Probably approximately correct high-dimensional causal effect estimation given a valid adjustment set

Davin Choo[*] DAVINCHOO@SEAS.HARVARD.EDU
Harvard University

Chandler Squires[†] CSQUIRES@ANDREW.CMU.EDU
Carnegie Mellon University

Arnab Bhattacharyya[‡]
University of Warwick

David Sontag
Massachusetts Institute of Technology

Editors: Biwei Huang and Mathias Drton

Abstract

Accurate estimates of causal effects play a key role in decision-making across applications such as healthcare, economics, and operations. In the absence of randomized experiments, a common approach to estimating causal effects uses *covariate adjustment*. In this paper, we study covariate adjustment for discrete distributions from the PAC learning perspective, assuming knowledge of a valid adjustment set $\mathbf{Z}$, which might be high-dimensional. Our first main result PAC-bounds the estimation error of covariate adjustment by a term that is exponential in the size of the adjustment set; it is known that such a dependency is unavoidable even if one only aims to minimize the mean squared error. Motivated by this result, we introduce the notion of an *ε-Markov blanket*, give bounds on the misspecification error of using such a set for covariate adjustment, and provide an algorithm for ε-Markov blanket discovery; our second main result upper bounds the sample complexity of this algorithm. Furthermore, we provide a misspecification error bound and a constraint-based algorithm that allow us to go beyond ε-Markov blankets to even smaller adjustment sets. Our third main result upper bounds the sample complexity of this algorithm, and our final result combines the first three into an overall PAC bound. Altogether, our results highlight that one does not need to perfectly recover causal structure in order to ensure accurate estimates of causal effects.

Keywords: Causality, covariate adjustment, PAC bounds, finite sample complexity

1. Introduction

Let $\mathbb{P}(\mathbf{V})$ be an *unknown* probability distribution over discrete random variables $\mathbf{V}$. Using i.i.d. samples from $\mathbb{P}(\mathbf{V})$, we wish to estimate the probability that $\mathbf{Y} \subset \mathbf{V}$ equals $\mathbf{y}$, in the interventional distribution where $\mathbf{X} \subset \mathbf{V}$ is set to $\mathbf{x}$. This problem, called *causal effect estimation*, can be formalized in either the Neyman-Rubin potential outcomes (PO) framework (Rubin, 1974; Splawa-Neyman et al., 1990; Sekhon, 2009) or in Pearl's graphical causality framework (Pearl, 2009). Depending on the framework, the desired estimand is written as $\mathbb{P}(\mathbf{Y}(\mathbf{x}) = \mathbf{y})$ or $\mathbb{P}_{\mathbf{x}}(\mathbf{y}) = \mathbb{P}(\mathbf{Y} = \mathbf{y} \mid \mathrm{do}(\mathbf{X} = \mathbf{x}))$, and has several important downstream applications such as estimating treatment

[*] Equal contribution. Part of work was done while author was affiliated with National University of Singapore.
[†] Equal contribution. Part of work was done while author was affiliated with Massachusetts Institute of Technology.
[‡] Part of work was done while author was affiliated with National University of Singapore.

effects. In this work, we consider the problem from the viewpoint of distribution learning (Kearns et al., 1994) under the Probably Approximately Correct (PAC) learning model (Valiant, 1984).

The PAC causal effect estimation (PAC-CEE) problem. Given (1) estimation tolerance $\lambda > 0$, (2) failure tolerance $\delta > 0$, (3) sample access to a distribution $\mathbb{P}(\mathbf{V})$, and (4) an interventional query $\mathbb{P}_{\mathbf{x}}(\mathbf{y})$, output an estimate $\widehat{\mathbb{P}}_{\mathbf{x}}(\mathbf{y})$ such that $\Pr\left(\left|\widehat{\mathbb{P}}_{\mathbf{x}}(\mathbf{y}) - \mathbb{P}_{\mathbf{x}}(\mathbf{y})\right| \leq \lambda\right) \geq 1 - \delta$.

For this problem to be well-posed, one must be able to relate the observational distribution $\mathbb{P}(\mathbf{V})$ to the interventional distribution $\mathbb{P}_{\mathbf{x}}(\mathbf{y})$ via some *identification formula*, i.e., $\mathbb{P}_{\mathbf{x}}(\mathbf{y})$ must be uniquely determined by $\mathbb{P}(\mathbf{V})$. Here, we focus on a commonly studied identification formula that involve a set of variables $\mathbf{Z} \subset \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$ such that $\mathbb{P}_{\mathbf{x}}(\mathbf{y}) = T_{\mathbf{Z},\mathbf{x},\mathbf{y}}$, where

$$T_{\mathbf{Z},\mathbf{x},\mathbf{y}} := \sum\nolimits_{\mathbf{z} \in \mathbf{\Sigma_Z}} \mathbb{P}(\mathbf{Y} = \mathbf{y} \mid \mathbf{Z} = \mathbf{z}, \mathbf{X} = \mathbf{x}) \cdot \mathbb{P}(\mathbf{Z} = \mathbf{z}) = \sum\nolimits_{\mathbf{z}} \mathbb{P}(\mathbf{y} \mid \mathbf{z}, \mathbf{x}) \cdot \mathbb{P}(\mathbf{z}), \qquad (1)$$

with $\mathbf{\Sigma_Z}$ denoting the alphabet of the variables $\mathbf{Z}$. For instance, in the PO framework, Eq. (1) holds under the assumptions of *consistency* and *conditional ignorability* of $\mathbf{X}$ with respect to $\mathbf{Z}$; see Appendix C.1 for a simple derivation. Meanwhile, in the graphical framework, Eq. (1) can be shown to hold if $\mathbf{Z}$ satisfies certain graphical criterion with respect to $\mathbf{X}$ and $\mathbf{Y}$, such as the *(generalized) backdoor criterion* or the *(generalized) adjustment criteria* (Pearl, 1995; Shpitser et al., 2010; Maathuis and Colombo, 2015; Perković et al., 2018). Following the latter viewpoint, we call $\mathbf{Z}$ a *valid adjustment set* for $\mathbb{P}_{\mathbf{x}}(\mathbf{y})$ if $\mathbf{Z}$ satisfies $\mathbb{P}_{\mathbf{x}}(\mathbf{y}) = T_{\mathbf{Z},\mathbf{x},\mathbf{y}}$ in Eq. (1), but we emphasize that our results are framework-agnostic, i.e., they do not depend on how Eq. (1) is derived.

In particular, we establish our PAC guarantees by directly analyzing the sample complexity required to produce an estimate $\widehat{T}_{\mathbf{Z},\mathbf{x},\mathbf{y}}$ of $T_{\mathbf{Z},\mathbf{x},\mathbf{y}}$. A recent work of Zeng et al. (2024) shows that $\Omega\left(\frac{1}{\lambda^2 \alpha_{\mathbf{Z}}} + \frac{|\mathbf{\Sigma_Z}|}{\lambda \alpha_{\mathbf{Z}}}\right)$ samples are sufficient to ensure an expectation bound of $\mathbb{E}\left(|T_{\mathbf{Z},\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{Z},\mathbf{x},\mathbf{y}}|\right) \leq \lambda$, where $\alpha_{\mathbf{Z}}$ is a *positivity* (a.k.a. *overlap*) parameter that is common in causal effect estimation; we translate their actual stated bound into the form we describe here in Appendix B.1. Zeng et al. (2024) also presented a minimax lower bound showing that linear dependency on $|\mathbf{\Sigma_Z}|$ is unavoidable. Since $|\mathbf{\Sigma_Z}|$ grows exponentially with the size of $\mathbf{Z}$ (e.g. when all variables are binary, we have $|\mathbf{\Sigma_Z}| = 2^{|\mathbf{Z}|}$), it is critical to use *small* adjustment sets whenever possible. The importance of using small adjustment sets brings up an interesting but subtle distinction between the non-asymptotic setting considered in this paper, and asymptotic setting considered in works such as Rotnitzky and Smucler (2020) and Brookhart et al. (2006). Such works have found that adding certain variables to the adjustment set can decrease the asymptotic variance (i.e. increase the precision) of the covariate adjustment estimator. We emphasize that our focus on using small adjustment sets is not at odds with these results, but reflects a fundamental limitation of asymptotic results that is widely known within statistics. In particular, the variance of an estimator typically dominates in the asymptotic setting, since it is associated with errors of order $\mathcal{O}(\sqrt{n})$, but in the non-asymptotic setting, lower-order terms with large coefficients play a significant role; c.f. the error term scaling in Theorem 7.[1]

In our work, given a valid adjustment set $\mathbf{Z} \subseteq \mathbf{V}$ as an initial input, we explore the possibility of searching for smaller adjustment sets with the objective of using less total samples than directly producing a λ-good estimate $\widehat{T}_{\mathbf{Z},\mathbf{x},\mathbf{y}}$. We are able to obtain lower sample complexities because of the adage from the property testing literature that "testing can be cheaper than learning". In particular,

[1] For example, when estimating the entropy of a discrete distribution, the maximum likelihood estimator is asymptotically efficient, but its sample complexity has strictly suboptimal dependence on alphabet size (Jiao et al., 2017).

we develop testing-based algorithms to find a candidate adjustment set $\mathbf{S} \subseteq \mathbf{Z}$, then estimate $\widehat{T}_{\mathbf{S},\mathbf{x},\mathbf{y}}$ and bound its error from $\mathbb{P}_{\mathbf{x}}(\mathbf{y}) = T_{\mathbf{Z},\mathbf{x},\mathbf{y}}$ via triangle inequality:

$$\left|\mathbb{P}_{\mathbf{x}}(\mathbf{y}) - \widehat{T}_{\mathbf{S},\mathbf{x},\mathbf{y}}\right| = \left|T_{\mathbf{Z},\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{S},\mathbf{x},\mathbf{y}}\right| \leq |T_{\mathbf{Z},\mathbf{x},\mathbf{y}} - T_{\mathbf{S},\mathbf{x},\mathbf{y}}| + \left|T_{\mathbf{S},\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{S},\mathbf{x},\mathbf{y}}\right| \leq \varepsilon_1 + \varepsilon_2 = \lambda$$

The overall error bound (λ) is at most the sum of the misspecification bias error term (ε_1) and the estimation error term (ε_2). There is an inherent tradeoff between these two sources of error: using $\mathbf{S} \subseteq \mathbf{Z}$ for adjustment might introduce misspecification bias (if $\mathbf{S}$ is not a valid adjustment set), but this bias may dominated by a reduction in estimation error if $\mathbf{S}$ is much smaller than $\mathbf{Z}$. While our approach to selecting $\mathbf{S}$ is best appreciated through the lens of the graphical causality framework, it also applies in the PO setting, as we only rely on conditional independences of $\mathbb{P}(\mathbf{V})$.

1.1. Our main results

Our first main result extends the result of Zeng et al. (2024) to the PAC setting by bounding the *estimation error* $|T_{\mathbf{A},\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{A},\mathbf{x},\mathbf{y}}|$ for arbitrary subsets $\mathbf{A} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$, where $\widehat{T}_{\mathbf{A},\mathbf{x},\mathbf{y}}$ is the estimate of $T_{\mathbf{A},\mathbf{x},\mathbf{y}}$ obtained using empirical sample estimates of $\mathbb{P}(\mathbf{y} \mid \mathbf{a}, \mathbf{x})$ and $\mathbb{P}(\mathbf{a})$ for all $\mathbf{a} \in \Sigma_{\mathbf{A}}$. Throughout the paper, for any $\mathbf{A} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$ arbitrary, we let

$$\alpha_{\mathbf{A}} = \min_{\mathbf{a} \in \mathbf{\Sigma}_{\mathbf{A}}} \mathbb{P}(\mathbf{x} \mid \mathbf{a}) \tag{2}$$

Theorem 1 (Estimation error) *Suppose we are given (1) estimation tolerance $\varepsilon > 0$, (2) failure tolerance $\delta > 0$, (3) sample access to $\mathbb{P}(\mathbf{V})$, and (4) a subset $\mathbf{A} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$. Then, there is an algorithm that uses $\widetilde{\mathcal{O}}\left(\left(\frac{|\mathbf{\Sigma}_{\mathbf{A}}|}{\varepsilon \alpha_{\mathbf{A}}} + \frac{1}{\varepsilon^2 \alpha_{\mathbf{A}}} + \frac{|\mathbf{\Sigma}_{\mathbf{A}}|}{\varepsilon^2}\right) \cdot \log \frac{1}{\delta}\right)$ samples and produces an estimate $\widehat{T}_{\mathbf{A},\mathbf{x},\mathbf{y}}$ such that $\Pr(|\widehat{T}_{\mathbf{A},\mathbf{x},\mathbf{y}} - T_{\mathbf{A},\mathbf{x},\mathbf{y}}| \leq \varepsilon) \geq 1 - \delta$.*

Note that, up to logarithmic factors and the additional $\frac{|\mathbf{\Sigma}_{\mathbf{A}}|}{\varepsilon^2}$ factor, the sample complexity of the PAC bound matches the sample complexity of the expectation bound. Here, we switched from λ to ε and from $\mathbf{Z}$ to $\mathbf{A}$ to emphasize that the estimation error is only one part of our overall bound. Surprisingly, although covariate adjustment is one of the simplest and most widely-used estimation techniques in causality, this result is (to the best of our knowledge) the first PAC bound on causal effect estimation for discrete variables. In particular, previous works either focus on different estimands (under additional assumptions such as knowing a causal graph) or consider continuous variables and primarily provide only asymptotic results; we discuss related works in Appendix A.3.

Importantly, the sample complexity depends exponentially on $|\mathbf{A}|$, hence it is crucial to keep $\mathbf{A}$ small. As a paradigmatic example, consider the causal graph given in Fig. 1: instead of directly using $\mathbf{Z} = \{A_1, \ldots, A_k, B\}$, there are two possible smaller subsets within $\mathbf{Z}$ itself that satisfy the (generalized) backdoor adjustment criterion (Pearl, 1995; Shpitser et al., 2010; Maathuis and Colombo, 2015; Perković et al., 2018), and that therefore also serve as valid adjustment sets.

As a first approach for obtaining smaller adjustment sets, we consider Markov blankets of the treatments $\mathbf{X}$. In particular, we say that $\mathbf{S} \subseteq \mathbf{Z}$ is a *Markov blanket* of $\mathbf{X}$ with respect to $\mathbf{Z}$ when $\mathbf{X} \perp\!\!\!\perp \mathbf{Z} \setminus \mathbf{S} \mid \mathbf{S}$; in this case, one can show that $T_{\mathbf{S},\mathbf{x},\mathbf{y}}$ = $T_{\mathbf{Z},\mathbf{x},\mathbf{y}}$. In the example given in Fig. 1, the parent set $\text{Pa}(X) = \{A_1, \ldots, A_k\}$ satisfies this condition: $X \perp\!\!\!\perp \{B\} \mid \text{Pa}(X)$. To adapt this notion in the finite sample setting, we consider an approximate version of conditional independence and define a parameter $\Delta_{\mathbf{X} \perp\!\!\!\perp \mathbf{Z} \setminus \mathbf{S} \mid \mathbf{S}}$ that quantifies how much the conditional independence condition is violated. When $\Delta_{\mathbf{X} \perp\!\!\!\perp \mathbf{Z} \setminus \mathbf{S} \mid \mathbf{S}} \leq \varepsilon$, we say that $\mathbf{S}$ is an ε-*Markov blanket* of $\mathbf{X}$ with respect to $\mathbf{Z}$.

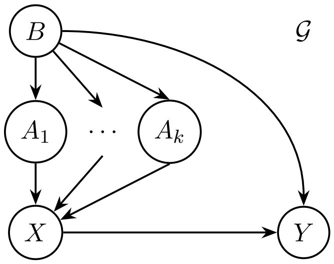


Figure 1: Consider the graphical causality framework and $\mathbb{P}_x(y)$ in $\mathcal{G}$ with $\mathbf{Z} = \{A_1, \ldots, A_k, B\}$ as a valid adjustment set. Both the parental set $\text{Pa}(X) = \{A_1, \ldots, A_k\}$ and the singleton set $\{B\}$ satisfy the backdoor adjustment criterion (Pearl, 1995) and are also valid.

Definition 2 (Approximate conditional independence) *For disjoint sets* $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$, *we define* $\Delta_{\mathbf{A} \perp\!\!\!\perp \mathbf{B} | \mathbf{C}} = \sum_{\mathbf{a},\mathbf{b},\mathbf{c}} \mathbb{P}(\mathbf{c}) \cdot |\mathbb{P}(\mathbf{a}, \mathbf{b} \mid \mathbf{c}) - \mathbb{P}(\mathbf{a} \mid \mathbf{c}) \cdot \mathbb{P}(\mathbf{b} \mid \mathbf{c})|$. *If* $\Delta_{\mathbf{A} \perp\!\!\!\perp \mathbf{B} | \mathbf{C}} \leq \varepsilon$, *we write* $\mathbf{A} \perp\!\!\!\perp_\varepsilon \mathbf{B} \mid \mathbf{C}$.

Definition 3 ((Approximate) Markov blanket) *Consider an arbitrary subset* $\mathbf{A} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$. *A subset* $\mathbf{S} \subseteq \mathbf{A}$ *is called a* Markov blanket *of* $\mathbf{X}$ *with respect to* $\mathbf{A}$ *if* $\mathbf{X} \perp\!\!\!\perp \mathbf{A} \setminus \mathbf{S} \mid \mathbf{S}$ *and an* ε-Markov blanket *if* $\mathbf{X} \perp\!\!\!\perp_\varepsilon \mathbf{A} \setminus \mathbf{S} \mid \mathbf{S}$.

We show that $|T_{\mathbf{A},\mathbf{x},\mathbf{y}} - T_{\mathbf{S},\mathbf{x},\mathbf{y}}| \leq \frac{\varepsilon}{\alpha_{\mathbf{S}}}$ whenever $\mathbf{S}$ is an ε-Markov blanket of $\mathbf{A}$. In particular, if $\mathbf{A} = \mathbf{Z}$ is a valid adjustment set, then this bound applies to the misspecification bias mentioned above. Our next result bounds the sample complexity for discovering an ε-Markov blanket.

Theorem 4 (Approximate Markov blanket discovery) *Suppose we are given (1)* $\varepsilon > 0$, *(2)* $\delta > 0$, *(3) sample access to a distribution* $\mathbb{P}(\mathbf{V})$, *and (4) an arbitrary subset* $\mathbf{A} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$. *Suppose that there is a Markov blanket of* $\mathbf{X}$ *with respect to* $\mathbf{A}$ *with* k *variables. Then, there is an algorithm that uses* $\widetilde{\mathcal{O}}\left(\frac{|\mathbf{S}|}{\varepsilon^2} \cdot \sqrt{|\Sigma_{\mathbf{X}}| \cdot |\Sigma_{\mathbf{A}}|} \cdot \log \frac{1}{\delta}\right)$ *samples and produces a subset* $\mathbf{S} \subseteq \mathbf{A}$ *such that* $|\mathbf{S}| \leq k$, $\Pr\left(\Delta_{\mathbf{X} \perp\!\!\!\perp \mathbf{A} \setminus \mathbf{S} | \mathbf{S}} \leq \varepsilon\right) \geq 1 - \delta$, *and* $\Pr\left(|T_{\mathbf{S},\mathbf{x},\mathbf{y}} - T_{\mathbf{A},\mathbf{x},\mathbf{y}}| \leq \frac{\varepsilon}{\alpha_{\mathbf{S}}}\right) \geq 1 - \delta$.

Now, suppose that Theorem 4 outputs $\mathbf{S} \subseteq \mathbf{Z}$ when given a valid adjustment set $\mathbf{Z}$. While $|\mathbf{S}|$ may be smaller than $|\mathbf{Z}|$, it may still be much larger than the smallest valid adjustment set for $\mathbb{P}_{\mathbf{x}}(\mathbf{y})$. For example, we see that $|\mathbf{Z}| = k + 1 > k = |\text{Pa}(X)| \gg |\{B\}| = 1$ in Fig. 1 where $\mathbf{Z}$, $\text{Pa}(X)$, and $\{B\}$ are all valid adjustment sets. Our next result aims to find an adjustment set $\mathbf{S}' \subseteq \mathbf{Z}$ of *minimal size* given a valid adjustment set $\mathbf{Z}$ and an ε-Markov blanket $\mathbf{S} \subseteq \mathbf{Z}$ of it. To this end, we introduce the more general concept of a *screening set* of an arbitrary subset $\mathbf{A} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$.

Definition 5 ((Approximate) Screening set) *Let* $\mathbf{A} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$ *and* $\mathbf{S} \subseteq \mathbf{A}$. *A subset* $\mathbf{B} \subseteq \mathbf{A}$ *is called a* screening set *for* $(\mathbf{S}, \mathbf{A}, \mathbf{X}, \mathbf{Y})$ *if* $\mathbf{Y} \perp\!\!\!\perp \mathbf{S} \setminus \mathbf{B} \mid \mathbf{X} \cup \mathbf{B}$ *and* $\mathbf{X} \perp\!\!\!\perp \mathbf{B} \setminus \mathbf{S} \mid \mathbf{S}$. *Meanwhile, the subset* $\mathbf{B}$ *is called an* ε-screening set *for* $(\mathbf{S}, \mathbf{A}, \mathbf{X}, \mathbf{Y})$ *if* $\mathbf{Y} \perp\!\!\!\perp_\varepsilon \mathbf{S} \setminus \mathbf{B} \mid \mathbf{X} \cup \mathbf{B}$ *and* $\mathbf{X} \perp\!\!\!\perp_\varepsilon \mathbf{B} \setminus \mathbf{S} \mid \mathbf{S}$.

As a technical side note (see the exposition after Lemma 10), given an adjustment set $\mathbf{S}$, the screening set condition for $(\mathbf{S}, \mathbf{A}, \mathbf{X}, \mathbf{Y})$ is sound for $\mathbf{B}$ to be a valid adjustment set, but it is incomplete in general, in that sense that there may exist valid adjustment sets that do not satisfy the screening set condition. In the worst case, our algorithm in Theorem 6 will output $\mathbf{S}' = \mathbf{S}$.

Theorem 6 (Beyond approximate Markov blankets) *Suppose we are given (1) $\varepsilon > 0$, (2) $\delta > 0$, (3) sample access to $\mathbb{P}(\mathbf{V})$, (4) an arbitrary subset $\mathbf{A} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$, and (5) an ε-Markov blanket $\mathbf{S} \subseteq \mathbf{A}$. Suppose there is a screening set $\mathbf{B}$ for $(\mathbf{S}, \mathbf{A}, \mathbf{X}, \mathbf{Y})$ such that $|\mathbf{B}| = k'$ and $|\Sigma_{\mathbf{B}}| \leq |\Sigma_{\mathbf{S}}|$. There is an algorithm that uses $\widetilde{\mathcal{O}}\left(\frac{|\mathbf{S}'|}{\varepsilon^2} \cdot \sqrt{|\boldsymbol{\Sigma}_{\mathbf{X}}| \cdot |\boldsymbol{\Sigma}_{\mathbf{Y}}| \cdot |\boldsymbol{\Sigma}_{\mathbf{A}}|} \cdot \log \frac{1}{\delta}\right)$ samples and produces a subset $\mathbf{S}' \subseteq \mathbf{A}$ such that $|\mathbf{S}'| \leq k'$, $|\boldsymbol{\Sigma}_{\mathbf{S}'}| \leq |\boldsymbol{\Sigma}_{\mathbf{S}}|$ and $\Pr\left(|T_{\mathbf{S}',\mathbf{x},\mathbf{y}} - T_{\mathbf{A},\mathbf{x},\mathbf{y}}| \leq \frac{2\varepsilon}{\alpha_{\mathbf{S}}}\right) \geq 1 - \delta$.*

As we shall see, unlike many existing causal discovery methods, e.g. the PC algorithm (Spirtes et al., 2000), which perform a sequence of dependent conditional independence checks, our algorithms for Theorem 4 and Theorem 6 use a *non-dependent* collection of conditional independence tests, allowing us to avoid error propagation and control the sample complexity of our procedures.

