Neyman-Pearson Causal Inference

Joni Shaska
University of Southern California
shaska@usc.edu

Urbashi Mitra
University of Southern California
ubli@usc.edu

***Abstract*—Motivated by controlling the errors of individual edges in the causal graph discovery problem, we propose a novel framework for causal discovery inspired by the Neyman-Pearson formulation of hypothesis testing. In particular, our formulation requires that the false negative rate is minimized while simultaneously ensuring that the false positive rate is held below a specified tolerance level. This allows us to call on techniques from binary hypothesis testing. Specifically, we derive the optimal rule for our problem, which consists of a likelihood ratio test on the edges, and derive a series of matching upper and lower bounds on the false negative rate, characterized by the Rényi divergence, which can be used as benchmarks for current discovery algorithms.**

***Index Terms*—causal inference, Neyman-Pearson hypothesis testing, Rényi divergence, converse bounds**

I. Introduction

Understanding the underlying causes of phenomena affected by multiple variables can often by 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]), impact of medications, and optimizing the impact of advertising [4].

In general causal discovery, the goal is to learn the underlying directed acyclic graph from observational data. While some methods focus on recovering the direction of a subset of edges [5], others aim to recover the full directed graph [6]–[8]. A challenge with this prior art, in the current context, is an inability to analyze the edge detection performance. Given a key equivalence class (Markov equivalence), the results of [9] imply the asymptotic consistency of greedy search algorithms, in the number of observations, though no finite data results are given. However, a challenge of greedy search, in general, is the associated computational complexity.

Full causal graph recovery is provably hard, even with access to large observational sets. Within the class of recoverable graphs, the prohibitively large search space challenges even greedy algorithms. As a result, there has been a focus on computationally feasible algorithms that exploit sparsity assumptions in the graph as well as a focus on support recovery (is an edge present or not), versus full recovery. The following [10], [11] examine sparse networks and provide sample complexity results. Causal discovery is posed as a matrix completion problem in [12], and consistency results are provided. We underscore that we seek problem frameworks where graph sparsity is not necessarily present.

This work has been funded in part by one or more of the following grants: NSF CCF-1817200, ARO W911NF1910269, DOE DE-SC0021417, Swedish Research Council 2018-04359, NSF CCF-2008927, NSF CCF-2200221, ONR 503400-78050, ONR N00014-15-1-2550.

Recent work has examined controlling error rates for individual edge detection such as [13]. An algorithm based on *causation entropy* is proposed in [14] and the *false positive* and *false negative* rates are empirically analyzed. In turn, [15] derives derives converse bounds on the achievable false negative and false positive error rates for the framework of [14]. One of our main contributions in this work, is an improvement on the bound derived in [15] (as adapted to our context) specifically for the low false positive rate regime.

We control edge error rates by formulating the support recovery problem as a constrained optimization problem similar to the Neyman-Pearson formulation of binary hypothesis testing. This allows us to use techniques from hypothesis testing and information-theoretic measures, such as the Rényi divergence, which has been used to study M-ary hypothesis testing bounds, [16], [17] and even network information flow in neural networks [18]. Our formulation has the flavor of an M-ary hypothesis testing problem if one considers all possible realizations of edges being absent and present as different hypotheses. We observe that our framework is not identical to those previously considered (*e.g.* [15]. Given our Neyman-Pearson based framework, we can tradeoff between the two error types, versus prior work which considered all types of errors equally. Our contributions are as follows:

1) We pose support detection in a causal graph 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 propose the optimal support detector, upper and lower bounds on the false negative rate, and by leveraging tools from [15], a genie-aided lower bound.
3) We investigate the performance of the optimal detector compared to other causal discovery methods, as well as our derived lower bound, and the converse bound derived in [15]. In particular, we shall see that the optimal detector outperforms: the Bayesian information criterion (BIC) [6], NOTEARS [7], and LASSO neighborhood selection [10] method.
4) We show through the numerical examples that our lower bound is tighter than the bound proposed in [15] when the false positive rate is small.

