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.

Figure 1: Three causal relationships between $X$ and $Y$.

If the causal structure is $X \rightarrow Y$, then $(X,Y)$ is generated by the assignments: $X := N_X$ and $Y := f_Y(X,N_Y)$ 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 $X \leftarrow Y$. If the causal structure is $X \leftarrow U \rightarrow Y$, then $(X,Y)$ is generated as: $U := N_U$, $X := f_X(U,N_X)$, and $Y := f_Y(U,N_Y)$, 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 $X \rightarrow Y$ 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 $X \rightarrow Y$, 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 $X \leftarrow Y$. On the other hand, if the causal structure was $X \leftarrow U \rightarrow Y$, 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 inter-
ventional 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); Shannugam 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 if $m_1 \approx k^c$ for $2/3 \leq c \leq 1$, then $m_2 \approx k^{1-c/2}$ 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 $\Sigma$ 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

$$\rho_{TV}(X, Y) := d_{TV}(P[X, Y], P[X]P[Y]) = \mathbb{E}_X[|d_{TV}(P[Y], P[Y | X])|]$$

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

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

Let $X$ and $Y$ be two correlated random variables with $\rho_{TV}(X, Y) \geq \varepsilon$ 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 $P_x[Y] = P[Y | \text{do}(X = x)]$, 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

$$P_x[Y] = f_y(X, N_Y) = P[Y | X]$$

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 $\Sigma$ of size $k$ satisfies $\rho_{TV}(X, Y) \geq \varepsilon$. Given $m_1$ observational and $m_2$ interventional samples, an algorithm solves Causal Structure Identification problem $(\text{CSI}(k, \varepsilon))$ if with probability at least $2/3$, the algorithm outputs:

- $X \rightarrow Y$ if the true causal structure is Figure 1(a)
- $Y \rightarrow X$ if the true causal structure is Figure 1(b)
- $X \leftarrow U \rightarrow Y$ 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 $\log(1/\delta)$ times and outputting the median response which incurs only a logarithmic increase in the sample complexity.
We provide sample complexity bounds for solving \( \text{CSI}(k, \varepsilon) \). We present our results in the following four regimes: (1) \( k \), (2) Sublinear observational samples: \( m \), (3) Superlinear observational samples: \( m \), and (4) Sufficient observational samples: \( m \).

Theorem 1.1. There exists an algorithm that uses zero observational samples and \( m_2 = O \left( \max(k^{2/3}/\varepsilon \sqrt[4]{m_1}, \sqrt{k}/\varepsilon^2) \right) \) samples from interventions to solve \( \text{CSI}(k, \varepsilon) \). 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; Bhattacharyya 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 \( \text{CSI}(k, \varepsilon) \). 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 \( m_2 = \Omega(\max(k/(\sqrt{m_1}^2), \sqrt{k}/\varepsilon^2)) \) interventional samples to solve 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 \( d_{TV}(P[Y], P[Y \mid x]) = \Omega(\varepsilon) \). 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 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 \( k^{2/\varepsilon^2} \) does not help reducing the intervention complexity up to logarithmic factors.

Theorem 1.5. Any algorithm that solves 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 Hori (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 \( d_{TV}(p, q) > \varepsilon \) 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 \( m_2 = O\left(\max\left\{ k\log k \over \sqrt{m_1}, \sqrt{k\log k} \over \varepsilon^2 \right\}\right) \) 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 \( m_2 = O\left(\max\left\{ k\over \sqrt{m_1}\varepsilon, \sqrt{k}\over \varepsilon^2 \right\}\right) \), 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 \( p_1, p_2, \ldots, p_r \), the goal is to test whether they are identical or there doesn’t exist a distribution \( p \) such that \( (1/s)\sum_{i=1}^s d_{TV}(p, p_i) \leq \varepsilon \). 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 \( X \rightarrow Y \), which corresponds to Figure 1(a) or \( X \notightarrow Y \), which corresponds to Figure 1(b) and Figure 1(c). By symmetry, whether \( Y \rightarrow X \) or \( Y \notightarrow X \) can be distinguished similarly. By definition, the testing problem reduces to distinguishing the following two cases:

\[
X \rightarrow Y \text{ if and only if } \forall x \in \Sigma, P_x[Y] = P[Y | X = x].
\]

\[
X \notightarrow Y \text{ if and only if } \forall x \in \Sigma, P_x[Y] = P[Y].
\]

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

\[
E_{x \sim P[X]} [d_{TV}(P[Y | X = x], P[Y])] > \varepsilon. \tag{1}
\]

Hence there must exist \( x \in \Sigma \) such that \( d_{TV}(P[Y | X = x], P[Y]) > \varepsilon \). 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 \( d_{TV}(P_x[Y], P[Y]) > \varepsilon \). 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) \( m_1 = O(k^2/\varepsilon^2) \), (2) \( m_1 = \Omega(k^2/\varepsilon^2) \).