Finally, one can combine the PAC bound results above to yield an overall PAC bound guarantee for solving the causal effect estimation problem as follows. Since Lemma 8 tells us that $T_{\mathbf{S},\mathbf{x},\mathbf{y}}$ and $T_{\mathbf{Z},\mathbf{x},\mathbf{y}}$ are close whenever $\mathbf{S}$ is an ε-Markov blanket of $\mathbf{X}$ with respect to $\mathbf{Z}$, we can employ the algorithm in Theorem 4 to find a subset $\mathbf{S} \subseteq \mathbf{Z}$ such that $T_{\mathbf{S},\mathbf{x},\mathbf{y}} \approx T_{\mathbf{Z},\mathbf{x},\mathbf{y}}$. Using the ε-Markov blanket $\mathbf{S}$, we can further use the algorithm in Theorem 6 to find a subset $\mathbf{S}' \subseteq \mathbf{Z}$ such that $T_{\mathbf{S}',\mathbf{x},\mathbf{y}} \approx T_{\mathbf{Z},\mathbf{x},\mathbf{y}}$. Depending on whether $|\boldsymbol{\Sigma}_{\mathbf{S}}|$ or $|\boldsymbol{\Sigma}_{\mathbf{S}'}|$ is smaller, we can employ Theorem 1 to obtain an estimate $\widehat{T}_{\mathbf{S},\mathbf{x},\mathbf{y}}$ or $\widehat{T}_{\mathbf{S}',\mathbf{x},\mathbf{y}}$, and use that as an estimate for $T_{\mathbf{Z},\mathbf{x},\mathbf{y}} = \mathbb{P}_{\mathbf{x}}(\mathbf{y})$.

In practical situations where one is given a fixed number of samples, we can re-express the results of Theorem 1, Theorem 4 and Theorem 6 in terms of an error upper bound. Then, one can derive a condition under which a combined approach based on above results estimates $\widehat{\mathbb{P}}_{\mathbf{x}}(\mathbf{y})$ via $\widehat{T}_{\mathbf{S}',\mathbf{x},\mathbf{y}}$, for some $\mathbf{S}' \subseteq \mathbf{Z}$, and provably achieves a smaller asymptotic error than directly estimating $\widehat{\mathbb{P}}_{\mathbf{x}}(\mathbf{y})$ via $\widehat{T}_{\mathbf{Z},\mathbf{x},\mathbf{y}}$. The condition relies on the positivity of $\alpha_{\mathbf{S}}$ for subsets $\mathbf{S}, \mathbf{S}' \subseteq \mathbf{Z}$ which are unknown a priori. However, if one is willing to make lower bound assumptions on these α values, possibly due to background knowledge, then one can obtain a result in the same vein as Theorem 7.

Theorem 7 (PAC causal effect estimation with positivity) *Suppose we are given (1) $\varepsilon > 0$, (2) $\delta > 0$, (3) n i.i.d. samples from $\mathbb{P}(\mathbf{V})$, (4) an interventional query $\mathbb{P}_{\mathbf{x}}(\mathbf{y})$, (5) a valid adjustment set $\mathbf{Z} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$, and (6) guaranteed that $\alpha_{\mathbf{S}} \geq \alpha \in (0, 1)$ for any $\mathbf{S} \subseteq \mathbf{Z}$. Then, there is an algorithm that outputs a subset $\mathbf{S}^* \subseteq \mathbf{Z}$ and an estimate $\widehat{\mathbb{P}}_{\mathbf{x}}(\mathbf{y}) = \widehat{T}_{\mathbf{S}^*,\mathbf{x},\mathbf{y}}$ such that $\Pr\left(\left|\widehat{\mathbb{P}}_{\mathbf{x}}(\mathbf{y}) - \mathbb{P}_{\mathbf{x}}(\mathbf{y})\right| \leq \varepsilon\right) \geq 1 - \delta$ for some error term*

$$\varepsilon \in \widetilde{\mathcal{O}}\left(\frac{1}{n} \cdot \frac{|\boldsymbol{\Sigma}_{\mathbf{S}^*}|}{\alpha} + \frac{1}{\sqrt{n}} \cdot \left(\frac{\sqrt{|\mathbf{Z}|} \cdot (|\boldsymbol{\Sigma}_{\mathbf{X}}| \cdot |\boldsymbol{\Sigma}_{\mathbf{Y}}| \cdot |\boldsymbol{\Sigma}_{\mathbf{Z}}|)^{\frac{1}{4}}}{\alpha} + \frac{1}{\sqrt{\alpha}} + \sqrt{|\boldsymbol{\Sigma}_{\mathbf{S}^*}|}\right)\right).$$

Moreover, if there exists a Markov blanket $\mathbf{S}$ of $\mathbf{X}$ such that $|\mathbf{S}| \cdot \sqrt{\frac{|\boldsymbol{\Sigma}_{\mathbf{X}}|}{|\boldsymbol{\Sigma}_{\mathbf{Z}}|}} < \max\left\{\frac{|\boldsymbol{\Sigma}_{\mathbf{Z}}|}{n}, \frac{\alpha_{\mathbf{S}}}{|\boldsymbol{\Sigma}_{\mathbf{Z}}|}, \alpha_{\mathbf{S}}^2\right\}$, then $|\mathbf{S}^*| \leq |\mathbf{S}|$. Note that we hide a dependency on $\log(1/\delta)$ using the $\widetilde{\mathcal{O}}(\cdot)$ notation for readability.

For binary systems with constant α, $|\mathbf{X}|$, and $|\mathbf{Y}|$, the error term in Theorem 7 reduces to $\varepsilon \in \widetilde{\mathcal{O}}\left(\frac{2^{|\mathbf{S}^*|}}{n} + \frac{1}{\sqrt{n}} \cdot \left(2^{\frac{|\mathbf{Z}|}{4} + \log_2 |\mathbf{Z}|} + 2^{\frac{|\mathbf{S}^*|}{2}}\right)\right)$. Recall that $\mathbf{S}^* \subseteq \mathbf{Z}$ and so on would expect substantial improvements over directly adjusting using $\mathbf{Z}$ when $|\mathbf{S}^*| \ll |\mathbf{Z}|/2$. This can happen when there exists a relatively small covariate adjustment set with respect to a high-dimensional covariate setting; see related work in Appendix A.3.

1.2. Overview of technical results

We now review some additional technical results used to establish our main results, though we note that these results may be of independent interest for future work. While our notation and language is closer to Pearl's graphical causal modeling framework (Pearl, 2009), all of our results are compatible with both the PO and graphical frameworks as long as Eq. (1) holds for the given $\mathbf{Z}$. This is because our analysis is purely probabilistic in nature, with the causal interpretation always going back to assuming that $\mathbb{P}_{\mathbf{x}}(\mathbf{y}) = T_{\mathbf{Z},\mathbf{x},\mathbf{y}}$ as a starting point.

The sample complexity bound of Theorem 1 heavily relies on a common technique in the property testing literature known as Poissonization, e.g. see (Valiant, 2008, Section 4.3), (Canonne, 2020b, Appendix D.3), and (Canonne, 2022, Appendix C). The high level idea is that instead of drawing n i.i.d. samples, we will draw $N_{\text{Pois}} \sim \text{Pois}(n)$ i.i.d samples, where N_{Pois} is a random Poisson variable, so that the random count for each realized value will be independent. Section 2.1 describes the Poissonization sampling technique in further detail.

In our error analyses in Theorem 4 and Theorem 6, we manipulate approximate conditional independence terms $\Delta_{\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}}$ (from Definition 2) and adjustment terms $T_{\mathbf{A},\mathbf{x},\mathbf{y}}$ (from Eq. (1)) for various subsets $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$. By standard probability manipulations, one can easily obtain the following alternative representation of $\Delta_{\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}}$ for arbitrary disjoint subsets $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$.

$$\begin{aligned} \Delta_{\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}} &= \sum\nolimits_{\mathbf{a},\mathbf{b},\mathbf{c}} \mathbb{P}(\mathbf{c}) \cdot \left| \mathbb{P}(\mathbf{a}, \mathbf{b} \mid \mathbf{c}) - \mathbb{P}(\mathbf{a} \mid \mathbf{c}) \cdot \mathbb{P}(\mathbf{b} \mid \mathbf{c}) \right| \\ &= \sum\nolimits_{\mathbf{a},\mathbf{b},\mathbf{c}} \mathbb{P}(\mathbf{a}, \mathbf{c}) \cdot \left| \mathbb{P}(\mathbf{b} \mid \mathbf{a}, \mathbf{c}) - \mathbb{P}(\mathbf{b} \mid \mathbf{c}) \right| \leq \varepsilon \end{aligned} \tag{3}$$

Meanwhile, for any arbitrary disjoint subsets $\mathbf{A}, \mathbf{B} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$, $T_{\mathbf{A},\mathbf{x},\mathbf{y}}$ can be re-expressed in multiple ways (depending on the desired analytical use case) using law of total probability as

$$T_{\mathbf{A},\mathbf{x},\mathbf{y}} = \sum\nolimits_{\mathbf{a}} \mathbb{P}(\mathbf{y} \mid \mathbf{a}, \mathbf{x}) \cdot \mathbb{P}(\mathbf{a}) = \sum\nolimits_{\mathbf{a},\mathbf{b}} \mathbb{P}(\mathbf{y} \mid \mathbf{a}, \mathbf{x}) \cdot \mathbb{P}(\mathbf{a}) \cdot \mathbb{P}(\mathbf{b} \mid \mathbf{a}) \tag{4}$$

The correctness of Theorem 4 follows from the following result that $T_{\mathbf{S},\mathbf{x},\mathbf{y}}$ and $T_{\mathbf{A},\mathbf{x},\mathbf{y}}$ are close whenever $\mathbf{S}$ is an ε-Markov blanket of $\mathbf{S}$ with respect to $\mathbf{A}$, and that there is a sample efficient way to obtain such an ε-Markov blanket.

Lemma 8 (Misspecification error) *If* $\mathbf{S} \subseteq \mathbf{A} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$ *such that* $\mathbf{X} \perp\!\!\!\perp_{\varepsilon} \mathbf{A} \setminus \mathbf{S} \mid \mathbf{S}$, *then* $|T_{\mathbf{S},\mathbf{x},\mathbf{y}} - T_{\mathbf{A},\mathbf{x},\mathbf{y}}| \leq \frac{\varepsilon}{\alpha_{\mathbf{S}}}$.

We additionally complement Lemma 8 with a hardness result of Lemma 9.

Lemma 9 (Misspecification error lower bound) *Let* $0 \leq \sqrt{\varepsilon} \leq \alpha \leq 1/2$. *There exists* $\mathbb{P}(\mathbf{V})$ *such that (i)* $\mathbf{Z}$ *is a valid adjustment set, (ii)* $\mathbf{S} \subset \mathbf{Z}$ *satisfies* $\mathbf{X} \perp\!\!\!\perp_{\varepsilon} \mathbf{Z} \setminus \mathbf{S} \mid \mathbf{S}$, *(iii)* $\alpha_{\mathbf{S}} \geq \alpha$, *and (iv)* $|T_{\mathbf{S},\mathbf{x},\mathbf{y}} - T_{\mathbf{Z},\mathbf{x},\mathbf{y}}| \geq \frac{\varepsilon}{16\alpha}$.

Similar in spirit to Theorem 4, the correctness of Theorem 6 relies on the relating $T_{\mathbf{S}',\mathbf{x},\mathbf{y}}$ and $T_{\mathbf{S},\mathbf{x},\mathbf{y}}$ via some conditional independence relations. Given an ε-Markov blanket $\mathbf{S} \subseteq \mathbf{A}$, we search for a minimal-sized screening set for $(\mathbf{S}, \mathbf{X}, \mathbf{Y})$. This is a sound approach because of Lemma 10.

Lemma 10 (Adjustment soundness) *Let* $\mathbf{A} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$ *be an arbitrary subset and* $\mathbf{S} \subseteq \mathbf{A}$. *If* $\mathbf{S}'$ *is a screening set for* $(\mathbf{S}, \mathbf{X}, \mathbf{Y})$, *then* $T_{\mathbf{S}',\mathbf{x},\mathbf{y}} = T_{\mathbf{S},\mathbf{x},\mathbf{y}}$.

Although this approach is sound, it may not find the smallest $\mathbf{S}'$ satisfying $T_{\mathbf{S}',\mathbf{x},\mathbf{y}} = T_{\mathbf{S},\mathbf{x},\mathbf{y}}$ for arbitrary $\mathbf{S} \subseteq \mathbf{A}$. Nevertheless, there are special scenarios in which this approach is also complete. We give a full statement of such a graphical condition and a completeness proof in Appendix B.4. Here, we provide some brief intuition: the necessity of the $\mathbf{Y} \perp\!\!\!\perp \mathbf{S} \setminus \mathbf{S}' \mid \mathbf{X} \cup \mathbf{S}'$ condition can be seen by setting $\mathbf{A} = \{A_1, \ldots, A_k, B\}$, $\mathbf{S} = \{A_1, \ldots, A_k\}$, and $\mathbf{S}' = \{B\}$ in Fig. 1. In this setup, we see that $\mathbf{S}$ is a valid backdoor adjustment set. Conditioning on X blocks any paths from $\mathbf{S}$ to Y that has a causal path from X to Y as a subpath. So, $\mathbf{Y} \perp\!\!\!\perp \mathbf{S} \setminus \mathbf{S}' \mid \mathbf{X} \cup \mathbf{S}'$ will imply that $\mathbf{S}'$ also blocks non-causal X to Y paths, as any such path passing through $\mathbf{S} \setminus \mathbf{S}'$ has to pass through $\mathbf{S}'$.

Finally, there is nothing technically special about Theorem 7 besides simply combining the results Theorem 1, Theorem 4, and Theorem 6 in a straightforward fashion.

1.3. Organization of the paper

The remainder of the paper is devoted to establishing and providing intuition behind our results. Related work, additional derivations, full proofs, and experimental results are deferred to the appendix. In Section 2, we establish our notation and describe necessary preliminaries. In Sections 3, 4.2 and 4.3, we prove our main results and associated technical results. In particular, we prove Theorem 1 in Section 3 and the other results (Theorem 4, Theorem 6, Lemma 8, Lemma 9, and Lemma 10) in Section 4. We end with a summary and a discussion of open problems in Section 5.

2. Preliminaries

Notation We use capital letters for random variables and lowercase letters for the realizations, e.g. $X = x$, $Y = y$, etc. We use bold letters for sets of variables and write $\mathbb{P}(\mathbf{A} = \mathbf{a})$ as $\mathbb{P}(\mathbf{a})$ as shorthand. We denote the alphabet of the variable V as $\mathbf{\Sigma}_V$, and extend this to sets by letting $\mathbf{\Sigma}_\mathbf{A} = \mathbf{\Sigma}_{V_1} \times \ldots \times \mathbf{\Sigma}_{V_k}$, where $\mathbf{A} = \{V_1, \ldots, V_k\}$ and × denotes the Cartesian product. To lighten notation, summations are always taken over the entire alphabet of the index, i.e., $\sum_\mathbf{a} f(\mathbf{a})$ denotes $\sum_{\mathbf{a} \in \Sigma_\mathbf{A}} f(\mathbf{a})$. We employ the standard asymptotic notations $\mathcal{O}(\cdot)$, $\Omega(\cdot)$ $\Theta(\cdot)$, and $\widetilde{\mathcal{O}}(\cdot)$.

Throughout this work, we will denote $\mathbf{X}$ as the intervened treatment variables and $\mathbf{Y}$ as the outcome variables of interest. For some $\mathbf{x} \in \mathbf{\Sigma}_\mathbf{X}$ and $\mathbf{y} \in \mathbf{\Sigma}$, our goal is to estimate $\mathbb{P}_\mathbf{x}(\mathbf{y})$, which denotes the probability that $\mathbf{Y}$ takes on the value $\mathbf{y}$ if we intervene to set $\mathbf{X}$ equal to $\mathbf{x}$.

A short note on valid adjustment sets In this work, we take as our starting point knowledge of some $\mathbf{Z} \subset \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$ that is a valid adjustment set, i.e., we assume that Eq. (1) holds for some known $\mathbf{Z}$. Instead, one may prefer to derive Eq. (1) from more basic assumptions. For example, in the potential outcomes (PO) framework, $\mathbb{P}_\mathbf{x}(\mathbf{y})$ is usually written as $\mathbb{P}(\mathbf{Y}(\mathbf{x}) = \mathbf{y})$, where $\mathbf{Y}(\mathbf{x})$ denotes the potential outcome when $\mathbf{X}$ is set to $\mathbf{x}$. Then, Eq. (1) is implied by the standard consistency assumption and conditional ignorability of $\mathbf{X}$ with respect to $\mathbf{Z}$; see Lemma 29. Alternatively, Eq. (1) can be derived in the graphical causality framework, which relates the distributions $\mathbb{P}(\mathbf{V})$ and $\mathbb{P}_\mathbf{x}(\mathbf{Y})$ to a (possibly unknown) *causal graph* $\mathcal{G}$ over the random variables $\mathbf{V}$. Typically, one might assume that $\mathcal{G}$ an *acyclic directed mixed graph* (ADMG), which can contain both directed edges and bidirected edges, but no directed cycles; if there are no bidirected edges, then $\mathcal{G}$ is called a *directed acyclic graph* (DAG). An ADMG (resp. DAGs) $\mathcal{G}$ defines a relation between subsets $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$ called *m-separation* (resp. *d-separation*), and the distributions $\mathbb{P}(\mathbf{V})$ and $\mathbb{P}_\mathbf{x}(\mathbf{y})$ are assumed to be related via m-separation in $\mathcal{G}$ and related graphs. Then, Eq. (1) can be derived from these assumptions and graphical conditions on $\mathbf{Z}$, see Appendix A for examples of such conditions.

2.1. Poissonization

We now describe a common technique known as Poissonization; e.g. see Valiant (2008); Canonne (2020b, 2022). When drawing n i.i.d. samples from an underlying distribution $\mathbb{P}(X)$ over a domain $\mathbf{\Sigma}_X = \{1, \ldots, k\}$, the vector of counts $(N_1, \ldots, N_k)$ follows a multinomial distribution with parameters n and $(\mathbb{P}(X = 1), \ldots, \mathbb{P}(X = k))$, where each random variable N_i is the number of times we observe $i \in [k]$ amongst the $n = N_1 + \ldots + N_k$ drawn samples. Oftentimes, in analysis, we would like that the random variables $N_1, \ldots, N_k$ are independent; unfortunately, this is false in the standard sampling setting since $N_1, \ldots, N_k$ are negatively correlated.

Instead of directly drawing n i.i.d. samples, the idea behind Poissonization is to modify the sampling process by first sampling a Poisson number $N_{\text{Pois}} \sim \text{Pois}(n)$ with mean n and then drawing N_{Pois} i.i.d samples. By standard Poisson concentration bounds (e.g., Lemma 12 below), N_{Pois} is of order $\mathcal{O}(n)$ with high probability; thus, PAC bounds for the Poissonized setting are interchangeable with those in the standard setting up to constant factors (see e.g., Lemmas C.1 and C.2 in Canonne (2022)). In the Poissonized setting, the resulting count vector has a few desirable properties.

Lemma 11 (Appendix C of Canonne (2022)) *Let $(N_1, \ldots, N_k)$ be the sample counts in the Poissonized sampling process such that $N_1 + \ldots + N_k = N_{\text{Pois}} \sim \text{Pois}(n)$. The following hold:*
(a) The random count variables $N_1, \ldots, N_k$ are mutually independent.
(b) For each $i \in [k]$, we have $N_i \sim \text{Pois}(n \cdot \mathbb{P}(X = i))$.
(c) For each $i \in [k]$ and natural number n', we have $(N_i \mid N_{\text{Pois}} = n') \sim \text{Bin}(n', \mathbb{P}(X = i))$.

2.2. Concentration bounds

The first two terms in Theorem 1 come from splitting the values of $\mathbf{A}$ into two sets according to some cutoff parameter τ that we later optimize. In Eq. (9), we define an event $\mathcal{E}^J_{\geq \tau}$; the $\widetilde{\mathcal{O}}\left(\frac{|\mathbf{\Sigma}_{\mathbf{A}}|}{\varepsilon \alpha_{\mathbf{A}}}\right)$ term captures the number of samples needed to ensure the $\mathcal{E}^J_{\geq \tau}$ holds with high probability, which we compute using a standard Poisson concentration bound (Lemma 12). Next, we define a random variable $J_{\geq \tau}$ (Eq. (7)); the $\widetilde{\mathcal{O}}\left(\frac{1}{\varepsilon^2 \alpha_{\mathbf{A}}}\right)$ term comes from a concentration bound on this term. In particular, the bulk of our analysis is devoted to showing that $J_{\geq \tau}$ is sub-Gaussian conditioned on $\mathcal{E}^J_{\geq \tau}$, for this result, we require Lemmas 14, 15, and 16 below.

Lemma 12 (Poisson concentration; e.g. see Theorem A.8 in Canonne (2022)) *Let $N \sim \text{Pois}(n)$ be a Poisson random variable with parameter n. Then, for any $0 < t < n$, we have $\Pr(N \leq n - t) \leq \exp\left(-\frac{t^2}{2(n+t)}\right)$. In particular, setting $t = n/2$, we have $\Pr(N \leq n/2) \leq \exp\left(-\frac{n}{12}\right)$.*

Definition 13 (Sub-Gaussian distribution; e.g. see Section 1.2 of Rigollet and Hütter (2023)) *A random variable X is* sub-Gaussian *with parameter σ^2 if $\mathbb{E}(X) = 0$ and $\mathbb{E}\left(e^{\lambda X}\right) \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right)$ for all $\lambda \in \mathbb{R}$. If $X \sim \text{subG}(\sigma^2)$, it is known that $\Pr(|X| \geq t) \leq 2 \exp\left(-\frac{t^2}{2\sigma^2}\right)$ for any $t \geq 0$.*

Lemma 14 (Sub-Gaussian additivity; e.g. see Corollary 1.7 of Rigollet and Hütter (2023)) *For $i \in [k]$, let $X_i \sim \text{subG}(\sigma_i^2)$ be an independent sub-Gaussian random variable with parameter σ_i^2. Then, for any set of real coefficients $a_1, \ldots, a_k \in \mathbb{R}$, we have $\left(\sum_{i=1}^k a_i X_i\right) \sim \text{subG}(\sum_{i=1}^k a_i^2 \sigma_i^2)$.*

Lemma 15 (See Appendix B.2) *Let X and Y be discrete random variables. If $(X \mid Y = y) \sim \text{subG}(\sigma_y^2)$ for every $y \in \mathbf{\Sigma}_Y$, then $X \sim \text{subG}(\max_{y \in \mathbf{\Sigma}_Y} \sigma_y^2)$.*

Lemma 16 (Hoeffding's lemma; Hoeffding (1994)) *Let X be any real-valued random variable in the range $[a,b]$. Then, for any $\lambda \in \mathbb{R}$, we have $\mathbb{E}\left(e^{\lambda(X-\mathbb{E}(X))}\right) \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right)$.*

2.3. Known sample complexity results for discrete distributions

The third term in Theorem 1 comes from reducing one subtask in our analysis to the problem of estimating $\mathbb{P}(\mathbf{A})$ well in TV distance, for some subset of variables $\mathbf{A} \subseteq \mathbf{V}$. In the distribution testing literature, this task is well-known to require $\widetilde{\Theta}\left(\frac{|\mathbf{\Sigma_A}|}{\varepsilon^2}\right)$ i.i.d. samples.