Much of our notation mimics that used in [15] and [14].

II. Problem Formulation

Consider a *directed acyclic graph* $\mathcal{G}$ with edge set $\mathcal{E}$, vertex set $\mathcal{V}$ with $|\mathcal{V}| = p$, and corresponding weighted adjacency

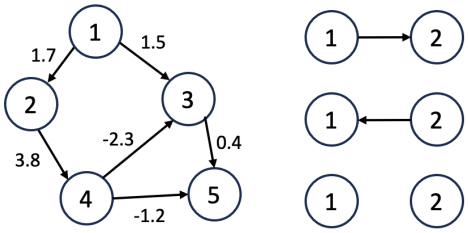


Fig. 1: (L) An example, five-node, directed, acyclic graph with edge weights. (R) Two-node graphs corresponding to the matrices $\boldsymbol{A}_0$, $\boldsymbol{A}_1$, and $\boldsymbol{A}_2$, respectively.

matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$. We assume a prior on the adjacency matrix, denoted by $\pi_{\boldsymbol{A}}$. We have a series of observations $\{X_k\}_{k=1}^n$ generated by the underlying distribution $P_{\boldsymbol{A}}$ conditioned on the realization of $\boldsymbol{A}$. For ease of exposition, we restrict ourselves to the linear, additive, exogenous inputs case, i.e.,

$$X_k \quad = \quad \boldsymbol{A}X_k + W_k, \tag{1}$$

where the exogeneous input terms $\{W_k\}_{k=1}^n$ are *i.i.d.*. 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. Neyman-Pearson Causal Discovery

We define the following error rates

$$\textit{false positive}: \quad \epsilon^+ = \frac{\mathbb{E}\sum_{i,j} \mathbb{1}\{\hat{\boldsymbol{\chi}}_{i,j} = 1, \boldsymbol{A}_{i,j} = 0\}}{\mathbb{E}\sum_{i,j} \mathbb{1}\{\boldsymbol{A}_{i,j} = 0\}}, \tag{2}$$

$$\textit{false negative}: \quad \epsilon^- = \frac{\mathbb{E}\sum_{i,j} \mathbb{1}\{\hat{\boldsymbol{\chi}}_{i,j} = 0, \boldsymbol{A}_{i,j} \neq 0\}}{\mathbb{E}\sum_{i,j} \mathbb{1}\{\boldsymbol{A}_{i,j} \neq 0\}}, \tag{3}$$

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

$$\inf_{\hat{\boldsymbol{\chi}}} \epsilon^- \quad \text{s.t. } \epsilon^+ \leq \epsilon, \tag{4}$$

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

$$H_0 : \boldsymbol{\chi}_{i,j} = 0, \qquad \{X_k\}_1^n \sim P_{i,j} \tag{5}$$

$$H_1 : \boldsymbol{\chi}_{i,j} = 1, \qquad \{X_k\}_1^n \sim Q_{i,j} \tag{6}$$

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 detection problem.

Definition 1. *We define the following probabilities and weights:*

$$P_{i,j}^+ = \mathbb{P}(\hat{\boldsymbol{\chi}}_{i,j} = 1 | \boldsymbol{\chi}_{i,j} = 0), \tag{7}$$

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

$$w_{i,j}^+ = \frac{\mathbb{P}(\boldsymbol{A}_{i,j} = 0)}{\sum_{k,l}\mathbb{P}(\boldsymbol{A}_{k,l} = 0)}, \tag{9}$$

$$w_{i,j}^- = \frac{\mathbb{P}(\boldsymbol{A}_{i,j} \neq 0)}{\sum_{k,l}\mathbb{P}(\boldsymbol{A}_{k,l} \neq 0)}. \tag{10}$$

□

With definitions 1, we can straightforwardly write out our error rates as,

$$\epsilon^+ = \sum_{i,j} w_{i,j}^+ P_{i,j}^+, \qquad \epsilon^- = \sum_{i,j} w_{i,j}^- Q_{i,j}^-, \tag{11}$$

III. Main Results

We first present the optimal detector for the Neyman-Pearson discovery problem. Due to space constraints, all proofs are relegated to the supplemental notes [19].