When \( m_1 = O(k^2/\varepsilon^2) \), our algorithm takes advantage of the fact that we can sample from \( P[X] \). Consider a simple case where \( E_{x \sim P[X]} [d_{TV}(P_x[Y], P[Y])] = \Theta(\varepsilon) \) and all \( x \) satisfy either \( d_{TV}(P[Y | X = x], P[Y]) = 0 \) (trivial element) or \( d_{TV}(P[Y | X = x], P[Y]) = \tau > \varepsilon \) (informative element). When we sample from \( P[X] \), we see an informative element with probability \( \Theta(\varepsilon/\tau) \) and can use \( \Theta(\max(k/\tau^2, \sqrt{k}/\varepsilon^2)) \) 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 \( \tau \) is large, we see an informative element less often and it takes less samples to test. While if \( \tau \) is small, we see an informative element more often but it also takes more samples to test. However, without knowing \( \tau \), 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 \( \tau \)'s that form a geometric sequence and do as well as if \( \tau \) is fixed and known. The upper bound we obtain recovers the \( O(\max(k/\varepsilon^2, \sqrt{k}/\varepsilon^2)) \) 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 \( \Omega \left( \min \left\{ \frac{k}{\sqrt{m_1} \varepsilon}, \frac{\sqrt{k}}{\varepsilon \tau^2} \right\} \right) \) samples from \( q \) to distinguish \( p = q \) versus \( d_{TV}(p, q) \). 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 \( \Sigma \). The marginal probabilities \( P[X = 0] = P[X = 1] = 1/2 \). The conditional distributions are \( P[Y | X = 0] = q \) and \( P[Y | X = 1] = q^- \), 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 \( P[Y | X = 0], P[Y | X = 1], P_{X=0}[Y] \) 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 \( X \rightarrow Y \) versus \( X \notightarrow Y \). 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 \( \rho_{TV}(X, Y) > \varepsilon \),

\[
E_{x_1, x_2 \sim P[X]} [d_{TV}(P[Y | X = x_1], P[Y | X = x_2])] > \varepsilon.
\]

While we do not have access to the observational marginal \( P[X] \), if \( X \rightarrow Y \), we can instead simulate \( P[X] \) by sampling from \( P_y[X] \) for an arbitrary \( y \). Note that if \( X \notightarrow Y \), 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 \( \Omega(k^2/\varepsilon^2) \) 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 \( X \rightarrow Y \) or \( X \notightarrow Y \). By a similar reasoning as Section 1.4.1 we want to find a \( x \in \Sigma \) such that \( d_{TV}(P[Y | X = x], P[Y]) > \varepsilon \).

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 \( \varepsilon \) is \( \theta \left( \frac{k}{\varepsilon^2} \right) \). Canonne (2020a). Therefore, given \( \Omega(k^2/\varepsilon^2) \) 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 \( \varepsilon \)-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 \( \varepsilon \)-close estimates of \( P[Y \mid X = x^*] \) and \( P[Y] \) and ensure \( d_{\text{TV}} \left( \tilde{P}[Y \mid X = x^*], \tilde{P}[Y] \right) > \varepsilon \) which in turn implies \( d_{\text{TV}} (P[Y \mid X = x^*], P[Y]) > \varepsilon \). Therefore, a simple hypothesis test with \( O \left( \frac{1}{\varepsilon^2} \right) \) 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 CSG(k, \( \varepsilon \)) for \( m_1 = \Omega \left( k^{2/3}/\varepsilon^{1/3} \right) \). For \( m_1 = O \left( k^{2/3}/\varepsilon^{1/3} \right) \), 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, \( q : \text{supp}(D) \rightarrow \{0, 1\} \), and \( \varepsilon \in (0, 1] \). Suppose that \( \mathbb{E}[q(s)] > \varepsilon \), and let \( \ell = \lceil \log_2(2/\varepsilon) \rceil \). Then, there exists \( j \in \lceil \ell \rceil \) such that \( \mathbb{P}_{x \sim D}[q(s) > 2^{-j}] > 2^j \varepsilon/(\ell + 5 - j)^2 \).

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 \( CT \left( m_1, m_2, \varepsilon, \delta \right) \) draws \( m_1 \) samples from \( p \), \( m_2 \) samples from \( q \) and outputs “YES” if \( p = q \) and “NO” if \( d_{\text{TV}}(p, q) > \varepsilon \) with probability at least \( 1 - \delta \) 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 \( CT \left( m_1, m_2, \varepsilon, \delta \right) \) where \( m_1 \geq k^{2/3} \log(1/\delta)^{1/3} / \varepsilon^{1/3} \) and

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

**Theorem 1.2.** When \( m_1 = \Omega \left( k^{2/3}/\varepsilon^{1/3} \right) \), there exists an algorithm that takes \( m_1 \) observational samples and \( m_2 = O \left( \max \left( k \sqrt{k}, \sqrt{k}/\varepsilon^2 \right) \right) \) intervention samples and solves CSG(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 \( X \rightarrow Y \) or \( X \not\rightarrow Y \). If this test doesn’t return \( X \rightarrow Y \), we move to the second part and use essentially the same steps to test whether \( Y \rightarrow X \) or \( Y \not\rightarrow X \). Finally, if this test doesn’t return \( Y \rightarrow X \), we return \( X \not\rightarrow Y \).