Lemma 17 (Estimating well in TV; e.g. see Canonne (2020a)) *Given tolerance parameters $\varepsilon, \delta > 0$ and sample access to a distribution $\mathbb{P}(\mathbf{A})$, the empirical distribution $\widehat{\mathbb{P}}(\mathbf{A})$ constructed from $\mathcal{O}\left(\frac{|\mathbf{\Sigma_A}|+\log\frac{1}{\delta}}{\varepsilon^2}\right)$ i.i.d. samples has the property that $\Pr\left(\sum_{\mathbf{a}\in\mathbf{\Sigma_A}} |\mathbb{P}(\mathbf{a}) - \widehat{\mathbb{P}}(\mathbf{a})| \leq \varepsilon\right) \geq 1-\delta$.*

Meanwhile, for our proofs of Theorems 4 and 6, several methods have been developed which satisfy the requirements of the ε-approximate conditional independence tester for Definition 2. In this work, we call our ε-approximate conditional independence tester ApproxCondInd. Assuming that ε^{-1} is sufficiently large[2] compared to $|\mathbf{\Sigma_A}|$, $|\mathbf{\Sigma_B}|$, and $|\mathbf{\Sigma_C}|$, Canonne et al. (2018) proposes a test based on total variation distance that uses $\widetilde{\mathcal{O}}\left(\frac{1}{\varepsilon^2}\cdot\sqrt{|\mathbf{\Sigma_A}|\cdot|\mathbf{\Sigma_B}|\cdot|\mathbf{\Sigma_C}|}\right)$ samples from $\mathbb{P}$; see their Theorem 1.3 and Lemma 2.2. There is also a simpler test based on the empirical mutual information, proposed by Bhattacharyya et al. (2021), that uses $\widetilde{\mathcal{O}}\left(\frac{1}{\varepsilon^2}\cdot|\mathbf{\Sigma_A}|\cdot|\mathbf{\Sigma_B}|\cdot|\mathbf{\Sigma_C}|\right)$ samples from $\mathbb{P}$, though we use the former to obtain optimal dependence on the alphabet sizes.

Lemma 18 (Using Canonne et al. (2018) for ApproxCondInd) *Given $\varepsilon, \delta > 0$ and sample access to distribution $\mathbb{P}(\mathbf{V})$, ApproxCondInd uses $\widetilde{\mathcal{O}}\left(\frac{1}{\varepsilon^2}\cdot\sqrt{|\mathbf{\Sigma_A}|\cdot|\mathbf{\Sigma_B}|\cdot|\mathbf{\Sigma_C}|}\cdot\log\frac{1}{\delta}\right)$ samples and correctly determines whether $\Delta_{\mathbf{A}\perp\!\!\!\perp\mathbf{B}|\mathbf{C}} = 0$ (outputs YES) or $\Delta_{\mathbf{A}\perp\!\!\!\perp\mathbf{B}|\mathbf{C}} > \varepsilon$ (outputs NO) with probability at least $1-\delta$, for any disjoint sets $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$.*

Note that when $0 < \Delta_{\mathbf{A}\perp\!\!\!\perp\mathbf{B}|\mathbf{C}} \leq \varepsilon$, ApproxCondInd is allowed to output arbitrarily. In particular, when ApproxCondInd outputs YES, then we have $\Pr(\Delta_{\mathbf{A}\perp\!\!\!\perp\mathbf{B}|\mathbf{C}} \leq \varepsilon) \geq 1-\delta$.

3. Sample complexity for empirical estimation

In this section, we prove Theorem 1, our upper bound on the sample complexity of estimating $T_{\mathbf{A},\mathbf{x},\mathbf{y}}$ given any $\mathbf{A} \subseteq \mathbf{V}$. For analysis purposes, we will use the Poissonization sampling process (Section 2.1) so that we invoke Lemma 11 to obtain PAC style bounds.

Proof sketch of Theorem 1: By definition, we have

$$\begin{aligned} T_{\mathbf{A},\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{A},\mathbf{x},\mathbf{y}} &= \sum\nolimits_{\mathbf{a}} \left(\mathbb{P}(\mathbf{a})\cdot\mathbb{P}(\mathbf{y}\mid\mathbf{x},\mathbf{a}) - \frac{N_{\mathbf{a}}}{N_{\text{Pois}}}\cdot\frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}}\right) \\ &= \sum\nolimits_{\mathbf{a}} \mathbb{P}(\mathbf{a})\cdot\left(\mathbb{P}(\mathbf{y}\mid\mathbf{x},\mathbf{a}) - \frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}}\right) + \sum\nolimits_{\mathbf{a}} \left(\mathbb{P}(\mathbf{a}) - \frac{N_{\mathbf{a}}}{N_{\text{Pois}}}\right)\cdot\frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}} \end{aligned}$$

[2] For instance, $\frac{1}{\varepsilon} > |\mathbf{\Sigma_C}|^{\frac{1}{4}}\cdot(\max\{|\mathbf{\Sigma_A}|,|\mathbf{\Sigma_B}|,|\mathbf{\Sigma_C}|\})^{\frac{1}{4}}$ would suffice.

where $N_{\mathbf{a}}$, N_{Pois}, $N_{\mathbf{y},\mathbf{x},\mathbf{a}}$, and $N_{\mathbf{x},\mathbf{a}}$ are random variables from the Poissonization process with $N_{\text{Pois}} \sim \text{Pois}(n)$ for some parameter n; see Section 2.1. Since $N_{\text{Pois}} = \sum_{\mathbf{a}} N_{\mathbf{a}} = \sum_{\mathbf{a},\mathbf{x}} N_{\mathbf{x},\mathbf{a}} = \sum_{\mathbf{a},\mathbf{x},\mathbf{y}} N_{\mathbf{y},\mathbf{x},\mathbf{a}}$, we see that $0 \leq \frac{N_{\mathbf{a}}}{N_{\text{Pois}}} \leq 1$ and $0 \leq \frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}} \leq 1$ for each of these fractional terms.

We can define a threshold $\tau > 0$ and partition the values of $\mathbf{A}$ accordingly, and then define three summations $J_{<\tau}$, $J_{\geq\tau}$, and K so that $T_{\mathbf{A},\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{A},\mathbf{x},\mathbf{y}} = J_{<\tau} + J_{\geq\tau} + K$:

$$\begin{aligned}
\mathbf{\Sigma}_{\mathbf{A}\geq\tau} &= \{\mathbf{a} \in \mathbf{\Sigma}_{\mathbf{A}} : \mathbb{P}(\mathbf{x},\mathbf{a}) \geq \tau\} \\
J_{<\tau} &= \sum\nolimits_{\mathbf{a}\notin\mathbf{\Sigma}_{\mathbf{A}\geq\tau}} \mathbb{P}(\mathbf{a}) \cdot \left(\mathbb{P}(\mathbf{y} \mid \mathbf{x},\mathbf{a}) - \frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}} \right) \\
J_{\geq\tau} &= \sum\nolimits_{\mathbf{a}\in\mathbf{\Sigma}_{\mathbf{A}\geq\tau}} \mathbb{P}(\mathbf{a}) \cdot \left(\mathbb{P}(\mathbf{y} \mid \mathbf{x},\mathbf{a}) - \frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}} \right) \\
K &= \sum\nolimits_{\mathbf{a}} \left(\mathbb{P}(\mathbf{a}) - \frac{N_{\mathbf{a}}}{N_{\text{Pois}}} \right) \cdot \frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}}
\end{aligned}$$

Since $\alpha_{\mathbf{A}} = \min_{\mathbf{a}\in\mathbf{\Sigma}_{\mathbf{A}}} \mathbb{P}(\mathbf{x} \mid \mathbf{a})$, we see that $\mathbb{P}(\mathbf{a}) \leq \frac{\tau}{\alpha_{\mathbf{A}}}$ for $\mathbf{a} \notin \mathbf{\Sigma}_{\mathbf{A}\geq\tau}$. One can show that $|J_{<\tau}| \leq \frac{\tau \cdot |\mathbf{\Sigma}_{\mathbf{A}}|}{\alpha_{\mathbf{A}}}$ using triangle inequality and the definition of $\mathbf{\Sigma}_{\mathbf{A}\geq\tau}$. As for $|J_{\geq\tau}|$ and $|K|$, we condition on the concentration event $\mathcal{E}^{J}_{\geq\tau} = \bigcap_{\mathbf{a}\in\mathbf{\Sigma}_{\mathbf{A}\geq\tau}} \left\{ N_{\mathbf{x},\mathbf{a}} > \frac{n \cdot \mathbb{P}(\mathbf{x},\mathbf{a})}{2} \right\}$ which holds with probability at least $1 - |\mathbf{\Sigma}_{\mathbf{A}}| \cdot \exp\left(-\frac{n\tau}{12}\right)$. Under the event $\mathcal{E}^{J}_{\geq\tau}$, one can show that $\Pr\left(|J_{\geq\tau}| > t \mid \mathcal{E}^{J}_{\geq\tau}\right) \leq 2\exp\left(-n\alpha_{\mathbf{A}} t^2\right)$ and we can reduce the analysis of $|K|$ to the problem of producing an ε-close estimate of $\mathbb{P}(\mathbf{A})$ by observing that $0 \leq \frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}} \leq 1$ and $\frac{N_{\mathbf{a}}}{N_{\text{Pois}}}$ is the empirical estimate of $\mathbb{P}(\mathbf{a})$ for each $\mathbf{a} \in \mathbf{\Sigma}_{\mathbf{A}}$. That is, $|K| \leq \sum_{\mathbf{a}} \left| \mathbb{P}(\mathbf{a}) - \widehat{\mathbb{P}}(\mathbf{a}) \right|$. The claim follows by putting together the above discussed bounds appropriately. See Appendix C.2 for details. ■

4. Finding small adjustment sets via approximate Markov blankets

As discussed in Section 1.1, the bound in Theorem 1 motivates the use of small adjustment sets whenever possible. In this section, we focus on searching for such sets. In Section 4.1, we describe our proofs of Lemmas 8 and 9 on the misspecification error of performing covariate adjustment with an approximate Markov blanket. In Section 4.2, we describe our proof of Theorem 4 on the sample complexity of approximate Markov blanket discovery; similarly, Section 4.3 describes our proof of Theorem 6, which searches for even smaller adjustment sets. Section 4.4 concludes by putting these results together into a combination algorithm and describing the proof of our final result, Theorem 7.

4.1. Misspecification error for approximate Markov blankets

We begin with Lemma 8, which extends the equality $T_{\mathbf{S},\mathbf{x},\mathbf{y}} = T_{\mathbf{A},\mathbf{x},\mathbf{y}}$ for exact Markov blankets to a bound on the misspecification error $|T_{\mathbf{S},\mathbf{x},\mathbf{y}} - T_{\mathbf{A},\mathbf{x},\mathbf{y}}|$ for approximate Markov blankets. To accompany this result, we also give a matching lower bound on the misspecification error in Lemma 9.

Proof sketch of Lemma 8: By Eq. (4) and $\mathbf{S} \subseteq \mathbf{A}$, we see that

$$|T_{\mathbf{S},\mathbf{x},\mathbf{y}} - T_{\mathbf{A},\mathbf{x},\mathbf{y}}| = \left| \sum\nolimits_{\mathbf{a}} \mathbb{P}(\mathbf{y} \mid \mathbf{a},\mathbf{x}) \cdot \mathbb{P}(\mathbf{a} \setminus \mathbf{s} \mid \mathbf{s},\mathbf{x}) \cdot \mathbb{P}(\mathbf{s}) - \sum\nolimits_{\mathbf{a}} \mathbb{P}(\mathbf{y} \mid \mathbf{a},\mathbf{x}) \cdot \mathbb{P}(\mathbf{a}) \right|$$

Then by triangle inequality, non-negativity of probabilities, and since $\mathbb{P}(\mathbf{y} \mid \mathbf{a},\mathbf{x}) \leq 1$, Eq. (2), and Eq. (3), one can show that this is at most $\frac{\varepsilon}{\alpha_{\mathbf{S}}}$. See Appendix C.3 for a step-by-step derivation. ■

Proof sketch of Lemma 9: One can construct a distribution $\mathbb{P}$ defined over binary variables $\{X, Y, A, B\}$ with (conditional) probabilities below, which implies the four properties of the claim.

a	b	$\mathbb{P}(b \mid a)$	$\mathbb{P}(X = 0 \mid a, b)$	$\mathbb{P}(X = 0 \mid a)$	$\sum_x \lvert\mathbb{P}(x \mid a, b) - \mathbb{P}(x \mid a)\rvert$
0	0	$\sqrt{\varepsilon}/2$	$1 - \alpha + \sqrt{\varepsilon}/2$	$1 - \alpha + \varepsilon/4$	$\sqrt{\varepsilon} - \varepsilon/2$
0	1	$1 - \sqrt{\varepsilon}/2$	$1 - \alpha$	$1 - \alpha + \varepsilon/4$	$\varepsilon/2$
1	0	$1 - \sqrt{\varepsilon}/2$	α	$\alpha - \varepsilon/4$	$\varepsilon/2$
1	1	$\sqrt{\varepsilon}/2$	$\alpha - \sqrt{\varepsilon}/2$	$\alpha - \varepsilon/4$	$\sqrt{\varepsilon} - \varepsilon/2$

where $\mathbf{S} = \{A\} \subset \{A, B\} = \mathbf{Z}$. See Appendix C.3 for detailed calculations and derivations. ■

4.2. Approximate Markov blankets: discovery and adjustment

We now prove Theorem 4, our upper bound on the sample complexity of finding an ε-Markov blanket of $\mathbf{X}$ with respect to an arbitrary set $\mathbf{A} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$ using Algorithm 1.

Algorithm 1 APPROXIMATEMARKOVBLANKETADJUSTMENT (AMBA)

```
Data: $\varepsilon, \delta > 0$, dataset $\mathbf{D}$ of $n$ i.i.d. samples from $\mathbb{P}(\mathbf{V})$, and subset $\mathbf{A} \subseteq \mathbf{V}$
Result: $\mathbf{S} \subseteq \mathbf{A}$
for $k = 0, 1, 2, \ldots, |\mathbf{A}|$ do
    Let $w_k = \left(|\mathbf{A}| \cdot \binom{|\mathbf{A}|}{k}\right)^{-1}$
    Let $\mathbf{C}_k = \{\mathbf{S} \subseteq \mathbf{A} : |\mathbf{S}| = k$, where
                  APPROXCONDIND$(\mathbf{X} \perp\!\!\!\perp \mathbf{A} \setminus \mathbf{S} \mid \mathbf{S}, \varepsilon, \delta w_k, \mathbf{D})$ outputs YES$\}$
    if $|\mathbf{C}_k| > 0$ then
        return any $\mathbf{S} \in \mathbf{C}_k$
return $\mathbf{A}$
```

Proof of Theorem 4: Suppose AMBA (Algorithm 1) terminates at iteration $|\mathbf{S}| \in \{0, 1, \ldots, |\mathbf{A}|\}$.

Correctness. Suppose all calls to APPROXCONDIND succeed, then Lemma 18 tells us that any produced $\mathbf{S} \subseteq \mathbf{A}$ satisfies the property that $\Delta_{\mathbf{X} \perp\!\!\!\perp \mathbf{A} \setminus \mathbf{S} \mid \mathbf{S}} \leq \varepsilon$. Lemma 8 then further tells us that $\Delta_{\mathbf{X} \perp\!\!\!\perp \mathbf{A} \setminus \mathbf{S} \mid \mathbf{S}} \leq \varepsilon$ implies $|T_{\mathbf{S},\mathbf{x},\mathbf{y}} - T_{\mathbf{A},\mathbf{x},\mathbf{y}}| \leq \frac{\varepsilon}{\alpha_{\mathbf{S}}}$.

Failure rate. There are at most $\binom{|\mathbf{A}|}{k}$ possible candidate sets in $\mathbf{C}_k$ for each $k \in \{0, 1, \ldots, |\mathbf{A}|\}$. Since we invoked each call to APPROXCONDIND with δw_k in iteration k, union bound tells us that the probability of *any* call failing across all calls is at most

$$\sum_{k=0}^{|\mathbf{S}|} \delta w_k \cdot \binom{|\mathbf{A}|}{k} = \sum_{k=0}^{|\mathbf{S}|} \delta \cdot \frac{1}{|\mathbf{A}| \cdot \binom{|\mathbf{A}|}{k}} \cdot \binom{|\mathbf{A}|}{k} = \sum_{k=0}^{|\mathbf{S}|} \frac{\delta}{|\mathbf{A}|} \leq \frac{\delta \cdot |\mathbf{S}|}{|\mathbf{A}|} \leq \delta$$

Sample complexity. Since we use a union bound to bound our overall failure probability, we can reuse samples in all our calls to APPROXCONDIND. Thus, the total sample complexity is dominated by the final call (where $k = |\mathbf{S}|$), which uses $\widetilde{\mathcal{O}}\left(\frac{1}{\varepsilon^2} \cdot \sqrt{|\mathbf{\Sigma}_{\mathbf{X}}| \cdot |\mathbf{\Sigma}_{\mathbf{A} \setminus \mathbf{S}}| \cdot |\mathbf{\Sigma}_{\mathbf{S}}|} \cdot \log \frac{1}{\delta w_k}\right)$ samples according to Lemma 18. Plugging in $w_k = \left(|\mathbf{A}| \cdot \binom{|\mathbf{A}|}{k}\right)^{-1}$, we obtain total sample complexity $\widetilde{\mathcal{O}}\left(\frac{|\mathbf{S}|}{\varepsilon^2} \cdot \sqrt{|\mathbf{\Sigma}_{\mathbf{X}}| \cdot |\mathbf{\Sigma}_{\mathbf{A}}|} \cdot \log \frac{1}{\delta}\right)$, where we omit $\log |\mathbf{A}|$ within $\widetilde{\mathcal{O}}(\cdot)$ because $|\mathbf{A}| \leq |\mathbf{\Sigma}_{\mathbf{A}}|$. ■

4.3. Beyond approximate Markov blankets

Motivated by Fig. 1, which shows that the Markov blanket of $\mathbf{X}$ with respect to $\mathbf{Z}$ may still be large compared to the smallest adjustment set, we study in this section an approach for finding smaller adjustment sets than the Markov blanket. We prove Lemma 10 in Appendix C.4, which establishes conditions on sets $\mathbf{S}' \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$ and $\mathbf{S} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$ such that $T_{\mathbf{S}',\mathbf{x},\mathbf{y}} = T_{\mathbf{S},\mathbf{x},\mathbf{y}}$; this result suggest an approach for going beyond adjustment by Markov blankets. This allows us to show Theorem 6, our upper bound on the sample complexity of finding a set $\mathbf{S}'$ that approximately satisfies the conditions of Lemma 10 with respect to an ε-Markov blanket $\mathbf{S}$.

Algorithm 2 BEYONDAPPROXIMATEMARKOVBLANKETADJUSTMENT (BAMBA)

```
Data: $\varepsilon, \delta > 0$, $n$ i.i.d. samples $\mathbf{D}$ from $\mathbb{P}(\mathbf{V})$, subset $\mathbf{A} \subseteq \mathbf{V}$, and $\varepsilon$-Markov blanket $\mathbf{S} \subseteq \mathbf{A}$
Result: $\mathbf{S}' \subseteq \mathbf{A}$ such that $|\mathbf{\Sigma}_{\mathbf{S}'}| \leq |\mathbf{\Sigma}_{\mathbf{S}}|$
for $k = 0, 1, 2, \ldots, |\mathbf{A}|$ do
    Let $w_k = \left(|\mathbf{A}| \cdot \binom{|\mathbf{A}|}{k}\right)^{-1}$
    Let $\mathbf{C}_k = \{\mathbf{S}' \subseteq \mathbf{A} : |\mathbf{S}'| = k$ and $|\mathbf{\Sigma}_{\mathbf{S}'}| \leq |\mathbf{\Sigma}_{\mathbf{S}}|$, where
                  APPROXCONDIND$(\mathbf{Y} \perp\!\!\!\perp \mathbf{S} \setminus \mathbf{S}' \mid \mathbf{X} \cup \mathbf{S}', \varepsilon, \frac{\delta w_k}{2}, \mathbf{D})$ outputs YES,
                  APPROXCONDIND$(\mathbf{X} \perp\!\!\!\perp \mathbf{S}' \setminus \mathbf{S} \mid \mathbf{S}, \varepsilon, \frac{\delta w_k}{2}, \mathbf{D})$ outputs YES$\}$
    if $|\mathbf{C}_k| > 0$ then
        return any $\mathbf{S}' \in \mathbf{C}_k$
return $\mathbf{S}$
```

Proof sketch of Theorem 6: The proof follows the same structure as that of Theorem 4. The key difference is that in the correctness analysis, we apply triangle inequality $\left|T_{\mathbf{S},\mathbf{x},\mathbf{y}} - T_{\mathbf{S}',\mathbf{x},\mathbf{y}}\right| \leq \left|T_{\mathbf{S},\mathbf{x},\mathbf{y}} - Z_{\mathbf{x},\mathbf{y}}\right| + \left|Z_{\mathbf{x},\mathbf{y}} - T_{\mathbf{S}',\mathbf{x},\mathbf{y}}\right|$ for some intermediate term $Z_{\mathbf{x},\mathbf{y}} = \sum_{\mathbf{s} \cup \mathbf{s}'} \frac{1}{\mathbb{P}(\mathbf{x}|\mathbf{s})} \cdot \mathbb{P}(\mathbf{x}, \mathbf{s} \cup \mathbf{s}') \cdot \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s}')$ which allows us to relate the two approximate conditional independences to $\alpha_{\mathbf{S}}$. The claim follows after upper bounding each of the terms by $\frac{\varepsilon}{\alpha_{\mathbf{S}}}$. See Appendix C.4 for details. ■

4.4. A combination algorithm

In Appendix C.7, we re-express the results of Theorem 1, Theorem 4 and Theorem 6 in terms of an upper bound on error for a fixed number of samples n. After which, there are a couple of ways one could attempt to estimate $\mathbb{P}_{\mathbf{x}}(\mathbf{y})$ when given a valid adjustment set $\mathbf{Z} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$: (1) directly estimate using $\mathbf{Z}$, (2) use AMBA on $\mathbf{Z}$ to produce a subset $\mathbf{S} \subseteq \mathbf{Z}$ and estimate using $\mathbf{S}$, or (3) use AMBA on $\mathbf{Z}$ to produce a subset $\mathbf{S} \subseteq \mathbf{Z}$, then use BAMBA to further produce subset $\mathbf{S}'$, and then estimate using $\mathbf{S}'$. In Appendix C.7, we show the second approach yields an asymptotically smaller error when $|\mathbf{S}| \cdot \sqrt{\frac{|\mathbf{\Sigma}_{\mathbf{X}}|}{|\mathbf{\Sigma}_{\mathbf{Z}}|}} < \max\left\{\frac{|\mathbf{\Sigma}_{\mathbf{Z}}|}{n}, \frac{\alpha_{\mathbf{S}}}{|\mathbf{\Sigma}_{\mathbf{Z}}|}, \alpha_{\mathbf{S}}^2\right\}$; a similar condition holds for the third approach. The proof of Theorem 7 follows by a careful combination of these insights.