Proposition 1. *Let U be a uniform random variable on $[0,1]$. Then, the optimal detector $\hat{\boldsymbol{\chi}}^*$ for the Neyman-Pearson causal discovery problem is given as follows, where the per edge detector is given by,*

$$\hat{\boldsymbol{\chi}}_{i,j}^* = \begin{cases} 0, & \frac{dP_{i,j}}{dQ_{i,j}} > \frac{w_{i,j}^-}{w_{i,j}^+}\gamma \\ 0, & \frac{dP_{i,j}}{dQ_{i,j}} = \frac{w_{i,j}^-}{w_{i,j}^+}\gamma \textit{ and } U \leq \eta \\ 1, & \frac{dP_{i,j}}{dQ_{i,j}} < \frac{w_{i,j}^-}{w_{i,j}^+}\gamma \\ 1, & \frac{dP_{i,j}}{dQ_{i,j}} = \frac{w_{i,j}^-}{w_{i,j}^+}\gamma \textit{ and } U > \eta \end{cases} \tag{12}$$

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

Similar to the optimal detector in Neyman-Pearson hypothesis testing, the random variable U controls our randomization to achieve the false negative rate exactly.

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 $\hat{\boldsymbol{\chi}}^*$.

Proposition 2. *For the detector $\hat{\boldsymbol{\chi}}^*$ given in Proposition 1, we have, for any $\lambda \in [0,1]$*

$$\epsilon^- \leq \sum_{i,j} (w_{i,j}^-)^{1-\lambda}(w_{i,j}^+)^{\lambda}\frac{1}{\gamma^{\lambda}} e^{-(1-\lambda)D_{\lambda}(P_{i,j}||Q_{i,j})}, \tag{13}$$

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 the Appendix. □

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

Proposition 3. *For any detector $\hat{\boldsymbol{\chi}}$ that satisfies $\epsilon^+ \leq \epsilon$, we have that for any $\lambda \in [0,1]$.*

$$\begin{aligned}\epsilon^- \geq \frac{1}{2}\sum_{i,j} w_{i,j}^- \exp\big\{ &-(1-\lambda)D_\lambda(P_{i,j}||Q_{i,j}) \\ &-\lambda D'_\lambda(P_{i,j}||Q_{i,j}) - \lambda\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}\big\} \\ &- \epsilon \max_{i,j}\Bigg\{\frac{w_{i,j}^-}{w_{i,j}^+}\exp\big\{ - D'_\lambda(P_{i,j}||Q_{i,j}) \\ &+ (1-2\lambda)\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}\big\}\Bigg\},\end{aligned} \tag{14}$$

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

Propositions 2 and 3 show that the information-theoretic quantity that controls the false negative rate for a given tolerance ϵ is the Rényi divergence. Their proofs are provided in the Supplemental material [19].

A. Genied Aided Lower Bound

Given that our optimal detector necessitates the averaging over all possible edge combinations, it is computationally expensive to compute. We leverage the technique proposed in [15] to create a further lower bound. Specifically, we introduce a switching variable that acts as a genie, and lower bound the achievable false negative error rate for a specified tolerance level ϵ. This lower bound is considerably easier to compute compared to the optimal detector.

In the sequel, we make the following assumption.

