} = \sum_{i,j} w^+_{i,j}\left(P_{i,j}\left(\frac{dP_{i,j}}{dQ_{i,j}} < \frac{w^-_{i,j}}{w^+_{i,j}}\gamma\right) \right.$$
$$\left. + \eta P_{i,j}\left(\frac{dP_{i,j}}{dQ_{i,j}} = \frac{w^-_{i,j}}{w^+_{i,j}}\gamma\right)\right) \tag{78}$$

$$= \sum_{i,j} w^+_{i,j} P_{i,j}\left(\frac{dP_{i,j}}{dQ_{i,j}} < \frac{w^-_{i,j}}{w^+_{i,j}}\gamma\right)$$
$$+ \eta\sum_{i,j} w^+_{i,j} P_{i,j}\left(\frac{dP_{i,j}}{dQ_{i,j}} = \frac{w^-_{i,j}}{w^+_{i,j}}\gamma\right) = \epsilon. \tag{79}$$

We next show that threshold rules are optimal. We use the Lagrange multiplier $\lambda_0 > 0$ (if the optimal estimator is such that $\lambda_0 = 0$ then the constraint is inactive, and so the optimal estimator should always declare $\hat{\chi}_{i,j} = 1$) and seek to minimize $\epsilon^- + \lambda_0 \epsilon^+$. Then, we have

$$\epsilon^- + \lambda_0 \epsilon^+ = \sum_{i,j} w^-_{i,j}\mathbb{P}(\hat{\chi}_{i,j} = 0|\chi_{i,j} = 1)$$
$$+ \lambda_0 \sum_{i,j} w^+_{i,j}\mathbb{P}(\hat{\chi}_{i,j} = 1|\chi_{i,j} = 0) \tag{80}$$

11

$$= \sum_{i,j} \int_{X_1,...,X_n} w_{i,j}^- \mathbb{P}(\hat{\chi}_{i,j} = 0 | X_1,...,X_n) Q_{i,j}(X_1,..,X_n)$$
$$+ \lambda_0 w_{i,j}^+ \mathbb{P}(\hat{\chi}_{i,j} = 1 | X_1,...,X_n) P_{i,j}(X_1,..,X_n).$$
$$(81)$$

Then, to minimize $\epsilon^- + \lambda_0 \epsilon^+$, the estimator $\hat{\chi}$ should be such that $\hat{\chi}_{i,j} = 0$ if $w_{i,j}^+ Q_{i,j} \leq \lambda_0 w_{i,j}^- P_{i,j}(X_1,...,X_n)$ and $\hat{\chi}_{i,j} = 1$ otherwise. Since threshold rules are optimal, it remains to be seen how to select the threshold γ. Observe that the probabilities $P_{i,j}\left(\frac{dP_{i,j}}{dQ_{i,j}} < \frac{w_{i,j}^-}{w_{i,j}^+}\gamma\right)$ are increasing in γ for all pairs (i, j), and that the probabilities $Q_{i,j}\left(\frac{dP_{i,j}}{dQ_{i,j}} > \frac{w_{i,j}^-}{w_{i,j}^+}\gamma\right)$ are decreasing in γ. Hence, ϵ^+ is increasing in γ, and ϵ^- is decreasing in γ. Then, to minimize ϵ^-, γ should be chosen to make ϵ^+ as large as possible. $\square$

E. Proof of Theorem 2.

Proof. First, recall that $\epsilon^+ = \sum_{i,j} w_{i,j}^+ P_{i,j}^+$, and so if it holds that for all i, j, $P_{i,j}^+ \leq \epsilon$, we get $\epsilon^+ = \sum_{i,j} w_{i,j}^+ P_{i,j}^+ \leq \epsilon \sum_{i,j} w_{i,j}^+ = \epsilon$. Then, it suffices to show that for all i, j, $P_{i,j}^+ \leq \epsilon$. We begin by writing