Proof sketch of Theorem 7: Consider the following algorithm:

1. Run AMBA to obtain $\mathbf{S} \subseteq \mathbf{Z}$
2. Check if $|\mathbf{S}| \cdot \sqrt{\frac{|\mathbf{\Sigma}_{\mathbf{X}}|}{|\mathbf{\Sigma}_{\mathbf{Z}}|}} < \max\left\{\frac{|\mathbf{\Sigma}_{\mathbf{Z}}|}{n}, \frac{\alpha_{\mathbf{S}}}{|\mathbf{\Sigma}_{\mathbf{Z}}|}, \alpha_{\mathbf{S}}^2\right\}$
3. If so, run BAMBA to obtain $\mathbf{S}' \subseteq \mathbf{Z}$ and produce estimate $\widehat{\mathbb{P}}_{\mathbf{x}}(\mathbf{y}) = \widehat{T}_{\mathbf{S}',\mathbf{x},\mathbf{y}}$
4. Otherwise, produce estimate $\widehat{\mathbb{P}}_{\mathbf{x}}(\mathbf{y}) = \widehat{T}_{\mathbf{Z},\mathbf{x},\mathbf{y}}$

That is, depending on the check, we decide to perform estimation based on $\mathbf{S}^* = \mathbf{S}'$ or $\mathbf{S}^* = \mathbf{Z}$. One can show that the bound holds for each case separately while noting that $\alpha_{\mathbf{S}}, \alpha_{\mathbf{S}'}, \alpha_{\mathbf{Z}} \geq \alpha$. $\blacksquare$

5. Conclusion

We now provide a brief final summary of our contributions. In this paper, we have focused on the problem of estimating the causal effect $\mathbb{P}_{\mathbf{x}}(\mathbf{y})$ in the PAC setting, given access to a valid adjustment set $\mathbf{Z}$, i.e., $\mathbf{Z}$ such that $\mathbb{P}_{\mathbf{x}}(\mathbf{y}) = T_{\mathbf{Z},\mathbf{x},\mathbf{y}}$, defined in Eq. (1).

(1) In Section 3, we established a PAC bound for estimation of $T_{\mathbf{A},\mathbf{x},\mathbf{y}}$ for an arbitrary set $\mathbf{A} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$, with linear dependence on $|\mathbf{\Sigma}_{\mathbf{A}}|$, the alphabet size of $\mathbf{A}$.

(2) In Section 4.2, we established a bound on the misspecification error $|T_{\mathbf{S},\mathbf{x},\mathbf{y}} - T_{\mathbf{A},\mathbf{x},\mathbf{y}}|$ for $\mathbf{S}$ which is an ε-Markov blanket of $\mathbf{X}$ with respect to $\mathbf{Z}$, and a PAC bound for discovering such a set $\mathbf{S}$; leading to a new estimator of $T_{\mathbf{A},\mathbf{x},\mathbf{y}}$ with reduced sample complexity.

(3) In Section 4.3, we established conditions under which $T_{\mathbf{S}',\mathbf{x},\mathbf{y}} = T_{\mathbf{S},\mathbf{x},\mathbf{y}}$, and gave a PAC bound for discovering a set $\mathbf{S}'$ under which these conditions approximately hold; leading to a new estimator of $T_{\mathbf{A},\mathbf{x},\mathbf{y}}$ which goes beyond using the Markov blanket for adjustment.

Furthermore, in Appendix A, we review related work, give further interpretations of our results under the graphical causality framework, and connect our results to existing results in this line of work. These results pave the way for future connections between causal discovery and causal effect estimation, while each standing alone as results of independent interest. In Appendix D, we provide experimental validation of our approach; code for replication is available at https://github.com/csquires/amba-bamba-clear2025. We conclude with a non-exhaustive list of open problems raised by our work, which we expect to be of immediate future interest.

(1) Compared to the expectation bound of Zeng et al. (2024), our PAC bounds contains an additional $\widetilde{\mathcal{O}}\left(\frac{|\mathbf{\Sigma}_{\mathbf{A}}|}{\varepsilon^2}\right)$ term. Can this term be eliminated or shown to be necessary?

(2) Our AMBA algorithm for ε-Markov blanket discovery performs an exhaustive search over subsets of increasing size. Fortunately, this search is "embarrassingly parallel", but is computationally prohibitive without access to parallel computing. Is there a more computationally efficient algorithm for this problem with (nearly) the same sample complexity?

(3) The basic positivity/overlap assumption states that $\mathbb{P}(\mathbf{x} \mid \mathbf{z}) > 0$; whereas our results (and indeed, finite-sample results in general) require bounding *away* from 0, i.e. $\mathbb{P}(\mathbf{x} \mid \mathbf{z}) > \alpha$ for some $\alpha > 0$ that can be estimated from samples. More "local" notions of overlap have also been considered in the literature such as requiring $\mathbb{P}(\mathbf{x} \mid \mathbf{z}) > \alpha$ for those $\mathbf{z}$ such that $\mathbb{P}(\mathbf{z}) > \beta$ for some $\beta > 0$ (Oberst et al., 2020). Such assumptions are *stronger* than ours, so our bounds also apply to those settings, but may be overly pessimistic. Strengthening our bounds under such assumptions would be very interesting for future work.

Acknowledgments

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-PhD/2021-08-013). This study was supported in part by Office of Naval Research Award No. N00014-21-1-2807. The authors would like to thank Karthikeyan Shanmugam, Murat Kocaoglu, Raghavendra Addanki, Emilija Perković, and Bijan Mazaheri for useful discussions, correspondence, and feedback on earlier draft versions.

References

Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: Algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11(1), 2010a.

Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. Local causal and Markov blanket induction for causal discovery and feature selection for classification part II: Analysis and extensions. *Journal of Machine Learning Research*, 11(1), 2010b.

Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):597–623, 2018.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650, 2014.

Arnab Bhattacharyya, Sutanu Gayen, Eric Price, and NV Vinodchandran. Near-optimal learning of tree-structured distributions by Chow-Liu. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 147–160, 2021.

Arnab Bhattacharyya, Sutanu Gayen, Saravanan Kandasamy, Vedant Raval, and Vinodchandran N Variyam. Efficient interventional distribution learning in the PAC framework. In *International Conference on Artificial Intelligence and Statistics*, pages 7531–7549. PMLR, 2022.

Jelena Bradic, Stefan Wager, and Yinchu Zhu. Sparsity double robust inference of average treatment effects. *arXiv preprint arXiv:1905.00744*, 2019.

M Alan Brookhart, Sebastian Schneeweiss, Kenneth J Rothman, Robert J Glynn, Jerry Avorn, and Til Stürmer. Variable selection for propensity score models. *American journal of epidemiology*, 163(12):1149–1156, 2006.

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

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

Clément L Canonne. Topics and techniques in distribution testing: A biased but representative sample. *Foundations and Trends® in Communications and Information Theory*, 19(6):1032–1198, 2022.

Clément L Canonne, Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Testing conditional independence of discrete distributions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 735–748, 2018.

Ting-Hsuan Chang, Zijian Guo, and Daniel Malinsky. Post-selection inference for causal effects after causal discovery. *arXiv preprint arXiv:2405.06763*, 2024.

Moses Charikar, Kirankumar Shiragur, and Aaron Sidford. A general framework for symmetric property estimation. *Advances in Neural Information Processing Systems*, 32, 2019.

Debo Cheng, Jiuyong Li, Lin Liu, Kui Yu, Thuc Duy Le, and Jixue Liu. Toward unique and unbiased causal effect estimation from data with hidden variables. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

Davin Choo and Kirankumar Shiragur. Adaptivity complexity for causal graph discovery. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI '23, 2023a.

Davin Choo and Kirankumar Shiragur. New metrics and search algorithms for weighted causal DAGs. In *International Conference on Machine Learning*, pages 5868–5903. PMLR, 2023b.

Davin Choo and Kirankumar Shiragur. Subset verification and search algorithms for causal DAGs. In *International Conference on Artificial Intelligence and Statistics*, 2023c.

Davin Choo, Kirankumar Shiragur, and Arnab Bhattacharyya. Verification and search algorithms for causal DAGs. *Advances in Neural Information Processing Systems*, 35, 2022.

Davin Choo, Themistoklis Gouleakis, and Arnab Bhattacharyya. Active causal structure learning with advice. In *International Conference on Machine Learning*, pages 5838–5867. PMLR, 2023.

Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

Hao Dong and Yuedong Wang. Nonparametric neighborhood selection in graphical models. *Journal of Machine Learning Research*, 23(317):1–36, 2022.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

Frederick Eberhardt. Causation and intervention. *Unpublished doctoral dissertation, Carnegie Mellon University*, page 93, 2007.

Frederick Eberhardt, Clark Glymour, and Richard Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among N variables. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pages 178–184, 2005.

Frederick Eberhardt, Clark Glymour, and Richard Scheines. N-1 experiments suffice to determine the causal relations among N variables. In *Innovations in Machine Learning*, pages 97–112. Springer, 2006.

Doris Entner, Patrik Hoyer, and Peter Spirtes. Data-driven covariate selection for nonparametric estimation of causal effects. In *Artificial Intelligence and Statistics*, pages 256–264. PMLR, 2013.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

Lewis Frey, Douglas Fisher, Ioannis Tsamardinos, Constantin F Aliferis, and Alexander Statnikov. Identifying Markov blankets with decision tree induction. In *Third IEEE International Conference on Data Mining*, pages 59–66. IEEE, 2003.

Shunkai Fu and Michel C Desmarais. Fast Markov blanket discovery algorithm via local learning within single pass. In *Advances in Artificial Intelligence: 21st Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2008 Windsor, Canada, May 28-30, 2008 Proceedings 21*, pages 96–107. Springer, 2008.

Ming Gao and Bryon Aragam. Efficient Bayesian network structure learning via local Markov boundary search. *Advances in Neural Information Processing Systems*, 34:4301–4313, 2021.

Ming Gao, Yi Ding, and Bryon Aragam. A polynomial-time algorithm for learning nonparametric causal graphs. *Advances in Neural Information Processing Systems*, 33:11599–11611, 2020.

Ming Gao, Wai Ming Tai, and Bryon Aragam. Optimal estimation of Gaussian DAG models. *arXiv preprint arXiv:2201.10548*, 2022.

Tian Gao and Qiang Ji. Local causal discovery of direct causes and effects. *Advances in Neural Information Processing Systems*, 28, 2015.

Tian Gao and Qiang Ji. Efficient Markov blanket discovery and its application. *IEEE transactions on Cybernetics*, 47(5):1169–1179, 2016.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524, 2019.

Kristjan Greenewald, Dmitriy Katz, Karthikeyan Shanmugam, Sara Magliacane, Murat Kocaoglu, Enric Boix-Adserà, and Guy Bresler. Sample efficient active learning of causal trees. *Advances in Neural Information Processing Systems*, 32, 2019.

Kristjan Greenewald, Karthikeyan Shanmugam, and Dmitriy Katz. High-dimensional feature selection for sample efficient treatment effect estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 2224–2232. PMLR, 2021.

Shantanu Gupta, David Childers, and Zachary Chase Lipton. Local causal discovery for estimating causal effects. In *Conference on Causal Learning and Reasoning*, pages 408–447. PMLR, 2023.

Alain Hauser and Peter 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.

Christina Heinze-Deml, Marloes H Maathuis, and Nicolai Meinshausen. Causal structure learning. *Annual Review of Statistics and Its Application*, 5(1):371–391, 2018.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *The Collected Works of Wassily Hoeffding*, pages 409–426, 1994.

Huining Hu, Zhentao Li, and Adrian Vetta. Randomized experimental design for causal graph discovery. *Advances in Neural Information Processing Systems*, 27, 2014.

Yimin Huang and Marco Valtorta. Pearl's calculus of intervention is complete. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 217–224, 2006.

Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. *Advances in Neural Information Processing Systems*, 33:9551–9561, 2020.

Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30(1):175–193, 1906.

Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Maximum likelihood estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 63(10):6774–6798, 2017.

Markus Kalisch and Peter Bühlman. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In *Conference on Learning Theory*, pages 1066–1100. PMLR, 2015.

Michael Kearns, Yishay Mansour, Dana Ron, Ronitt Rubinfeld, Robert E Schapire, and Linda Sellie. On the learnability of discrete distributions. In *Proceedings of the Twenty-Sixth Annual ACM symposium on Theory of Computing*, pages 273–282, 1994.

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

Daphne Koller and Mehran Sahami. Toward optimal feature selection. Technical report, Stanford InfoLab, 1996.

Wai-yin Lam. Causal razors. *arXiv preprint arXiv:2302.10331*, 2023.

Erik M. Lindgren, Murat Kocaoglu, Alexandros G. Dimakis, and Sriram Vishwanath. Experimental design for cost-aware learning of causal graphs. *Advances in Neural Information Processing Systems*, 31, 2018.

Zhaolong Ling, Kui Yu, Hao Wang, Lei Li, and Xindong Wu. Using feature selection for local causal structure learning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(4):530–540, 2020.

Jacqueline Maasch, Weishen Pan, Shantanu Gupta, Volodymyr Kuleshov, Kyra Gan, and Fei Wang. Local discovery by partitioning: polynomial-time causal discovery around exposure-outcome pairs. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, pages 2350–2382, 2024.

Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright. *Handbook of graphical models*. CRC Press, 2018.

Marloes H Maathuis and Diego Colombo. A generalized back-door criterion. *The Annals of Statistics*, 43(3):1060, 2015.

Marloes H Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, pages 3133–3164, 2009.

Daniel Malinsky. A cautious approach to constraint-based causal model selection. *arXiv preprint arXiv:2404.18232*, 2024.

Subramani Mani and Gregory F Cooper. Causal discovery using a Bayesian local causal discovery algorithm. In *MEDINFO 2004*, pages 731–735. IOS Press, 2004.

Michael Oberst, Fredrik Johansson, Dennis Wei, Tian Gao, Gabriel Brat, David Sontag, and Kush Varshney. Characterization of overlap in observational studies. In *International Conference on Artificial Intelligence and Statistics*, pages 788–798. PMLR, 2020.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3(none):96 – 146, 2009. doi: 10.1214/09-SS057. URL https://doi.org/10.1214/09-SS057.

Jose M Pena, Roland Nilsson, Johan Björkegren, and Jesper Tegnér. Towards scalable and data efficient learning of Markov boundaries. *International Journal of Approximate Reasoning*, 45(2): 211–232, 2007.

Emilija Perkovic. Identifying causal effects in maximally oriented partially directed acyclic graphs. In *Conference on Uncertainty in Artificial Intelligence*, pages 530–539. PMLR, 2020.

Emilija Perković, Johannes Textor, and Markus Kalisch. Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. *Journal of Machine Learning Research*, 18:1–62, 2018.

Joe Ramsey. A PC-style Markov blanket search for high dimensional datasets. Technical report, Carnegie Mellon University, 2006. Technical Report, CMU-PHIL-177.

Thomas Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

Thomas S Richardson, Robin J Evans, James M Robins, and Ilya Shpitser. Nested Markov properties for acyclic directed mixed graphs. *The Annals of Statistics*, 51(1):334–361, 2023.

Philippe Rigollet and Jan-Christian Hütter. High-dimensional statistics. *arXiv preprint arXiv:2310.19244*, 2023.

Andrea Rotnitzky and Ezequiel Smucler. Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *The Journal of Machine Learning Research*, 21 (1):7642–7727, 2020.

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

Mátyás Schubert, Tom Claassen, and Sara Magliacane. Snap: Sequential non-ancestor pruning for targeted causal effect estimation with an unknown graph. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Megan S Schuler and Sherri Rose. Targeted maximum likelihood estimation for causal inference in observational studies. *American Journal of Epidemiology*, 185(1):65–73, 2017.

Jasjeet Sekhon. The Neyman-Rubin model of causal inference and estimation via matching methods. *The Oxford Handbook of Political Methodology*, pages 271–299, 2009.

Abhin Shah, Karthikeyan Shanmugam, and Kartik Ahuja. Finding valid adjustments under non-ignorability with minimal DAG knowledge. In *International Conference on Artificial Intelligence and Statistics*, pages 5538–5562. PMLR, 2022.

Abhin Shah, Karthikeyan Shanmugam, and Murat Kocaoglu. Front-door adjustment beyond Markov equivalence with limited graph knowledge. *Advances in Neural Information Processing Systems*, 2023.

Karthikeyan Shanmugam, Murat Kocaoglu, Alexandros G. Dimakis, and Sriram Vishwanath. Learning causal graphs with small interventions. *Advances in Neural Information Processing Systems*, 28, 2015.

Susan M Shortreed and Ashkan Ertefaie. Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4):1111–1122, 2017.

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1219. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

Ilya Shpitser, Tyler VanderWeele, and James M Robins. On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 527–536, 2010.

Ezequiel Smucler and Andrea Rotnitzky. A note on efficient minimum cost adjustment sets in causal graphical models. *Journal of Causal Inference*, 10(1):174–189, 2022.

Liam Solus, Yuhao Wang, and Caroline Uhler. Consistency guarantees for greedy permutation-based causal inference algorithms. *Biometrika*, 108(4):795–814, 2021.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, Prediction, and Search*. MIT press, 2000.

Jerzy Splawa-Neyman, Dorota M Dabrowska, and Terrence P Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990.

Chandler Squires and Caroline Uhler. Causal structure learning: A combinatorial perspective. *Foundations of Computational Mathematics*, 23(5):1781–1815, 2023.

Chandler Squires, Sara Magliacane, Kristjan Greenewald, Dmitriy Katz, Murat Kocaoglu, and Karthikeyan Shanmugam. Active structure learning of causal DAGs via directed clique trees. *Advances in Neural Information Processing Systems*, 33:21500–21511, 2020.

David Strieder and Mathias Drton. Confidence in causal inference under structure uncertainty in linear causal models with equal variances. *Journal of Causal Inference*, 11(1):20230030, 2023.

David Strieder and Mathias Drton. Dual likelihood for causal inference under structure uncertainty. In *Causal Learning and Reasoning*, pages 1–17. PMLR, 2024.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Eighteenth national conference on Artificial intelligence*, pages 567–573, 2002.

Ioannis Tsamardinos, Constantin F Aliferis, Alexander R Statnikov, and Er Statnikov. Algorithms for large scale Markov blanket discovery. In *FLAIRS*, volume 2, pages 376–81, 2003.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65:31–78, 2006.

Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, pages 436–463, 2013.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Paul Valiant. Testing symmetric properties of distributions. In *Proceedings of the Fortieth Annual ACM symposium on Theory of Computing*, pages 383–392, 2008.

Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

Benito van der Zander, Maciej Liśkiewicz, and Johannes Textor. Constructing separators and adjustment sets in ancestral graphs. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI '14, 2014.

Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like DAGs? A survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36, 2022.

Samir Wadhwa and Roy Dong. On the sample complexity of causal discovery and the value of domain expertise. *arXiv preprint arXiv:2102.03274*, 2021.

Changzhang Wang, You Zhou, Qiang Zhao, and Zhi Geng. Discovering and orienting the edges connected to a target variable in a DAG via a sequential local learning approach. *Computational Statistics & Data Analysis*, 77:252–266, 2014.

Yuhao Wang and Rajen D Shah. Debiased inverse propensity score weighting for estimation of average treatment effects with high-dimensional confounders. *arXiv preprint arXiv:2011.08661*, 2020.

Yuhao Wang, Liam Solus, Karren Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. *Advances in Neural Information Processing Systems*, 30, 2017.

Janine Witte and Vanessa Didelez. Covariate selection strategies for causal inference: Classification and comparison. *Biometrical Journal*, 61(5):1270–1289, 2019.

Steve Yadlowsky, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. *The Annals of Statistics*, 50(5):2587–2615, 2022.

Jianxin Yin, You Zhou, Changzhang Wang, Ping He, Cheng Zheng, and Zhi Geng. Partial orientation and local structural learning of causal networks for prediction. In *Causation and Prediction Challenge*, pages 93–105. PMLR, 2008.

Kui Yu, Xianjie Guo, Lin Liu, Jiuyong Li, Hao Wang, Zhaolong Ling, and Xindong Wu. Causality-based feature selection: Methods and evaluations. *ACM Computing Surveys (CSUR)*, 53(5):1–36, 2020.

Zhenghao Zeng, Sivaraman Balakrishnan, Yanjun Han, and Edward H Kennedy. Causal inference with high-dimensional discrete covariates. *arXiv preprint arXiv:2405.00118*, 2024.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Contents of Appendix

Appendix A. Related work

For sake of clarity, we have positioned this work from the perspective of causal effect estimation, and emphasized how our primary assumption (knowledge of a valid adjustment set $\mathbf{Z}$) is compatible with both the potential outcomes (PO) and graphical frameworks for causality. Now, we further explore connections between our work and existing work from both of these perspectives, beginning with the graphical perspective.

In Appendix A.1, we review graphical characterizations of valid adjustment sets given a causal graph (or an equivalence class of graphs) as input. In some domains, these causal graphs may be constructed from expert knowledge, but when $\mathbf{V}$ is large or the system under consideration is not well-studied, practitioners may be unable to specify an accurate causal graph. Thus, we also review conditions for causal effect estimation which require minimal graphical knowledge. In Appendix A.2, we review a different approach to the unspecified graph setting; in particular, we discuss methods for learning all or part of a causal graph from data, and relate these to our work by providing graphical interpretations of our AMBA and BAMBA methods. Finally, we pivot to the PO perspective in Appendix A.3, focusing on existing results on the statistical aspects of causal effect estimation.

A.1. Causal effect identification in the graphical setting

In the graphical framework, several classes of graphs have been used to formally define causal assumptions about a system, with the nodes of these graphs corresponding to the observed variables $\mathbf{V}$. Here, we focus our discussion on acyclic directed mixed graphs (ADMGs), which consist of both directed edges (of the form $V_1 \to V_2$) and bidirected edges (of the form $V_1 \leftrightarrow V_2$). Such graphs can be used to model systems that are subject latent (a.k.a. unobserved) confounding, but which are not subject to selection bias. As a special case, an ADMG with no bidirected edges is called a directed acyclic graph (DAG), representing a system with no latent confounding. To distinguish between these cases, we use the term *causally sufficient* to refer to settings where latent confounding is assumed to be absent; thus *causally insufficient* refers to settings where latent confounding is allowed to exist. Finally, given an ADMG $\mathcal{G}$, an intervention that sets the variables $\mathbf{X} \subseteq \mathbf{V}$ equal to the values $\mathbf{x}$ can be represented by a new ADMG, the *mutilated graph* $\mathcal{G}_{\overline{\mathbf{X}}}$, which is obtained by copying $\mathcal{G}$ and then removing all edges of the form $V \to X$ or $V \leftrightarrow X$ for an $X \in \mathbf{X}$ and $V \in \mathbf{V}$.