Assumption 1. *For each (i,j), there exists a random variable S with distribution α_S such that for all X,*

$$P_{i,j}(X) = \sum_s \alpha_s P_{i,j}(X|S=s) = \sum_s \alpha_s P_{i,j,s}, \tag{15}$$

$$Q_{i,j}(X) = \sum_s \alpha_s Q_{i,j}(X|S=s) = \sum_s \alpha_s Q_{i,j,s}, \tag{16}$$

$$\text{where} \quad \begin{aligned} P_{i,j,s} &= P_{i,j}(X|S=s), \\ Q_{i,j,s} &= Q_{i,j}(X|S=s) \end{aligned}$$

□

The assumption suggests a mixture structure of the distributions $P_{i,j}$ and $Q_{i,j}$. We exploit this assumption to determine the following lower bound on the achievable false negative rate.

Proposition 4. *Suppose Assumption 1 holds. Then, for a given tolerance level $\epsilon \in (0,1)$ and any detector $\hat{\boldsymbol{\chi}}$ that satisfies $\epsilon^+ \leq \epsilon$, we have*

$$\sum_{i,j} w_{i,j}^- \sum_s \alpha_s Q_{i,j,s}\left(\frac{dP_{i,j,s}}{dQ_{i,j,s}} > \frac{w_{i,j}^-}{w_{i,j}^+}\gamma\right) \leq \epsilon^-, \tag{17}$$

where γ is chosen so that

$$\sum_{i,j} w_{i,j}^+ \sum_s \alpha_s P_{i,j,s}\left(\frac{dP_{i,j,s}}{dQ_{i,j,s}} < \frac{w_{i,j}^-}{w_{i,j}^+}\gamma\right) = \epsilon. \tag{18}$$

□

Proof Sketch of Theorem 4. Consider a detector $\hat{\boldsymbol{\chi}}^s$ that receives not only $\{X_k\}_1^n$, but also the genie information, S. Then, consider the following hypothesis testing problem for each edge,

$$H_0 : \boldsymbol{\chi}_{i,j} = 0, \qquad \{X_k\}_1^n \sim P_{i,j,s} \tag{19}$$

$$H_1 : \boldsymbol{\chi}_{i,j} = 1, \qquad \{X_k\}_1^n \sim Q_{i,j,s}. \tag{20}$$

If we define the probabilities

$$\epsilon_s^- = \sum_{i,j} w_{i,j}^- \sum_s \alpha_s \mathbb{P}(\hat{\boldsymbol{\chi}}_{i,j}^s = 0 | \boldsymbol{\chi}_{i,j} = 1, S = s), \tag{21}$$

$$\epsilon_s^+ = \sum_{i,j} w_{i,j}^+ \sum_s \alpha_s \mathbb{P}(\hat{\boldsymbol{\chi}}_{i,j}^s = 1 | \boldsymbol{\chi}_{i,j} = 0, S = s), \tag{22}$$

we have the following for any detector $\hat{\boldsymbol{\chi}}^s$ that does not use S,

$$\epsilon_s^- = \sum_{i,j} w_{i,j}^- \sum_s \alpha_s \mathbb{P}(\hat{\boldsymbol{\chi}}_{i,j}^s = 0 | \boldsymbol{\chi}_{i,j} = 1, S = s) \tag{23}$$

$$\overset{(a)}{=} \sum_{i,j} w_{i,j}^- \sum_s \alpha_s \mathbb{P}(\hat{\boldsymbol{\chi}}_{i,j} = 0 | \boldsymbol{\chi}_{i,j} = 1, S = s) \tag{24}$$

$$= \sum_{i,j} w_{i,j}^- \mathbb{P}(\hat{\boldsymbol{\chi}}_{i,j} = 0 | \boldsymbol{\chi}_{i,j} = 1) \tag{25}$$

$$\overset{(b)}{=} \sum_{i,j} w_{i,j}^- Q_{i,j}^- \overset{(c)}{=} \epsilon^-, \tag{26}$$

where (a) follows since $\hat{\boldsymbol{\chi}}^s$ is not a function of S, (b) follows from Definition 1, and (c) follows from 11. Similarly, $\epsilon_s^+ = \epsilon^+$. Then, any detector $\hat{\boldsymbol{\chi}}$ that does not use the genie information and satisfies $\epsilon^+ \leq \epsilon$ is also a valid detector for the following optimization problem,

$$\inf_{\hat{\boldsymbol{\chi}}^s} \epsilon_s^- \quad \text{s.t. } \epsilon_s^+ \leq \epsilon. \tag{27}$$

