Sample Complexity of Distinguishing Cause from Effect

Jayadev Acharya
Cornell University

Sourbh Bhadane
Cornell University

Arnab Bhattacharyya
National University of Singapore

Saravanan Kandasamy
Cornell University

Ziteng Sun
Google Research

Abstract

We study the sample complexity of causal structure learning on a two-variable system with observational and experimental data. Specifically, for two variables X and Y, we consider the classical scenario where either X causes Y, Y causes X, or there is an unmeasured confounder between X and Y. Let m_1 be the number of observational samples of (X, Y), and let m_2 be the number of interventional samples where either X or Y has been subject to an external intervention. We show that if X and Y are over a finite domain of size k and are significantly correlated, the minimum m_2 needed is sublinear in k. Moreover, as m_1 grows, the minimum m_2 needed to identify the causal structure decreases. In fact, we can give a tight characterization of the tradeoff between m_1 and m_2 when $m_1 = O(k)$ or is sufficiently large. We build upon techniques for closeness testing when m_1 is small (e.g., sublinear in k), and for non-parametric density estimation when m_1 is large. Our hardness results are based on carefully constructing causal models whose marginal and interventional distributions form hard instances of canonical results on property testing.

1 Introduction

Reichenbach's Common Cause Principle states that if two variables X and Y are correlated, then either X causes Y, or Y causes X, or there is a common hidden variable U that causes both. We focus on the three situations depicted in Figure 1, with the goal being to discover which of the three alternatives is true.

Proceedings of the 26[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

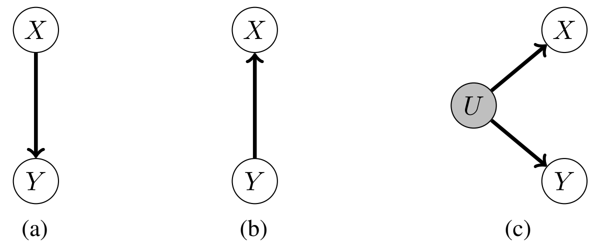

Figure 1: Three causal relationships between X and Y.

If the causal structure is $X \longrightarrow Y$, then (X, Y) is generated by the assignments: $X \coloneqq N_X$ and $Y \coloneqq f_Y(X, N_Y)$ where N_X, N_Y are independent random variables and f_Y is some (deterministic) function. The roles of X and Y are interchanged if the structure is $X \longleftarrow Y$. If the causal structure is $X \longleftarrow U \longrightarrow Y$, then (X, Y) is generated as: $U \coloneqq N_U$, $X \coloneqq f_X(U, N_X)$, and $Y \coloneqq f_Y(U, N_Y)$, where f_X, f_Y are functions and N_X, N_Y, N_U are independent random variables. In this formalism, an *intervention* corresponds to setting one of the variables to a fixed value and examining the distribution of the other. For example, if the structure is $X \longrightarrow Y$ and the intervention of fixing X to x is performed, then Y is generated as $f_Y(x, N_Y)$. We denote this intervention by $do(X = x)$.

How would one distinguish between the three possibilities shown in Figure 1? Observationally, they are impossible to distinguish as the joint distribution of (X, Y) may be exactly the same in all three cases. But a fundamental insight of Fisher (1925) was that they *can* be distinguished if interventions are allowed. For example, if the true causal structure was $X \longrightarrow Y$, then intervening on X should have an effect on Y, while intervening on Y should have no effect on X. The situation is vice versa for $X \longleftarrow Y$. On the other hand, if the causal structure was $X \longleftarrow U \longrightarrow Y$, intervention on neither X nor Y would affect the other.

In this work, we revisit the problem of recovering the correct causal structure from a quantitative point of view. While it is clear that a nonzero number of samples from interventional distributions are necessary, what is the minimum number of such samples needed? More precisely, given m_1, what is the minimum m_2 such that m_1 observations and m_2 samples from interventions suffice to distinguish between the possibilities in Figure 1? While there is a long line of work on causal structure learning while minimizing the number of experiments (e.g., Eberhardt (2007, 2008); Hauser and Bühlmann (2012); Shanmugam et al. (2015); Kocaoglu et al. (2017); Greenewald et al. (2019); Squires et al. (2020)), most of these works ignore the issue of finite sample complexity that this work addresses.

We uncover a non-trivial tradeoff between m_1 and m_2 that shows that as the number of observations increases, we need fewer and fewer (but of course, positive) number of samples from interventions. Our study holds in the setting where X and Y are random variables over a finite domain of size k and are known to be "significantly correlated". For example, we show that if $m_1 \sim k^c$ for $2/3 \leqslant c \leqslant 1$, then $m_2 \sim k^{1-c/2}$ samples are sufficient, while if m_1 is sufficiently large, m_2 can be completely independent of k. Furthermore, the tradeoffs we establish are nearly tight, in several interesting parameter regimes.

Organization We will define our problem statement in Section 1.1 and state our results precisely in Section 1.2. Related work will be discussed in Section 1.3. We present an overview of our technique in Section 1.4. Due to the page limit, we focus on the regime where $m_1 = O(k)$ in the main paper and present the algorithm and lower bound in Section 2 and Section 3 respectively. We present detailed analysis of other regimes in the supplementary material.

1.1 Problem formulation

We will work in the semantic framework of structural causal models (SCMs) (Pearl, 2009). Below we provide a complete formulation of the problem statement by tailoring SCMs to our setting.

Our goal is to test causal relationships between two correlated discrete random variables X and Y over a domain Σ of size k. We use a distance from independence as a notion of correlation below.

Definition 1 (TV-correlation)**.** *For discrete random variables X and Y define their total variation correlation as*

$$\begin{aligned}\rho_{\mathrm{TV}}(X,Y) &:= d_{\mathrm{TV}}\left(P[X,Y], P[X]P[Y]\right) \\ &= \mathrm{E}_X[d_{\mathrm{TV}}\left(P[Y], P[Y \mid X]\right)]\end{aligned}$$

where $d_{\mathrm{TV}}\left(.,.\right)$ is the total variation distance defined for any discrete distributions p and q over a domain Σ of size k as

$$d_{\mathrm{TV}}\left(p,q\right) := \sup_{S\subseteq\Sigma} p(S) - q(S) = \frac{1}{2}\sum_{x\in\Sigma} |p_x - q_x| \,.$$

Let X and Y be two correlated random variables with $\rho_{\mathrm{TV}}(X,Y) \geqslant \varepsilon$ with a causal structure given in Figure 1. We are given access to two types of samples:

- **Observational samples.** These are draws from the joint distribution $P[X,Y]$.
- **Interventional samples.** We can intervene by setting $X = x$ for some $x \in \Sigma$ (resp. $Y = y$) and observe samples of Y (resp. X) under the intervention. We denote the interventional distribution as

  $$P_x[Y] = P[Y \mid do\,(X = x)\,],$$

  and resp. as $P_y[X]$.

Note the distinction between "intervention" and "interventional samples"; the former corresponds to fixing, say X, to a certain value x whereas the latter corresponds to drawing multiple samples from the interventional distribution $P_x[Y]$. In practice, an intervention corresponds to setting up a certain medical trial and the number of interventional samples correspond to the number of people participating in the trial.

As discussed in the introduction, suppose we intervene $do\,(X = x)$, if we are in Figure 1(a), the samples of Y we obtain satisfies

$$P_x[Y] = f_Y(X, N_Y) = P[Y \mid X]$$

whereas if we are in Figure 1(b) or Figure 1(c) then we just obtain samples from $P[Y]$ as there is no causal influence from X to Y.

We now define our Causal Structure Identification problem.

Definition 2 (Causal Structure Identification)**.** *Suppose an SCM on two observable random variables X, Y supported over Σ of size k satisfies $\rho_{\mathrm{TV}}(X,Y) \geqslant \varepsilon$. Given m_1 observational and m_2 interventional samples, an algorithm solves Causal Structure Identification problem (*$\mathrm{CSI}(k,\varepsilon)$*) if with probability at least 2/3, the algorithm outputs:*

$X \longrightarrow Y$ if the true causal structure is Figure 1(a),

$Y \longrightarrow X$ if the true causal structure is Figure 1(b), and

$X \longleftarrow U \longrightarrow Y$ if the true causal structure is Figure 1(c).

Note that while the above definition is for the case of constant probability, we can boost the success of our algorithms to $1 - \delta$ for an arbitrary $\delta > 0$ by the median trick, i.e., repeating a testing algorithm $\log\left(1/\delta\right)$ times and outputting the median response which incurs only a logarithmic increase in the sample complexity.

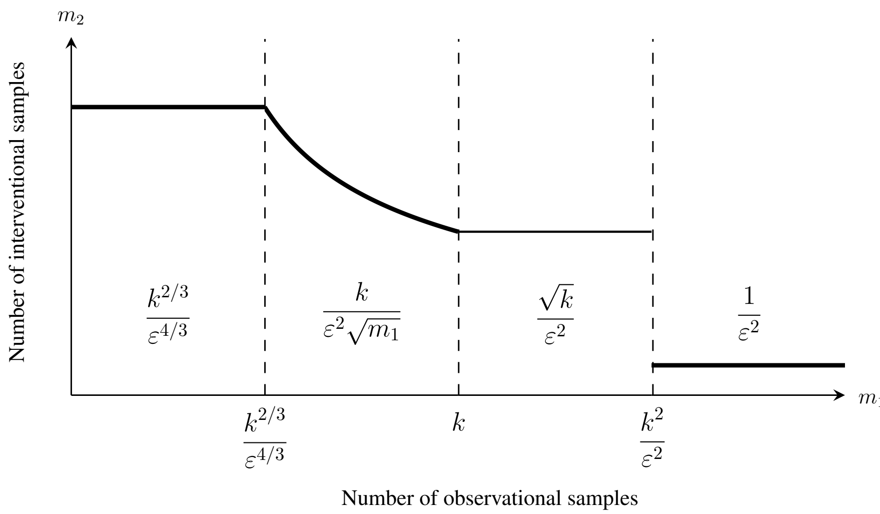

Figure 2: An illustration of the tradeoffs between m_1 and m_2 for $\mathrm{CSI}(k, \varepsilon)$ for $\varepsilon > 1/k^{1/4}$. The curves in bold are tight up to logarithmic factors. For $\varepsilon < 1/k^{1/4}$, the curve is flat at $\sqrt{k}/\varepsilon^2$ until m_1 approaches k^2/ε^2 and then becomes $O\left(1/\varepsilon^2\right)$. The x-axis is scaled such that all regimes of interest appear equal in length.

1.2 Our results

We provide sample complexity bounds for solving $\mathrm{CSI}(k, \varepsilon)$ with the number of interventional samples sublinear in k. The upper bounds we obtain have interesting phase transitions on the trade-offs between the number of interventional and observational samples in different regimes. We also show that the obtained upper bounds are tight when $m_1 = O(k)$ and $m_1 > k^2/\varepsilon^2$. While our main tradeoffs are in terms of the interventional samples, we view it as a positive that the number of distinct interventions that our algorithms need is independent of the domain size.