**Algorithm 1:**

```
Input: \( \varepsilon > 0 \), sample access to \( P[X, Y], P_{x}[Y], P_{y}[X] \).
Output: Return the underlying graph in \( \{X \rightarrow Y, Y \rightarrow X, X \leftarrow U \rightarrow Y\} \).
```

Let \( \ell = \log \left( \frac{2}{\varepsilon} \right) \), \( \ell_j = (\ell + 5 - j) \), \( \delta_j = \frac{2^{\ell_j}}{2^{\ell_j} \cdot s_j} = \frac{2^\ell}{\ell^2} \).

Let \( n_1 = m_1 \cdot 2^{4(j-\ell)/3} \), \( n_2 = \frac{k}{2 - 2\varepsilon \sqrt{n_1}} \); for \( j \in [\ell] \) do

for \( i \in \left[ \frac{20}{s_j} \right] \) do

| Sample \( x_i \sim P[X] \); |
| For distributions \( P[Y], P_{x}[Y], P_{y}[X], \) if |
| \( CT \left(n_1, n_2, \sqrt{\log(1/\delta_j)}, 2^{-j}, \delta_j \right) = "NO" \) |
| then return \( X \rightarrow Y \)

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 \( P[X], P_{y}[X], \) if |
| \( CT \left(n_1, n_2, \sqrt{\log(1/\delta_j)}, 2^{-j}, \delta_j \right) = "NO" \) |
| then return \( X \leftarrow U \rightarrow Y \)