So, for any tolerance ϵ we have that

$$\inf_{\hat{\boldsymbol{\chi}}^s} \epsilon_s^- \leq \inf_{\hat{\boldsymbol{\chi}}} \epsilon^-. \tag{28}$$

Hence, solving (27) lower bounds the achievable false negative rate for Problem (4). We can use the same procedure used to derive the optimal detector in Proposition 1 to prove Proposition 4. ■

In the numerical results, we shall compare the derived optimal detector to various lower bounds. In [15], the performance of graph discovery is investigated through a Bayesian error metric (probability of error). However, we can adapt their results to provide a comparable bound for our Neyman-Pearson detection problem. To this end, we modify Proposition 2 in [15] as follows.

Proposition 5 (adapted from [15])**.** *For any detector $\hat{\boldsymbol{\chi}}$ that satisfies $\epsilon^+ \leq \epsilon$, we have for any $\lambda \in (0,1)$*

$$\epsilon^- \geq \frac{1}{2\lambda}\sum_{i,j}\left(\sqrt{1-4\lambda(1-\lambda)\rho_{i,j}^2}\right)\min\{w_{i,j}^-, w_{i,j}^+\} - \frac{1-\lambda}{\lambda}\epsilon \tag{29}$$

where

$$\rho_{i,j} = \int_{\mathcal{X}} \sqrt{\frac{dP_{i,j}}{dQ_{i,j}}} dQ_{i,j} \tag{30}$$

is the Bhattacharyya coefficient between $P_{i,j}$ and $Q_{i,j}$. □

In the numerical results, we will refer to this lower bound as the *Bhattacharyya lower bound.*

IV. Numerical Results

We consider a numerical example to illustrate our results. Our numerical results a suggest that our proposed lower bound is tighter than the one proposed in [15] when the allowed tolerance ϵ goes to zero. We compare the optimal detector to other algorithms, such as the Bayesian information criterion (BIC) [6], NOTEARS [7], and LASSO neighborhood selection [10]. We briefly summarize each method. We underscore that each method below necessitates hyper-parameter tuning; however, our method does not have any hyper-parameters. These methods optimize an objective function either via brute force search over the possible active edges in $\boldsymbol{A}$ or via a low complexity approximation.

1) *BIC*: The BIC method is a penalized likelihood detector. Given jointly Gaussian observations conditioned on the matrix $\boldsymbol{A}$, the score function is given by

$$\frac{np}{2}\log(2\pi\sigma^2)+\frac{n}{2\sigma^2}\text{Trace}\big((I-\boldsymbol{A})^\top(I-\boldsymbol{A})\hat{\Sigma}\big)+\beta\|\boldsymbol{A}\|_0, \tag{31}$$

where $\hat{\Sigma}$ is the empirical covariance matrix, β is a regularizer term, and $\|\boldsymbol{A}\|_0$ is the l_0 norm of $\boldsymbol{A}$ (the number of non-zero entries). BIC employs exhaustive search.

2) *NOTEARS*: The NOTEARS algorithm seeks to minimize the penalized squared loss,

$$\frac{1}{2n}\sum_{k=1}^{n}\|X_k-\boldsymbol{A}X_k\|_2^2+\beta\|\boldsymbol{A}\|_1, \tag{32}$$

where β 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 $\boldsymbol{A}$. It is shown in [7] that a weighted adjacency matrix $\boldsymbol{A}$ represents a directed acyclic graph if and only if

$$\text{Trace}(e^{\boldsymbol{A}\circ\boldsymbol{A}})=p, \tag{33}$$

where $\circ$ denotes the Hadamard product (element-wise multiplication). Hence, the NOTEARS algorithm converts the combinatorial optimization problem into a continuous problem, and an algorithm used to solve this continuous problem is given in [7]. The elements of the continuous-valued estimate $\hat{\boldsymbol{A}}$ are thresholded to determine the inactive edges.