Based on the number of observational samples available, we present our results in the following four regimes: (1) Zero (few) observational samples: $m_1 = O(k^{2/3}/\varepsilon^{4/3})$ (2) Sublinear observational samples: $m_1 = \Omega(k^{2/3}/\varepsilon^{4/3})$ and $m_1 = O(k)$; (3) Superlinear observational samples: $m_1 = \Omega(k)$ and $m_1 = O(k^2/\varepsilon^2)$; (4) Sufficient observational samples: $m_1 = \Omega(k^2/\varepsilon^2)$. The bounds we obtain are shown in Figure 2. The above regimes are separated when $\varepsilon \geqslant 1/k^{1/4}$, which is the regime where interesting phase transitions happen. We note that all described results hold for all regimes of ε but regimes (1)-(3) will overlap and lead to the same sample for m_2 when $\varepsilon < 1/k^{1/4}$.

Zero (few) observational samples: $m_1 = O(k^{2/3}/\varepsilon^{4/3})$. Here we discuss the case when the number of observational samples is small. The next theorem shows that we can solve $\mathrm{CSI}(k, \varepsilon)$ with $m_2 = O(k^{2/3}/\varepsilon^{4/3})$ even when the number of observational samples m_1 is zero.

Theorem 1.1. *There exists an algorithm that uses zero observational samples and* $m_2 = O\left(\max(k^{2/3}/\varepsilon^{4/3}, \sqrt{k}/\varepsilon^2)\right)$ *samples from interventions to solve* $\mathrm{CSI}(k, \varepsilon)$. *Moreover, the number of distinct interventions for the algorithm is* $O\left((1/\varepsilon)\log^2(1/\varepsilon)\right)$.

Interestingly, as we will see in Theorem 1.3, the requirement on m_2 cannot be improved even when $m_1 = \Theta(k^{2/3}/\varepsilon^{4/3})$. This shows that the above interventional complexity is optimal.

Sublinear observational samples: $m_1 = O(k)$. In this case, we show that the tradeoff between observational samples and interventional samples largely resembles the tradeoff for asymmetric closeness testing (see Definition 3)(Acharya et al., 2014a; Bhattacharya and Valiant, 2015; Diakonikolas and Kane, 2016; Diakonikolas et al., 2021).

Theorem 1.2. *When* $m_1 = \Omega(k^{2/3}/\varepsilon^{4/3})$, *there exists an algorithm that takes* m_1 *observational samples and* $m_2 = O\left(\max(k/(\sqrt{m_1}\varepsilon^2), \sqrt{k}/\varepsilon^2)\right)$ *interventional samples and solves* $\mathrm{CSI}(k, \varepsilon)$. *The number of distinct interventions the algorithm makes is* $O\left((1/\varepsilon)\log^2(1/\varepsilon)\right)$.

The algorithm we use relies on asymmetric closeness testing between conditional distributions $P[Y \mid x]$ (or $P[X \mid y]$) and interventional distributions $P_x[Y]$ (or $P_y[X]$). While asymmetric closeness testing only deals with two distributions, the causal structure identification problem involves k conditional distributions and interventional distributions, which makes it trickier to handle. To resolve this, we use Levin's investment strategy(Levin, 1985; Goldreich, 2014) to select a sequence of conditional and interventional distributions to conduct closeness tests. See Section 1.4 for an overview of the technique.

We also prove a lower bound showing that the result is tight in the sublinear regime $m_1 = O(k)$.

Theorem 1.3. *For $m_1 = \Omega(k^{2/3}/\varepsilon^{4/3})$ and $m_1 = O(k)$, any algorithm that takes m_1 observational samples must take $m_2 = \Omega(\max(k/(\sqrt{m_1}\varepsilon^2), \sqrt{k}/\varepsilon^2))$ interventional samples to solve* $\mathrm{CSI}(k, \varepsilon)$.

Superlinear observational samples: $m_1 = \Omega(k)$ and $m_1 = O(k^2/\varepsilon^2)$. In this regime, we obtain a sample complexity upper bound of $O(\sqrt{k}/\varepsilon^2)$, which doesn't improve as m_1 increases. The result follows immediately from Theorem 1.2 as the upper bound in Theorem 1.2 doesn't improve when $m_1 > k$. As we will see when m_1 is sufficiently large $\Omega(k^2/\varepsilon^2)$, the interventional complexity can be further reduced. We leave improving the interventional complexity in the superlinear regime or proving hardness results as future work.

Sufficient observational samples: $m_1 = \tilde{\Omega}(k^2/\varepsilon^2)$. In this regime, we have enough observational samples to get near-"perfect" estimates of $P[Y]$ and $P[Y \mid x]$ for some x with $d_{\mathrm{TV}}(P[Y], P[Y \mid x]) = \Omega(\varepsilon)$. Hence the problem can be solved using simple hypothesis testing between $P_x[Y] = P[Y \mid x]$ or $P_x[Y] = P[Y]$, for which $O(1/\varepsilon^2)$ samples would be enough. The result is stated below.

Theorem 1.4. *When $m_1 = \tilde{\Omega}(k^2/\varepsilon^2)$, there exists an algorithm that takes m_1 observational samples and $m_2 = O(1/\varepsilon^2)$ interventional samples and solves* $\mathrm{CSI}(k, \varepsilon)$. *Moreover, the algorithm only needs to make* one *distinct intervention.*

We also show that this simple hypothesis testing approach is optimal with the following lower bound, showing that increasing m_1 beyond k^2/ε^2 does not help reducing the intervention complexity up to logarithmic factors.

Theorem 1.5. *Any algorithm that solves* $\mathrm{CSI}(k, \varepsilon)$ *requires $m_2 = \Omega(1/\varepsilon^2)$ interventional samples.*

1.3 Related Work

Causal discovery from observational and experimental data has been subject to intense study, both from the potential outcomes (Rubin, 1974; Rosenbaum and Rubin, 1983) and the graphical model (Pearl, 2009) schools of causality. For the particular case of two variables, there is also a newer line of research (Peters et al., 2017) that constrain the mechanisms underlying parent-child relationships in the causal model, allowing the causal direction to be identifiable solely from observations. A high-level account of different approaches to learn the causal direction from observations can be found in Guyon et al. (2019). The effect of sample size on causal structure discovery has been empirically studied in several contexts Mooij et al. (2016). Compton et al. (2022) obtain finite-sample results for the two-variable system under the assumption of causal sufficiency and an assumption on the entropy of the exogenous variable. Wadhwa and Dong (2021) study the sample complexity of causal discovery with multiple nodes by applying finite-sample conditional independence testers Canonne et al. (2018) to the inferred causation algorithm Pearl and Verma (1991). Bello and Honorio (2018) study the sample complexity of causal discovery for discrete causal Bayesian networks but have a negative dependence on a parameter quantifying *the minimal causal effect* that can be arbitrarily small for our setting. With access to interventions Eberhardt et al. (2010) and Yang et al. (2018) experimentally demonstrated how the sample complexity affects structure learning. Both these studies compared perfect interventions (the notion used here) with *soft interventions*; in the future, we hope to extend our theory also to soft interventions.

Some of the techniques we use were developed in the context of distribution property testing; see Canonne (2020b) for an excellent survey. Specifically, we rely on existing work for the *asymmetric closeness testing problem*, where given sample access to two distributions p and q, the question is for a given m_1 number of samples from p, how many samples m_2 are required from q such that the hypothesis $p = q$ can be distinguished from $d_{\mathrm{TV}}(p, q) > \varepsilon$ with probability at least $2/3$. The problem interpolates between the case when p is known (*identity testing*) and when p is not known (*closeness testing*). Sample complexity bounds for asymmetric closeness testing were first studied by Acharya et al. (2014b) who showed that it is sufficient to have $m_2 = O\left(\max\left\{\frac{k \log k}{\varepsilon^3 \sqrt{m_1}}, \frac{\sqrt{k \log k}}{\varepsilon^2}\right\}\right)$ samples from q, where k is the size of the support of p and q. The relation between m_1 and m_2 was made tight by the work of Bhattacharya and Valiant (2015); Diakonikolas et al. (2021), where they showed that it is sufficient to have $m_2 = O\left(\max\left\{\frac{k}{\sqrt{m_1}\varepsilon^2}, \frac{\sqrt{k}}{\varepsilon^2}\right\}\right)$, and that in fact, this is optimal.

Our setting is also related to the problem of testing against a collection of distributions Levi et al. (2013); Diakonikolas and Kane (2016), where given sample access to a collection of distributions $p_1, p_2, \ldots, p_s$, the goal is to test whether they are identical or there doesn't exist a distribution p such that $(1/s)\sum_{i=1}^{s} d_{\mathrm{TV}}(p, p_i) \leqslant \varepsilon$. Our algorithm for the zero observation case (Theorem 1.1) can be viewed as a modified version the algorithm proposed in Diakonikolas and Kane (2016) in the query model in the setting where the distribution mixture is not necessarily uniform.

1.4 Our technique

In this section, we give an overview of our algorithms and the lower bound constructions. We will focus mostly on the sublinear regime where $m_2 = O(k)$ as it exhibits the main intuition on how sublinear interventional complexity can be achieved. We briefly elaborate on the techniques employed in the zero(few) and sufficient observational samples cases and refer the reader to the supplementary material for details.

1.4.1 Sublinear Observational Samples

Our algorithm. Consider the testing problem of whether $X \longrightarrow Y$, which corresponds to Figure 1(a), or $X \not\longrightarrow Y$, which corresponds to Figure 1(b) and Figure 1(c). By symmetry, whether $Y \longrightarrow X$ or $Y \not\longrightarrow X$ can be distinguished similarly. By definition, the testing problem reduces to distinguishing the following two cases.

$X \longrightarrow Y$ if and only if $\forall x \in \Sigma, P_x[Y] = P[Y \mid X = x]$.

$X \not\longrightarrow Y$ if and only if $\forall x \in \Sigma, P_x[Y] = P[Y]$.

Observe that $\rho_{TV}(X,Y) = d_{\mathrm{TV}}\left(P[X,Y], P[X]P[Y]\right) > \varepsilon$ implies

$$\mathrm{E}_{x\sim P[X]}\left[d_{\mathrm{TV}}\left(P[Y \mid X = x], P[Y]\right)\right] > \varepsilon. \tag{1}$$

Hence there must exist $x \in \Sigma$ such that $d_{\mathrm{TV}}\left(P[Y \mid X = x], P[Y]\right) > \varepsilon$. A naive algorithm would be to intervene on all $x \in \Sigma$ and use existing techniques on asymmetric closeness testing to test whether $P_x[Y] = P[Y]$ or $d_{\mathrm{TV}}\left(P_x[Y], P[Y]\right) > \varepsilon$. Clearly, this would result in a sample complexity at least linear in k. We resolve this issue by using different methods in two regimes: (1) $m_1 = O\left(k^2/\varepsilon^2\right)$; (2) $m_1 = \Omega\left(k^2/\varepsilon^2\right)$.