end

end

return \( X \leftarrow U \rightarrow Y \).
```

**Test whether \( X \rightarrow Y \) or \( X \not\rightarrow Y \).**

1. If \( X \rightarrow Y \), then \( P_{x}[Y] = P[Y \mid x] \) and \( P_{y}[X] = P[X] \).

2. If \( X \not\rightarrow Y \), then \( P_{x}[Y] = P[Y] \).

Define \( q(x) := d_{\text{TV}} \left( P[Y \mid x], P[Y] \right) \) for \( x \in [k] \). Then, \( E_{x \sim P[X]}[q(x)] = \rho_{\text{TV}}(X, Y) > \varepsilon \).

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

\[
\mathbb{P}_{x \sim P[X]} (d_{\text{TV}} \left( P[Y \mid x], P[Y] \right) > 2^{-j^*}) \geq s_j.
\]

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 \( d_{\text{TV}} (P[Y \mid x], P[Y]) > 2^{-j^*} \).

If \( X \rightarrow Y \), there exists \( j^* \in [\ell] \) and a sample \( x_i \) that satisfies \( d_{\text{TV}} (P_{x}[Y], P[Y]) > 2^{-j^*} \) with probability \( 1 - e^{-10} \). In contrast, if \( X \not\rightarrow Y \), \( P_{x}[Y] = P[Y] \) for every \( x \).
Consider the following test: for every \( j \in [\ell] \), Algorithm \( \mathcal{A} \) samples \( x_i, 20/s_j \) times from \( P[X] \) to distinguish
\[
P_x[Y] = P[Y] \quad \text{and} \quad d_{TV}(P_x[Y], P[Y]) > 2^{-j}.
\]
For \( n_1 = \Omega \left( k^{2/3}/2^{-4j/3} \right) \) and \( n_2 = O(k/\sqrt{n_1}2^{-2j}) \), each closeness test requires \( n_1 \) samples from \( P[Y] \) and \( n_2 \sqrt{\log(1/\delta_j)} \) samples from \( P_x[Y] \) to succeed with probability \( 1 - \delta_j \). If any of the tests output “NO”, then the algorithm returns \( X \rightarrow Y \).

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

Indeed, if \( X \leftarrow U \rightarrow Y \), \( P_x[Y] = P[Y] \) 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
\[
\sum_{j \in [\ell]} \frac{20\ell_j^2}{2^{j\varepsilon}} m_1 2^{4(j-\ell)/3} \leq 2m_1 \sum_{j \in [\ell]} \frac{\ell_j^2}{2^{j\varepsilon}} 2^{2(j-\ell)/3} = 2m_1 \sum_{j \in [\ell]} \ell_j^2 2^{j\varepsilon} = O(m_1).
\]
Similarly, the total number of interventional samples taken by the algorithm in the first stage is
\[
\sum_{j \in [\ell]} \frac{20\ell_j^2}{2^{j\varepsilon}} \frac{k}{2^{-2j\varepsilon} n_1^j} \sqrt{\log(1/\delta_j)} \leq \frac{2k}{\sqrt{m_1\varepsilon^2}} \cdot \sum_{j \in [\ell]} \ell_j 2^{4(j-\ell)/3} \log(20(\ell_j + 5)2^{-\ell_j}) = O\left( \frac{k}{\sqrt{m_1\varepsilon^2}} \right).
\]

**Error analysis.** Now we analyze the error probability. The total number of tests performed is at most
\[
O\left( \sum_{j=1}^{\ell} 20/s_j \right) = 100 \cdot \sum_{j=1}^{\ell} O\left( \frac{1}{(\ell + 5 - j)^2} \right) < 1/300.
\]
When \( X \rightarrow Y \), the probability of the algorithm failing to find a sample \( x \) satisfying \( d_{TV}(P_x[Y], P[Y]) > 2^{-j} \) is at most 1/300. The analysis is the same for graph \( X \leftarrow Y \). 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:
\[
\sum_{j \in [\ell]} \frac{20\ell_j^2}{2^{j\varepsilon}} \leq \frac{40}{\varepsilon} \cdot \sum_{j \in [\ell]} \ell_j^2 2^{-j\varepsilon} = O\left( \frac{\ell_j^2}{\varepsilon} \right) = O\left( \frac{\log(1/\varepsilon)^2}{\varepsilon} \right).
\]

---

**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)](https://example.com).

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)](https://example.com)’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 \( x_{1:4} \) 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
\[
m(a_{1:4}) := \left( \prod_{i=1}^{4} n_{a_i}^{n_i} \right) \sum_{i=1}^{k} p_{i_1}^{n_1} (i) p_{i_2}^{n_2} (i) p_{i_3}^{n_3} (i) p_{i_4}^{n_4} (i).
\]

**Proposition 1.** Given integers \( n_{1:4} > 0 \) and two distribution quadruplets \( (p^{+}_{1:4}) \) and \( (p^{-}_{1:4}) \), where \( p_i^+ \) and \( p_i^- \) have frequencies at most \( \frac{1}{n_i} \), for constant \( c > 0 \). If \( m^+ \) and \( m^- \) are the \( n_{1:4} \)-based moments of \( (p^{+}_{1:4}) \) and \( (p^{-}_{1:4}) \) respectively that satisfy
\[
\sum_{a_1 + a_2 + a_3 + a_4 > 0} \frac{|m^+ (a_{1:4}) - m^- (a_{1:4})|}{\sqrt{1 + \max \{m^+ (a_{1:4}), m^- (a_{1:4})\}}} < \frac{1}{c},
\]
then the algorithm returns the correct graph with error probability at most 1/150.
for some constant $c' > 0$, then $(p_{a_i:4}^-)$ cannot be distinguished from $(\tilde{p}_{a_i:4}^-)$ with probability greater than 0.5 using a tester that takes Poi$(n_i)$ samples from $(p_{a_i:4}^+, p_{a_i:4}^-)$ 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 CSI$(k, \varepsilon)$.