$$P_{i,j}^+ = \int_{\boldsymbol{A}} \mathbb{P}_{\boldsymbol{A}}(\hat{\boldsymbol{\chi}}_{i,j} = 1 | \boldsymbol{\chi}_{i,j} = 0) \mathbb{P}(\boldsymbol{A} | \boldsymbol{\chi}_{i,j} = 0)$$
$$\overset{(a)}{=} \int_{\boldsymbol{A}} \mathbb{P}_{\boldsymbol{A}}(\hat{\boldsymbol{\chi}}_{i,j} = 1) \mathbb{P}(\boldsymbol{A} | \boldsymbol{\chi}_{i,j} = 0)$$
$$= \int_{\boldsymbol{A}} \int_{\boldsymbol{X}^{\backslash i,j}} \mathbb{P}_{\boldsymbol{A}}(\hat{\boldsymbol{\chi}}_{i,j} = 1 | \boldsymbol{X}^{\backslash i,j}) \mathbb{P}_{\boldsymbol{A}}(\boldsymbol{X}^{\backslash i,j}) \mathbb{P}(\boldsymbol{A} | \boldsymbol{\chi}_{i,j} = 0)$$
$$(82)$$

where $\boldsymbol{X}^{\backslash i,j}$ denotes the set of all measurements except those from the ith and jth vertex. That is, $\boldsymbol{X}^{\backslash i,j} = \{X_k^{\backslash i,j}\}_{k=1}^n$ where $X_k^{\backslash i,j} = [X_k(1),...,X_k(i-1), X_k(i+1),...,X_k(j-1), X_k(j+1),...,X_k(d)]^\top$. (a) holds since $\boldsymbol{\chi}_{i,j}$ is a deterministic function of $\boldsymbol{A}$ which gives us

$$\mathbb{P}_{\boldsymbol{A}}(\hat{\boldsymbol{\chi}}_{i,j} = 1 | \boldsymbol{\chi}_{i,j} = 0) = \frac{\mathbb{P}(\hat{\boldsymbol{\chi}}_{i,j} = 1, \boldsymbol{\chi}_{i,j} = 0, \boldsymbol{A})}{\mathbb{P}(\boldsymbol{A}, \boldsymbol{\chi}_{i,j} = 0)} \quad (83)$$

$$= \frac{\mathbb{P}(\boldsymbol{\chi}_{i,j} = 0 | \hat{\boldsymbol{\chi}}_{i,j} = 1, \boldsymbol{A}) \mathbb{P}_{\boldsymbol{A}}(\hat{\boldsymbol{\chi}}_{i,j} = 1) \pi_{\boldsymbol{A}}}{\mathbb{P}(\boldsymbol{\chi}_{i,j} = 0 | \boldsymbol{A}) \pi_{\boldsymbol{A}}} \quad (84)$$

$$= \frac{\mathbb{P}(\boldsymbol{\chi}_{i,j} = 0 | \boldsymbol{A}) \mathbb{P}_{\boldsymbol{A}}(\hat{\boldsymbol{\chi}}_{i,j} = 1) \pi_{\boldsymbol{A}}}{\mathbb{P}(\boldsymbol{\chi}_{i,j} = 0 | \boldsymbol{A}) \pi_{\boldsymbol{A}}} = \mathbb{P}_{\boldsymbol{A}}(\hat{\boldsymbol{\chi}}_{i,j} = 1). \quad (85)$$

Now, if we can say that for any $\boldsymbol{A}$ and $\boldsymbol{X}^{\backslash i,j}$ we have $\mathbb{P}_{\boldsymbol{A}}(\hat{\boldsymbol{\chi}}_{i,j} = 1 | \boldsymbol{X}^{\backslash i,j}) \leq \epsilon$, then the proof is complete since that would give us

$$\int_{\boldsymbol{A}} \int_{\boldsymbol{X}^{\backslash i,j}} \mathbb{P}_{\boldsymbol{A}}(\hat{\boldsymbol{\chi}}_{i,j} = 1 | \boldsymbol{X}^{\backslash i,j}) \mathbb{P}_{\boldsymbol{A}}(\boldsymbol{X}^{\backslash i,j}) \mathbb{P}(\boldsymbol{A} | \boldsymbol{\chi}_{i,j} = 0)$$
$$\leq \epsilon \int_{\boldsymbol{A}} \int_{\boldsymbol{X}^{\backslash i,j}} \mathbb{P}_{\boldsymbol{A}}(\boldsymbol{X}^{\backslash i,j}) \mathbb{P}(\boldsymbol{A} | \boldsymbol{\chi}_{i,j} = 0) = \epsilon.$$
$$(86)$$