When $m_1 = O\left(k^2/\varepsilon^2\right)$, our algorithm takes advantage of the fact that we can sample from $P[X]$. Consider a simple case where $\mathrm{E}_{x\sim P[X]}\left[d_{\mathrm{TV}}\left(P_x[Y], P[Y]\right)\right] = \Theta(\varepsilon)$ and all x satisfy either $d_{\mathrm{TV}}\left(P[Y \mid X = x], P[Y]\right) = 0$ (trivial element) or $d_{\mathrm{TV}}\left(P[Y \mid X = x], P[Y]\right) = \tau > \varepsilon$ (informative element). When we sample from $P[X]$, we see an informative element with probability $\Theta(\varepsilon/\tau)$ and can use $\Theta(\max\{k/\tau^2\sqrt{m_1}, \sqrt{k}/\tau^2\})$ interventional samples from $P_x[Y]$ to test whether $P_x[Y] = P[Y]$ or not using existing algorithms on closeness testing.

This shows that if τ is large, we see an informative element less often and it takes less samples to test. While if τ is small, we see an informative element more often but it also takes more samples to test. However, without knowing τ, it is hard to "invest" the right amount of samples to test $P_x[Y]$ for each x. In the general case, we resolve this by using Levin's investment strategy (Lemma 1) which shows that it is sufficient to design a collection of τ's that form a geometric sequence and do as well as if τ is fixed and known . The upper bound we obtain recovers the $O(\max\{k/\varepsilon^2\sqrt{m_1}, \sqrt{k}/\varepsilon^2\})$ rate similar to asymmetric closeness testing. See details in Section 2.

Lower bound construction. Our construction for $m_1 = O(k)$ uses the lower bound construction of Bhattacharya and Valiant (2015) for asymmetric closeness testing as a primitive. They showed that there exist distributions p and q such that given access to $O(m_1)$ samples from p any asymmetric closeness tester, requires $\Omega\left(\min\left\{\frac{k}{\sqrt{m_1}\varepsilon^2}, \frac{\sqrt{k}}{\varepsilon^2}\right\}\right)$ samples from q to distinguish $p = q$ versus $d_{\mathrm{TV}}\left(p, q\right)$. We construct q^-, a slight modification of q, such that a uniform mixture of q and q^- is p.

Using p, q, and q^- we construct SCMs with marginal and conditional distributions as follows. For simplicity, we take X as a binary random variable while Y takes values from Σ. The marginal probabilities $P[X = 0] = P[X = 1]$ are $1/2$. The conditional distributions are $P[Y \mid X = 0] = q$ and $P[Y \mid X = 1] = q^-$, thus obtaining $P[X,Y]$. Note that the marginal distribution $P[Y]$ is p by construction. Note that there exists two SCMs, under Figure 1(a) and Figure 1(c) that could generate $P[X,Y]$. We show that it is impossible to distinguish these two figures using the available samples from $P[Y \mid X = 0], P[Y \mid X = 1], P_{X=0}[Y]$ and $P_{X=1}[Y]$. To do this, we extend the wishful thinking theorem Valiant (2011) to distinguishing a collection of four distributions and show that their fourth-order moments are close. See Section 3 for details.

1.4.2 Zero(few) observational samples

Again consider testing $X \longrightarrow Y$ versus $X \not\longrightarrow Y$. We use a similar strategy as that of the sublinear observational samples algorithm with a modification. Instead of finding symbols to intervene by sampling x from $P[X]$ to test whether $P_x[Y]$ and $P[Y]$ are far, we show that $\rho_{TV}(X,Y) > \varepsilon$ implies that it is sufficient to test if two distinct interventional distributions $P_{x_1}[Y]$ and $P_{x_2}[Y]$ are far apart where (x_1, x_2) are sampled i.i.d. from $P[X]$. This holds since if $\rho_{TV}(X,Y) > \varepsilon$,

$$\mathrm{E}_{x_1,x_2\sim P[X]}\left[d_{\mathrm{TV}}\left(P[Y \mid X = x_1], P[Y \mid X = x_2]\right)\right] > \varepsilon.$$

While we do not have access to the observational marginal $P[X]$, if $X \longrightarrow Y$, we can instead simulate $P[X]$ by sampling from $P_y[X]$ for an arbitrary y. Note that if $X \not\longrightarrow Y$, then the interventional distributions $P_{x_1}[Y]$ and $P_{x_2}[Y]$ are identical.

1.4.3 Sufficient observational samples

We show that a single intervention is sufficient when we observe $\tilde{\Omega}(k^2/\varepsilon^2)$ observational samples. The crux of the algorithm in this regime is to identify the symbol to intervene upon. Assume that we want to test if $X \longrightarrow Y$ or $X \not\longrightarrow Y$. By a similar reasoning as Section 1.4.1, we want to find a $x \in \Sigma$ such that $d_{\mathrm{TV}}\left(P[Y \mid X = x], P[Y]\right) > \varepsilon$.

A folklore result states that for a discrete distribution of domain size k, the optimal sample complexity of estimating the distribution up to total variation distance ε is $\theta\left(\frac{k}{\varepsilon^2}\right)$ Canonne (2020a). Therefore, given $\tilde{\Omega}(k^2/\varepsilon^2)$ samples from the joint distribution $P[X,Y]$, empirical estimates of the marginals are $O(\varepsilon)$-close in total variation distance. However, not all conditional distributions are guaranteed to be ε-close since it is possible that if $P[X]$ is small enough, the number of samples to empirically estimate $P[Y \mid X]$ might not be sufficient. We claim that there exists a symbol $x^* \in \Sigma$ such that with a high probability there are sufficient samples to obtain ε-close estimates of $P[Y \mid X = x^*]$ and $P[Y]$ and ensure $d_{\mathrm{TV}}\left(\widehat{P}[Y \mid X = x^*], \widehat{P}[Y]\right) > \varepsilon$ which in turn implies $d_{\mathrm{TV}}\left(P[Y \mid X = x^*], P[Y]\right) > \varepsilon$. Therefore, a simple hypothesis test with $O\left(1/\varepsilon^2\right)$ samples from $P_{x^*}[Y]$ suffices.

2 Sublinear Observational Samples: Algorithm

Here we discuss trade-offs between m_1 and m_2 when m_1 is sublinear in k. We show Theorem 1.2 in this section, where we present an almost optimal algorithm that solves $\mathrm{CSI}(k, \varepsilon)$ for $m_1 = \Omega\left(k^{2/3}/\varepsilon^{4/3}\right)$. For $m_1 = O\left(k^{2/3}/\varepsilon^{4/3}\right)$, a modification of this algorithm combined with further analysis results in Theorem 1.1, which is included in supplementary material.

The following lemma is critical to our analysis.

Lemma 1. *Levin (1985); Goldreich (2014) Let D be a probability distribution, $q\colon \mathrm{supp}(D) \mapsto [0,1]$, and $\varepsilon \in (0,1]$. Suppose that $\mathrm{E}[q(s)] > \varepsilon$, and let $\ell = \lceil \log_2(2/\varepsilon) \rceil$. Then, there exists $j \in [\ell]$ such that $\mathbf{Pr}_{s\sim D}[q(s) > 2^{-j}] > 2^j \varepsilon/(\ell + 5 - j)^2$.*

Our algorithm uses the asymmetric closeness tester in Diakonikolas et al. (2021) as a primitive.

Definition 3 (Closeness Testing)**.** *Given sample access to unknown discrete distributions p and q over domain $[k]$, a closeness tester $\mathcal{CT}\left(m_1, m_2, \varepsilon, \delta\right)$ draws m_1 samples from p, m_2 samples from q and outputs "YES" if $p = q$ and "NO" if $d_{\mathrm{TV}}\left(p, q\right) > \varepsilon$ with probability atleast $1 - \delta$ where $\varepsilon, \delta > 0$.*

Lemma 2 (Theorem 1.5 in Diakonikolas et al. (2021))**.** *For discrete distributions p and q over $[k]$, there exists a computationally efficient closeness tester $\mathcal{CT}\left(m_1, m_2, \varepsilon, \delta\right)$ where $m_1 \geqslant \frac{k^{2/3}\log(1/\delta)^{1/3}}{\varepsilon^{4/3}}$ and*

$$m_2 = O\left(\frac{k\sqrt{\log(1/\delta)}}{\sqrt{m_1}\varepsilon^2} + \frac{\sqrt{k\log(1/\delta)}}{\varepsilon^2} + \frac{\log(1/\delta)}{\varepsilon^2}\right).$$

Theorem 1.2. *When $m_1 = \Omega(k^{2/3}/\varepsilon^{4/3})$, there exists an algorithm that takes m_1 observational samples and $m_2 = O\left(\max(k/(\sqrt{m_1}\varepsilon^2), \sqrt{k}/\varepsilon^2)\right)$ interventional samples and solves $\mathrm{CSI}(k, \varepsilon)$. The number of distinct interventions the algorithm makes is $O\left((1/\varepsilon)\log^2(1/\varepsilon)\right)$.*

Proof. We analyze Algorithm 1 in two parts to prove the theorem. In the first part, we test whether $X \longrightarrow Y$ or $X \not\longrightarrow Y$. If this test doesn't return $X \longrightarrow Y$, we move to the second part and use essentially the same steps to test

Algorithm 1:

Input : $\varepsilon > 0$, sample access to $P[X, Y], P_x[Y], P_y[X]$.
Output : Return the underlying graph in $\{X \to Y, Y \to X, X \leftarrow U \to Y\}$.

Let $\ell = \log\left(\frac{2}{\varepsilon}\right),\ \ell_j = (\ell + 5 - j),\ \delta_j = \frac{2^{\ell-j}}{20\ell_j^4}, s_j = \frac{2^j\varepsilon}{\ell_j^2}$;
Let $n_1^j = m_1 \cdot 2^{4(j-\ell)/3},\ n_2^j = \frac{k}{2^{-2j}\sqrt{n_1^j}}$;
for $j \in [\ell]$ **do**
  for $i \in \left[\frac{20}{s_j}\right]$ **do**
    Sample $x_i \sim P[X]$;
    For distributions $P[Y], P_{x_i}[Y]$, if $\mathcal{CT}\left(n_1^j, n_2^j\sqrt{\log(1/\delta_j)}, 2^{-j}, \delta_j\right) =$ "NO" then **return** $X \longrightarrow Y$
  end
end
for $j \in [\ell]$ **do**
  for $i \in \left[\frac{20}{s_j}\right]$ **do**
    Sample $y_i \sim P[Y]$.;
    For distributions $P[X], P_{y_i}[X]$, if $\mathcal{CT}\left(n_1^j, n_2^j\sqrt{\log(1/\delta_j)}, 2^{-j}, \delta_j\right) =$ "NO" then **return** $X \longleftarrow Y$
  end
end
return $X \longleftarrow U \longrightarrow Y$.

whether $Y \longrightarrow X$ or $Y \not\longrightarrow X$. Finally, if this test doesn't return $Y \longrightarrow X$, we return $X \longleftarrow U \longrightarrow Y$.

Test whether $X \longrightarrow Y$ or $X \not\longrightarrow Y$.

1. If $X \longrightarrow Y$, then $P_x[Y] = P[Y \mid x]$ and $P_y[X] = P[X]$.
2. If $X \not\longrightarrow Y$, then $P_x[Y] = P[Y]$.