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

**Some definitions.** Let $\pi$ be a permutation of $[k]$ chosen uniformly at random. Let $A = \{\pi(1), \pi(2), \ldots, \pi(a)\}$ and $B = \{\pi(a+1), \pi(a+2), \ldots, \pi(a+b)\}$ 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 $S$ be the indicator function of set $S$. For permutation $\pi$, define distributions $p, q$ and $q^-$ over $[k]$ as follows:

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

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

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

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

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

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

**Analysis.** Let $n_1 = n_2 = cm_1$, $n_3 = n_4 = ckm\varepsilon^{-2}/\sqrt{m_1}$ for a sufficiently small $c$. Let $(p_{1:4}^+) := (q, q^-, q, q^-)$ and $(\tilde{p}_{1:4}^-) := (q, q^-, q, q^-)$. Suppose, for contradiction, there exists an algorithm $A$ that solves CSI$(k, \varepsilon)$ that uses $n_1$ observational samples and $n_3$ interventional samples. A single sample from each of the two conditions 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 $(p_{1:4}^+)$ cannot be distinguished from $(\tilde{p}_{1:4}^-)$ with probability greater than 0.5 by any tester that takes Poi$(n_i)$ samples from $(p_{1:4}^+, p_{1:4}^-)$. We show this in the supplementary material by bounding the difference in moments

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

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 tradeoff 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 $\Omega(k^2/\varepsilon^2)$. 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. Sourabh 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


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 $\rho_{TV}(X, Y) \geq \varepsilon$. Then,

$$E_{(x_1, x_2) \sim (P[X], P[X])}[d_{TV}(P[Y \mid x_1], P[Y \mid x_2])] \geq \varepsilon,$$

$$E_{(y_1, y_2) \sim (P[Y], P[Y])}[d_{TV}(P[X \mid y_1], P[X \mid y_2])] \geq \varepsilon.$$

**Proof:**

$$E_{(x_1, x_2) \sim (P[X], P[X])}[d_{TV}(P[Y \mid x_1], P[Y \mid x_2])] \geq E_{x_1 \sim P[X]}[d_{TV}(P[Y \mid x_1], E_{x_2 \sim P[X]}[P[Y \mid x_2]])]$$ (By Jensen’s inequality)

$$= E_{x_1 \sim P[X]}[d_{TV}(P[Y \mid x_1], \sum_{x_2} P[x_2]P[Y \mid x_2])]$$

$$= E_{x_1}[d_{TV}(P[Y \mid x_1], P[Y])]$$

$$\geq \varepsilon.$$  

Similarly, $E_{(y_1, y_2) \sim (P[Y], P[Y])}[d_{TV}(P[X \mid y_1], P[X \mid y_2])] \geq \varepsilon.$

**Algorithm 2:**

**Input:** $\varepsilon > 0$, sample access to $P[X, Y], P_x[Y], P_y[X]$.

**Output:** Return the underlying graph in $\{X \rightarrow Y, Y \rightarrow X, X \leftarrow U \rightarrow Y\}$.

Let $\ell = \log \left(\frac{2}{\varepsilon}\right)$, $\delta_j = \frac{2^{\ell-j}}{20(\ell + 5 - j)^2}, s(j) = \frac{20(\ell + 5 - j)^2}{\varepsilon 2^j}$;

Let $n_j = \left(\frac{k^{2/3}}{2^{(s(j))}} + \frac{k^{1/2}}{2^{(s(j))}}\right) \sqrt{\log \left(\frac{2}{\varepsilon}\right)}$;

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 $CT\left(n_j, n_j, 2^{-j}, \delta_j\right) = \text{"NO"}$;
    then return $X \rightarrow Y$;
  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 $CT\left(n_j, n_j, 2^{-j}, \delta_j\right) = \text{"NO"}$;
    then return $Y \rightarrow X$;
  end

end

return $X \leftarrow U \rightarrow Y$.

We now use this result prove that $\text{CSI}(k, \varepsilon)$ can be solved using $O\left(k^{2/3}/\varepsilon^{4/3}\right)$ samples from interventions even with zero samples from observations.
We apply Levin’s investment strategy (Lemma 1), for the above choice of \( s \) correct graph.

Consider the following test. For every \( j \in [\ell] \), we analyze Algorithm 3 in two parts. In the first part, the algorithm tests whether \( X \rightarrow Y \) or \( X \not\rightarrow Y \). If this test doesn’t return \( X \rightarrow Y \), the second part tests between \( Y \rightarrow X \) or \( Y \not\rightarrow X \). If this test doesn’t return \( Y \rightarrow X \), then \( X \leftarrow U \rightarrow Y \) is returned.

**Proof.** For \( x_1 \in [k] \) and \( x_2 \in [k] \), define

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

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

\[
\mathbb{E}_{(x_1, x_2) \sim (P[X], P[X])} [q(x_1, x_2)] > \varepsilon.
\]

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

\[