We focus on the event $\hat{\boldsymbol{\chi}}_{i,j} = 1$. Notice that our algorithm only declares an edge between the vertices i and j if for *all*

sets of potential parents, which we denote as $\hat{\mathcal{Z}}(i)$ and $\hat{\mathcal{Z}}(j)$, respectively, we have that $|\hat{\sigma}_i^2 - \hat{\sigma}_j^2| > \tau_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)}$, where

$$\hat{\sigma}_i^2 = \sum_{k=1}^n \left(X_k(i) - \hat{\alpha}^\top \hat{Z}_k(i)\right)^2, \quad (87)$$

$$\hat{\sigma}_j^2 = \sum_{k=1}^n \left(X_k(j) - \hat{\beta}^\top \hat{Z}_k(j)\right)^2, \quad (88)$$

and the vectors $\hat{Z}_k(i)$ and $\hat{Z}_k(i)$ contain the kth measurements of the vertices in $\hat{\mathcal{Z}}(i)$ and $\hat{\mathcal{Z}}(j)$, respectively. The vectors $\hat{\alpha}$ and $\hat{\beta}$ are the resulting coefficient vectors obtained by performing least squares of vertices i and j on the potential parent sets $\hat{\mathcal{Z}}(i)$ and $\hat{\mathcal{Z}}(j)$, respectively. $\tau_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)}$ is chosen so that

$$2 - F_{n-p}(\tau'_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)} + n - p) + F_{n-p}(-\tau'_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)} + n - p)$$
$$- F_{n-q}(\tau'_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)} + n - q) + F_{n-q}(-\tau'_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)} + n - q)$$
$$= \epsilon,$$
$$(89)$$

where $\tau'_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)} = (\tau_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)} - |q-p|\sigma^2)/2\sigma^2$, $p = |\hat{\mathcal{Z}}(i)|$ and $q = |\hat{\mathcal{Z}}|(j)$ and F_l is the cdf of the chi-squared distributions with l-degrees of freedom.

Then, we have that

$$\mathbb{P}_{\boldsymbol{A}}(\hat{\boldsymbol{\chi}}_{i,j} = 1 | \boldsymbol{X}^{\backslash i,j}) \quad (90)$$

$$= \mathbb{P}_{\boldsymbol{A}}\left(\bigcap_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)} |\hat{\sigma}_i^2 - \hat{\sigma}_j^2| > \tau_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)} | \boldsymbol{X}^{\backslash i,j}\right) \quad (91)$$

$$\leq \mathbb{P}_{\boldsymbol{A}}\left(|\hat{\sigma}_i^{*2} - \hat{\sigma}_j^{*2}| > \tau_{\mathcal{Z}(i),\mathcal{Z}(j)} | \boldsymbol{X}^{\backslash i,j}\right) \quad (92)$$

$$\overset{(a)}{=} \mathbb{P}_{\boldsymbol{A}}\left(|\hat{\sigma}_i^{*2} - \hat{\sigma}_j^{*2} - (n-p)\sigma^2 + (n-q)\sigma^2 + (q-p)\sigma^2|\right.$$
$$\left. > \tau_{\mathcal{Z}(i),\mathcal{Z}(j)} | \boldsymbol{X}^{\backslash i,j}\right) \quad (93)$$

$$\overset{(b)}{\leq} \mathbb{P}_{\boldsymbol{A}}\left(|\hat{\sigma}_i^{*2} - (n-p)\sigma^2| + |\hat{\sigma}_j^{*2} - (n-q)\sigma^2|\right.$$
$$\left. + |(q-p)\sigma^2| > \tau_{\mathcal{Z}(i),\mathcal{Z}(j)} | \boldsymbol{X}^{\backslash i,j}\right) \quad (94)$$

$$= \mathbb{P}_{\boldsymbol{A}}\left(|\hat{\sigma}_i^{*2} - (n-p)\sigma^2| + |\hat{\sigma}_j^{*2} - (n-q)\sigma^2|\right.$$
$$\left. > \tau_{\mathcal{Z}(i),\mathcal{Z}(j)} - |q-p|\sigma^2 | \boldsymbol{X}^{\backslash i,j}\right) \quad (95)$$