A.1.1. CAUSAL EFFECT IDENTIFICATION GIVEN A GRAPH

The graph $\mathcal{G}$ and $\mathcal{G}_{\overline{\mathbf{X}}}$ can be used to model the behavior of the system, and to derive relationships between the observational distribution $\mathbb{P}(\mathbf{V})$ and the interventional distribution $\mathbb{P}_{\mathbf{x}}(\mathbf{V})$. The details of these definitions are not necessary for our discussion; instead, we describe some of the major results which have been shown when taking these definitions as a starting point. Most importantly for our discussion, these definitions can be used to derive *identification formulas*, which express interventional queries $\mathbb{P}_{\mathbf{x}}(\mathbf{y})$ in terms of equations which only involve $\mathbb{P}(\mathbf{V})$, and thus permit causal effects to be estimated from only observational data. These identification formulas can be derived algorithmically, for example using the *ID Algorithm* (Tian and Pearl, 2002), which is both sound and complete (Shpitser and Pearl, 2006; Huang and Valtorta, 2006). PAC bounds have also been established for the ID algorithm in (Bhattacharyya et al., 2022).

Importantly, the ID Algorithm may be able to construct an identification formula even if the adjustment formula (Eq. (1)) does not hold for any set $\mathbf{A} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$. However, in practice, the adjustment formula remains one of the most widely-used and well-studied identification approaches, due in part to its simplicity and its familiarity in the potential outcomes literature (see Appendix C.1). Particular attention has been given to developing graphical criteria for determining whether a set $\mathbf{A} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$ is a valid adjustment set for $\mathbb{P}_{\mathbf{x}}(\mathbf{y})$, and algorithmically finding such a set if one exists. A simple and intuitive condition for adjustment validity is the *backdoor criterion* (Pearl, 1995) in DAGs, which is sound, but not complete. This criterion has been refined by long line of work on sound and complete conditions (Shpitser et al., 2010; van der Zander et al., 2014; Maathuis and Colombo, 2015; Perković et al., 2018; Perkovic, 2020) for different classes (and equivalence classes) of causal graphs. Our results further contribute to this line of work: as we discuss in Section 1, Lemma 10 directly implies a graphical condition that is sound for determining whether a subset is an adjustment set given a valid adjustment set; Appendix B.4 shows that under additional assumptions, this condition is also complete.

A.1.2. CAUSAL EFFECT IDENTIFICATION WITHOUT A GRAPH

While the criteria above are stated in terms of a known causal graph $\mathcal{G}$, they can also be used in our setting to derive conditions under which Eq. (1) holds, even when the graph is an unknown. Indeed, using Eq. (1) requires quite minimal background knowledge of $\mathcal{G}$, as we now discuss. For simplicity, we limit our discussion to a single treatment variable X. In the case of DAGs, the backdoor criterion implies that $\mathbf{Z} = \text{ND}(X)$ is a valid adjustment set, where $\text{ND}(X)$ denotes the set of non-descendants of X in $\mathcal{G}$. Thus, assuming causal sufficiency, our method can be employed given only knowledge of $\text{ND}(X)$, a quite common setting in applications such as healthcare, where a doctor's treatment assignment X can only depend on pre-treatment patient covariates. Under causal sufficiency and $\mathbf{Z} = \text{ND}(X)$, the Markov blanket of X with respect to $\mathbf{Z}$ is the set $\mathbf{S} = \text{Pa}(X)$, and our AMBA algorithm can be interpreted as searching for the parents of X. Similar results hold in the more general case of ADMGs under an additional assumption on the graph $\mathcal{G}$, as we discuss in Appendix B.5.

In light of these connections, our results fit into a recent line of work establishing identifiability of causal effects with minimal graphical background knowledge. Entner et al. (2013) consider a setting that matches ours in the DAG setting with $\mathbf{Z} = \text{ND}(X)$, and establish a condition similar to Lemma 10 to determine whether $\mathbf{A} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$ is a valid adjustment set. While our condition is sound, their condition is both sound *and* complete, but relies on conditional dependence checks instead of only conditional independence checks. Furthermore, in contrast with our work, where statistical guarantees are a primary focus, their work does not provide any guarantees outside of the oracle setting, though it would be interesting to study their approach in the finite-sample setting.

Follow-up works in this space have extended this problem to the causally insufficient setting by incorporating additional background knowledge on $\mathcal{G}$; all of the works discussed assume knowledge of $\mathbf{Z} = \text{ND}(X)$. For example, Cheng et al. (2022) assumes knowledge of some variable A that is a "cause or spouse of treatment only (COSO)" variable, i.e. that A is adjacent to X but not to Y in $\mathcal{G}$, and establishes a sound condition for determining whether $\mathbf{S} \subseteq \mathbf{Z}$ is an adjustment set. Relatedly, Shah et al. (2022) assumes knowledge of some variable A that is a parent of X and establishes a similar condition. Both conditions are sound, but not complete; in contrast, we show in Appendix B.4 that the BAMBA approach is both sound and complete in the causally sufficient setting when $\mathbf{Z} = \text{ND}(X)$. Finally, Shah et al. (2023) goes beyond using the adjustment formula for identification, in particular studying when background knowledge is sufficient to identify the causal effect using *frontdoor* adjustment.

A.2. Causal graph discovery

This work is strongly motivated by our recognition of the pressing need for better connections between the areas of causal effect estimation and causal structure learning. In a typical *causal discovery* (a.k.a. *causal structure learning*) task, one takes data on the observed variables $\mathbf{V}$ as input, and seeks to return a causal graph $\mathcal{G}$ (or an equivalence class of graphs) that provides an accurate causal model of the system. Traditionally, this goal is (implicitly or explicitly) motivated by the utility of such a model for generating causal predictions, e.g., predicting $\mathbb{P}_{\mathbf{x}}(\mathbf{y})$ as discussed in this work.

A.2.1. Causal discovery and faithfulness

The field of causal discovery is quite well-developed, and has been the subject of several surveys, e.g. see (Heinze-Deml et al., 2018; Glymour et al., 2019; Vowels et al., 2022; Squires and Uhler, 2023). Various approaches address settings such as learning from observational data in the causally sufficient setting (Spirtes et al., 2000; Chickering, 2002; Zheng et al., 2018; Solus et al., 2021) and in the causally insufficient setting (Spirtes et al., 2000; Colombo et al., 2012), as well as learning from interventional data, possibly involving actively chosen interventions (Eberhardt et al., 2005, 2006; Eberhardt, 2007; Hauser and Bühlmann, 2012; Hu et al., 2014; Shanmugam et al., 2015; Wang et al., 2017; Kocaoglu et al., 2017; Lindgren et al., 2018; Greenewald et al., 2019; Jaber et al., 2020; Squires et al., 2020; Choo et al., 2022; Choo and Shiragur, 2023c; Choo et al., 2023; Choo and Shiragur, 2023b,a). Many of these algorithms enjoy theoretical guarantees in the well-specified setting, i.e. under the assumption that the system is correctly described by some (unknown) causal graph $\mathcal{G}^*$. In this setting, an algorithm is called *consistent* if it recovers $\mathcal{G}^*$ (or an appropriate equivalence class) with probability one in the limit of infinite data.

Significant attention has been devoted to finding conditions under which various causal discovery algorithms are consistent. For example, the well-known *faithfulness* assumption requires that if $\mathbf{A}$ and $\mathbf{B}$ and not d-separated by $\mathbf{C}$ in $\mathcal{G}^*$, then $\mathbf{A} \not\perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}$ in $\mathbb{P}(\mathbf{V})$. Although faithfulness is a sufficient condition for the consistency of many causal discovery algorithms, it is often a stronger condition than necessary, and many weaker conditions have been established, see (Lam, 2023) for a recent review and comparison of such conditions. The search for weaker consistency conditions is motivated by a practical issue: although the consistency of an algorithm may depend only on there being no violations of faithfulness, *near* violations of faithfulness (where the conditional independence $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}$ nearly holds, e.g. $\Delta_{\mathbf{A} \perp\!\!\!\perp \mathbf{B} | \mathbf{C}} \leq \varepsilon$ for some small ε) can significantly affect its finite sample properties. Therefore, finite sample guarantees for graph recovery (Kalisch and Bühlman, 2007; Maathuis et al., 2009; Gao et al., 2020; Wadhwa and Dong, 2021; Gao et al., 2022) often depend on assumptions such as *strong* faithfulness, which may be significantly more restrictive in practice (Uhler et al., 2013).

In this work, we avoid making any such assumptions. Indeed, since our goal is causal effect estimation, rather than graph recovery, faithfulness conditions are unnecessary, and existing sample complexity guarantees for causal discovery are pessimistic for our purposes. Within the graphical framework, a main message of our work is that accurate causal effect estimation does not require learning the correct causal graph $\mathcal{G}^*$. For example, if $\mathcal{G}^*$ has “weak” edges, these may be hard to distinguish from missing edges, but those edges are also exactly those that do not significantly impact causal effects; in pragmatic terms, whether an edge is weak or missing is “a difference that doesn’t make a difference”. We provide a concrete example of this phenomenon in Appendix B.6. Nonetheless, such conditions may be useful in improving the sample complexity and/or the computational complexity of our approach, as we discuss in Appendix B.7.

A.2.2. CAUTIOUS APPROACHES AND LOCAL CAUSAL DISCOVERY

To better align theory and practice, a few recent works have focused on new kinds of theoretical guarantees. A number of recent and contemporaneous works (Malinsky, 2024; Chang et al., 2024; Strieder and Drton, 2023, 2024) explicitly consider the interplay between causal discovery and causal effect estimation. As in our work, Malinsky (2024) advocates the use of conditional dependence tests (as opposed to conditional independence tests) to control model misspecification, an approach they call "cautious" causal discovery, where Chang et al. (2024) advocate a bootstrap-style approach. However, their guarantees are for the asymptotic setting, rather than the PAC setting considered in this work, and their approaches aim to recover an entire causal graph, unlike our approach. Similarly, Strieder and Drton (2023) and Strieder and Drton (2024) use asymptotic techniques to derive a valid confidence region over the causal effect when graph structure is unknown.

More closely related to our approach are methods for *local causal discovery*, which aim to recover only part of a causal graph. Indeed, one of the canonical problems in local discovery is Markov blanket recovery (Koller and Sahami, 1996; Frey et al., 2003; Tsamardinos et al., 2003; Ramsey, 2006; Pena et al., 2007; Fu and Desmarais, 2008; Aliferis et al., 2010a,b; Gao and Ji, 2016; Ling et al., 2020; Dong and Wang, 2022), potentially combined with partial edge orientation (Yin et al., 2008; Wang et al., 2014; Gao and Ji, 2015; Gupta et al., 2023) often used in the context of full causal discovery algorithms (Mani and Cooper, 2004; Tsamardinos et al., 2006; Solus et al., 2021; Gao and Aragam, 2021), and sometimes targeted to specific cause-effect pairs (Maasch et al., 2024; Schubert et al., 2025). A number of these algorithms employ greedy search, adding variables to the Markov blanket one at a time, e.g. (Tsamardinos et al., 2003; Fu and Desmarais, 2008; Gao and Ji, 2016). However, greedy search is not guaranteed to return a correct Markov blanket without additional assumptions, such as those in (Gao and Aragam, 2021), in which the authors also provide finite sample guarantees. In contrast, many non-greedy algorithms do enjoy consistency guarantees (i.e. recovery of a correct Markov blanket in the infinite data limit), but thus far lack finite sample guarantees.

Thus, our finite sample guarantees for the (non-greedy) AMBA algorithm contribute to this important line of work, and may be of independent interest beyond the context of causal effect estimation. Furthermore, our BAMBA highlights that using only local structure may be suboptimal for some estimation problems. This fact suggests that we extend from local causal discovery to the more general problem of *targeted* causal discovery, i.e., causal discovery tailored to specific estimation problems, analogous to techniques such as targeted maximum likelihood estimation (Van Der Laan and Rubin, 2006; Schuler and Rose, 2017).

A.3. Causal effect estimation via covariate adjustment

Now, we relate our results to existing statistical results on causal effect estimation, focusing on estimation using the adjustment formula. Existing results are largely written in terms of potential outcomes but, as with our result, are usually applicable as long as Eq. (1) holds and are thus independent of framework choice.[3] In many domains such as healthcare and econometrics, Eq. (1) can

[3] When the random variables are continuous or mixed, Eq. (1) is written as $T_{\mathbf{s},\mathbf{x},\mathbf{y}} = \mathbb{E}_{\mathbf{S}}[\mathbb{P}(\mathbf{Y} = \mathbf{y} \mid \mathbf{X} = \mathbf{x}, \mathbf{S})]$.

be justified by domain knowledge. For example, in healthcare, where $\mathbf{X}$ and $\mathbf{Y}$ may represent medical treatments and patient outcomes, respectively, it is sufficient for $\mathbf{Z}$ to contain all information that doctors may be using to assign treatment, e.g. patient demographic information and past medical history. In such domains, $\mathbf{Z}$ are often referred to as a set of *covariates*; we adopt this terminology here.

As datasets become larger and richer, causal effect estimation is increasingly being applied to problems with high-dimensional covariates. These problems present novel challenges, including violations of the overlap assumption (D'Amour et al., 2021) and the breakdown of traditional asymptotic results. Dimensionality reduction techniques such as feature selection are often crucial to addressing the challenges. However, in the context of treatment effect estimation, naïve usage of feature selection methods such as the Lasso can introduce substantial misspecification bias. Several works aim to address this issue; here, we focus on methods based on feature selection, pointing readers to Yadlowsky et al. (2022) as a starting point for methods using other forms of dimensionality reduction, and to Witte and Didelez (2019) and Yu et al. (2020) for a more complete review and comparison of methods based on feature selection.

Whereas our work focuses on discrete covariates, with no additional assumptions on $\mathbb{P}(\mathbf{Z})$, $\mathbb{P}(\mathbf{X} \mid \mathbf{Z})$ and $\mathbb{P}(\mathbf{Y} \mid \mathbf{X}, \mathbf{Z})$, the majority of prior works consider *continuous* covariates $\mathbf{Z}$, and thus require additional assumptions, such as parametric or smoothness assumptions. When X is a binary treatment, a common assumption is that $\mathbb{P}(X \mid \mathbf{Z})$ follows a logit model, so that $\mathbb{P}(X \mid \mathbf{Z})$ is parameterized by a vector $\boldsymbol{\beta} \in \mathbb{R}^{|\mathbf{Z}|}$. Similarly, when Y is a scalar outcome, a common assumption is that $\mathbb{P}(Y \mid X, \mathbf{Z})$ follows a linear model, i.e. it is parameterized by a vector $\boldsymbol{\gamma} \in \mathbb{R}^{|\mathbf{Z}|}$. Sparsity assumptions may be imposed on one or both of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$; for example, Shortreed and Ertefaie (2017) and Wang and Shah (2020) assume sparsity on $\boldsymbol{\beta}$, Bradic et al. (2019) and Athey et al. (2018) assume sparsity on $\boldsymbol{\gamma}$, and Greenewald et al. (2021) assumes sparsity on both. Other common assumptions include semiparametric restrictions, e.g. partially linear models (Belloni et al., 2014; Chernozhukov et al., 2018), and smoothness assumptions (Farrell et al., 2021).

In these works, sparse regression methods (e.g. Lasso and its variants) play a role similar to our search for a smaller adjustment set $\mathbf{S} \subseteq \mathbf{Z}$, and the choice of regularization parameter plays a role similar to our choice of ε in balancing between misspecification bias and estimation error. In comparison to these methods, our focus on discrete variables obviates the need for additional assumptions, and allows us to establish deeper connections between causal effect estimation and fields such as distribution testing (Canonne, 2020b) and property estimation (Charikar et al., 2019). These connections make the problem accessible to a wider audience and provide access to a broader range of tools: in particular, we note that most of these prior results (e.g. Shortreed and Ertefaie (2017), Athey et al. (2018), Bradic et al. (2019), Wang and Shah (2020), Belloni et al. (2014), Chernozhukov et al. (2018), and Farrell et al. (2021)) are of an asymptotic nature, with Greenewald et al. (2021) being a key exception.

Appendix B. Additional results

B.1. Derivation of expectation bound

Here, we translate the result of Zeng et al. (2024) into our language, showing that $\mathcal{O}\left(\frac{|\mathbf{\Sigma_Z}|}{\lambda \alpha_{\mathbf{Z}}} + \frac{1}{\lambda^2 \alpha_{\mathbf{Z}}}\right)$ samples suffice to obtain an expectation bound of $\mathbb{E}\left(\left|T_{\mathbf{Z},\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{Z},\mathbf{x},\mathbf{y}}\right|\right) \leq \lambda$, for $T_{\mathbf{Z},\mathbf{x},\mathbf{y}}$ defined as in Eq. (1).

Zeng et al. (2024) studies the setting where one is given n i.i.d. copies of (Y, X, A) where $Y \in \{0,1\}$ is the binary outcome, $X \in \{0,1\}$ is the binary treatment, and $A \in [d] = \{1, \dots, d\}$ is a multivariate covariate. Under their positivity assumption (Zeng et al., 2024, Assumption 2), $\mathbb{P}(X = 1 \mid A = k) \in [\varepsilon, 1 - \varepsilon]$ holds for some constant $\varepsilon \in (0, 1/2)$ and any $k \in [d]$. Then, for $\psi_1 = \sum_{k=1}^{d} \mathbb{P}(A = k) \cdot \mathbb{P}(Y = 1 \mid X = 1, A = k)$ and plug-in estimator $\widehat{\psi_1}$, Theorem 1 of Zeng et al. (2024) states that $\mathbb{E}[\psi_1 - \widehat{\psi_1}] \leq \frac{|\mathbf{\Sigma_Z}|^2}{\alpha_{\mathbf{Z}}^2 n^2} + \frac{C}{\alpha_{\mathbf{Z}} n}$ when $\widehat{\psi_1}$ is computed using n i.i.d. samples from $\mathbb{P}(Y, X, A)$, for the worst case distribution $\mathbb{P}(Y, X, A)$ satisfying their positivity assumption.

To adapt their result to our setting, let us define $Y' = \mathbb{1}_{\mathbf{Y}=\mathbf{y}}$, $X' = \mathbb{1}_{\mathbf{X}=\mathbf{x}}$, and A' as a flattened version of $\mathbf{Z}$. Relating (Y', X', A') to their (Y, X, A) setup, we see that $\psi_1 = T_{\mathbf{Z},\mathbf{x},\mathbf{y}}$, $d = |\mathbf{\Sigma_Z}|$, and $\alpha_{\mathbf{Z}} = \varepsilon$. So,

$$\mathbb{E}\left[\left(T_{\mathbf{Z},\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{Z},\mathbf{x},\mathbf{y}}\right)^2\right] \leq \frac{|\mathbf{\Sigma_Z}|^2}{\alpha_{\mathbf{Z}}^2 n^2} + \frac{C}{\alpha_{\mathbf{Z}} n}, \tag{5}$$

for some absolute constant $C > 0$, where we have replaced ε by $\alpha_{\mathbf{Z}}$, d by $|\mathbf{\Sigma_Z}|$, and used that $(1 - \alpha_{\mathbf{Z}})^2 \leq 1$.

To translate this bound into our desired form, we first apply Jensen's inequality Jensen (1906):

$$\begin{aligned}
\left(\mathbb{E}\left[\left|T_{\mathbf{Z},\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{Z},\mathbf{x},\mathbf{y}}\right|\right]\right)^2 &\leq \mathbb{E}\left[\left(\left|T_{\mathbf{Z},\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{Z},\mathbf{x},\mathbf{y}}\right|\right)^2\right] && \text{(Jensen's inequality)} \\
&= \mathbb{E}\left[\left(T_{\mathbf{Z},\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{Z},\mathbf{x},\mathbf{y}}\right)^2\right] \\
&\leq 2 \max\left(\frac{|\mathbf{\Sigma_Z}|^2}{\alpha_{\mathbf{Z}}^2 n^2}, \frac{C}{\alpha_{\mathbf{Z}} n}\right) && \text{(By Eq. (5))}
\end{aligned}$$

Thus, to obtain that $\mathbb{E}\left(\left|T_{\mathbf{Z},\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{Z},\mathbf{x},\mathbf{y}}\right|\right) \leq \lambda$, it suffices to have $2 \max\left(\frac{|\mathbf{\Sigma_Z}|^2}{\alpha_{\mathbf{Z}}^2 n^2}, \frac{C}{\alpha_{\mathbf{Z}} n}\right) \leq \lambda^2$. Then, solving for n yields $n \in \mathcal{O}\left(\frac{|\mathbf{\Sigma_Z}|}{\lambda \alpha_{\mathbf{Z}}} + \frac{1}{\lambda^2 \alpha_{\mathbf{Z}}}\right)$ as stated.

B.2. Derivation for conditional sub-Gaussian

For completeness, we present the following proof of Lemma 15.
Proof By iterated expectation,

$$\begin{aligned}
\mathbb{E}\left(e^{\lambda X}\right) &= \mathbb{E}\left(\mathbb{E}\left(e^{\lambda X} \mid Y\right)\right) \\
&\leq \mathbb{E}\left(\exp\left(\frac{\lambda^2 \sigma_Y^2}{2}\right)\right) \\
&\leq \mathbb{E}\left(\exp\left(\frac{\lambda^2 \max_{y \in \Sigma_Y} \sigma_y^2}{2}\right)\right) \\
&\leq \exp\left(\frac{\lambda^2 \max_{y \in \Sigma_Y} \sigma_y^2}{2}\right),
\end{aligned}$$

i.e., $X \in \text{subG}(\max_{y \in \Sigma_Y} \sigma_y^2)$, as desired. ■

B.3. Additional results in the graphical framework

In Appendix A, we described a special case of our setting in the graphical framework. In particular, assuming that $\mathcal{G}$ is a DAG and considering only a single treatment variable X, it is easy to see that $\mathbf{Z} = \mathrm{ND}(X)$ is a valid adjustment set, and that $\mathbf{S} = \mathrm{Pa}(X)$ is a Markov blanket of X with respect to $\mathbf{Z}$. We now discuss two other applications of our results. First, in Appendix B.4, we show that BAMBA is complete in this special case, i.e. the two conditional independences in Lemma 10 are not just sufficient to ensure that $\mathbf{S}'$ is an adjustment set, they are also *necessary*. Second, in Appendix B.5 we introduce an assumption under which we can extend the special case of $\mathbf{Z} = \mathrm{ND}(X)$ from DAGs to ADMGs, and describe the Markov blanket blanket of X with respect to $\mathbf{Z}$.

We assume that the interested reader is familiar with definitions related to d-separation (in DAGs) and m-separation (in ADMGs); e.g. concepts such as colliders, active paths, and blocked paths; see e.g. Richardson (2003) for a detailed overview. In particular, we will make use of the following Markov property for DAGs:

Definition 19 *A probability distribution $\mathbb{P}(\mathbf{V})$ is said to be* Markov *with respect to a DAG $\mathcal{G}$ if, whenever $\mathbf{A}$ and $\mathbf{B}$ are d-separated by $\mathbf{C}$ in $\mathcal{G}$, then $\mathbf{A}$ is conditionally independent from $\mathbf{B}$ given $\mathbf{C}$ in $\mathbb{P}(\mathbf{V})$.*

B.4. Completeness of BAMBA for a special case

In this section, we show that BAMBA is not just sound (Lemma 10) but also complete in a special setting. In particular, Lemma 20 implies that searching for a minimal sized $\mathbf{S}' \subseteq \mathbf{Z}$ that satisfies both $\mathbf{Y} \perp\!\!\!\perp \mathbf{S} \setminus \mathbf{S}' \mid \mathbf{X} \cup \mathbf{S}'$ and $\mathbf{X} \perp\!\!\!\perp \mathbf{S}' \setminus \mathbf{S} \mid \mathbf{S}$ necessarily produces a minimal sized adjustment set.