Define $q(x) \coloneqq d_{\mathrm{TV}}\left(P[Y \mid x], P[Y]\right)$ for $x \in [k]$. Then, $\mathrm{E}_{x\sim P[X]}[q(x)] = \rho_{\mathrm{TV}}(X, Y) > \varepsilon$.

We apply Levin's investment strategy (Lemma 1), for the above choice of q. Let $\ell_j \coloneqq (\ell + 5 - j)^2$ and $s_j \coloneqq \left(2^j\varepsilon\right)/\left(\ell_j\right)^2$. Lemma 1 guarantees the existence of $j^* \in [\ell]$ such that:

$$\mathop{\mathbf{Pr}}_{x\sim P[X]}(d_{\mathrm{TV}}\left(P[Y \mid x], P[Y]\right) > 2^{-j^*}) \geqslant s_{j^*}.$$

Therefore, in $20/s_{j^*}$ samples from $P[X]$, by Chernoff bound, with probability at least $1 - e^{-10}$, there exists a sample x_i that satisfies $d_{\mathrm{TV}}\left(P[Y \mid x], P[Y]\right) > 2^{-j^*}$.

If $X \longrightarrow Y$, there exists $j^* \in [\ell]$ and a sample x_i that satisfies $d_{\mathrm{TV}}\left(P_{x_i}[Y], P[Y]\right) > 2^{-j^*}$ with probability $1 - e^{-10}$. In contrast, if $X \not\longrightarrow Y$, $P_x[Y] = P[Y]$ for every x.

Consider the following test: for every $j \in [\ell]$, Algorithm 1 samples x_i, $20/s_j$ times from $P[X]$ to distinguish

$$P_{x_i}[Y] = P[Y] \text{ and } d_{\mathrm{TV}}\left(P_{x_i}[Y], P[Y]\right) > 2^{-j}.$$

For $n_1^j = \Omega\left(k^{2/3}/2^{-4j/3}\right)$ and $n_2^j = O(k/\sqrt{n_1^j}2^{-2j})$, each closeness test requires n_1^j samples from $P[Y]$ and $n_2^j\sqrt{\log(1/\delta_j)}$ samples from $P_x[Y]$ to succeed with probability $1-\delta_j$. If any of the tests output "NO", then the algorithm returns $X \longrightarrow Y$.

Test whether $Y \longrightarrow X$ **or** $Y \not\longrightarrow X$**.** If all tests return "YES" then with a high probability, $X \not\longrightarrow Y$ and the algorithm proceeds to distinguish $Y \longrightarrow X$ and $Y \not\longrightarrow X$ using the same steps as before. Similar to the previous part, each individual test requires n_1^j samples from $P[X]$ and $n_2^j\sqrt{\log(1/\delta_j)}$ samples from $P_y[X]$ to succeed with probability $1-\delta_j$. The algorithm returns $Y \longrightarrow X$ if one of the tests outputs "NO". If all tests return "YES", then the algorithm returns $X \longleftarrow U \longrightarrow Y$.

Indeed, if $X \longleftarrow U \longrightarrow Y$, $P_x[Y] = P[Y]$ and $P_y[X] = P[X]$, and hence all of the previous closeness tests are "YES" instances.

Sample complexity. The number of samples we take from $P[Y]$ or $P[X]$ is

$$\sum_{j\in\ell}\frac{20\ell_j^2}{2^j\varepsilon}m_1 2^{4(j-\ell)/3} \leqslant 2m_1\sum_{j\in[\ell]}\frac{\ell_j^2}{2^{j-\ell}}2^{(j-\ell)/3} = 2m_1\sum_{j\in[\ell]}\ell_j^2 2^{\frac{j-\ell}{3}} = O(m_1).$$

Similarly, the total number of interventional samples taken by the algorithm in the first stage is

$$\begin{aligned}&\sum_{j\in\ell}\frac{20\ell_j^2}{2^j\varepsilon}\frac{k}{2^{-2j}\sqrt{n_1^j}}\sqrt{\log(1/\delta_j)}\\ &\leqslant \frac{2k}{\sqrt{m_1}\varepsilon^2}\cdot\sum_{j\in[\ell]}\ell_j^2 2^{\frac{4(j-\ell)}{3}}\log\frac{20\ell_j^4}{2^{j-\ell}}\\ &\leqslant \frac{2k}{\sqrt{m_1}\varepsilon^2}\cdot\sum_{j'\in[\ell]}(j'+5)^2 2^{-\frac{4j'}{3}}\log(20(j'+5)^4 2^{j'})\\ &= O\left(\frac{k}{\sqrt{m_1}\varepsilon^2}\right).\end{aligned}$$

Error analysis. Now we analyze the error probability. The total number of tests performed is at most $O\left(\sum_{j=1}^{\ell}(20/s_j)\right)$. Hence by union bound the probability of failure of these tests is at most

$$O\left(\sum_{j=1}^{\ell}\frac{20\delta_j}{s_j}\right) = \frac{1}{100}\cdot\sum_{j=1}^{\ell}O\left(\frac{1}{(\ell+5-j)^2}\right) < 1/300.$$

When $X \longrightarrow Y$, the probability of the algorithm failing to find a sample x satisfying $d_{\mathrm{TV}}\left(P_x[Y], P[Y]\right) > 2^{-j}$ is at most $1/300$. The analysis is the same for graph $X \longleftarrow Y$. Hence the algorithm returns the correct graph with error probability at most $1/150$.

Number of interventions. The number of interventions taken by the algorithm is upper bounded by the number of x_i's and y_i's drawn, which is:

$$\sum_{j\in\ell}\frac{20\ell_j^2}{2^j\varepsilon} \leqslant \frac{40}{\varepsilon}\cdot\sum_{j\in[\ell]}\ell_j^2 2^{-j} = O\left(\frac{\ell^2}{\varepsilon}\right) = O\left(\frac{\log(1/\varepsilon)^2}{\varepsilon}\right).$$

□

3 Sublinear Observational Samples: Hardness

We now prove Theorem 1.3, which establishes an almost optimal lower bound on the tradeoff between m_1 and m_2 when m_1 is $O(k)$, through a reduction to canonical results on property testing. We construct causal models under different structures (see Figure 1) with the same observational distribution. We base our construction on the hard instance for asymmetric closeness testing in Bhattacharya and Valiant (2015).

For testing causal models, extra care is needed to prove hardness since the adversary has sample access to multiple interventional distributions. To handle this, we extend Valiant (2008, 2011)'s wishful thinking theorem (see Theorem 4.6.9 in Valiant (2008)) which distinguishes two distribution pairs to the case of distinguishing two distribution quadruplets. While the extension is immediate, we remark that the constants become much larger for the quadruplet case. For completeness, we state the extension below and defer the proofs to the supplementary material.

3.1 Wishful thinking for quadruplets

For the rest of this section we denote the sequence x_1, x_2, x_3, x_4 by $x_{1:4}$.

Definition 4. *For integers $n_{1:4} > 0$, the $(n_{1:4})$-based moment $m(a_{1:4})$ of the distribution quadruplet $(p_{1:4})$ is defined as*

$$m(a_{1:4}) := \left(\prod_{i=1}^{4} n_i^{a_i}\right)\sum_{i=1}^{k} p_1^{a_1}(i)\, p_2^{a_2}(i)\, p_3^{a_3}(i)\, p_4^{a_4}(i).$$

Proposition 1. *Given integers $n_{1:4} > 0$, and two distribution quadruplets $\left(p_{1:4}^{+}\right)$ and $\left(p_{1:4}^{-}\right)$, where p_i^{+}, p_i^{-} have frequencies at most $\frac{1}{cn_i}$, for constant $c > 0$. If m^{+} and m^{-} are the $(n_{1:4})$-based moments of $\left(p_{1:4}^{-}\right)$ and $\left(p_{1:4}^{-}\right)$ respectively that satisfy*

$$\sum_{a_1+a_2+a_3+a_4>0}\frac{\left|m^{+}(a_{1:4}) - m^{-}(a_{1:4})\right|}{\sqrt{1+\max\left\{m^{+}(a_{1:4}), m^{-}(a_{1:4})\right\}}} < \frac{1}{c'}$$

for some constant $c' > 0$, then $\left(p_{1:4}^{-}\right)$ cannot be distinguished from $\left(p_{1:4}^{-}\right)$ with probability greater than 0.5 using a tester that takes $\mathrm{Poi}(n_i)$ samples from $\left(p_i^{+}, p_i^{-}\right)$ for each $i \in [4]$.

3.2 Lower Bound

Theorem 1.3. *For $m_1 = \Omega(k^{2/3}/\varepsilon^{4/3})$ and $m_1 = O(k)$, any algorithm that takes m_1 observational samples must take $m_2 = \Omega(\max(k/(\sqrt{m_1}\varepsilon^2), \sqrt{k}/\varepsilon^2))$ interventional samples to solve $\mathrm{CSI}(k, \varepsilon)$.*

Proof Sketch. We present a proof sketch and defer the details to the supplementary material.

Some definitions. Let π be a permutation of $[k]$ chosen uniformly at random. Let $A = \{\pi(1), \pi(2), \dots, \pi(a)\}$ and $B = \{\pi(a+1), \pi(a+2), \dots, \pi(a+b)\}$ be disjoint subsets of $[k]$ of size $a = (3/4)m_1$ and $b = k/C$, where C is a constant. For $S \subseteq [k]$, let $\mathbb{1}_S$ be the indicator function of set S. For permutation π, define distributions p, q and q^- over $[k]$ as following:

$$p(i) := \quad (1/m_1)\mathbb{1}_A + (1/4)(1/b)\mathbb{1}_B \qquad \forall i \in [k].$$

$$q(i) := \begin{cases} (1/m_1)\mathbb{1}_A + (1/4)(1/b)(1 + 4\varepsilon)\mathbb{1}_B & \forall i \text{ even} \\ (1/m_1)\mathbb{1}_A + (1/4)(1/b)(1 - 4\varepsilon)\mathbb{1}_B & \forall i \text{ odd.} \end{cases}$$

$$q^-(i) := \begin{cases} (1/m_1)\mathbb{1}_A + (1/4)(1/b)(1 - 4\varepsilon)\mathbb{1}_B & \forall i \text{ even} \\ (1/m_1)\mathbb{1}_A + (1/4)(1/b)(1 + 4\varepsilon)\mathbb{1}_B & \forall i \text{ odd.} \end{cases}$$

Construction. Consider an SCM on variables X and Y over support $\{0, 1\}$ and $[k]$ resp. with

1. $X \sim \mathrm{Bernoulli}\,(0.5)$
2. $P[Y \mid X = 0] = q$ and $P[Y \mid X = 1] = q^-$.

Note that $\rho_{\mathrm{TV}}(X, Y) = \varepsilon$ because each pair of the three distributions p, q, q^- has TV distance ε and the marginal distribution $P[Y]$ is p.