$$\overset{(c)}{\leq} \mathbb{P}_{\boldsymbol{A}}\left(|\hat{\sigma}_i^{*2} - (n-p)\sigma^2| > \frac{\tau_{\mathcal{Z}(i),\mathcal{Z}(j)} - |q-p|\sigma^2}{2}\right.$$
$$\left.\bigcup |\hat{\sigma}_j^{*2} - (n-q)\sigma^2| > \frac{\tau_{\mathcal{Z}(i),\mathcal{Z}(j)} - |q-p|\sigma^2}{2} | \boldsymbol{X}^{\backslash i,j}\right) \quad (96)$$

$$\overset{(d)}{\leq} \mathbb{P}_{\boldsymbol{A}}\left(\left|\frac{\hat{\sigma}_i^{*2}}{\sigma^2} - (n-p)\right| > \tau'_{\mathcal{Z}(i),\mathcal{Z}(j)} | \boldsymbol{X}^{\backslash i,j}\right)$$
$$+ \mathbb{P}_{\boldsymbol{A}}\left(\left|\frac{\hat{\sigma}_j^{*2}}{\sigma^2} - (n-q)\right| > \tau'_{\mathcal{Z}(i),\mathcal{Z}(j)} | \boldsymbol{X}^{\backslash i,j}\right). \quad (97)$$

Where (a) holds since

$$-(n-p)\sigma^2 + (n-q)\sigma^2 + (q-p)\sigma^2 = 0. \quad (98)$$

12

(b) holds since by the triangle inequality we have

$$
\begin{aligned}
&|\hat{\sigma}_i^{*2} - \hat{\sigma}_j^{*2} - (n-p)\sigma^2 + (n-q)\sigma^2 + (q-p)\sigma^2| \\
&\leq |\hat{\sigma}_i^{*2} - (n-p)\sigma^2| + |\hat{\sigma}_j^{*2} - (n-q)\sigma^2| + |(q-p)\sigma^2|,
\end{aligned} \tag{99}
$$

and so if the left side is greater than $\tau_{\mathcal{Z}(i),\mathcal{Z}(j)}$, this implies that right side is as well. Hence,

$$
\begin{aligned}
&\Big\{ \tau_{\mathcal{Z}(i),\mathcal{Z}(j)} < \\
&|\hat{\sigma}_i^{*2} - \hat{\sigma}_j^{*2} - (n-p)\sigma^2 + (n-q)\sigma^2 + (q-p)\sigma^2| \Big\} \\
&\subseteq \Big\{ \tau_{\mathcal{Z}(i),\mathcal{Z}(j)} < |\hat{\sigma}_i^{*2} - (n-p)\sigma^2| \\
&\qquad\qquad + |\hat{\sigma}_j^{*2} - (n-q)\sigma^2| + |(q-p)\sigma^2| \Big\}.
\end{aligned} \tag{100}
$$

(c) holds due to the following fact: for any real numbers a, b, and c,

$$
a + b > c \implies a > \frac{c}{2} \text{ or } b > \frac{c}{2}, \tag{101}
$$

and (d) holds by the union bound and the definition $\tau'_{\mathcal{Z}(i),\mathcal{Z}(j)} = (\tau_{\mathcal{Z}(i),\mathcal{Z}(j)} - |q-p|\sigma^2)/2\sigma^2$. For simplicity, we turn our attention to

$$
\mathbb{P}_{\boldsymbol{A}}\left(\left| \frac{\hat{\sigma}_i^{*2}}{\sigma^2} - (n-p) \right| > \tau'_{\mathcal{Z}(i),\mathcal{Z}(j)} \Big| \boldsymbol{X}^{\backslash i,j} \right). \tag{102}
$$