\Pr_{(x_1, x_2) \sim (P[X], P[X])} \left( d_{TV} (P[Y \mid x_1], P[Y \mid x_2]) > 2^{-j^*} \right) \geq s(j^*)
\]

where \( s(j) := (2^j / (\ell + 5 - j))^2 \), 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,

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

**Part I.** The first half of the algorithm considers testing whether \( X \rightarrow Y \) or \( X \not\rightarrow Y \).

1. \( \mathcal{H}_0 \): \( X \not\rightarrow Y \), which implies \( P_x[Y] = P[Y] \), for all \( x \);
2. \( \mathcal{H}_1 \): \( X \rightarrow Y \), 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

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

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 \( d_{TV} (P_{x_1}[Y], P_{x_2}[Y]) \) is zero versus greater than \( 2^{-j} \), with probability \( 1 - \delta_j \), is

\[
n_j := O \left( k^{(2/3)j/3} \log (1/\delta_j) + \left( \frac{k^{1/2} \log (1/\delta_j)}{\log (1/\delta_j)} \right)^{2j/3} \right).
\]

For \( \mathcal{H}_0 \), \( P_y[X] = P[X] \) 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

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

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

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

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 \( X \rightarrow Y \) in Part I, then the underlying causal graph is either \( X \leftarrow Y \) or \( X \leftarrow U \rightarrow Y \). The algorithm performs local tests similar to Part I to test whether \( Y \rightarrow X \) or \( Y \not\rightarrow X \), 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

$$\sum_{j \in \ell} \frac{20 \cdot t(j)^2}{2^j \varepsilon} = 20 \sum_{j \in \ell} \frac{(\ell + 5 - j)^2}{2^j \varepsilon}$$

$$\leq \frac{20}{\varepsilon} \sum_{j \in \ell} (\ell + 5 - j)^2 2^{-j}$$

$$= O\left(\frac{\log^2 \varepsilon}{\varepsilon}\right).$$

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:

$$\sum_{j \in \ell} \frac{20 \cdot t(j)^2}{2^j \varepsilon} k^{2/3} 2^{-4j/3} \sqrt{\log(1/\delta_j)} + \sum_{j \in \ell} \frac{20 \cdot t(j)^2}{2^j \varepsilon} k^{1/2} 2^{-7j/2} \sqrt{\log(1/\delta_j)}$$

$$= \frac{20k^{2/3}}{\varepsilon} \sum_{j \in \ell} \frac{t(j)^2}{2^{-4j/3}} \sqrt{\log(20(\ell + 5 - j)^4/2^j)} + \frac{20k^{1/2}}{\varepsilon} \sum_{j \in \ell} \frac{t(j)^2}{2^{-7j/2}} \sqrt{\log(20(\ell + 5 - j)^4/2^j)}$$

$$\leq \frac{20k^{2/3}}{\varepsilon} \sum_{j \in \ell} \frac{(j' + 5)^2}{2^{j'} 2^{j'}} \sqrt{\log(20(j' + 5)^4/2^{-j'})} + \frac{20k^{1/2}}{\varepsilon} \sum_{j \in \ell} \frac{(j' + 5)^2}{2^{j'} 2^{j'}} \sqrt{\log(20(j' + 5)^4/2^{-j'})}$$

$$= O\left(\frac{k^{2/3}}{\varepsilon^{4/3}} + \frac{k^{1/2}}{\varepsilon^2}\right)$$

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

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

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

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 $1/150$. \qed

### B Super-quadratic Observations

Here we analyze the tradeoff between $m_1$ and $m_2$ when $m_1$ is $\tilde{\Omega}(k^2/\varepsilon^2)$. We show that $O(1/\varepsilon^2)$ interventional samples are sufficient and necessary to solve CSI($k, \varepsilon$).
Algorithm 3:

**Input**: \( \varepsilon > 0 \), sample access to \( P[X, Y], P_x[Y], P_y[X] \).

**Output**: Return the underlying graph in \( \{ X \rightarrow Y, Y \rightarrow X, X \leftarrow U \rightarrow Y \} \).

Let \( m_1 \geq 20 \log k \cdot \frac{k^2}{\varepsilon^2} \) and \( S \) be \( m_1 \) samples drawn from \( P[X, Y] \);
Let \( S_x \leftarrow \{(x^*, y^*) \in S : x^* = x\} \) and \( S_y \leftarrow \{(x^*, y^*) : y^* = y\} \);
Let \( \hat{P}[X] \) and \( \hat{P}[Y] \) be the empirical distributions of \( P[X] \) and \( P[Y] \) from \( S \);

for \( j \in [k] : |S_{y_j}| \) is \( \tilde{\Omega} \left( \frac{k}{\varepsilon^2} \right) \) do
  if \( \mathcal{TT} \left( \hat{P}[Y], P[Y \mid x_j], 8\varepsilon/10, 9\varepsilon/10, S_{x_j}, 1/(10k) \right) \) is 'NO' then
    Let \( \hat{P}[Y \mid x_j] \) be the empirical distribution of \( P[Y \mid x_j] \) from \( S_{x_j} \);
    Find \( T \) such that \( |\hat{P}[Y \in T] - \hat{P}[Y \in T \mid x_j]| > 6\varepsilon/10 \);
    \( R \leftarrow \) empirical distribution of \( O \left( \frac{1}{\varepsilon^2} \right) \) samples from \( P_{x_j}[Y] \) with accuracy \( \varepsilon/100 \);
    if \( |R[Y \in T] - \hat{P}[Y \in T \mid x_j]| \leq 3\varepsilon/10 \) then