3) *LASSO*: For a given observed node $X_{i,k}$, let $Y_k^i = [X_{1,k},...,X_{i-1,k},X_{i+1,k},...,X_{p,k}]^\top$ for $k=1,2,...,n$. Then, assuming a linear model for the observations, $\hat{X}_{i,k}=A_i^\top Y_{i,k}$, where A_i is a vector of coefficients, LASSO [10] minimizes the following sparsity penalized loss over the coefficients of A_i,

$$\frac{1}{2n}\sum_{k=1}^{n}\|X_{i,k}-A_i^\top Y_k^i\|_2^2+\beta\|A_i\|_1, \tag{34}$$

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

Consider a system with two nodes, see Figure 1. We receive observation $\{X_k\}_1^n$ with $X_k=[X_{1,k},X_{2,k}]^\top$, $k=1,2,...,n$. As stated before, we assume a linear system,

$$\begin{bmatrix}X_{1,k}\\X_{2,k}\end{bmatrix}=\boldsymbol{A}\begin{bmatrix}X_{1,k}\\X_{2,k}\end{bmatrix}+\begin{bmatrix}W_{1,k}\\W_{2,k}\end{bmatrix}, \tag{35}$$

where the $[W_{1,k},W_{2,k}]^\top$ 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,

$$\boldsymbol{A}_0=\begin{bmatrix}0&0\\0&0\end{bmatrix},\qquad \boldsymbol{A}_1=\begin{bmatrix}0&0\\a&0\end{bmatrix},\qquad \boldsymbol{A}_2=\begin{bmatrix}0&a\\0&0\end{bmatrix}, \tag{36}$$

where $a\in\mathbb{R}\setminus\{0\}$. For simplicity, we assume $a=1$. The corresponding graph structures are given in Figure 1. Furthermore, we assume that the prior $\pi_{\boldsymbol{A}}$ selects from $\boldsymbol{A}_0$, $\boldsymbol{A}_1$, and $\boldsymbol{A}_2$ uniformly at random. Then, it is not difficult to see that

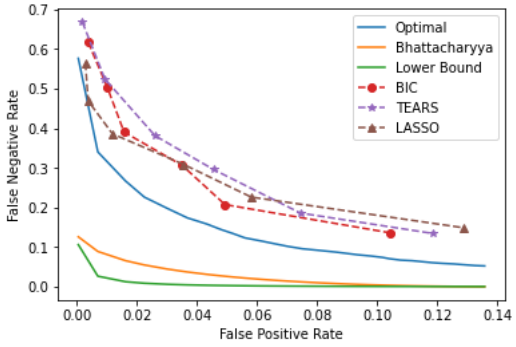
$$w_{i,j}^+=w_{i,j}^-=1. \tag{37}$$

In order to implement $\hat{\boldsymbol{\chi}}^*$ we must first compute the conditional distributions $P_{1,2}$ and $Q_{1,2}$. 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 $X_{1,k}$ and $X_{2,k}$ 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 $\boldsymbol{A}_2$; each occur with equal probability. In either case, $X_{1,k}$ and $X_{2,k}$ are jointly Gaussian random variables. Under $\boldsymbol{A}_1$ they have zero mean and covariance matrix $\sigma^2(I-\boldsymbol{A}_1)^{-1}(I-\boldsymbol{A}_1)^{-\top}$. Under $\boldsymbol{A}_2$ they have zero mean and covariance matrix $\sigma^2(I-\boldsymbol{A}_2)^{-1}(I-\boldsymbol{A}_2)^{-\top}$. Then, we have that