First, begin by noticing (102) can be written as

$$
\begin{aligned}
\mathbb{P}_{\boldsymbol{A}}\Big(&\frac{\hat{\sigma}_i^{*2}}{\sigma^2} - (n-p) > \tau'_{\mathcal{Z}(i),\mathcal{Z}(j)} \\
&\bigcup \frac{\hat{\sigma}_i^{*2}}{\sigma^2} - (n-p) < -\tau'_{\mathcal{Z}(i),\mathcal{Z}(j)} \Big| \boldsymbol{X}^{\backslash i,j} \Big),
\end{aligned} \tag{103}
$$

which is further upper bounded by

$$
\begin{aligned}
&\mathbb{P}_{\boldsymbol{A}}\Big(\frac{\hat{\sigma}_i^{*2}}{\sigma^2} - (n-p) > \tau'_{\mathcal{Z}(i),\mathcal{Z}(j)} \Big| \boldsymbol{X}^{\backslash i,j} \Big) \\
&+ \mathbb{P}_{\boldsymbol{A}}\Big(\frac{\hat{\sigma}_i^{*2}}{\sigma^2} - (n-p) < -\tau'_{\mathcal{Z}(i),\mathcal{Z}(j)} \Big| \boldsymbol{X}^{\backslash i,j} \Big).
\end{aligned} \tag{104}
$$

We turn our attention to the first-term

$$
\mathbb{P}_{\boldsymbol{A}}\Big(\frac{\hat{\sigma}_i^{*2}}{\sigma^2} - (n-p) > \tau'_{\mathcal{Z}(i),\mathcal{Z}(j)} \Big| \boldsymbol{X}^{\backslash i,j} \Big). \tag{105}
$$

From Lemma 2, $\frac{\hat{\sigma}_i^{*2}}{\sigma^2}$ follows a chi-squared distribution with $n-p$ degrees of freedom conditioned on $\boldsymbol{X}^{\backslash i,j}$. We have that (105) is equal to

$$
1 - F_{n-p}(\tau'_{\mathcal{Z}(i),\mathcal{Z}(j)} + n - p), \tag{106}
$$

where F_{n-p} is the cumulative distribution function (cdf) of a chi-squared distribution with $n-p$ degrees of freedom. Following a similar argument, we also get that

$$
\begin{aligned}
&\mathbb{P}_{\boldsymbol{A}}\Big(\frac{\hat{\sigma}_i^{*2}}{\sigma^2} < -\tau'_{\mathcal{Z}(i),\mathcal{Z}(j)} + n - p \Big| \boldsymbol{X}^{\backslash i,j} \Big) \\
&= F_{n-p}(-\tau'_{\mathcal{Z}(i),\mathcal{Z}(j)} + n - p).
\end{aligned} \tag{107}
$$

Following the same argument, we have that

$$
\begin{aligned}
&\mathbb{P}_{\boldsymbol{A}}\Big(\Big| \frac{\hat{\sigma}_j^{*2}}{\sigma^2} - (n-q) \Big| > \tau'_{\mathcal{Z}(i),\mathcal{Z}(j)} \Big| \boldsymbol{X}^{\backslash i,j} \Big) \\
&\leq 1 - F_{n-q}(\tau'_{\mathcal{Z}(i),\mathcal{Z}(j)} + n - q) \\
&\qquad + F_{n-q}(-\tau'_{\mathcal{Z}(i),\mathcal{Z}(j)} + n - q).
\end{aligned} \tag{108}
$$

Hence, from (97) we have

$$
\begin{aligned}
&\mathbb{P}_{\boldsymbol{A}}(\hat{\chi}_{i,j} = 1 | \boldsymbol{X}^{\backslash i,j}) \leq 2 - F_{n-q}(\tau'_{\mathcal{Z}(i),\mathcal{Z}(j)} + n - q) \\
&- F_{n-p}(\tau'_{\mathcal{Z}(i),\mathcal{Z}(j)} + n - p) + F_{n-q}(-\tau'_{\mathcal{Z}(i),\mathcal{Z}(j)} + n - q) \\
&\qquad\qquad + F_{n-p}(-\tau'_{\mathcal{Z}(i),\mathcal{Z}(j)} + n - p).
\end{aligned} \tag{109}
$$

From the definition of $\tau_{\mathcal{Z}(i),\mathcal{Z}(j)}$, the right-hand side of (109) is equal to ϵ, which completes the proof. $\square$

F. Proof of Lemma 1

For notational simplicity and readability, we restate the lemma with simplified notation.