      return \( X \rightarrow Y \)
  end
end

for \( j \in [k] : |S_{y_j}| \) is \( \tilde{\Omega} \left( \frac{k}{\varepsilon^2} \right) \) do
  if \( \mathcal{TT} \left( \hat{P}[X], P[X \mid y_j], 8\varepsilon/10, 9\varepsilon/10, S_{y_j}, 1/(10k) \right) \) is 'NO' then
    Let \( \hat{P}[X \mid y_j] \) be the empirical distribution of \( P[X \mid y_j] \) from \( S_{y_j} \);
    Find \( T \) such that \( |\hat{P}[X \in T] - \hat{P}[X \in T \mid y_j]| > 6\varepsilon/10 \);
    \( R \leftarrow \) empirical distribution of \( O \left( \frac{1}{\varepsilon^2} \right) \) samples from \( P_{y_j}[X] \) with accuracy \( \varepsilon/100 \);
    if \( |R[X \in T] - \hat{P}[X \in T \mid y_j]| \leq 3\varepsilon/10 \) then
      return \( Y \rightarrow X \)
  end
end

return \( X \leftarrow U \rightarrow Y \).

**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 \( \tau_x := d_{TV}(P[Y], P[Y \mid x]) \). First we prove a claim that assures the existence of \( x \in \Sigma \) such that we get sufficient samples on \( X = x \) and also \( \tau_x \) is small.

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

1. \( \tau_x \) is at least \( \varepsilon/10 \);

2. \( |S_x| = \Omega \left( m_1 \cdot \frac{\varepsilon^2}{k^2\tau_x^2} \right) \) for \( m_1 = \tilde{\Omega} \left( \frac{k^2}{\varepsilon^2} \right) \).

**Proof.** By Cauchy-Schwarz inequality,

\[
\sum_x P(x) \tau_x^2 = \left( \sum_x P(x) \right) \cdot \left( \sum_x P(x) \tau_x^2 \right) \geq \left( \sum_x P(x) \cdot \tau_x \right)^2 \geq \varepsilon^2. \tag{3}
\]

Also,

\[
2 \sum_x P(x) \cdot \tau_x^2 \cdot 1_{\{\tau_x \leq \varepsilon/10\}} \leq \frac{\varepsilon^2}{100}. \tag{4}
\]
Combining Equations (3) and (4),
\[ \sum_x P(x) (\tau_x)^2 1_{\{\tau_x > \varepsilon/10\}} > 99\varepsilon^2/100. \]

Hence, there exists \( x \in [k] \) that satisfies (i) \( \tau_x \geq \frac{\varepsilon}{10} \) and
\[ (ii) P(x) \cdot \tau_x^2 \geq \frac{99\varepsilon^2}{100k} \quad \Rightarrow \quad P(x) \geq \frac{99\varepsilon^2}{100k \cdot \tau_x^2}. \]

Applying Chernoff bound,
\[ \Pr \left[ |S_x| \leq (1 + C) m_1 \frac{99\varepsilon^2}{100k \cdot \tau_x^2} \right] \leq \Pr \left[ |S_x| \leq (1 + C) \frac{k \log k}{\varepsilon^2} \cdot \frac{99\varepsilon^2}{100k \cdot \tau_x^2} \right] \leq \exp \left( -\frac{99(C)^2 \log k}{2 \cdot 100 \tau_x^2} \right) \leq \frac{1}{10^k}. \] (For any \( C > 5 \).)

This proves the claim. \( \square \)

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

When \( \tau_x > \varepsilon/10 \), by triangle inequality,
\[ d_{TV} \left( \hat{P}[Y], P[Y \mid x'] \right) \geq d_{TV} \left( P[Y], P[Y \mid x'] \right) - d_{TV} \left( \hat{P}[Y], P[Y] \right) \geq \varepsilon/10 - \varepsilon/100 \geq 9\varepsilon/100. \]

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

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

Hence we can find one such \( x_{x'} \) that satisfies the two conditions by filtering all \( x \in \Sigma \) with large \( |S_x| = \tilde{\Omega} (k/\varepsilon^2) \). The tolerant test \( \mathcal{T}_T \left( \hat{P}[Y], P[Y \mid x_j], 8\varepsilon/10, 9\varepsilon/10, S_x, 1/(10k) \right) \) uses \( S_x \) and outputs

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

with probability \( 1 - 1/(10k) \).

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

We now have the following:
\[ d_{TV} \left( \hat{P}[Y], P[Y \mid X] \right) > 9\varepsilon/10 \quad \text{(5)} \]
\[ d_{TV} \left( \hat{P}[Y], P[Y] \right) \leq \varepsilon/100 \quad \text{(6)} \]
\[ d_{TV} \left( \hat{P}[Y \mid x], P[Y \mid x] \right) \leq \varepsilon/100. \quad \text{(7)} \]

---

\(^1\)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,

\[ d_{TV}(P[Y], P[Y | X]) > 8\varepsilon/100. \]  

(8)

Also, by triangle inequality:

\[ d_{TV}(\tilde{P}[Y], \tilde{P}[Y | x]) \geq d_{TV}(P[Y], P[Y | X]) - d_{TV}(\tilde{P}[Y], P[Y]) - d_{TV}(\tilde{P}[Y | x], P[Y | x]) \geq 6\varepsilon/100. \]  

(9)

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

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

and \( T \) satisfies:

\[
|P[Y \in T] - P[Y \in T | x]| \geq |\tilde{P}[Y \in T] - \tilde{P}[Y \in T | x]| - |\tilde{P}[Y \in T] - P[Y \in T]| \\
\geq 6\varepsilon/100 - \varepsilon/100 = 4\varepsilon/100
\]

We also have the following:

1. If \( X \rightarrow Y \), \( P_{x,i}[Y] = P[Y | x_i] \).
2. If \( X \not\rightarrow Y \), \( P_{x,i}[Y] = P[Y] \).