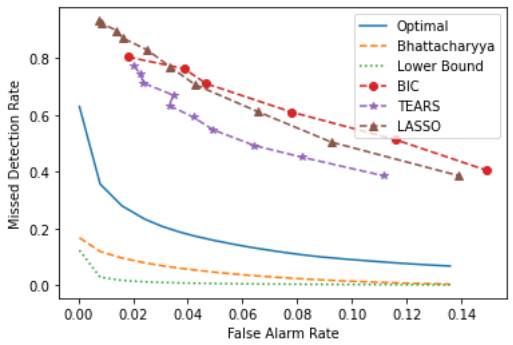
$$\begin{aligned}Q_{1,2}(X_k)=&\frac{1}{2}\frac{e^{-\frac{1}{2\sigma^2}X_k^\top(I-\boldsymbol{A}_1)^\top(I-\boldsymbol{A}_1)X_k}}{2\pi|\sigma^2(I-\boldsymbol{A}_1)^{-1}(I-\boldsymbol{A}_1)^{-\top}|}\\&+\frac{1}{2}\frac{e^{-\frac{1}{2\sigma^2}X_k^\top(I-\boldsymbol{A}_2)^\top(I-\boldsymbol{A}_2)X_k}}{2\pi|\sigma^2(I-\boldsymbol{A}_2)^{-1}(I-\boldsymbol{A}_2)^{-\top}|}.\end{aligned} \tag{38}$$

In addition to the two-node case described above, we consider the same system with four nodes. That is, all possible four-node directed acyclic graphs are equally likely to be selected, and all edge weights are equal to one.

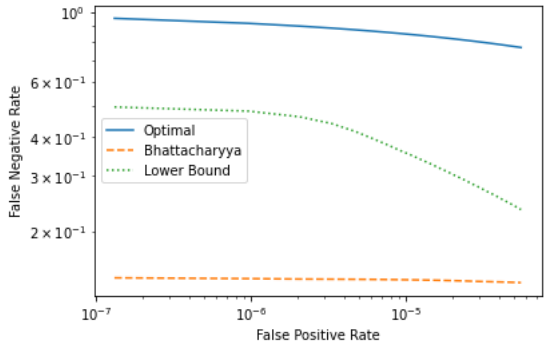
In examining Figures 2a and 2b, we see, unsurprisingly, that the derived optimal detector strongly outperforms the existing strategies with respect to minimizing the false negative rate, and this performance gap seems to increase as the number of nodes increases. We also see that our proposed lower bound is slightly looser than that adapted from the bound in [15] when the allowed tolerance ϵ is high. However, as can be seen in Figures 2c and 2d, our bound is tighter as ϵ goes to



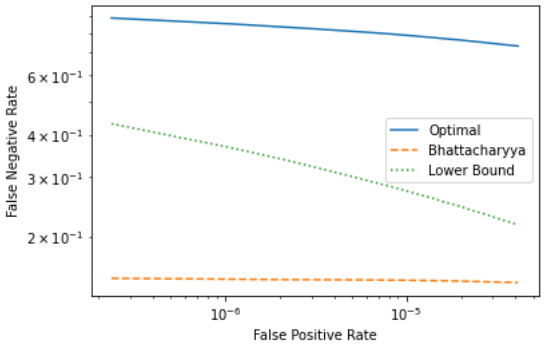
(a) Performance curves for p=2; comparing various detectors and bounds.



(b) Performance curves for p=4; comparing various detectors and bounds.



(c) Comparisons of lower bounds for $p = 2$.



(d) Comparisons of lower bounds for $p = 4$.