Lemma 2. *Let $\{X_k\}_{k=1}^n$ be a **fixed** dataset of p-dimensional vectors. Let $\boldsymbol{X}$ be the matrix constructed by taking the kth row as $X_k^\top$. Let y_k be defined as*

$$
y_k = X_k^\top \alpha + w_k,
$$

where α is a vector of coefficients and w_k is zero mean Gaussian noise with variance σ^2. Furthermore, let $\hat{\alpha} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top Y$, i.e., the ordinary least squares estimate of α, where $Y = [y_1, y_2, ..., y_n]^\top$. Then,

$$
\frac{1}{\sigma^2} \sum_{k=1}^n (y_k - X_k \hat{\alpha})^2,
$$

follows a chi-squared distribution with $n - p$ degrees of freedom.

Proof of Lemma 2. We begin by defining the vector of residuals as

$$
\hat{W} = Y - \boldsymbol{X}\hat{\alpha} = \boldsymbol{Q}Y = \boldsymbol{Q}(\boldsymbol{X}\alpha + W), \tag{110}
$$

where $\boldsymbol{Q} = I - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$ and $W = [w_1, w_2, ..., w_n]^\top$. Notice that $\boldsymbol{Q}\boldsymbol{X} = 0$, and that $\boldsymbol{Q}$ is a projection matrix. Hence, $\boldsymbol{Q}^2 = \boldsymbol{Q}$ which implies $\hat{W}$ is a Gaussian vector with zero mean and variance $\sigma^2 \boldsymbol{Q}$. Then, there exists some unitary matrix $\boldsymbol{U}$ that diagonalizes $\boldsymbol{Q}$ Moreover, notice that

$$
\text{Trace}(Q) \overset{(a)}{=} \text{Trace}(I - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top) = n - p \tag{111}
$$

where (a) holds since the trace is invariant to cyclic shifts. Then, if we define $\hat{W}' = \hat{W}^\top \boldsymbol{U}$, we see that $\hat{W}'$ is a Gaussian vector with zero mean and covariance $\sigma^2 \boldsymbol{U}^\top \boldsymbol{Q}^\top \boldsymbol{Q} \boldsymbol{U}$ which from the previous calculations is equal to $\sigma^2 \boldsymbol{U}^\top \boldsymbol{Q} \boldsymbol{U} = \sigma^2 \Lambda$, where Λ is the diagonal matrix of the eigenvalues of $\boldsymbol{Q}$, which can only be 0 or 1 due to $\boldsymbol{Q}$ being a projection matrix. Then, we have that p of the entries of $\hat{W}'$ are equal to zero, and so $\|\hat{W}'\|_2^2/\sigma^2$ is the sum of the squares $n - p$ independent Gaussian random variables each with mean zero and variance one, which is exactly a chi-squared distribution with $n - p$ degrees of freedom. Finally, notice that $\|\hat{W}\|_2^2 = \|\boldsymbol{U}^\top \hat{W}\| = \|\hat{W}'\|_2^2$. $\square$

13

References

[1] J. Pearl, *Causality*. Cambridge University Press, 2009.

[2] E. Testi and A. Giorgetti, "Blind wireless network topology inference," *IEEE Transactions on Communications*, vol. 69, no. 2, pp. 1109–1120, 2020.

[3] G. L. Lozano, J. I. Bravo, M. F. G. Diago, H. B. Park, A. Hurley, S. B. Peterson, E. V. Stabb, J. M. Crawford, N. A. Broderick, and J. Handelsman, "Introducing THOR, a model microbiome for genetic dissection of community behavior," *mBio*, vol. 10, no. 2, pp. e02 846–18, 2019. [Online]. Available: https://journals.asm.org/doi/abs/10.1128/mBio.02846-18

[4] E. Van den Broeck, K. Poels, and M. Walrave, "An experimental study on the effect of ad placement, product involvement and motives on facebook ad avoidance," *Telematics and Informatics*, vol. 35, no. 2, pp. 470–479, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0736585317307645

[5] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*. MIT press, 2000.

[6] J. Peters, D. Janzing, and B. Schlkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

[7] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, "DAGS with no tears: Continuous optimization for structure learning," *Advances in neural information processing systems*, vol. 31, 2018.