Thus we can estimate \( P_{x,i}[Y] \) to test between \( P[Y \in T] \) and \( P[Y \in T | x] \) on \( T \) using \( O(1/\varepsilon^2) \) 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 \( \tilde{P}[Y] \) and \( \tilde{P}[Y | x] \) 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 \( \Omega(1/\varepsilon^2) \) samples from interventions to solve \( CSI(k, \varepsilon) \).

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

\[
P(2i+1) = \frac{1 - 3 \cdot z_i \varepsilon}{k} \quad P(2i) = \frac{1 + 3 \cdot z_i \varepsilon}{k}
\]

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 \( P[Y | X = i] = q_i \). Here \( E[d_{TV}(\text{unif}(k), q_i)] > \varepsilon \). Also the marginal \( P[Y] \) is \text{unif}[k] because \((1/2)q_0 + (1/2)q_1 = \text{unif}[k]\).

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

1. \( \mathcal{H}_0 \): When \( X \rightarrow Y \) then \( P_{X=i}[Y] = q_i \).
2. \( \mathcal{H}_1 \): When \( X \leftarrow U \rightarrow Y \), \( P_{X=i}[Y] = \text{unif}[k] \).
by taking samples from $P_{X=1}[Y]$. Let $A$ 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 $P_{X=1}[Y]$. This is equivalent to distinguishing $\text{unif}[k]$ versus $q_0$, which requires $\Omega(1/\varepsilon^2)$ 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. \hfill \Box

## 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 $r, s, t, u \geq 0$, $m^+(r, s, t, u)$ is

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

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

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

where

$$h(\varepsilon, r, s, t, u) = \frac{1 + \varepsilon^r (1 - \varepsilon)^s + (1 - \varepsilon)^r (1 + \varepsilon)^s}{2}$$

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,

$$\frac{|m^+(r, s, t, u) - m^-(r, s, t, u)|}{\sqrt{1 + \max\{m^+(n_1, n_2, n_3), m^-(n_1, n_2, n_3)\}}} \leq \frac{n_1 r s t u (1/4)^{r+s+t+u} (C/k)^{r+s+t+u-1} (h'(\varepsilon, r, s, t, u) - h(\varepsilon, s, t))}{\sqrt{n_1 r s t u (3/4) (1/m_1)^{r+s+t+u-1} + (1/4)^{r+s+t+u} (C/k)^{r+s+t+u-1}}}$$

$$\leq \frac{n_1 r s t u (1/4)^{r+s+t+u} (C/k)^{r+s+t+u-1}}{\sqrt{n_1 r s t u (3/4) (1/m_1)^{r+s+t+u-1} + (1/4)^{r+s+t+u} (C/k)^{r+s+t+u-1}}}$$

for $C' > C$

$$\leq \frac{n_1 r s t u (1/4)^{r+s+t+u} (C/k)^{r+s+t+u-1}}{\sqrt{n_1 r s t u (3/4) (1/m_1)^{r+s+t+u-1}}}$$

(because $m_1 < k$)

$$\leq \frac{c^{(r+s+t+u)/2} m_1^{(r+s+t+u-1)/2} (C/k)^{r+s+t+u-1}}{m_1^{(t+u)/4} (\varepsilon^2)^{t/2+u/2}}$$

(used $n_1, n_2, n_3, n_4$)

$$\leq \frac{c^{(r+s+t+u)/2} m_1^{r+s} (\sqrt{m_1^{r+s}} \varepsilon^{2+k})^{t/2+u/2+1}}{m_1^{r+s} (\sqrt{m_1^{r+s}} \varepsilon^{2+k})^{t/2+u/2+1}}$$

(for small $c$)

$$\leq \frac{c^{(r+s+t+u)/2} m_1^{r+s} \sqrt{m_1^{r+s}} \varepsilon^{2+k}}{m_1^{r+s}}$$

(substituting $\varepsilon^2 > k^{-1/2}$)

which is small when $c$ 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 \((n_1, n_2, n_3, n_4)\)-Poisonized tester that correctly classifies a symmetric property on a distribution quadruplet \((p_1, p_2, p_3, p_4)\) with probability \(\frac{1}{200}\) 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 \(M^\rho\) where \(\rho\) 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 \(\rho\), letting \(\vec{\lambda}(a) = \sum_i \rho(i, a)\) be the vector of column sums,

\[
d_{TV} \left( M^\rho, \text{Poi}(\vec{\lambda}) \right) \leq 8.8 \sum_a \frac{\sum_i \rho(i, a)^2}{\sum_i \rho(i, a)}.
\]

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 \(0 < c \leq 0.5\), such that for all \(i \in [k], j \in [4]\), \(p_j(i) \leq \frac{c}{n_j}\), if \(\rho\) is the matrix with \((i, (a_1, a_2, a_3, a_4))\) entry \(\prod_j \rho_{ij}(a_j; k_jp_j(i))\), then

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

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 \(\{p_j^+\}_{j=1}^4\) and \(\{p_j^-\}_{j=1}^4\), where \(p_j^+, p_j^-\) have frequencies at most \(\frac{1}{30000n_j}\) for \(j \in [4]\), \(\sum_j \rho_{ij}(a_j; k_jp_j^+(i))\) and \(\sum_j \rho_{ij}(a_j; k_jp_j^-(i))\) for \(a_1 + a_2 + a_3 + a_4 > 0\) and if

\[
\sum_{a_1+a_2+a_3+a_4 > 0} \frac{\left| \sum_j \rho_{ij}(a_j; k_jp_j^+(i)) \right|}{\sqrt{1 + \max \left\{ \sum_j \rho_{ij}(a_j; k_jp_j^+(i)), \sum_j \rho_{ij}(a_j; k_jp_j^-(i)) \right\}}} < \frac{1}{200}
\]

then it is impossible to test any symmetric property that is true for \(\{p_j^+\}_{j=1}^4\) and false for \(\{p_j^-\}_{j=1}^4\) 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 4.