Analysis. Let $n_1 = n_2 = cm_1$, $n_3 = n_4 = ck\varepsilon^{-2}/\sqrt{m_1}$ for a sufficiently small c. Let $\left(p_{1:4}^{+}\right) := (q, q^-, p, p)$ and $\left(p_{1:4}^{-}\right) := (q, q^-, q, q^-)$. Suppose, for contradiction, there exists an algorithm $\mathcal{A}$ that solves $\mathrm{CSI}(k, \varepsilon)$ that uses n_1 observational samples and n_3 interventional samples. A single sample from each of the two conditionals simulates one sample of $P[X, Y]$. Hence, we consider quadruplets of the form $(P[Y \mid X = 0], P[Y \mid X = 1], P_{X=0}[Y], P_{X=1}[Y])$. It is sufficient to show that $\left(p_{1:4}^{+}\right)$ cannot be distinguished from $\left(p_{1:4}^{-}\right)$ with probability greater than 0.5 by any tester that takes $\mathrm{Poi}(n_i)$ samples from $\left(p_i^{+}, p_i^{-}\right)$. We show this in the supplementary material by bounding the difference in moments

$$\sum_{a_1+a_2+a_3+a_4>0} \frac{\left|m^{+}\left(a_{1:4}\right) - m^{-}\left(a_{1:4}\right)\right|}{\sqrt{1 + \max\left\{m^{+}\left(a_{1:4}\right), m^{-}\left(a_{1:4}\right)\right\}}}$$

for the choices of n_1, n_2, n_3, n_4, which by Proposition 1 prove the theorem.

□

4 Discussion

We view our work as the first to provide finite sample complexity guarantees for the simplest causal structure identification problem, namely categorical two-variable systems. We parameterize the system using a quantitative notion of correlation between the two variables and quantify the trade-off between observational and interventional data. Our sample complexity bounds on interventional samples, m_2, as a function of observation samples, m_1, exhibit interesting phase transitions and are tight when m_1, as a function of the domain size k, is either sublinear $O(k)$ or sufficient $\tilde{\Omega}\left(k^2/\varepsilon^2\right)$. In addition the number of interventional samples is always sublinear in k.

There are several directions for future work. The most immediate is improving upon the sample complexity of the superlinear regime and proving hardness results for the same. While the simplest property testing algorithms have only recently been studied for synthetic data Gupta and Price (2022), validating our proposed algorithms on real-world data is an important next step.

5 Acknowledgements

Jayadev Acharya and Saravanan Kandasamy are supported in part by the grant NSF-CCF-1846300 (CAREER), and a fellowship from Google. Sourbh Bhadane is supported by grants NSF-CCF-1846300 (CAREER) and NSF-CCF-1934985. Arnab Bhattacharyya is supported in part by an NRF Fellowship for AI, and part of this work was done while he was visiting the Simons Institute for the Theory of Computing.

References

J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh. Sublinear algorithms for outlier detection and generalized closeness testing. In *2014 IEEE International Symposium on Information Theory*, pages 3200–3204. IEEE, 2014a.

J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh. Sublinear algorithms for outlier detection and generalized closeness testing. In *2014 IEEE International Symposium on Information Theory*, pages 3200–3204. IEEE, 2014b.

Z. Bar-Yossef. *The complexity of massive data set computations*. University of California, Berkeley, 2002.

K. Bello and J. Honorio. Computationally and statistically efficient learning of causal bayes nets using path queries. In *Advances in Neural Information Processing Systems*, volume 31, page 10931–10941, 2018.

B. Bhattacharya and G. Valiant. Testing closeness with unequal sized samples. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL `https://proceedings.neurips.cc/paper/2015/file/5cce8dede893813f879b873962fb669f-Paper.pdf`.

C. L. Canonne. A short note on learning discrete distributions. *arXiv preprint arXiv:2002.11457*, 2020a.

C. L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Theory of Computing*, pages 1–100, 2020b.

C. L. Canonne, I. Diakonikolas, D. M. Kane, and A. Stewart. Testing conditional independence of discrete distributions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, page 735–748, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355599.

S. Compton, K. H. Greenewald, D. A. Katz, and M. Kocaoglu. Entropic causal inference: Graph identifiability. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 4311–4343. PMLR, 2022.

I. Diakonikolas and D. M. Kane. A new approach for testing properties of discrete distributions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 685–694. IEEE, 2016.

I. Diakonikolas, T. Gouleakis, D. M. Kane, J. Peebles, and E. Price. *Optimal Testing of Discrete Distributions with High Probability*, page 542–555. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450380539. URL `https://doi.org/10.1145/3406325.3450997`.

F. Eberhardt. *Causation and intervention*. PhD thesis, Carnegie Mellon University, 2007.

F. Eberhardt. Almost optimal intervention sets for causal discovery. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 161–168, 2008.

F. Eberhardt, P. Hoyer, and R. Scheines. Combining experiments to discover linear cyclic models with latent variables. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 185–192. JMLR Workshop and Conference Proceedings, 2010.

R. A. Fisher. *Statistical methods for research workers*. Oliver and Boyd, 1925.

O. Goldreich. On Multiple Input Problems in Property Testing. In K. Jansen, J. D. P. Rolim, N. R. Devanur, and C. Moore, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2014)*, volume 28 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 704–720, Dagstuhl, Germany, 2014. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-939897-74-3. doi: 10.4230/LIPIcs.APPROX-RANDOM.2014.704. URL `http://drops.dagstuhl.de/opus/volltexte/2014/4733`.

K. Greenewald, D. Katz, K. Shanmugam, S. Magliacane, M. Kocaoglu, E. Boix Adsera, and G. Bresler. Sample efficient active learning of causal trees. *Advances in Neural Information Processing Systems*, 32, 2019.

S. Gupta and E. Price. Sharp constants in uniformity testing via the huber statistic. In *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 3113–3192. PMLR, 2022.

I. Guyon, A. Statnikov, and B. B. Batu. *Cause effect pairs in machine learning*. Springer, 2019.

A. Hauser and P. Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13(1):2409–2464, 2012.

M. Kocaoglu, A. Dimakis, and S. Vishwanath. Cost-optimal learning of causal graphs. In *International Conference on Machine Learning*, pages 1875–1884. PMLR, 2017.

R. Levi, D. Ron, and R. Rubinfeld. Testing properties of collections of distributions. *Theory of Computing*, 9(8):295–347, 2013. doi: 10.4086/toc.2013.v009a008. URL `https://theoryofcomputing.org/articles/v009a008`.

L. A. Levin. One-way functions and pseudorandom generators. In *Proceedings of the Seventeenth Annual ACM Symposium on Theory of Computing*, STOC '85, page 363–365, New York, NY, USA, 1985. Association for Computing Machinery. ISBN 0897911512. doi: 10.1145/22145.22185. URL `https://doi.org/10.1145/22145.22185`.

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204, 2016.

J. Pearl. *Causality*. Cambridge university press, 2009.

J. Pearl and T. Verma. A theory of inferred causation. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, KR'91, page 441–452, San Francisco, CA, USA, 1991. Morgan Kaufmann Publishers Inc. ISBN 1558601651.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

B. Roos. On the rate of multivariate poisson convergence. In *Journal of Multivariate Analysis 69*, pages 120–134, 1999.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

K. Shanmugam, M. Kocaoglu, A. G. Dimakis, and S. Vishwanath. Learning causal graphs with small interventions. *Advances in Neural Information Processing Systems*, 28, 2015.

C. Squires, S. Magliacane, K. Greenewald, D. Katz, M. Kocaoglu, and K. Shanmugam. Active structure learning of causal dags via directed clique trees. *Advances in Neural Information Processing Systems*, 33:21500–21511, 2020.

G. Valiant and P. Valiant. The power of linear estimators. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*, pages 403–412. IEEE, 2011.

P. Valiant. *Testing symmetric properties of distributions*. PhD thesis, Massachusetts Institute of Technology, 2008.

P. Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing*, 40(6):1927–1968, 2011.

S. Wadhwa and R. Dong. On the sample complexity of causal discovery and the value of domain expertise. *CoRR*, abs/2102.03274, 2021.

K. Yang, A. Katcoff, and C. Uhler. Characterizing and learning equivalence classes of causal dags under interventions. In *International Conference on Machine Learning*, pages 5541–5550. PMLR, 2018.

A Zero(few) Observational Samples

Here we look at the case when the number of observational samples m_1 is small. The algorithmic strategy is similar to the one discussed in Section 2 for sublinear observations. For Algorithm 1 because we had access to a large number of observational samples, we were able to estimate the marginals $P[X]$ and $P[Y]$ and test if the marginals are far from interventions. Since m_1 is small here, we may not have access to the marginals. The idea here is to check if there exists interventions that are far apart, which is aided by the following lemma.

Lemma 3. *Let X and Y be two random variables with joint distribution $P[X, Y]$ such that $\rho_{\mathrm{TV}}(X, Y) \geqslant \varepsilon$. Then,*

$$\mathrm{E}_{(x_1,x_2)\sim(P[X],P[X])}[d_{\mathrm{TV}}\left(P[Y \mid x_1], P[Y \mid x_2]\right)] \geqslant \varepsilon,$$
$$\mathrm{E}_{(y_1,y_2)\sim(P[Y],P[Y])}[d_{\mathrm{TV}}\left(P[X \mid y_1], P[X \mid y_2]\right)] \geqslant \varepsilon.$$

Proof.

$$\begin{aligned}
&\mathrm{E}_{(x_1,x_2)\sim(P[X],P[X])}\left[d_{\mathrm{TV}}\left(P[Y \mid x_1], P[Y \mid x_2]\right)\right] \\
&\quad \geqslant \mathrm{E}_{x_1\sim P[X]}\left[d_{\mathrm{TV}}\left(P[Y \mid x_1], \mathrm{E}_{x_2\sim P[X]}[P[Y \mid x_2]]\right)\right] \quad \text{(By Jensen's inequality)} \\
&\quad = \mathrm{E}_{x_1\sim P[X]}\left[d_{\mathrm{TV}}\left(P[Y \mid x_1], \sum_{x_2} P[x_2]P[Y \mid x_2]\right)\right] \\
&\quad = \mathrm{E}_{x_1}\left[d_{\mathrm{TV}}\left(P[Y \mid x_1], P[Y]\right)\right] \\
&\quad = \rho_{\mathrm{TV}}(X, Y) \\
&\quad \geqslant \varepsilon.
\end{aligned}$$

Similarly, $\mathrm{E}_{(y_1,y_2)\sim(P[Y],P[Y])}[d_{\mathrm{TV}}\left(P[X \mid y_1], P[X \mid y_2]\right)] \geqslant \varepsilon$. □

Algorithm 2:

Input : $\varepsilon > 0$, sample access to $P[X, Y], P_x[Y], P_y[X]$.
Output : Return the underlying graph in $\{X \longrightarrow Y, Y \longrightarrow X, X \longleftarrow U \longrightarrow Y\}$.