[8] K. Bello, B. Aragam, and P. Ravikumar, "Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 8226–8239. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/36e2967f87c3362e37cf988781a887ad-Paper-Conference.pdf

[9] D. Waxman, K. Butler, and P. M. Djurić, "Dagma-dce: Interpretable, non-parametric differentiable causal discovery," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 393–401, 2024.

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

[11] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the Lasso," *The Annals of Statistics*, vol. 34, no. 3, pp. 1436 – 1462, 2006. [Online]. Available: https://doi.org/10.1214/009053606000000281

[12] I. Ng, A. Ghassami, and K. Zhang, "On the role of sparsity and dag constraints for learning linear dags," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 943–17 954, 2020.

[13] K. Butler, G. Feng, and P. M. Djurić, "On causal discovery with convergent cross mapping," *IEEE Transactions on Signal Processing*, vol. 71, pp. 2595–2607, 2023.

[14] A. Bolstad, B. D. Van Veen, and R. Nowak, "Causal network inference via group sparse regularization," *IEEE Transactions on Signal Processing*, vol. 59, no. 6, pp. 2628–2641, 2011.

[15] C. Cui, P. Banelli, and P. M. Djurić, "Topology inference of directed graphs by Gaussian processes with sparsity constraints," *IEEE Transactions on Signal Processing*, vol. 72, pp. 2147–2159, 2024.

[16] V. Elvira and Chouzenoux, "Graphical inference in linear-gaussian state-space models," *IEEE Transactions on Signal Processing*, vol. 70, pp. 4757–4771, 2022.

[17] A. M. Alaa and M. van der Schaar, "Bayesian nonparametric causal inference: Information rates and learning algorithms," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 5, pp. 1031–1046, 2018.

[18] D. Strieder and M. Drton, "Confidence in causal inference under structure uncertainty in linear causal models with equal variances," *Journal of Causal Inference*, vol. 11, no. 1, p. 20230030, 2023. [Online]. Available: https://doi.org/10.1515/jci-2023-0030

[19] X. Kang and B. Hajek, "Lower bounds on information requirements for causal network inference," in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 754–759.

[20] A. Bellot and M. van der Schaar, "Linear deconfounded score method: Scoring dags with dense unobserved confounding," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 4, pp. 4948–4962, 2024.

[21] W. Chen, M. Drton, and Y. S. Wang, "On causal discovery with an equal-variance assumption," *Biometrika*, vol. 106, no. 4, p. 973–980, Sep. 2019. [Online]. Available: http://dx.doi.org/10.1093/biomet/asz049

[22] Z. Chen, J. Qiao, F. Xie, R. Cai, Z. Hao, and K. Zhang, "Testing conditional independence between latent variables by independence residuals," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2024.

[23] H. Zhang, S. Zhou, and J. Guan, "Measuring conditional independence by independent residuals: Theoretical results and application in causal discovery," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11555

[24] J. Shaska and U. Mitra, "Neyman-Pearson causal inference," in *2024 IEEE International Symposium on Information Theory (ISIT)*, 2024, pp. 1269–1274.

[25] ——, "Causal discovery with unequal edge error tolerance," *Presented at the Asilomar Conference of Signals, Systems, and Computers (2024), Pacific Grove, CA, USA*, 2024.

[26] J. Sun, D. Taylor, and E. M. Bollt, "Causal network inference by optimal causation entropy," *SIAM Journal on Applied Dynamical Systems*, vol. 14, no. 1, pp. 73–106, 2015.

[27] H. V. Poor, *An introduction to signal detection and estimation*. Springer Science & Business Media, 2013.

[28] H. L. Van Trees, *Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory*. John Wiley & Sons, 2004.

[29] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels. I," *Information and Control*, vol. 10, no. 1, pp. 65–103, Jan. 1967.

[30] D. M. Chickering, "Learning equivalence classes of Bayesian-network structures," *The Journal of Machine Learning Research*, vol. 2, pp. 445–498, 2002.

[31] A. Agarwal, M. Dahleh, D. Shah, and D. Shen, "Causal matrix completion," in *Proceedings of Thirty Sixth Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, G. Neu and L. Rosasco, Eds., vol. 195. PMLR, 12–15 Jul 2023, pp. 3821–3826. [Online]. Available: https://proceedings.mlr.press/v195/agarwal23c.html