Fig. 2: Numerical results showcasing the performance of the optimal detector as well as the tightness of the lower bound for low false positive rates. For all the cases, $\sigma^2 = 1$ and $n = 10$. Performance curves are averaged over 1000 iterations.

zero. Moreover, the algorithms considered all perform very poorly in this low ϵ regime. In particular, if one increases the regularizers for the LASSO and BIC methods, which helps control the false positive rate, then the algorithms will begin heavily biasing graphs with no connections, effectively making the false negative rate equal to one. Although our estimator is optimal, it can become computationally infeasible to compute even for a modest number of nodes. For instance, for a ten-node graph, the number of possible directed acyclic graphs is on the order of 10^{18}. These observations facilitate the need to design computationally efficient algorithms that are robust to low false positive rates. As an example, the algorithm in [13] seeks to control critical errors (false negatives in our context) by only deleting the minimum number of edges so that the resulting graph is a directed acyclic graph. It is computationally efficient as it seeks to learn subsets of the graph.

V. Conclusions

In this paper, we have considered the problem of Neyman-Pearson causal inference, which seeks to minimize the aggregated false negative rate subject to a pre-defined tolerance on the aggregated false positive rate. We have also derived a lower bound on the achievable false negative rate for a given tolerance. Our lower bound is tighter in the regime where the false positive rate tends to zero, and current algorithms also perform poorly in this regime, motivating future research into methods that are robust to low false positive tolerance. Although not presented herein due to space limitations, we compare the genie bound to discovery algorithms on time-series data as was also considered in [15].

Appendix

A. Additional Definitions

We state classical definitions for measures that arise in our performance bounds.

Definition 2. *The Rényi divergence of order λ between two probability measures P and Q and its first two Radon-Nikodym derivatives are given by, respectively,*

$$D_\lambda(P||Q) \doteq \frac{1}{\lambda - 1} \log \int_{\mathcal{X}} \left(\frac{dP}{dQ}\right)^\lambda dQ \tag{39}$$

$$D'_\lambda(P||Q) \doteq \int_{\mathcal{X}} F_\lambda(x; P, Q) \log \frac{dP}{dQ} \tag{40}$$

$$D''_\lambda(P||Q) \doteq \int_{\mathcal{X}} F_\lambda(x; P, Q) \left(\log \frac{dP}{dQ}\right)^2 - \left(D'_\lambda(P||Q)\right)^2 \tag{41}$$

$$\textit{where} \quad F_\lambda(x; P, Q) \doteq \frac{f(x)^\lambda g(x)^{1-\lambda}}{\int_{\mathcal{X}} f(x)^\lambda g(x)^{1-\lambda} d\mu} \quad \textit{for} \quad \lambda \in [0, 1] \tag{42}$$

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

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 and P. Bühlmann, "Identifiability of gaussian structural equation models with equal error variances," *Biometrika*, vol. 101, no. 1, pp. 219–228, 2014.

[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] 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.

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

[10] 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

[11] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso)," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2183–2202, 2009.

[12] 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

[13] C. Peng, D. Zhang, and U. Mitra, "Graph identification and upper confidence evaluation for causal bandits with linear models," in *Interational Conference on Acoustics, Speech and Signal Processing*, 2024.

[14] 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.

[15] 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.

[16] I. Sason and S. Verdú, "Arimoto–rényi conditional entropy and bayesian m -ary hypothesis testing," *IEEE Transactions on Information Theory*, vol. 64, no. 1, pp. 4–25, 2018.

[17] R. Venkataramanan and O. Johnson, "A strong converse bound for multiple hypothesis testing, with applications to high-dimensional estimation," *Electronic Journal of Statistics*, vol. 12, no. 1, pp. 1126 – 1149, 2018. [Online]. Available: https://doi.org/10.1214/18-EJS1419

[18] S. Yu, K. Wickstrøm, R. Jenssen, and J. C. Príncipe, "Understanding convolutional neural networks with information theory: An initial exploration," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 435–442, 2021.

[19] J. Shaska and U. Mitra, "Neyman-pearson causal inference - supplemental notes," in *Interational Symposium on Information Theory*, 2024. [Online]. Available: https://github.com/shaskajo/NP-Causal-Inference