Let $\ell = \log\left(\frac{2}{\varepsilon}\right)$, $\delta_j = \frac{2^{\ell-j}}{20(\ell+5-j)^4}$, $s(j) = \dfrac{20(\ell + 5 - j)^2}{\varepsilon 2^j}$;
Let $n_j = \left(\frac{k^{2/3}}{2^{-(4j/3)}} + \frac{k^{1/2}}{2^{-(2j)}}\right)\sqrt{\log\left(\frac{1}{\delta_j}\right)}$;
for $j \in [\ell]$ **do**
  for $i \in [s(j)]$ **do**
    For an arbitrary y, sample $(x_1, x_2) \sim (P_y[X], P_y[X])$;
    For distributions $P_{x_1}[Y], P_{x_2}[Y]$, if $\mathcal{CT}\left(n_j, n_j, 2^{-j}, \delta_j\right)$ = "NO" ;
    then **return** $X \to Y$;
  end
end
for $j \in [\ell]$ **do**
  for $i \in [s(j)]$ **do**
    For an arbitrary x, sample $(y_1, y_2) \sim (P_x[Y], P_x[Y])$ for arbitrary x
    For distributions $P_{y_1}[X], P_{y_2}[X]$, if $\mathcal{CT}\left(n_j, n_j, 2^{-j}, \delta_j\right)$ = "NO" ;
    then **return** $Y \to X$;
  end
end
return $X \leftarrow U \to Y$.

We now use this result prove that $\mathrm{CSI}(k, \varepsilon)$ can be solved using $O\left(k^{2/3}/\varepsilon^{4/3}\right)$ samples from interventions even with zero samples from observations.

Theorem 1.1. *There exists an algorithm that uses zero observational samples and* $m_2 = O\left(\max(k^{2/3}/\varepsilon^{4/3}, \sqrt{k}/\varepsilon^2)\right)$ *samples from interventions to solve* $\mathrm{CSI}(k, \varepsilon)$. *Moreover, the number of distinct interventions for the algorithm is* $O\left((1/\varepsilon)\log^2(1/\varepsilon)\right)$.

Similar to Algorithm 1, we analyze Algorithm 3 in two parts. In the first part, the algorithm tests whether $X \longrightarrow Y$ or $X \not\longrightarrow Y$. If this test doesn't return $X \longrightarrow Y$, the second part tests between $Y \longrightarrow X$ or $Y \not\longrightarrow X$. If this test doesn't return $Y \longrightarrow X$, then $X \longleftarrow U \longrightarrow Y$ is returned.

Proof. For $x_1 \in [k]$ and $x_2 \in [k]$, define

$$q(x_1, x_2) := d_{\mathrm{TV}}\left(P[Y \mid x_1], P[Y \mid x_2]\right).$$

Because $\rho_{\mathrm{TV}}(X, Y) \geqslant \varepsilon$, by Lemma 3, we get,

$$\mathrm{E}_{(x_1,x_2)\sim(P[X],P[X])}\left[q(x_1, x_2)\right] > \varepsilon.$$

We apply Levin's investment strategy (Lemma 1), for the above choice of q. Therefore, there exists $j^* \in [\ell]$ such that:

$$\mathop{\mathbf{Pr}}_{(x_1,x_2)\sim(P[X],P[X])}\left(d_{\mathrm{TV}}\left(P[Y \mid x_1], P[Y \mid x_2]\right) > 2^{-j^*}\right) \geqslant s(j^*)$$

where $s(j) := (2^j\varepsilon)/(\ell + 5 - j)^2$, for all $j \in [\ell]$. For such j^*, if we sample (x_1, x_2) from $(P[X], P[X])$, $20/s(j^*)$ times, by Chernoff bound, with probability at least $1 - e^{-10}$, there exists (x_1, x_2) in the samples satisfying,

$$d_{\mathrm{TV}}\left(P[Y \mid x_1], P[Y \mid x_2]\right) > 2^{-j^*}. \tag{2}$$

Part I. The first half of the algorithm considers testing whether $X \longrightarrow Y$ or $X \not\longrightarrow Y$.

1. $\mathcal{H}_0$: $X \not\longrightarrow Y$, which implies $P_x[Y] = P[Y]$, for all x;
2. $\mathcal{H}_1$: $X \longrightarrow Y$, which implies $P_x[Y] = P[Y \mid x]$ and $P_y[X] = P[X]$, for all x, y.

Consider the following test. For every $j \in [\ell]$, the algorithm repeatedly samples (x_1, x_2), $20/s(j)$ times, from $(P_y[X], P_y[X])$ for an arbitrary $y \in [k]$, and tests whether

$$d_{\mathrm{TV}}\left(P_{x_1}[Y], P_{x_2}[Y]\right) = 0 \quad \text{versus} \quad d_{\mathrm{TV}}\left(P_{x_1}[Y], P_{x_2}[Y]\right) > 2^{-j}.$$

It is shown in Diakonikolas et al. (2021) (see Lemma 2 and set $m_1 = m_2$) that the total number of (interventional) samples to test whether $d_{\mathrm{TV}}\left(P_{x_1}[Y], P_{x_2}[Y]\right)$ is zero versus greater than 2^{-j}, with probability $1 - \delta_j$, is

$$n_j := O\left(k^{(2/3)} 2^{4j/3} \log^{(1/3)}(1/\delta_j) + \left(k^{1/2}\log^{1/2}(1/\delta) + \log(1/\delta)\right) 2^{2j}\right).$$

For $\mathcal{H}_0$, $P_y[X]$ is $P[X]$ and $P_{x_i}[Y] = P[Y \mid x_i]$ for both $i \in \{1, 2\}$. Hence, with probability $1 - e^{-10}$, there exists $j^* \in [\ell]$ and a sample (x_1, x_2) that satisfies

$$d_{\mathrm{TV}}\left(P_{x_1}[Y], P_{x_2}[Y]\right) = d_{\mathrm{TV}}\left(P[Y \mid x_1], P[Y \mid x_2]\right) > 2^{-j^*}.$$

For $\mathcal{H}_1$, $P_{x_i}[Y] = P[Y]$. Hence, for any (x_1, x_2),

$$d_{\mathrm{TV}}\left(P_{x_1}[Y], P[x_2 \mid Y]\right) = d_{\mathrm{TV}}\left(P[Y], P[Y]\right) = 0.$$

If $\mathcal{H}_0$, then the algorithm outputs X causes Y. Otherwise, the algorithm proceeds to Part II.

Part II If the algorithm does not output $X \longrightarrow Y$ in Part I, then the underlying causal graph is either $X \longleftarrow Y$ or $X \longleftarrow U \longrightarrow Y$. The algorithm performs local tests similar to Part I to test whether $Y \longrightarrow X$ or $Y \not\longrightarrow X$, to return the correct graph.

Number of interventions. The number of interventions performed by the algorithm corresponds to the number of samples (x_1, x_2) drawn from $P_x[Y], P_y[X]$ is at most

$$\begin{aligned}\sum_{j\in\ell}\frac{20}{s(j)} &= 20\sum_{j\in\ell}\frac{(\ell+5-j)^2}{2^j\varepsilon}\\ &= \frac{20}{\varepsilon}\sum_{j\in\ell}(\ell+5-j)^2 2^{-j}\\ &\simeq O\left(\frac{\log^2\varepsilon}{\varepsilon}\right).\end{aligned}$$

Sample Analysis We now analyze the total number of interventional samples. Let $t\,(j) = (\ell+5-j)$. The total number of samples taken from interventions to perform the closeness tests in each part is:

$$\begin{aligned}&\sum_{j\in\ell}\frac{20\cdot t\,(j)^2}{2^j\varepsilon}\frac{k^{2/3}}{2^{-4j/3}}\sqrt{\log(1/\delta_j)}+\sum_{j\in\ell}\frac{20\cdot t\,(j)^2}{2^j\varepsilon}\frac{k^{1/2}}{2^{-2j}}\sqrt{\log(1/\delta_j)}\\
&=\sum_{j\in\ell}\frac{20t\,(j)^2}{2^j\varepsilon}\frac{k^{2/3}}{2^{-4j/3}}\sqrt{\log\frac{20\,(\ell+5-j)^4}{2^{j-\ell}}}+\sum_{j\in\ell}\frac{20\cdot t\,(j)^2}{2^j\varepsilon}\frac{k^{1/2}}{2^{-2j}}\sqrt{\log\frac{20\,(\ell+5-j)^4}{2^{j-\ell}}}\\
&\leqslant\frac{20k^{2/3}}{\varepsilon}\sum_{j\in\ell}\frac{t\,(j)^2}{2^{-j/3}}\sqrt{\log\frac{20\,(\ell+5-j)^4}{2^{j-\ell}}}+\leqslant\frac{20k^{1/2}}{\varepsilon}\sum_{j\in\ell}\frac{t\,(j)^2}{2^{-j}}\sqrt{\log\frac{20\,(\ell+5-j)^4}{2^{j-\ell}}}\\
&\leqslant\frac{20k^{2/3}}{\varepsilon}\sum_{j'\in\ell}\frac{(j'+5)^2}{2^{(j'-\ell)/3}}\sqrt{\log\frac{20\,(j'+5)^4}{2^{-j'}}}+\frac{20k^{1/2}}{\varepsilon}\sum_{j'\in\ell}\frac{(j'+5)^2}{2^{(j'-\ell)}}\sqrt{\log\frac{20\,(j'+5)^4}{2^{-j'}}}\\
&\leqslant\frac{20k^{2/3}2^{(\ell/3)}}{\varepsilon}\sum_{j'\in\ell}\frac{(j'+5)^2}{2^{(j')/3}}\sqrt{\log\frac{20\,(j'+5)^4}{2^{-j'}}}+\frac{20k^{1/2}2^{\ell}}{\varepsilon}\sum_{j'\in\ell}\frac{(j'+5)^2}{2^{(j')}}\sqrt{\log\frac{20\,(j'+5)^4}{2^{-j'}}}\\
&\leqslant\frac{20k^{2/3}}{\varepsilon^{4/3}}\sum_{j'\in\ell}\frac{(j'+5)^2}{2^{(j')/3}}\sqrt{\log\frac{20\,(j'+5)^4}{2^{-j'}}}+\frac{20k^{1/2}}{\varepsilon^2}\sum_{j'\in\ell}\frac{(j'+5)^2}{2^{(j')}}\sqrt{\log\frac{20\,(j'+5)^4}{2^{-j'}}}\\
&=O\left(\frac{k^{2/3}}{\varepsilon^{4/3}}+\frac{k^{1/2}}{\varepsilon^2}\right)\end{aligned}$$

which implies m_2 is $O\left(\frac{k^{2/3}}{\varepsilon^{4/3}}+\frac{k^{1/2}}{\varepsilon^2}\right)$.

Error Analysis. The total number of tests performed at Part I is at most $O\left(\sum_{j=1}^{\ell} s_j\right)$ where $s(j) = 20(\ell+5-j)^2/2^j\varepsilon$. Hence, by union bound, the probability of failure of these tests is at most

$$O\left(\sum_{j=1}^{\ell} s_j\frac{1}{\delta_j}\right) = \frac{1}{100}\cdot\sum_{j=1}^{\ell} O\left(\frac{1}{(\ell+5-j)^2}\right) < 1/300.$$

Similarly, for Part II, the error probability of the algorithm is at most $1/300$. Hence the algorithm returns the correct graph with error probability at most $1/150$. □

B Super-quadratic Observations

Here we analyze the tradeof between m_1 and m_2 when m_1 is $\tilde{\Omega}\left(k^2/\varepsilon^2\right)$. We show that $O(1/\varepsilon^2)$ interventional samples are sufficient and necessary to solve CSI(k,ε).