Lemma 20 *Consider the graphical causal framework in the causally sufficient setting, where variables in $\mathbf{X}$ are non-ancestors of each other. Let $\mathbf{Z} = \mathrm{ND}(\mathbf{X}) \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$ be the set of non-descendants of $\mathbf{X}$ and $\mathbf{S} = \mathrm{Pa}(\mathbf{X}) = \bigcup_{X \in \mathbf{X}} \mathrm{Pa}(X) \subseteq \mathrm{ND}(\mathbf{X}) = \mathbf{Z}$ are the parents of $\mathbf{X}$. Then, any subset $\mathbf{S}' \subseteq \mathrm{ND}(\mathbf{X}) = \mathbf{Z}$ such that $T_{\mathbf{S}',\mathbf{x},\mathbf{y}} = T_{\mathbf{S},\mathbf{x},\mathbf{y}}$ must satisfy both (i) $\mathbf{Y} \perp\!\!\!\perp \mathbf{S} \setminus \mathbf{S}' \mid \mathbf{X} \cup \mathbf{S}'$ and (ii) $\mathbf{X} \perp\!\!\!\perp \mathbf{S}' \setminus \mathbf{S} \mid \mathbf{S}$.*

Proof We know that $\mathbf{S} = \mathrm{Pa}(\mathbf{X})$ is a valid adjustment set and so it must block any non-causal paths between $\mathbf{X}$ and $\mathbf{Y}$ (Perković et al., 2018). Then, since $T_{\mathbf{S}',\mathbf{x},\mathbf{y}} = T_{\mathbf{S},\mathbf{x},\mathbf{y}}$, it must be the case that $\mathbf{S}'$ is also a valid adjustment set.

Condition (i) $\quad : \mathbf{Y} \perp\!\!\!\perp \mathbf{S} \setminus \mathbf{S}' \mid \mathbf{X} \cup \mathbf{S}'$

Suppose, for a contradiction, that $\mathbf{Y} \not\perp\!\!\!\perp \mathbf{S} \setminus \mathbf{S}' \mid \mathbf{X} \cup \mathbf{S}'$. By contrapositive of the Markov property (Theorem 19), there is an active d-connected path in $\mathcal{G}$ from some $Y \in \mathbf{Y}$ to some $A \in \mathbf{S} \setminus \mathbf{S}'$, when $\mathbf{X} \cup \mathbf{S}'$ is conditioned upon. Let $\mathbf{P}_{Y,A}$ denote such an active path of minimal length. By minimality of $\mathbf{P}_{Y,A}$, there are no internal vertices from $\mathbf{S} \setminus \mathbf{S}'$ within the path $\mathbf{P}_{Y,A}$. We will argue that such a path $\mathbf{P}_{Y,A}$ *cannot* exist by considering the two cases of whether the path $\mathbf{P}_{Y,A}$ contains some vertex from $\mathbf{X}$ internally.

Case 1: Suppose $\mathbf{P}_{Y,A}$ contains some vertex from $\mathbf{X}$, i.e. $\mathbf{V}(\mathbf{P}_{Y,A}) \cap \mathbf{X} \neq \emptyset$. Let $X \in \mathbf{X}$ be the vertex in $\mathbf{V}(\mathbf{P}_{Y,A}) \cap \mathbf{X}$ that is closest to Y, i.e. there are no other vertices between X and Y along the path $\mathbf{P}_{Y,A}$. Let $\mathbf{Q}_{Y,X}$ denote this subpath of $\mathbf{P}_{Y,A}$. Since $\mathbf{P}_{Y,A}$ is active with respect to $\mathbf{X} \cup \mathbf{S}'$, X must appear as a collider on $\mathbf{P}_{Y,X}$. That is, $\mathbf{Q}_{Y,X}$ is a non-causal path from X to Y that does not contain any internal $\mathbf{X}$ vertices.

Case 2: Suppose $\mathbf{P}_{Y,A}$ does *not* contain any vertex from $\mathbf{X}$, i.e. $\mathbf{V}(\mathbf{P}_{Y,A}) \cap \mathbf{X} = \emptyset$. Since $A \in \mathbf{S} \setminus \mathbf{S}' \subseteq \mathbf{S} = \text{Pa}(\mathbf{X})$, there must be an edge $A \to X$ for some $X \in \mathbf{X}$. Therefore, the extended path $\mathbf{Q}_{Y,X} = \mathbf{P}_{Y,A} \cup \{A \to X\}$ is a non-causal path from X to Y that does not contain any internal $\mathbf{X}$ vertices.

In either case, we have some non-causal path from X to Y that does not contain any internal $\mathbf{X}$ vertices denoted by $\mathbf{Q}_{Y,X}$. Since $\mathbf{S}'$ is a valid adjustment set, $\mathbf{S}'$ must block $\mathbf{Q}_{Y,X}$, which implies that $\mathbf{P}_{Y,A}$ will be blocked by $\mathbf{X} \cup \mathbf{S}'$. This is a contradiction to the existence of such an active path $\mathbf{P}_{Y,A}$ in the first place.

Condition (ii) : $\mathbf{X} \perp\!\!\!\perp \mathbf{S}' \setminus \mathbf{S} \mid \mathbf{S}$

Suppose, for a contradiction, that $\mathbf{X} \not\perp\!\!\!\perp \mathbf{S}' \setminus \mathbf{S} \mid \mathbf{S}$. By contrapositive of the Markov property (Theorem 19), there is an active d-connected path from some $X \in \mathbf{X}$ to some $B \in \mathbf{S}' \setminus \mathbf{S}$, when $\mathbf{S}$ is being conditioned upon. Let $\mathbf{P}_{X,B}$ denote such an active path. Note that $\mathbf{P}_{X,B}$ *cannot* begin with an incoming edge into X. This is because otherwise $\mathbf{P}_{X,B}$ has the form $X \leftarrow C - \ldots$ for some $C \in \text{Pa}(\mathbf{X}) = \mathbf{S}$ and so would be not be active when $\mathbf{S}$ is being conditioned upon. So, it must be the case that $\mathbf{P}_{X,B}$ begins with an outgoing edge from X. Then, there must be a collider on $\mathbf{P}_{X,B}$ involving a descendant of X because $B \in \mathbf{S}' \setminus \mathbf{S} \subseteq \text{ND}(\mathbf{X})$. However, the conditioning set $\mathbf{S} \subseteq \text{ND}(\mathbf{X})$ would not include this descendant, so $\mathbf{P}_{X,B}$ would not be active. This contradicts the existence of such an active path $\mathbf{P}_{X,B}$ in the first place. ■

B.5. Valid adjustment by non-descendants in ADMGs

Consider an ADMG with single treatment variable X and single outcome variable Y, and let $\mathbf{Z} = \text{ND}(X)$. Unfortunately, $\mathbf{Z}$ is not necessarily a valid adjustment set for $\mathbb{P}_{\mathbf{x}}(\mathbf{y})$, see Fig. 2 for a counterexample. To ensure that $\mathbf{Z}$ is a valid adjustment set for $\mathbb{P}_{\mathbf{x}}(\mathbf{y})$, we must introduce an additional assumption on the causal graph $\mathcal{G}$. In particular, let $\text{Dis}(X)$ denote the *district* (a.k.a. *c-component*) of X, i.e. the set of all nodes in $\mathcal{G}$ that are connected to X by only bidirected edges. Following Definition 17 of Richardson et al. (2023), we say that X is *fixable* if $\text{Dis}(X) \cap \text{De}(X) = \{X\}$, i.e. if the district of X does not contain any of its descendants. Further, we say that a path from X to Y is *causal* if it is of the form $X \to \ldots \to Y$, i.e. if it contains only directed edges pointing away from X; otherwise, we call a path *non-causal*. Then, we have the following:

Assumption 21 *Assume that X is fixable in $\mathcal{G}$ (i.e.* $\text{Dis}(X) \cap \text{De}(X) = \varnothing$*) and* $Y \notin \text{Dis}(X)$.

Lemma 22 *Under Assumption 21,* $\mathbf{Z} = \text{ND}(X)$ *is a valid adjustment set for* $\mathbb{P}_{\mathbf{x}}(\mathbf{y})$.

Proof Shpitser et al. (2010) showed that a set $\mathbf{Z}$ is a valid adjustment set for $\mathbb{P}_{\mathbf{x}}(\mathbf{y})$ if it satisfies the following *adjustment criterion*:

(i) No $Z \in \mathbf{Z}$ is a descendant in $\mathcal{G}_{\overline{\mathbf{X}}}$ of any W on a causal path from X to Y.

(ii) All non-causal paths from X to Y are blocked by $\mathbf{Z}$.

Any Z violating Condition (i) must be a descendant of X; thus, Condition (i) is immediately satisfied for $\mathbf{Z} = \mathrm{ND}(X)$.

To see (ii), for any non-causal path from X to Y, let A be the node closest to X on that path which is not a collider; such a node must exist since we assume that $Y \notin \mathrm{Dis}(X)$. Then either $A \in \mathrm{Dis}(X)$ or $A \in \mathrm{Pa}(\mathrm{Dis}(X))$. Since X is fixable, both of these cases imply that $A \in \mathrm{ND}(X)$. Thus, any non-causal path is blocked by $\mathbf{Z} = \mathrm{ND}(X)$. ■

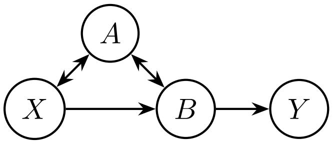


Figure 2: An ADMG in which $\mathrm{ND}(X) = \{A\}$ is not a valid adjustment set for $\mathbb{P}_{\mathbf{x}}(\mathbf{y})$. In particular, when conditioning on A, there is an m-connecting non-causal path $X \leftrightarrow A \leftrightarrow B \to Y$. Note that $\mathrm{Dis}(X) \cap \mathrm{De}(X) = \{B\}$, i.e., X is not a fixable node and thus does not satisfy Assumption 21.

Note that this result can also be obtained more directly using the conditional ADMG framework of Richardson et al. (2023); we have chosen to give this slightly longer proof to minimize the required background.

This result lets us extend the first part of our interpretation from DAGs to ADMGs: under Assumption 21, $\mathbf{Z} = \mathrm{ND}(X)$ is a valid adjustment set for $\mathbb{P}_{\mathbf{x}}(\mathbf{y})$, and hence a valid starting point for our algorithm. To generalize the second part of our interpretation, we note that $\mathrm{ND}(X)$ is an *ancestral set*, i.e. for any $Z \in \mathrm{ND}(\mathbf{X})$, $\mathrm{An}(Z) \subseteq \mathrm{ND}(X)$. From this fact, we rather directly have the following:

Lemma 23 *Let* $\mathbf{Z} = \mathrm{ND}(X)$ *and* $\mathbf{S} = \mathrm{Pa}(\mathrm{Dis}(X)) \cup \mathrm{Dis}(X) \setminus \{X\}$. *Under Assumption 21,* $\mathbf{S} \subseteq \mathbf{Z}$ *and* $\mathbf{S}$ *is a Markov blanket of* X *with respect to* $\mathbf{Z}$.

Proof The fact that $\mathbf{S} \subseteq \mathbf{Z}$ follows from directly from the fixability condition. Applying the *ordered local Markov property* (Richardson, 2003, Section 3) to the ancestral set $\mathrm{ND}(X)$, we obtain that $\mathbf{S}$ is a Markov blanket of X with respect to $\mathbf{Z}$. ■

B.6. Weak edges

In this section, we describe a simple concrete example whereby it is suboptimal to first learn a correct causal graph and then apply identifiability formulas to estimate $\mathbb{P}_{\mathbf{x}}(\mathbf{y})$. In particular, correctly learning the causal graph $\mathcal{G}^*$ may require taking a large number of samples, especially in the presence of "weak edges". However, one would expect such edges to contribute little to $\mathbb{P}_{\mathbf{x}}(\mathbf{y})$.

Suppose we have a probability distribution $\mathbb{P}$ on variables $\{X, Y, Z\}$ generated as follows:

$$\begin{aligned} Z &\leftarrow \mathrm{Bern}(1/2) \\ X &\leftarrow \begin{cases} Z & \text{with probability } \varepsilon > 0 \\ \mathrm{Bern}(1/2) & \text{with probability } 1 - \varepsilon \end{cases} \end{aligned}$$

$$Y \leftarrow X \oplus Z$$

The causal graph that exactly captures $\mathbb{P}$ is a complete DAG with edges $Z \to X \to Y$ and $Z \to Y$; see $\mathcal{G}_1$ in Fig. 3. However, for extremely small ε, one would require $\Omega(1/\varepsilon)$ samples to detect a dependency between X and Z. So, with small ε and insufficient samples, one may erroneously recover a subgraph without the $Z \to X$ arc; see $\mathcal{G}_2$ in Fig. 3.

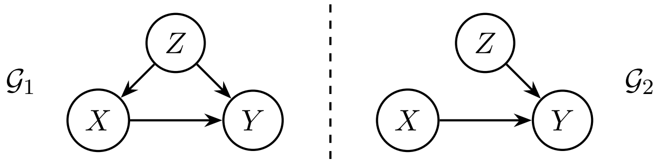


Figure 3: While it is hard to distinguish $\mathcal{G}_1$ from $\mathcal{G}_2$ for small ε with few samples from $\mathbb{P}$, estimating $\mathbb{P}_x(y)$ using $\mathcal{G}_2$ only incurs an additive error of $O(\varepsilon)$.

Now, suppose we are interested in estimating $\mathbb{P}_0(1) = \mathbb{P}(Y = 1 \mid \mathtt{do}(X = 0))$ from observational data. One can check that the correct answer is $\mathbb{P}(Y = 1 \mid \mathtt{do}(X = 0)) = 1/2$. Applying standard adjustment formulas under $\mathcal{G}_1$ yield $\mathbb{P}(Y = 1 \mid \mathtt{do}(X = 0)) = \sum_{z \in \{0,1\}} \mathbb{P}(Z = z) \cdot \mathbb{P}(Y = 1 \mid X = 0, Z = z) = 1/2$ as expected. Meanwhile, under $\mathcal{G}_2$, the estimation would simply by $\mathbb{P}(Y = 1 \mid X = 0) = (1 - \varepsilon)/2 = 1/2 - \varepsilon/2$. Thus, see that the estimation error is only an additive $O(\varepsilon)$ factor away from the ground truth.

B.7. Stronger results under causal faithfulness

Recall from AMBA (Algorithm 1) that we need to perform conditional independence checks of the form $\mathbf{X} \perp\!\!\!\perp_\varepsilon \mathrm{ND}(\mathbf{X}) \setminus \mathbf{S} \mid \mathbf{S}$, which could potentially involve up to $|\mathbf{V}|$ variables. Furthermore, we also know from Theorem 4 that the required sample complexity of conditional independence testing typically increases as the total number of variables involved increases. Thus, it would be preferable if we just check whether $\mathbf{X} \perp\!\!\!\perp_\varepsilon V \mid \mathbf{S}$ for each $V \in \mathrm{ND}(\mathbf{X}) \setminus \mathbf{S}$, and derive that $\mathbf{X} \perp\!\!\!\perp \mathrm{ND}(\mathbf{X}) \mid \mathbf{S}$. When $\varepsilon = 0$, this implication is a form *compositionality*, and is well-known to hold under the faithfulness assumption (since the set of d-separation statements in a graph is a *graphoid*, see e.g. (Maathuis et al., 2018, Chapter 1)), we provide an elementary proof below.

Lemma 24 *Let $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ be disjoint subsets of variables. Under the causal faithfulness assumption, if $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}$ and $\mathbf{A} \perp\!\!\!\perp \mathbf{D} \mid \mathbf{C}$, then $\mathbf{A} \perp\!\!\!\perp (\mathbf{B} \cup \mathbf{D}) \mid \mathbf{C}$.*

Proof Suppose, for a contradiction, that $\mathbf{A} \not\!\perp\!\!\!\perp (\mathbf{B} \cup \mathbf{D}) \mid \mathbf{C}$. Under the causal faithfulness assumption, this means that there is a d-connected path P from some $A \in \mathbf{A}$ to some $V \in \mathbf{B} \cup \mathbf{D}$ that is active with respect to $\mathbf{C}$. Without loss of generality, due to symmetry of the statement, suppose that $V \in \mathbf{B}$. That is, P is a path from $A \in \mathbf{A}$ to some $V \in \mathbf{V}$ that is active with respect to $\mathbf{C}$. But such an active path P contradicts the assumption that $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}$. Contradiction. ■

Note that Lemma 24 is *false* in general with respect to unfaithful distributions.

Example 1 *The simple 3-variable distribution $X = Z_1 \oplus Z_2$, where Z_1 and Z_2 are independent fair coin flips is unfaithful to any DAG on 3 nodes. To see why, observe that any two variables* *are unconditionally independent but completely dependent upon conditioning on the third. So, one would minimally have to use a v-structure, say $Z_1 \to X \leftarrow X_2$ to represent this. However, $Z_1 \to X$ is an active path which implies $Z_1 \not\perp\!\!\!\perp X$ under the causal faithfulness assumption, which is not true in $\mathbb{P}(Z_1, Z_2, X)$.*

Unfortunately, faithfulness alone is not sufficient to ensure the desired implication. As we demonstrate in the following example (a minor adaptation of the above example), a distribution $\mathbb{P}(\mathbf{V})$ may be faithful to a DAG, but fail to satisfy the desired compositionality-style property.

Lemma 25 *Let $0 < \varepsilon \leq 1/2$. Consider a probability distribution $\mathbb{P}$ over three binary variables (A, B, X) where $A \in \text{Bern}(1/2)$ and $B \in \text{Bern}(1/2)$ are two independent Bernoulli random variables, each with success probability $1/2$ and X is defined as follows:*

$$X = \begin{cases} A \oplus B & \textit{with probability } 1 - 2\varepsilon \\ A & \textit{with probability } \varepsilon \\ B & \textit{with probability } \varepsilon \end{cases}$$

We have $\Delta_{X \perp\!\!\!\perp A | \varnothing} = \Delta_{X \perp\!\!\!\perp B | \varnothing} = \varepsilon$ and $\Delta_{X \perp\!\!\!\perp (A,B) | \varnothing} = \frac{1}{2} - \varepsilon$.

Proof By construction, $\mathbb{P}(A = 0) = \mathbb{P}(B = 0) = \mathbb{P}(X = 0) = 1/2$. Meanwhile, one can check that $\mathbb{P}(X = 0, A = 0) = \mathbb{P}(X = 1, A = 1) = \frac{1}{4} + \frac{\varepsilon}{2}$ and $\mathbb{P}(X = 0, A = 1) = \mathbb{P}(X = 1, A = 0) = \frac{1}{4}$. For instance, $\mathbb{P}(X = 0, A = 0) = \mathbb{P}(X = 0 \mid A = 0) \cdot \mathbb{P}(A = 0) = \left((1 - \varepsilon) \cdot \frac{1}{2} + \varepsilon + \varepsilon \cdot \frac{1}{2}\right) \cdot \frac{1}{2} = \frac{1}{4} + \frac{\varepsilon}{2}$. So, $\sum_{x,a \in \{0,1\}} |\mathbb{P}(x, a) - \mathbb{P}(x) \cdot \mathbb{P}(a)| = \varepsilon$. By Definition 2, this establishes $\Delta_{X \perp\!\!\!\perp A | \varnothing} = \varepsilon$.

The analysis of $\Delta_{X \perp\!\!\!\perp B | \varnothing} = \varepsilon$ is symmetric by replacing the role of A by B in the above analysis.

Since A and B are independent Bernoulli random variables, we see that $\mathbb{P}(A = a, B = b) = \mathbb{P}(A = a) \cdot \mathbb{P}(B = b) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$ for any $a, b \in \{0, 1\}$. Meanwhile,

$$\begin{aligned} \mathbb{P}(X = 0 \mid A = 0, B = 0) &= 1 \\ \mathbb{P}(X = 0 \mid A = 0, B = 1) &= \varepsilon \\ \mathbb{P}(X = 0 \mid A = 1, B = 0) &= \varepsilon \\ \mathbb{P}(X = 0 \mid A = 1, B = 1) &= 1 - 2\varepsilon \end{aligned}$$

So,

$$\begin{aligned} \sum_{x,a,b \in \{0,1\}} |\mathbb{P}(x, a, b) - \mathbb{P}(x) \cdot \mathbb{P}(a, b)| &= \sum_{x,a,b \in \{0,1\}} \mathbb{P}(a, b) \cdot |\mathbb{P}(x \mid a, b) - \mathbb{P}(x)| \\ &= \frac{1}{4} \cdot \sum_{x,a,b \in \{0,1\}} \left| \mathbb{P}(x \mid a, b) - \frac{1}{2} \right| \\ &\qquad \text{(Since } \mathbb{P}(a, b) = \tfrac{1}{4} \text{ and } \mathbb{P}(x) = \tfrac{1}{2} \text{ always)} \\ &= \frac{1}{4} \cdot \left(\left|1 - \frac{1}{2}\right| + \left|\varepsilon - \frac{1}{2}\right| + \left|\varepsilon - \frac{1}{2}\right| + \left|1 - 2\varepsilon - \frac{1}{2}\right| \right) \\ &\qquad \text{(From above)} \\ &= \frac{1}{4} \cdot (2 - 4\varepsilon) \qquad \text{(Since } \varepsilon \leq \tfrac{1}{2}) \end{aligned}$$

$$= \frac{1}{2} - \varepsilon$$

By Definition 2, this establishes $\Delta_{X \perp\!\!\!\perp (A,B) \mid \varnothing} = \frac{1}{2} - \varepsilon$. ■

The above example demonstrates that the faithfulness assumption is insufficient for our purposes. Instead, we need an assumption of the following form; as we will show, this assumption is implied by a type of *strong faithfulness* assumption.

Definition 26 *We say that* $\mathbb{P}(\mathbf{V})$ *obeys* (ε, γ)-strong compositionality *if, for any disjoint sets* $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D} \subseteq \mathbf{V}$*, the following is true:*

$$(\mathbf{A} \perp\!\!\!\perp_{\varepsilon} \mathbf{B} \mid \mathbf{C}) \wedge (\mathbf{A} \perp\!\!\!\perp_{\varepsilon} \mathbf{D} \mid \mathbf{C}) \implies (\mathbf{A} \perp\!\!\!\perp_{\gamma\varepsilon} \mathbf{B} \cup \mathbf{D} \mid \mathbf{C})$$

Under (ε, γ)-strong compositionality, we can derive that $\mathbf{X} \perp\!\!\!\perp_{\varepsilon} \mathbf{Z} \setminus \mathbf{S} \mid \mathbf{S}$ from smaller conditional independence tests; in particular, using a bisection arguments, if $\mathbf{X} \perp\!\!\!\perp_{\varepsilon} V \mid \mathbf{S}$ for all $V \in \mathbf{Z} \setminus \mathbf{S}$, then $\mathbf{X} \perp\!\!\!\perp_{\gamma^k \varepsilon} \mathbf{Z} \setminus \mathbf{S} \mid \mathbf{S}$ for $k = \lceil \log_2(|\mathbf{Z} \setminus \mathbf{S}|) \rceil$. Finally, we relate can strong compositionality to the faithfulness assumption: strong faithfulness implies strong compositionality with $\gamma = 0$, as follows.