Algorithm 3:

Input : $\varepsilon > 0$, sample access to $P[X,Y], P_x[Y], P_y[X]$.
Output : Return the underlying graph in $\{X \longrightarrow Y, Y \longrightarrow X, X \longleftarrow U \longrightarrow Y\}$.

Let $m_1 \geqslant 20 \log k \cdot \frac{k^2}{\varepsilon^2}$ and S be m_1 samples drawn from $P[X,Y]$;
Let $S_x \leftarrow \{(x^*, y^*) \in S : x^* = x\}$ and $S_y \leftarrow \{(x^*, y^*) \in S : y^* = y\}$;
Let $\widehat{P}[X]$ and $\widehat{P}[Y]$ be the empirical distributions of $P[X]$ and $P[Y]$ from S;
for $j \in [k] : |S_{x_j}|$ *is* $\tilde{\Omega}\left(k/\varepsilon^2\right)$ **do**
  if $\mathcal{TT}\left(\widehat{P}[Y], P[Y \mid x_j], 8\varepsilon/10, 9\varepsilon/10, S_{x_j}, 1/(10k)\right)$ *is 'NO'* **then**
    Let $\widehat{P}[Y \mid x_j]$ be the empirical distribution of $P[Y \mid x_j]$ from S_{x_j};
    Find T such that $|\widehat{P}[Y \in T] - |\widehat{P}[Y \in T \mid x_j] > 6\varepsilon/10$;
    $R \leftarrow$ empirical distribution of $O\left(1/\varepsilon^2\right)$ samples from $P_{x_j}[Y]$ with accuracy $\varepsilon/100$;
    if $|R[Y \in T] - \widehat{P}[Y \in T \mid x_j]| \leqslant 3\varepsilon/10$ **then**
      return $X \longrightarrow Y$
    end
  end
end
for $j \in [k] : |S_{y_j}|$ *is* $\tilde{\Omega}\left(k/\varepsilon^2\right)$ **do**
  if $\mathcal{TT}\left(\widehat{P}[X], P[X \mid y_j], 8\varepsilon/10, 9\varepsilon/10, S_{y_j}, 1/(10k)\right)$ *is 'NO'* **then**
    Let $\widehat{P}[X \mid y_j]$ be the empirical distribution of $P[X \mid y_j]$ from S_{y_j};
    Find T such that $|\widehat{P}[X \in T] - |\widehat{P}[X \in T \mid y_j] > 6\varepsilon/10$;
    $R \leftarrow$ empirical distribution of $O\left(1/\varepsilon^2\right)$ samples from $P_{y_j}[X]$ with accuracy $\varepsilon/100$;
    if $|R[X \in T] - \widehat{P}[X \in T \mid y_j]| \leqslant 3\varepsilon/10$ **then**
      return $Y \longrightarrow X$
    end
  end
end
return $X \longleftarrow U \longrightarrow Y$.

B.1 Algorithm

Proof. Let S be a set of m_1 samples independently drawn from $P[X,Y]$. For $x \in \Sigma$, let $S_x := \{(x^*, y^*) \in S : x^* = x\}$ and $\tau_x := d_{\text{TV}}\left(P[Y], P[Y \mid x]\right)$. First we prove a claim that assures the existence of $x \in \Sigma$ such that we get sufficient samples on $X = x$ and also τ_x is small.

Claim 1. *With probability at least* $2/3$*, there exists* $x \in [k]$ *such that:*

1. τ_x *is at least* $\varepsilon/10$*;*
2. $|S_x|$ *is* $\Omega\left(m_1 \cdot \frac{\varepsilon^2}{k\tau_x^2}\right)$. *For* $m_1 = \tilde{\Omega}(k^2/\varepsilon^2)$, $|S_x|$ *is* $\tilde{\Omega}\left(\frac{k}{\tau_x^2}\right)$.

Proof. By Cauchy-Schwarz inequality,

$$\sum_x P(x)\tau_x^2 = \left(\sum_x P(x)\right) \cdot \left(\sum_x P(x)\tau_x^2\right) \geqslant \left(\sum_x P(x) \cdot \tau_x\right)^2 \geqslant \varepsilon^2. \tag{3}$$

Also,

$$2\sum_x P(x) \cdot \tau_x^2 \cdot \mathbb{1}_{\{\tau_x \leqslant \varepsilon/10\}} \leqslant \frac{\varepsilon^2}{100}. \tag{4}$$

Combining Equations (3) and (4),

$$\sum_x P(x)\,(\tau_x)^2\,\mathbb{1}_{\{\tau_x > \varepsilon/10\}} > 99\varepsilon^2/100.$$

Hence, there exists $x \in [k]$ that satisfies (i) $\tau_x \geqslant \frac{\varepsilon}{10}$ and

$$(ii) P(x) \cdot \tau_x^2 \geqslant \frac{99\varepsilon^2}{100k} \implies P(x) \geqslant \frac{99\varepsilon^2}{100k \cdot {\tau_x}^2}.$$

Applying Chernoff bound,

$$\begin{aligned}\mathbf{Pr}\left[|S_x| \leqslant (1+C)\,m_1 \frac{99\varepsilon^2}{100k\cdot\tau^2}\right] &\leqslant \mathbf{Pr}\left[|S_x| \leqslant (1+C)\cdot\frac{k\log k}{\varepsilon^2}\cdot\frac{99\varepsilon^2}{100k\cdot\tau^2}\right] \\ &\leqslant \exp\left(-\frac{99(C)^2\log k}{2\cdot 100\tau_x^2}\right) \leqslant \frac{1}{10k}. \qquad \text{(For any } C > 5.)\end{aligned}$$

This proves the claim. □

Let $\widehat{P}[Y]$ be the empirical distribution of $P[Y]$ and $\widehat{P}[Y \mid x]$ be the empirical conditional distribution $P[Y \mid X]$, using sample sets S and S_x both estimated upto accuracy up to $\varepsilon/100$ with error probability $1/10$.

When $\tau_{x'} > \varepsilon/10$, by triangle inequality,

$$d_{\mathrm{TV}}\left(\widehat{P}[Y], P[Y \mid x']\right) \geqslant d_{\mathrm{TV}}\left(P[Y], P[Y \mid x']\right) - d_{\mathrm{TV}}\left(\widehat{P}[Y], P[Y]\right) \geqslant \varepsilon/10 - \varepsilon/100 \geqslant 9\varepsilon/100.$$

Claim 1 indicates the existence of $x' \in \Sigma$ such that:

1. $\tau_{x'} = d_{\mathrm{TV}}\left(P[Y], P[Y \mid x']\right) > \varepsilon/10$; This implies $d_{\mathrm{TV}}\left(\widehat{P}[Y], P[Y \mid x']\right) > 9\varepsilon/100$.
2. $S_{x'}$ is $\tilde{\Omega}\left(k/{\tau_{x'}}^2\right)$.

Hence we can find one such $x_{i'}$ that satisfies the two conditions by filtering all $x \in \Sigma$ with large $|S_x| = \tilde{\Omega}\left(k/\varepsilon^2\right)$. The tolerant test $\mathcal{TT}\left(\widehat{P}[Y], P[Y \mid x_j], 8\varepsilon/10, 9\varepsilon/10, S_x, 1/(10k)\right)$ uses S_x and outputs

1. YES, if $d_{\mathrm{TV}}\left(\widehat{P}[Y], P[Y \mid x]\right) \leqslant 8\varepsilon/100$
2. NO, if $d_{\mathrm{TV}}\left(\widehat{P}[Y], P[Y \mid x]\right) > 9\varepsilon/100$

with probability $1 - 1/(10k)$[1]:

If $d_{\mathrm{TV}}\left(\widehat{P}[Y], P[Y \mid x]\right) > \frac{9}{100\varepsilon}$, then take $x' = x$.

We now have the following:

$$d_{\mathrm{TV}}\left(\widehat{P}[Y], P[Y \mid X]\right) > 9\varepsilon/10 \tag{5}$$

$$d_{\mathrm{TV}}\left(\widehat{P}[Y], P[Y]\right) \leqslant \varepsilon/100 \tag{6}$$

$$d_{\mathrm{TV}}\left(\widehat{P}[Y \mid x], P[Y \mid x]\right) \leqslant \varepsilon/100. \tag{7}$$

[1] See (Valiant and Valiant, 2011, Theorem 3 and 4) for a constant probability version and the small error probability can be obtained using the boosting trick.

Combining Equations 5 and 6 and applying triangle inequality,

$$d_{\mathrm{TV}}\left(P[Y], P[Y \mid X]\right) > 8\varepsilon/100. \tag{8}$$

Also, by triangle inequality:

$$\begin{aligned} d_{\mathrm{TV}}\left(\widehat{P}[Y], \widehat{P}[Y \mid x]\right) &\geqslant d_{\mathrm{TV}}\left(P[Y], P[Y \mid X]\right) - d_{\mathrm{TV}}\left(\widehat{P}[Y], P[Y]\right) - d_{\mathrm{TV}}\left(\widehat{P}[Y \mid x], P[Y \mid x]\right) \\ &\geqslant 6\varepsilon/100. \end{aligned} \tag{9}$$

This implies, we can compute $T \subseteq \Sigma$ such that

$$|\widehat{P}[Y \in T] - \widehat{P}[Y \in T \mid x]| > 6\varepsilon/100$$

and T satisfies:

$$\begin{aligned} |P[Y \in T] - P[Y \in T \mid x]| &\geqslant |\widehat{P}[Y \in T] - \widehat{P}[Y \in T \mid x]| - |\widehat{P}[Y \in T] - P[Y \in T]| \\ &\qquad - |\widehat{P}[Y \in T \mid x] - P[Y \in T \mid x]| \\ &\geqslant 6\varepsilon/100 - \varepsilon/100 - \varepsilon/100 = 4\varepsilon/100 \end{aligned}$$

We also have the following:

1. If $X \to Y$, $P_{x_{i'}}[Y] = P[Y \mid x_{i'}]$.
2. If $X \not\to Y$, $P_{s_{i'}}[Y] = P[Y]$.

Thus we can estimate $P_{x_{i'}}[Y]$ to test between $P[Y \in T]$ and $P[Y \in T \mid x]$ on T using $O\left(1/\varepsilon^2\right)$ samples (simple hypothesis testing).

There are at most k tolerant tests performed by the algorithm and the error probability of each of those tests is at most $1/10k$. The error probability of computing both empirical distributions $\widehat{P}[Y]$ and $\widehat{P}[Y \mid x]$ is $1/10$. Therefore, the total error probability is at most $3/10$.

Hardness. Next we show that even with infinite samples from observations, it requires at least $\Omega\left(1/\varepsilon^2\right)$ samples from interventions to solve CSI(k, ε).

Let $\mathcal{P}$ denote the set of all distributions $P : [k] \to [0, 1]$ of the form

$$\begin{aligned} P(2i-1) &= \frac{1 - 3 \cdot z_i \varepsilon}{k} \\ P(2i) &= \frac{1 + 3 \cdot z_i \varepsilon}{k} \end{aligned}$$

for all $i \in [k/2]$, where z_i is either -1 or $+1$. Let q_0 be uniformly chosen at random from $\mathcal{P}$. Let q_1 be the distribution obtained from q_0 by swapping the probabilities of odd and even coordinates (i.e., $q_1(2i) = q_0(2i-1)$ and $q_1(2i-1) = q_0(2i)$).

Let the marginal distribution $P[X]$ be $P[X = 0] = P[X = 1] = 1/2$ and the conditional distributions be $P[Y \mid X = i] = q_i$. Here $\mathrm{E}\left[d_{\mathrm{TV}}\left(\mathrm{unif}(k), q_i\right)\right] > \varepsilon$. Also the marginal $P[Y]$ is $\mathrm{unif}[k]$ because $(1/2)\, q_0 + (1/2)\, q_1 = \mathrm{unif}[k]$.

Note that it is possible to generate this joint distribution $P[X, Y] = P[X]P[Y \mid X]$ over SCMs defined on Figure 1(b) or Figure 1(c). In the former case $P_{X=i}[Y] = P[Y \mid X = i]$ is q_i, while in the later case $P_{X=i}[Y]$ is $P[Y]$ which is the uniform distribution $\mathrm{unif}[k]$. Hence we would like to distinguish the following problem: Given three distributions q_0 and q_1 and $\mathrm{unif}([k])$, distinguish

1. $\mathcal{H}_0$: When $X \to Y$ then $P_{X=i}[Y] = q_i$.
2. $\mathcal{H}_1$: When $X \leftarrow U \to Y$, $P_{X=i}[Y] = \mathrm{unif}([k])$.

by taking samples from $P_{X=i}[Y]$. Let $\mathcal{A}$ be an algorithm that distinguishes the above cases using the joint distribution and n samples from interventions. Observe that samples from $P_{X=0}[Y]$ can be simulated from $P_{X=1}[Y]$ by swapping the samples from the adjacent even and odd coordinates. Hence we can assume without loss of generality that A takes samples from exactly one of the interventions $P_{X=i}[Y]$. This is equivalent to distinguishing $\text{unif}[k]$ versus q_0, which requires $\Omega\left(1/\varepsilon^2\right)$ samples by standard lower bounds for hypothesis testing since the $H^2(\text{unif}[k], q_0) = \Theta(\varepsilon^2)$, where $H^2(\cdot,\cdot)$ denotes the squared Hellinger distance.[2]. $\square$

C Proof of Theorem 1.3

We complete the proof of Theorem 1.3 by bounding the difference in moments m^+ and m^-.

For all $r, s, t, u \geqslant 0$, $m^+(r,s,t,u)$ is

$$= n_1^r n_2^s n_3^t n_4^u \left(\left(\frac{3}{4}\right)\left(\frac{1}{m_1}\right)^{r+s+t+u-1} + h(\varepsilon,r,s)\left(\frac{1}{4}\right)^{r+s+t+u}\left(\frac{C}{k}\right)^{r+s+t+u-1}\right)$$

and similarly, $m^-(r,s,t,u)$ is

$$= n_1^r n_2^s n_3^t n_4^u \left(\left(\frac{3}{4}\right)(1/m_1)^{r+s+t+u-1} + h'(\varepsilon,r,s,t,u)\left(\frac{1}{4}\right)^{r+s+t+u}\left(\frac{C}{k}\right)^{r+s+t+u-1}\right)$$

where

$$h(\varepsilon,r,s) = \frac{(1+\varepsilon)^r(1-\varepsilon)^s + (1-\varepsilon)^r(1+\varepsilon)^s}{2} \qquad \text{and}$$

$$h'(\varepsilon,r,s,t,u) = \frac{(1+\varepsilon)^{r+t}(1-\varepsilon)^{s+u} + (1-\varepsilon)^{r+t}(1+\varepsilon)^{s+u}}{2}.$$

Then,

$$\begin{aligned}
&\frac{|m^+(n_1,n_2,n_3) - m^-(n_1,n_2,n_3)|}{\sqrt{1+\max\{m^+(n_1,n_2,n_3), m^-(n_1,n_2,n_3)\}}} \\
&\leqslant \frac{n_1^r n_2^s n_3^t n_4^u (1/4)^{r+s+t+u}(C/k)^{r+s+t+u-1}(h'(\varepsilon,r,s,t,u) - h(\varepsilon,s,t))}{\sqrt{n_1^r n_2^s n_3^t n_4^u\left((3/4)(1/m_1)^{r+s+t+u-1} + (1/4)^{r+s+t+u}(C/k)^{r+s+t+u-1}\right)}} \\
&\leqslant \frac{n_1^r n_2^s n_3^t n_4^u (1/4)^{r+s+t+u}(C'/k)^{r+s+t+u-1}}{\sqrt{n_1^r n_2^s n_3^t n_4^u\left((3/4)(1/m_1)^{r+s+t+u-1} + (1/4)^{r+s+t+u}(C/k)^{r+s+t+u-1}\right)}} \qquad \text{for } C' > C \\
&\leqslant \frac{n_1^r n_2^s n_3^t n_4^u (1/4)^{r+s+t+u}(C/k)^{r+s+t+u-1}}{\sqrt{n_1^r n_2^s n_3^t n_4^u\left((3/4)(1/m_1)^{r+s+t+u-1}\right)}} \qquad \text{(because } m_1 < k) \\
&\leqslant n_1^{r/2} n_2^{s/2} n_3^{t/2} n_4^{u/2} {m_1}^{(r+s+t+u-1)/2}(C/k)^{r+s+t+u-1} \\
&\leqslant c^{(r+s+t+u)/2}\cdot\frac{{m_1}^{(r+s)/2}}{{m_1}^{(t+u)/4}}\cdot\frac{{m_1}^{(r+s+t+u-1)/2}}{(\varepsilon^2)^{t/2+u/2}}\cdot\frac{C^{r+s+t+u-1}}{k^{r+s+t+u-1}}\cdot k^{(t+u)/2} \qquad \text{(using } n_1,n_2,n_3,n_4) \\
&\leqslant \hat{c}^{(r+s+t+u)/2}\frac{{m_1}^{r+s}}{k^{r+s}}\cdot\left(\frac{\sqrt{n_1}}{\varepsilon^2 k}\right)^{t/2+u/2+1} \qquad \text{(for small } \hat{c}) \\
&\leqslant \hat{c}^{(r+s+t+u)/2}\frac{{m_1}^{r+s}}{k^{r+s}}\cdot\left(\frac{\sqrt{n_1}}{\sqrt{k}}\right)^{t/2+u/2+1} \qquad \text{(substituting } \varepsilon^2 > k^{-1/2}) \\
&\leqslant \hat{c}^{(r+s+t+u)/2}
\end{aligned}$$

which is small when $\hat{c}$ is small.

[2]Mostly a folklore. See Bar-Yossef (2002) for a proof.

D Wishful thinking for quadruplets

For the sake of completeness, we state intermediate lemmas, that are immediate extensions of the results in Valiant (2008), for the case of distinguishing quadruplets of distributions.

Poissonization. For a symmetric property of the distributions, it suffices to analyze the distribution of fingerprint of samples from the quadruplets. We first consider a (n_1, n_2, n_3, n_4)-Poissonized tester that correctly classifies a symmetric property on a distribution quadruplet (p_1, p_2, p_3, p_4) with probability $\frac{49}{96}$ assuming Poisson sampling from each of the distributions. An extension of (Valiant, 2008, Lemma 4.6.4) for quadruplets establishes existence of a (n_1, n_2, n_3, n_4)-sample tester without assuming Poisson sampling.

Fingerprint distribution approximation by multivariate Poisson distributions. Like in (Valiant, 2008, Lemma 4.6.5), the distribution of fingerprints of Poi(n_1) samples from p_1, Poi(n_2) samples from p_2, Poi(n_3) samples from p_3 and Poi(n_4) samples from p_4 is a generalized multinomial distribution M^ρ where ρ is a matrix with k rows and columns indexed by fingerprint indices (a_1, a_2, a_3, a_4). We invoke Roos's theorem to approximate the multinomial distribution by multivariate Poisson distributions as in Valiant (2008).

Proposition 2. *(Roos's Theorem Roos (1999)) Given a matrix ρ, letting $\overrightarrow{\lambda}(a) = \sum_i \rho(i, a)$ be the vector of column sums,*

$$d_{\mathrm{TV}}\left(M^\rho, \textit{Poi}(\overrightarrow{\lambda})\right) \leqslant 8.8 \sum_a \frac{\sum_i \rho(i,(a))^2}{\sum_i \rho(i,(a))}.$$

For low-frequency distribution Lemma 4.6.6 in Valiant (2008) shows that the right-hand side above is small, thus enabling the approximation of a generalized multinomial distribution by multivariate Poisson distributions. The same is easily extended for quadruplets below.

Proposition 3 (Extension of Lemma 4.6.6 in Valiant (2008)). *Given p_1, p_2, p_3, p_4, integers n_1, n_2, n_3, n_4 and a real number $0 < c \leqslant 0.5$, such that for all $i \in [k], j \in [4]$, $p_j(i) \leqslant \frac{c}{n_j}$, if ρ is the matrix with $(i, (a_1, a_2, a_3, a_4))$ entry $\prod_j \textit{poi}(a_j; k_j p_j(i))$, then*

$$\sum_{a_1+a_2+a_3+a_4>0} \frac{\sum_i \rho(i,(a_1,a_2,a_3,a_4))^2}{\sum_i \rho(i,(a_1,a_2,a_3,a_4))} \leqslant 16c.$$

Moment-based bound. We can now extend (Valiant, 2008, Lemma 4.6.7) to show that if the total variation distance between the multivariate Poisson distributions of a pair of quadruplets with low-frequency elements is small, then no tester can distinguish between the pairs.

Proposition 4. *For two distribution quadruplets $\left\{p_j^+\right\}_{j=1}^4$ and $\left\{p_j^-\right\}_{j=1}^4$, where p_j^+, p_j^- have frequencies at most $\frac{1}{30000 n_j}$ for $j \in [4]$, if $\overrightarrow{\lambda}^+(a_1,a_2,a_3,a_4) = \sum_i \prod_j \textit{poi}(a_j; k_j p_j^+(i))$ and $\overrightarrow{\lambda}^-(a_1,a_2,a_3,a_4) = \sum_i \prod_j \textit{poi}(a_j; k_j p_j^-(i))$ for $a_1+a_2+a_3+a_4 > 0$ and if*

$$\sum_{a_1+a+2+a_3+a_4>0} \frac{\left|\overrightarrow{\lambda}^+(a_1,a_2,a_3,a_4) - \overrightarrow{\lambda}^-(a_1,a_2,a_3,a_4)\right|}{\sqrt{1+\max\left\{\overrightarrow{\lambda}^+(a_1,a_2,a_3,a_4), \overrightarrow{\lambda}^-(a_1,a_2,a_3,a_4)\right\}}} < \frac{1}{200},$$

then it is impossible to test any symmetric property that is true for $\left\{p_j^+\right\}_{j=1}^4$ and false for $\left\{p_j^-\right\}_{j=1}^4$ in (n_1,n_2,n_3,n_4) samples.

All that remains is converting the above expression in terms of the moments of distribution quadruplets (see Definition 4). By the same proof as that of Theorem 4.6.9 in Valiant (2008), adapted for quadruplets, we have Proposition 1.