Assumption 27 (TV Strong faithfulness) *If* $\mathbf{A}$ *is d-connected to* $\mathbf{B}$ *given* $\mathbf{C}$ *in* $\mathcal{G}^*$*, then*

$$\Delta_{\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}} > \beta$$

Equivalently, $\Delta_{\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}} \leq \beta \implies \mathbf{A}$ *is d-separated from* $\mathbf{B}$ *given* $\mathbf{C}$*.*

Lemma 28 (TV strong faithfulness implies strong compositionality) *Suppose* $\mathbb{P}(\mathbf{V})$ *is* β*-TV strong faithful to* $\mathcal{G}^*$*. Then* $\mathbb{P}(\mathbf{V})$ *is* $(\beta, 0)$*-compositional.*

Proof Suppose $\mathbf{A} \perp\!\!\!\perp_{\beta} \mathbf{B} \mid \mathbf{C}$ and $\mathbf{A} \perp\!\!\!\perp_{\beta} \mathbf{D} \mid \mathbf{C}$. Then, by β-TV strong faithfulness, $\mathbf{A}$ is is d-separated from $\mathbf{B}$ given $\mathbf{C}$, and $\mathbf{A}$ is d-separated from $\mathbf{D}$ given $\mathbf{C}$. Thus, $\mathbf{A}$ is d-separated from $\mathbf{B} \cup \mathbf{D} \mid \mathbf{C}$, so $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \cup \mathbf{D} \mid \mathbf{C}$. ■

Appendix C. Deferred proof details

C.1. Covariate adjustment in the potential outcomes framework

For simplicity, we describe the potential outcomes framework in the i.i.d. setting, i.e., we assume that n samples are drawn independently from a distribution $\mathbb{P}(\mathbf{V})$, though we note that the following result can be extended to weaker settings (e.g. when samples are exchangeable but not necessarily independent).

In the PO framework, the treatment variables $\mathbf{X}$ are considered to be given, along with a set $\mathbf{\Sigma_X}$ of possible values for $\mathbf{X}$. Given $\mathbf{X}$ and $\mathbf{\Sigma_X}$, one takes as their starting point an indexed set of random variables $\{\mathbf{Y}(\mathbf{x})\}_{\mathbf{x} \in \mathbf{\Sigma_X}}$, with $\mathbf{Y}(\mathbf{x})$ denoting the potential outcome associated with intervening to set $\mathbf{X}$ equal to $\mathbf{x}$. Then, the *factual* outcome $\mathbf{Y}$ is generated according to $\mathbf{X}$ and the potential outcomes; typically, one assumes *consistency*, i.e., that if $\mathbf{X} = \mathbf{x}$, then $\mathbf{Y} = \mathbf{Y}(\mathbf{x})$. Hence, under the PO framework, we have $\mathbb{P}_{\mathbf{x}}(\mathbf{y}) = \mathbb{P}(\mathbf{Y}(\mathbf{x}) = \mathbf{y})$ is the probability that $\mathbf{Y}$ takes on value $\mathbf{y}$ if $\mathbf{X}$ is set to $\mathbf{x}$.

Now, Eq. (1) can be derived as a consequence of consistency and an additional assumption about conditional independences. In particular, $\mathbf{X}$ is called *conditionally ignorable* with respect to $\mathbf{Z}$ if $\mathbf{Y}(\mathbf{x}) \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}$ for all $\mathbf{x} \in \Sigma_{\mathbf{X}}$.

Lemma 29 *Under consistency and conditional ignorability of $\mathbf{X}$ with respect to $\mathbf{Z}$, we have*

$$\mathbb{P}(\mathbf{Y}(\mathbf{x}) = \mathbf{y}) = \sum_{\mathbf{z} \in \mathbf{Z}} \mathbb{P}(\mathbf{Y} = \mathbf{y} \mid \mathbf{Z} = \mathbf{z}, \mathbf{X} = \mathbf{x}) \cdot \mathbb{P}(\mathbf{Z} = \mathbf{z})$$

Proof

$$\begin{aligned}
\mathbb{P}(\mathbf{Y}(\mathbf{x}) = y) &= \sum_{\mathbf{z} \in \mathbf{Z}} \mathbb{P}(\mathbf{Y}(\mathbf{x}) = \mathbf{y} \mid \mathbf{Z} = \mathbf{z}) \cdot \mathbb{P}(\mathbf{Z} = \mathbf{z}) && \text{(Law of total probability)} \\
&= \sum_{\mathbf{z} \in \mathbf{Z}} \mathbb{P}(\mathbf{Y}(\mathbf{x}) = \mathbf{y} \mid \mathbf{Z} = \mathbf{z}, \mathbf{X} = \mathbf{x}) \cdot \mathbb{P}(\mathbf{Z} = \mathbf{z}) && \text{(Since } \mathbf{Y}(\mathbf{x}) \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}\text{)} \\
&= \sum_{\mathbf{z} \in \mathbf{Z}} \mathbb{P}(\mathbf{Y} = \mathbf{y} \mid \mathbf{Z} = \mathbf{z}, \mathbf{X} = \mathbf{x}) \cdot \mathbb{P}(\mathbf{Z} = \mathbf{z}) && \text{(By consistency)}
\end{aligned}$$

■

C.2. Sample complexity for empirical estimation

The proof of Theorem 1 relies on the following lemma.

Lemma 30 *Suppose we have i.i.d. sample access to $\mathbb{P}(\mathbf{V})$. Given integer $n > 0$ as a sampling parameter, we take $N_{\text{Pois}} \sim \text{Pois}(n)$ samples. For any $\mathbf{U} \subseteq \mathbf{V}$, let the random variable $N_{\mathbf{u}}$ denote the number of times $\mathbf{u} \in \mathbf{\Sigma_U}$ was realized within the n_{Pois} samples. Then, the following statements hold:*

1. *Let $\mathbf{A}, \mathbf{B} \subseteq \mathbf{V}$ be disjoint sets of variables. For any $\mathbf{a}, \mathbf{a}' \in \mathbf{\Sigma_A}$ and $\mathbf{b}, \mathbf{b}' \in \mathbf{\Sigma_B}$ with $\mathbf{b} \neq \mathbf{b}'$, the ratios of random variables $\frac{N_{\mathbf{a},\mathbf{b}}}{N_{\mathbf{b}}}$ and $\frac{N_{\mathbf{a}',\mathbf{b}'}}{N_{\mathbf{b}'}}$ are independent.*

2. *Let $\mathbf{A}, \mathbf{B} \subseteq \mathbf{V}$ be disjoint sets of variables. For any $\mathbf{a} \in \mathbf{\Sigma_A}$, $\mathbf{b} \in \mathbf{\Sigma_B}$, and integer $k \geq 1$, we have $\left(\frac{N_{\mathbf{a},\mathbf{b}}}{N_{\mathbf{b}}} - \mathbb{P}(\mathbf{a} \mid \mathbf{b}) \mid N_{\mathbf{b}} \geq k\right) \sim \text{subG}\left(\frac{1}{4k}\right)$.*

Proof We prove each item one at a time.

1. By Lemma 11, the random variables $N_{\mathbf{b}}$ and $N_{\mathbf{b}'}$ are independent since $\mathbf{b} \neq \mathbf{b}'$. Then since $N_{\mathbf{a},\mathbf{b}}$ and $N_{\mathbf{a}',\mathbf{b}'}$ are subcounts of $N_{\mathbf{b}}$ and $N_{\mathbf{b}'}$ respectively, so the corresponding ratios are also independent.

2. By Lemma 11, we have $(N_{\mathbf{a},\mathbf{b}} \mid N_{\mathbf{b}} = k) \sim \text{Bin}(k, \mathbb{P}(\mathbf{a} \mid \mathbf{b}))$. Conditioned on $N_{\mathbf{b}} = k$, Lemma 16 implies that $(N_{\mathbf{a},\mathbf{b}} - \mathbb{E}(N_{\mathbf{a},\mathbf{b}})) = (N_{\mathbf{a},\mathbf{b}} - k \cdot \mathbb{P}(\mathbf{a} \mid \mathbf{b})) \sim \text{subG}(\frac{k}{4})$. Thus, $\left(\frac{N_{\mathbf{a},\mathbf{b}}}{N_{\mathbf{b}}} - \mathbb{P}(\mathbf{a} \mid \mathbf{b}) \mid N_{\mathbf{b}} = k\right) \sim \text{subG}(\frac{1}{4k})$. The claim follows via Lemma 15.

■

Theorem 1 (Estimation error) *Suppose we are given (1) estimation tolerance $\varepsilon > 0$, (2) failure tolerance $\delta > 0$, (3) sample access to $\mathbb{P}(\mathbf{V})$, and (4) a subset $\mathbf{A} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$. Then, there is an algorithm that uses $\widetilde{\mathcal{O}}\left(\left(\frac{|\mathbf{\Sigma_A}|}{\varepsilon \alpha_{\mathbf{A}}} + \frac{1}{\varepsilon^2 \alpha_{\mathbf{A}}} + \frac{|\mathbf{\Sigma_A}|}{\varepsilon^2}\right) \cdot \log \frac{1}{\delta}\right)$ samples and produces an estimate $\widehat{T}_{\mathbf{A},\mathbf{x},\mathbf{y}}$ such that $\Pr(|\widehat{T}_{\mathbf{A},\mathbf{x},\mathbf{y}} - T_{\mathbf{A},\mathbf{x},\mathbf{y}}| \leq \varepsilon) \geq 1 - \delta$.*

Proof [Proof sketch] By definition, we have

$$\begin{aligned} T_{\mathbf{A},\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{A},\mathbf{x},\mathbf{y}} &= \sum_{\mathbf{a}} \left(\mathbb{P}(\mathbf{a}) \cdot \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{a}) - \frac{N_{\mathbf{a}}}{N_{\text{Pois}}} \cdot \frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}} \right) \\ &= \sum_{\mathbf{a}} \mathbb{P}(\mathbf{a}) \cdot \left(\mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{a}) - \frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}} \right) + \sum_{\mathbf{a}} \left(\mathbb{P}(\mathbf{a}) - \frac{N_{\mathbf{a}}}{N_{\text{Pois}}} \right) \cdot \frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}} \end{aligned}$$

where $N_{\mathbf{a}}$, N_{Pois}, $N_{\mathbf{y},\mathbf{x},\mathbf{a}}$, and $N_{\mathbf{x},\mathbf{a}}$ are random Poisson variables from the Poissonization process with $N_{\text{Pois}} \sim \text{Pois}(n)$ for some parameter n; see Section 2.1. Since $N_{\text{Pois}} = \sum_{\mathbf{a}} N_{\mathbf{a}} = \sum_{\mathbf{a},\mathbf{x}} N_{\mathbf{x},\mathbf{a}} = \sum_{\mathbf{a},\mathbf{x},\mathbf{y}} N_{\mathbf{y},\mathbf{x},\mathbf{a}}$, we see that $0 \leq \frac{N_{\mathbf{a}}}{N_{\text{Pois}}} \leq 1$ and $0 \leq \frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}} \leq 1$ for each of these fractional terms.

Let us define a threshold $\tau > 0$ and partition the values of $\mathbf{A}$ accordingly:

$$\mathbf{\Sigma}_{\mathbf{A} \geq \tau} = \{\mathbf{a} \in \mathbf{\Sigma_A} : \mathbb{P}(\mathbf{x}, \mathbf{a}) \geq \tau\}$$

Since $\alpha_{\mathbf{A}} = \min_{\mathbf{a} \in \mathbf{\Sigma_A}} \mathbb{P}(\mathbf{x} \mid \mathbf{a})$, we see that $\mathbb{P}(\mathbf{a}) \leq \frac{\tau}{\alpha_{\mathbf{A}}}$ for $\mathbf{a} \not\in \mathbf{\Sigma}_{\mathbf{A} \geq \tau}$.

Let us define three summations $J_{<\tau}$, $J_{\geq\tau}$, and K so that $T_{\mathbf{A},\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{A},\mathbf{x},\mathbf{y}} = J_{<\tau} + J_{\geq\tau} + K$:

$$J_{<\tau} = \sum_{\mathbf{a} \not\in \mathbf{\Sigma}_{\mathbf{A} \geq \tau}} \mathbb{P}(\mathbf{a}) \cdot \left(\mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{a}) - \frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}} \right) \tag{6}$$

$$J_{\geq\tau} = \sum_{\mathbf{a} \in \mathbf{\Sigma}_{\mathbf{A} \geq \tau}} \mathbb{P}(\mathbf{a}) \cdot \left(\mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{a}) - \frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}} \right) \tag{7}$$

$$K = \sum_{\mathbf{a}} \left(\mathbb{P}(\mathbf{a}) - \frac{N_{\mathbf{a}}}{N_{\text{Pois}}} \right) \cdot \frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}} \tag{8}$$

We will proceed to bound each of $|J_{<\tau}|$, $|J_{\geq\tau}|$, and $|K|$.

The easiest is $|J_{<\tau}|$, which follows from the definition of $\mathbf{\Sigma}_{\mathbf{A} \geq \tau}$:

$$\begin{aligned} |J_{<\tau}| &= \left| \sum_{\mathbf{a} \not\in \mathbf{\Sigma}_{\mathbf{A} \geq \tau}} \mathbb{P}(\mathbf{a}) \cdot \left(\frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}} - \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{a}) \right) \right| && \text{(Definition of } |J_{<\tau}|\text{)} \\ &\leq \sum_{\mathbf{a} \not\in \mathbf{\Sigma}_{\mathbf{A} \geq \tau}} \mathbb{P}(\mathbf{a}) \cdot \left| \frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}} - \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{a}) \right| && \text{(By triangle inequality and } \mathbb{P}(\mathbf{a}) \geq 0\text{)} \\ &\leq \sum_{\mathbf{a} \not\in \mathbf{\Sigma}_{\mathbf{A} \geq \tau}} \mathbb{P}(\mathbf{a}) && \left(\text{Since } \left| \tfrac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}} - \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{a}) \right| \leq 1\right) \\ &\leq \frac{\tau \cdot |\mathbf{\Sigma_A}|}{\alpha_{\mathbf{A}}} && \left(\text{Since } \mathbb{P}(\mathbf{a}) \leq \tfrac{\tau}{\alpha_{\mathbf{A}}} \text{ for } \mathbf{a} \not\in \mathbf{\Sigma}_{\mathbf{A} \geq \tau} \text{ and } |\mathbf{\Sigma}_{\mathbf{A} \geq \tau}| \leq |\mathbf{\Sigma_A}|\right) \end{aligned}$$

To bound $|J_{\geq \tau}|$, consider the concentration event $\mathcal{E}^J_{\geq \tau}$ defined as follows:

$$\mathcal{E}^J_{\geq \tau} = \bigcap_{\mathbf{a} \in \mathbf{\Sigma}_{\mathbf{A} \geq \tau}} \left\{ N_{\mathbf{x},\mathbf{a}} > \frac{n \cdot \mathbb{P}(\mathbf{x}, \mathbf{a})}{2} \right\} \tag{9}$$

We first observe that the event $\mathcal{E}^J_{\geq \tau}$ holds with good probability.

$$\begin{aligned}
1 - \Pr(\mathcal{E}^J_{\geq \tau}) &\leq \sum_{\mathbf{a} \in \mathbf{\Sigma}_{\mathbf{A} \geq \tau}} \Pr\left(N_{\mathbf{x},\mathbf{a}} \leq \frac{n \cdot \mathbb{P}(\mathbf{x}, \mathbf{a})}{2} \right) && \text{(Union bound)} \\
&\leq \sum_{\mathbf{a} \in \mathbf{\Sigma}_{\mathbf{A} \geq \tau}} \exp\left(-\frac{n \cdot \mathbb{P}(\mathbf{x}, \mathbf{a})}{12} \right) \\
& && \text{(Using that } N_{\mathbf{x},\mathbf{a}} \sim \text{Pois}(n \cdot \mathbb{P}(\mathbf{x}, \mathbf{a})) \text{ and applying Lemma 12)} \\
&\leq |\mathbf{\Sigma}_{\mathbf{A}}| \cdot \exp\left(-\frac{n\tau}{12} \right) && \text{(Since } \mathbb{P}(\mathbf{x}, \mathbf{a}) \geq \tau \text{ for } \mathbf{z} \in \mathbf{\Sigma}_{\mathbf{A} \geq \tau} \subseteq \mathbf{\Sigma}_{\mathbf{A} \geq \tau})
\end{aligned}$$

Under the event $\mathcal{E}^J_{\geq \tau}$, we have $N_{\mathbf{x},\mathbf{a}} > \frac{n \cdot \mathbb{P}(\mathbf{x},\mathbf{a})}{2}$ for any $\mathbf{a} \in \mathbf{\Sigma}_{\mathbf{A} \geq \tau}$, so item 2 of Lemma 30 implies that

$$\left(\frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}} - \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{a}) \;\middle|\; N_{\mathbf{x},\mathbf{a}} \geq \frac{n \cdot \mathbb{P}(\mathbf{x}, \mathbf{a})}{2} \right) \sim \text{subG}\left(\frac{1}{4} \cdot \frac{2}{n \cdot \mathbb{P}(\mathbf{x}, \mathbf{a})} \right) = \text{subG}\left(\frac{1}{2n \cdot \mathbb{P}(\mathbf{x}, \mathbf{a})} \right) ,$$

for any $\mathbf{a} \in \mathbf{\Sigma}_{\mathbf{A} \geq \tau}$. For any two disjoint $\mathbf{a}, \mathbf{a}' \in \mathbf{\Sigma}_{\mathbf{A}}$, we see that $(\mathbf{x}, \mathbf{a})$ and $(\mathbf{x}, \mathbf{a}')$ are distinct values in the domain $\mathbf{\Sigma}_{\mathbf{X}} \times \mathbf{\Sigma}_{\mathbf{A}}$, so item 1 of Lemma 30 tells us that the terms $\frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}}$ and $\frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}'}}{N_{\mathbf{x},\mathbf{a}'}}$ are independent. Lemma 14 further tells us that $J_{\geq \tau} = \sum_{\mathbf{a} \in \mathbf{\Sigma}_{\mathbf{A}_{\geq \tau}}} \mathbb{P}(\mathbf{a}) \cdot \left(\frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}} - \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{a}) \right) \sim \text{subG}\left(\sum_{\mathbf{a} \in \mathbf{\Sigma}_{\mathbf{A} \geq \tau}} \frac{\mathbb{P}(\mathbf{a})^2}{2n \cdot \mathbb{P}(\mathbf{x},\mathbf{a})} \right)$ since coefficients $\{\mathbb{P}(\mathbf{a})\}_{\mathbf{a} \in \mathbf{A}}$ are just (unknown) real numbers. Then, for any $t > 0$, Definition 13 states that

$$\Pr\left(|J_{\geq \tau}| > t \mid \mathcal{E}^J_{\geq \tau} \right) \leq 2 \exp\left(-\frac{t^2}{2 \sum_{\mathbf{a} \in \mathbf{\Sigma}_{\mathbf{A} \geq \tau}} \frac{\mathbb{P}(\mathbf{a})^2}{2n \cdot \mathbb{P}(\mathbf{x},\mathbf{a})}} \right) \leq 2 \exp\left(-n \alpha_{\mathbf{A}} t^2 \right)$$

where the last inequality is because $\alpha_{\mathbf{A}} = \min_{\mathbf{a} \in \mathbf{\Sigma}_{\mathbf{A}}} \mathbb{P}(\mathbf{x} \mid \mathbf{a})$ and $\mathbf{\Sigma}_{\mathbf{A} \geq \tau} \subseteq \mathbf{\Sigma}_{\mathbf{A}}$:

$$\sum_{\mathbf{a} \in \mathbf{\Sigma}_{\mathbf{A} \geq \tau}} \frac{\mathbb{P}(\mathbf{a})^2}{2n \cdot \mathbb{P}(\mathbf{x}, \mathbf{a})} = \sum_{\mathbf{a} \in \mathbf{\Sigma}_{\mathbf{A} \geq \tau}} \frac{\mathbb{P}(\mathbf{a})}{2n \cdot \mathbb{P}(\mathbf{x} \mid \mathbf{a})} \leq \sum_{\mathbf{a} \in \mathbf{\Sigma}_{\mathbf{A} \geq \tau}} \frac{\mathbb{P}(\mathbf{a})}{2n \cdot \alpha_{\mathbf{A}}} \leq \frac{1}{2n \cdot \alpha_{\mathbf{A}}}$$

To bound $|K|$, we reduce to the analysis to the problem of producing an ε-close estimate of $\mathbb{P}(\mathbf{A})$ by observing that $0 \leq \frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}} \leq 1$ and $\frac{N_{\mathbf{a}}}{N_{\text{Pois}}}$ is the empirical estimate of $\mathbb{P}(\mathbf{a})$ for each $\mathbf{a} \in \mathbf{\Sigma}_{\mathbf{A}}$. That is,

$$\begin{aligned}
|K| &= \left| \sum_{\mathbf{a}} \left(\mathbb{P}(\mathbf{a}) - \frac{N_{\mathbf{a}}}{N_{\text{Pois}}} \right) \cdot \frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}} \right| && \text{(By Eq. (8))} \\
&= \sum_{\mathbf{a}} \left| \mathbb{P}(\mathbf{a}) - \frac{N_{\mathbf{a}}}{N_{\text{Pois}}} \right| \cdot \left| \frac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}} \right| && \text{(By triangle inequality)}
\end{aligned}$$

$$\leq \sum_{\mathbf{a}} \left| \mathbb{P}(\mathbf{a}) - \frac{N_{\mathbf{a}}}{N_{\text{Pois}}} \right| \qquad \left(\text{Since } 0 \leq \tfrac{N_{\mathbf{y},\mathbf{x},\mathbf{a}}}{N_{\mathbf{x},\mathbf{a}}} \leq 1\right)$$

$$\leq \sum_{\mathbf{a}} \left| \mathbb{P}(\mathbf{a}) - \widehat{\mathbb{P}}(\mathbf{a}) \right| \qquad \left(\text{By defining empirical distribution } \widehat{\mathbb{P}}(\mathbf{a}) = \tfrac{N_{\mathbf{a}}}{N_{\text{Pois}}}\right)$$

By Lemma 17, when $N_{\text{Pois}} \geq c_0 \cdot \left(\frac{|\mathbf{\Sigma_A}| + \log \frac{1}{\delta'}}{(\varepsilon')^2} \right)$ for some tolerance parameters $\varepsilon', \delta' > 0$ and absolute constant $c_0 > 0$, we will have $\Pr(|K| \leq \varepsilon') \leq \Pr\left(\sum_{\mathbf{a} \in \mathbf{\Sigma_A}} |\mathbb{P}(\mathbf{a}) - \widehat{\mathbb{P}}(\mathbf{a})| \leq \varepsilon' \right) \geq 1 - \delta'$.

Before we proceed to wrap up the proof, let us collect the proven bounds below:

$$|J_{<\tau}| \leq \frac{\tau \cdot |\mathbf{\Sigma_A}|}{\alpha_{\mathbf{A}}} \qquad \text{deterministically} \tag{10}$$

$$\Pr(\neg \mathcal{E}^J_{\geq \tau}) \leq |\mathbf{\Sigma_A}| \cdot \exp\left(-\frac{n\tau}{12} \right) \tag{11}$$

$$\Pr\left(|J_{\geq \tau}| > t \mid \mathcal{E}^J_{\geq \tau} \right) \leq 2 \exp\left(-n \alpha_{\mathbf{A}} t^2 \right) \qquad \text{for any } t > 0 \tag{12}$$

$$\Pr\left(|K| \leq \varepsilon' \right) \leq 1 - \delta' \qquad \text{for any } \varepsilon', \delta' > 0 \text{ when } N_{\text{Pois}} \in \mathcal{O}\left(\frac{|\mathbf{\Sigma}| + \log \frac{1}{\delta'}}{(\varepsilon')^2} \right) \tag{13}$$

Now, observe that $|J_{<\tau}| \leq \frac{\varepsilon}{3}$, $|J_{\geq \tau}| \leq \frac{\varepsilon}{3}$ and $|K| \leq \frac{\varepsilon}{3}$ jointly implies $|J_{<\tau} + J_{\geq \tau} + K| \leq |J_{<\tau}| + |J_{\geq \tau}| + |K| \leq \varepsilon$ by triangle inequality. So,

$$\begin{aligned}
&\Pr\left(\left| T_{\mathbf{A},\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{A},\mathbf{x},\mathbf{y}} \right| > \varepsilon \right) \\
= \ &\Pr\left(|J_{<\tau} + J_{\geq \tau} + K| > \varepsilon \right) && \text{(By definition)} \\
\leq \ &\Pr\left(|J_{<\tau}| > \frac{\varepsilon}{3} \right) + \Pr\left(|J_{\geq \tau}| > \frac{\varepsilon}{3} \right) + \Pr\left(|K| > \frac{\varepsilon}{3} \right) && \text{(Triangle inequality)} \\
\leq \ &0 + \Pr\left(|J_{\geq \tau}| > \frac{\varepsilon}{3} \right) + \Pr\left(|K| > \frac{\varepsilon}{3} \right) \\
& && \left(\text{If we set } \tfrac{\varepsilon}{3} = \tfrac{\tau \cdot |\mathbf{\Sigma_A}|}{\alpha_{\mathbf{A}}} \text{ in the deterministic bound of Eq. (10)}\right) \\
\leq \ &\Pr\left(|J_{\geq \tau}| > \frac{\varepsilon}{3} \right) + \frac{\delta}{3} && \left(\text{If we set } \varepsilon' = \tfrac{\varepsilon}{3} \text{ and } \delta' = \tfrac{\delta}{3} \text{ in Eq. (13) with } N_{\text{Pois}} \in \mathcal{O}\left(\tfrac{|\mathbf{\Sigma_A}| + \log \frac{1}{\delta'}}{(\varepsilon')^2} \right)\right) \\
\leq \ &\Pr(\neg \mathcal{E}^J_{\geq \tau}) + \Pr\left(|J_{\geq \tau}| > \frac{\varepsilon}{3} \mid \mathcal{E}^J_{\geq \tau} \right) + \frac{\delta}{3} && \left(\text{Conditioning on event } \mathcal{E}^J_{\geq \tau}\right) \\
\leq \ &|\mathbf{\Sigma_A}| \cdot \exp\left(-\frac{n\tau}{12} \right) + 2 \exp\left(-n \alpha_{\mathbf{A}} t^2 \right) + \frac{\delta}{3} \\
& && \left(\text{If we set } t = \tfrac{\varepsilon}{3} \text{ then apply Eq. (11) and Eq. (12)}\right)
\end{aligned}$$

Recall that we set $\frac{\varepsilon}{3} = \frac{\tau \cdot |\mathbf{\Sigma_A}|}{\alpha_{\mathbf{A}}} \iff \tau = \frac{\varepsilon \alpha_{\mathbf{A}}}{|3\mathbf{\Sigma_A}|}$ and $t = \frac{\varepsilon}{3}$ above. So, if we set

$$\begin{aligned}
n &= \frac{36 |\mathbf{\Sigma_A}|}{\varepsilon \alpha_{\mathbf{A}}} \log\left(\frac{3 |\mathbf{\Sigma_A}|}{\delta} \right) + \frac{9}{\varepsilon^2 \alpha_{\mathbf{A}}} \log\left(\frac{6}{\delta} \right) + \mathcal{O}\left(\frac{|\mathbf{\Sigma_A}| + \log \frac{1}{\delta}}{\varepsilon^2} \right) \\
&\in \widetilde{\mathcal{O}}\left(\left(\frac{|\mathbf{\Sigma_A}|}{\varepsilon \alpha_{\mathbf{A}}} + \frac{1}{\varepsilon^2 \alpha_{\mathbf{A}}} + \frac{|\mathbf{\Sigma_A}|}{\varepsilon^2} \right) \cdot \log\left(\frac{1}{\delta} \right) \right)
\end{aligned}$$

then $\Pr\left(\left| T_{\mathbf{A},\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{A},\mathbf{x},\mathbf{y}} \right| > \varepsilon \right) \leq \frac{\delta}{3} + \frac{\delta}{3} + \frac{\delta}{3} = \delta$. ■

C.3. Misspecification errors

Lemma 8 (Misspecification error) *If* $\mathbf{S} \subseteq \mathbf{A} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$ *such that* $\mathbf{X} \perp\!\!\!\perp_\varepsilon \mathbf{A} \setminus \mathbf{S} \mid \mathbf{S}$*, then* $|T_{\mathbf{S},\mathbf{x},\mathbf{y}} - T_{\mathbf{A},\mathbf{x},\mathbf{y}}| \leq \frac{\varepsilon}{\alpha_{\mathbf{S}}}$.

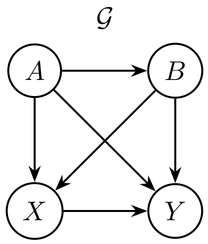
Proof Since $\mathbf{S} \subseteq \mathbf{A}$, we see that

$$\begin{aligned}
|T_{\mathbf{S},\mathbf{x},\mathbf{y}} - T_{\mathbf{A},\mathbf{x},\mathbf{y}}| &= \left| \sum_{\mathbf{a}} \mathbb{P}(\mathbf{y} \mid \mathbf{a}, \mathbf{x}) \cdot \mathbb{P}(\mathbf{a} \setminus \mathbf{s} \mid \mathbf{s}, \mathbf{x}) \cdot \mathbb{P}(\mathbf{s}) - \sum_{\mathbf{a}} \mathbb{P}(\mathbf{y} \mid \mathbf{a}, \mathbf{x}) \cdot \mathbb{P}(\mathbf{a}) \right| \\
&\qquad \text{(By Eq. (4) and } \mathbf{S} \subseteq \mathbf{A}\text{)} \\
&= \left| \sum_{\mathbf{a}} \mathbb{P}(\mathbf{y} \mid \mathbf{a}, \mathbf{x}) \cdot \mathbb{P}(\mathbf{s}) \cdot (\mathbb{P}(\mathbf{a} \setminus \mathbf{s} \mid \mathbf{s}, \mathbf{x}) - \mathbb{P}(\mathbf{a} \setminus \mathbf{s} \mid \mathbf{s})) \right| \\
&\qquad \text{(Pull out common terms)} \\
&= \left| \sum_{\mathbf{a}} \mathbb{P}(\mathbf{y} \mid \mathbf{a}, \mathbf{x}) \cdot \frac{\mathbb{P}(\mathbf{s}, \mathbf{x})}{\mathbb{P}(\mathbf{x} \mid \mathbf{s})} \cdot (\mathbb{P}(\mathbf{a} \setminus \mathbf{s} \mid \mathbf{s}, \mathbf{x}) - \mathbb{P}(\mathbf{a} \setminus \mathbf{s} \mid \mathbf{s})) \right| \\
&\leq \sum_{\mathbf{a}} \frac{\mathbb{P}(\mathbf{s}, \mathbf{x})}{\mathbb{P}(\mathbf{x} \mid \mathbf{s})} \cdot |\mathbb{P}(\mathbf{a} \setminus \mathbf{s} \mid \mathbf{s}, \mathbf{x}) - \mathbb{P}(\mathbf{a} \setminus \mathbf{s} \mid \mathbf{s})| \\
&\qquad \text{(Triangle inequality, non-negativity of probabilities, and since } \mathbb{P}(\mathbf{y} \mid \mathbf{a}, \mathbf{x}) \leq 1\text{)} \\
&\leq \frac{1}{\alpha_{\mathbf{S}}} \cdot \sum_{\mathbf{a}} \mathbb{P}(\mathbf{s}, \mathbf{x}) \cdot |\mathbb{P}(\mathbf{a} \setminus \mathbf{s} \mid \mathbf{s}, \mathbf{x}) - \mathbb{P}(\mathbf{a} \setminus \mathbf{s} \mid \mathbf{s})| \qquad \text{(By Eq. (2))} \\
&\leq \frac{\varepsilon}{\alpha_{\mathbf{S}}} \qquad \text{(Since } \mathbf{X} \perp\!\!\!\perp_\varepsilon \mathbf{A} \setminus \mathbf{S} \mid \mathbf{S} \text{ and using Eq. (3))}
\end{aligned}$$

■

Lemma 9 (Misspecification error lower bound) *Let* $0 \leq \sqrt{\varepsilon} \leq \alpha \leq 1/2$*. There exists* $\mathbb{P}(\mathbf{V})$ *such that (i)* $\mathbf{Z}$ *is a valid adjustment set, (ii)* $\mathbf{S} \subset \mathbf{Z}$ *satisfies* $\mathbf{X} \perp\!\!\!\perp_\varepsilon \mathbf{Z} \setminus \mathbf{S} \mid \mathbf{S}$*, (iii)* $\alpha_{\mathbf{S}} \geq \alpha$*, and (iv)* $|T_{\mathbf{S},\mathbf{x},\mathbf{y}} - T_{\mathbf{Z},\mathbf{x},\mathbf{y}}| \geq \frac{\varepsilon}{16\alpha}$.

Proof Consider the following probability distribution $\mathbb{P}$ defined over 4 binary variables $\{A, B, X, Y\}$ in a topological ordering of $A \prec B \prec X \prec Y$: see Fig. 4.



$$A = \begin{cases} 1 & \text{w.p. } \frac{\varepsilon}{4\alpha} \cdot \frac{\alpha - \varepsilon/4}{1 - \sqrt{\varepsilon}/2} \\ 0 & \text{else} \end{cases} \qquad X = \begin{cases} A & \text{w.p. } 1 - \alpha \\ 1 - A & \text{w.p. } \alpha - \sqrt{\varepsilon}/2 \\ B & \text{w.p. } \sqrt{\varepsilon}/2 \end{cases}$$

$$B = \begin{cases} 1 - A & \text{w.p. } 1 - \sqrt{\varepsilon} \\ 0 & \text{w.p. } \sqrt{\varepsilon}/2 \\ 1 & \text{w.p. } \sqrt{\varepsilon}/2 \end{cases} \qquad Y = \begin{cases} 1 & \text{if } X = 0, A = 1, B = 0 \\ 0 & \text{else} \end{cases}$$

Figure 4: Probability distribution $\mathbb{P}$ defined over 4 binary variables $\{A, B, X, Y\}$ in a topological ordering of $A \prec B \prec X \prec Y$ with parameters ε and α, where $0 < \sqrt{\varepsilon} \leq \alpha \leq 1/2$.

We show in Appendix C.6 that all the (conditional) probabilities of $\mathbb{P}$ are well-defined, and that we have the following conditional probabilities for $\mathbb{P}$:

a	b	$\mathbb{P}(b \mid a)$	$\mathbb{P}(X=0 \mid a,b)$	$\mathbb{P}(X=0 \mid a)$	$\sum_x \lvert\mathbb{P}(x \mid a,b) - \mathbb{P}(x \mid a)\rvert$
0	0	$\sqrt{\varepsilon}/2$	$1-\alpha+\sqrt{\varepsilon}/2$	$1-\alpha+\varepsilon/4$	$\sqrt{\varepsilon}-\varepsilon/2$
0	1	$1-\sqrt{\varepsilon}/2$	$1-\alpha$	$1-\alpha+\varepsilon/4$	$\varepsilon/2$
1	0	$1-\sqrt{\varepsilon}/2$	α	$\alpha-\varepsilon/4$	$\varepsilon/2$
1	1	$\sqrt{\varepsilon}/2$	$\alpha-\sqrt{\varepsilon}/2$	$\alpha-\varepsilon/4$	$\sqrt{\varepsilon}-\varepsilon/2$

Let us identify $\mathbf{Z}$ with $\{A,B\}$ and $\mathbf{S}$ with $\{A\}$, so $\mathbf{Z}\setminus\mathbf{S}=\{B\}$. We now show the four properties.

1. $\mathbf{Z}$ is a valid adjustment set

   This is true since $\{A,B\}$ satifies the backdoor adjustment criterion Pearl (1995).

2. $\mathbf{S}\subset\mathbf{Z}$ satisfies $\mathbf{X} \perp\!\!\!\perp_{\varepsilon} \mathbf{Z}\setminus\mathbf{S} \mid \mathbf{S}$

   Recall that $\mathbf{Z}=\{A,B\}$ and $\mathbf{S}=\{A\}$. To see that $X \perp\!\!\!\perp_{\varepsilon} B \mid A$, observe the following:

$$\begin{aligned}
&\sum_{x,a,b}\mathbb{P}(a)\cdot|\mathbb{P}(x,b \mid a)-\mathbb{P}(x \mid a)\cdot\mathbb{P}(b \mid a)|\\
&=\sum_{a,b}\mathbb{P}(a)\cdot\mathbb{P}(b \mid a)\cdot\sum_x|\mathbb{P}(x \mid a,b)-\mathbb{P}(x \mid a)|\\
&=\mathbb{P}(A=0)\cdot\mathbb{P}(B=0 \mid A=0)\cdot(\sqrt{\varepsilon}-\varepsilon/2)+\mathbb{P}(A=0)\cdot\mathbb{P}(B=1 \mid A=0)\cdot(\varepsilon/2)\\
&\quad+\mathbb{P}(A=1)\cdot\mathbb{P}(B=0 \mid A=1)\cdot(\varepsilon/2)+\mathbb{P}(A=1)\cdot\mathbb{P}(B=1 \mid A=1)\cdot(\sqrt{\varepsilon}-\varepsilon/2)\\
&=\mathbb{P}(A=0)\cdot(\sqrt{\varepsilon}/2)\cdot(\sqrt{\varepsilon}-\varepsilon/2)+\mathbb{P}(A=0)\cdot(1-\sqrt{\varepsilon}/2)\cdot(\varepsilon/2)\\
&\quad+\mathbb{P}(A=1)\cdot(1-\sqrt{\varepsilon}/2)\cdot(\varepsilon/2)+\mathbb{P}(A=1)\cdot(\sqrt{\varepsilon}/2)\cdot(\sqrt{\varepsilon}-\varepsilon/2)\\
&=(\sqrt{\varepsilon}/2)\cdot(\sqrt{\varepsilon}-\varepsilon/2)+(1-\sqrt{\varepsilon}/2)\cdot\varepsilon/2\\
&=\varepsilon
\end{aligned}$$

3. $\alpha_{\mathbf{S}}\geq\alpha$

   Since $\mathbf{S}=\{A\}$ and $\varepsilon\leq\alpha/2$, we have $\min_a\mathbb{P}(x \mid a)=\alpha-\varepsilon/4\geq\alpha/2$.

4. $|T_{\mathbf{S},\mathbf{x},\mathbf{y}}-T_{\mathbf{Z},\mathbf{x},\mathbf{y}}|\geq\frac{\varepsilon}{16\alpha}$.

$$|T_{\mathbf{S},\mathbf{x},\mathbf{y}}-T_{\mathbf{Z},\mathbf{x},\mathbf{y}}|=\left|\sum_a\mathbb{P}(a)\cdot\mathbb{P}(y \mid x,a)-\sum_{a,b}\mathbb{P}(a,b)\cdot\mathbb{P}(y \mid x,a,b)\right|$$

(Since $\mathbf{S}=\{A\}$, $\mathbf{Z}=\{A,B\}$, and by definition of $T_{\mathbf{S},\mathbf{x},\mathbf{y}}$ and $T_{\mathbf{Z},\mathbf{x},\mathbf{y}}$)

$$=\left|\sum_{a,b}\mathbb{P}(a)\cdot\mathbb{P}(y \mid x,a,b)\cdot(\mathbb{P}(b \mid a)-\mathbb{P}(b \mid x,a))\right|$$

(Since $\mathbb{P}(y \mid x,a)=\sum_b\mathbb{P}(y,b \mid x,a)=\sum_b\mathbb{P}(y \mid x,a,b)\cdot\mathbb{P}(b \mid x,a)$ and $\mathbb{P}(a,b)=\mathbb{P}(a)\cdot\mathbb{P}(b \mid a)$)

$$= \left| \sum_{a,b} \mathbb{P}(a) \cdot \mathbb{P}(y \mid x, a, b) \cdot \frac{\mathbb{P}(b \mid a)}{\mathbb{P}(x \mid a)} \cdot (\mathbb{P}(x \mid a) - \mathbb{P}(x \mid a, b)) \right|$$

(Since $\mathbb{P}(b \mid x, a) = \frac{\mathbb{P}(b|a) \cdot \mathbb{P}(x|a,b)}{\mathbb{P}(x|a)}$)

$$= \mathbb{P}(A = 1) \cdot \frac{\mathbb{P}(B = 0 \mid A = 1)}{\mathbb{P}(X = 0 \mid A = 1)} \cdot \left| \mathbb{P}(X = 0 \mid A = 1) - \mathbb{P}(X = 0 \mid A = 1, B = 0) \right|$$

(Since Y is an indicator variable for whether $(A, B, X) = (1, 0, 0)$)

$$= \frac{\varepsilon}{4\alpha} \cdot \frac{\alpha - \varepsilon/4}{1 - \sqrt{\varepsilon}/2} \cdot \frac{1 - \sqrt{\varepsilon}/2}{\alpha - \varepsilon/4} \cdot \frac{\varepsilon}{4} \qquad \text{(From construction in Fig. 4)}$$

$$= \frac{\varepsilon}{16\alpha}$$

■

C.4. Beyond approximate Markov blankets

Lemma 10 (Adjustment soundness) *Let $\mathbf{A} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$ be an arbitrary subset and $\mathbf{S} \subseteq \mathbf{A}$. If $\mathbf{S}'$ is a screening set for $(\mathbf{S}, \mathbf{X}, \mathbf{Y})$, then $T_{\mathbf{S}',\mathbf{x},\mathbf{y}} = T_{\mathbf{S},\mathbf{x},\mathbf{y}}$.*

Proof Consider arbitrary subsets $\mathbf{S} \subseteq \mathbf{A} \subseteq \mathbf{V}$ and $\mathbf{S}' \subseteq \mathbf{A} \subseteq \mathbf{V}$. Observe that

$$\begin{aligned}
T_{\mathbf{S},\mathbf{x},\mathbf{y}} &= \sum_{\mathbf{s},\mathbf{s}' \setminus \mathbf{s}} \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s}, \mathbf{s}' \setminus \mathbf{s}) \cdot \mathbb{P}(\mathbf{s}' \setminus \mathbf{s} \mid \mathbf{x}, \mathbf{s}) \cdot \mathbb{P}(\mathbf{s}) && \text{(By Eq. (4))}\\
&= \sum_{\mathbf{s},\mathbf{s}' \setminus \mathbf{s}} \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s}, \mathbf{s}' \setminus \mathbf{s}) \cdot \mathbb{P}(\mathbf{s}' \setminus \mathbf{s} \mid \mathbf{s}) \cdot \mathbb{P}(\mathbf{s}) && \text{(Since } \mathbf{X} \perp\!\!\!\perp \mathbf{S}' \setminus \mathbf{S} \mid \mathbf{S})\\
&= \sum_{\mathbf{s}',\mathbf{s} \setminus \mathbf{s}'} \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s}', \mathbf{s} \setminus \mathbf{s}') \cdot \mathbb{P}(\mathbf{s}' \setminus \mathbf{s} \mid \mathbf{s}) \cdot \mathbb{P}(\mathbf{s}) && \text{(Regrouping)}\\
&= \sum_{\mathbf{s}',\mathbf{s} \setminus \mathbf{s}'} \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s}') \cdot \mathbb{P}(\mathbf{s}' \setminus \mathbf{s} \mid \mathbf{s}) \cdot \mathbb{P}(\mathbf{s}) && \text{(Since } \mathbf{Y} \perp\!\!\!\perp \mathbf{S} \setminus \mathbf{S}' \mid \mathbf{X} \cup \mathbf{S}')\\
&= T_{\mathbf{S}',\mathbf{x},\mathbf{y}} && \text{(By Eq. (4))}
\end{aligned}$$

■

Theorem 6 (Beyond approximate Markov blankets) *Suppose we are given (1) $\varepsilon > 0$, (2) $\delta > 0$, (3) sample access to $\mathbb{P}(\mathbf{V})$, (4) an arbitrary subset $\mathbf{A} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$, and (5) an ε-Markov blanket $\mathbf{S} \subseteq \mathbf{A}$. Suppose there is a screening set $\mathbf{B}$ for $(\mathbf{S}, \mathbf{A}, \mathbf{X}, \mathbf{Y})$ such that $|\mathbf{B}| = k'$ and $|\Sigma_{\mathbf{B}}| \leq |\Sigma_{\mathbf{S}}|$. There is an algorithm that uses $\widetilde{\mathcal{O}}\left(\frac{|\mathbf{S}'|}{\varepsilon^2} \cdot \sqrt{|\mathbf{\Sigma_X}| \cdot |\mathbf{\Sigma_Y}| \cdot |\mathbf{\Sigma_A}|} \cdot \log \frac{1}{\delta} \right)$ samples and produces a subset $\mathbf{S}' \subseteq \mathbf{A}$ such that $|\mathbf{S}'| \leq k'$, $|\mathbf{\Sigma}_{\mathbf{S}'}| \leq |\mathbf{\Sigma_S}|$ and $\Pr\left(|T_{\mathbf{S}',\mathbf{x},\mathbf{y}} - T_{\mathbf{A},\mathbf{x},\mathbf{y}}| \leq \frac{2\varepsilon}{\alpha_{\mathbf{S}}} \right) \geq 1 - \delta$.*

Proof Suppose the BEYONDAPPROXIMATEMARKOVBLANKETADJUSTMENT algorithm (Algorithm 2) terminates at some iteration $|\mathbf{S}'| \in \{0, 1, \ldots, |\mathbf{A}|\}$.

Correctness. If BAMBA returns the ε-Markov blanket $\mathbf{S} \subseteq \mathbf{A}$ (e.g. in Line 8), then $|T_{\mathbf{S}',\mathbf{x},\mathbf{y}} - T_{\mathbf{A},\mathbf{x},\mathbf{y}}| = |T_{\mathbf{S},\mathbf{x},\mathbf{y}} - T_{\mathbf{A},\mathbf{x},\mathbf{y}}| \leq \frac{\varepsilon}{\alpha_{\mathbf{S}}} \leq \frac{2\varepsilon}{\alpha_{\mathbf{S}}}$ by Definition 3 and Lemma 8. Suppose all calls to APPROXCONDIND succeed across all iterations. Then, Lemma 18 tells us that $\Delta_{\mathbf{Y} \perp\!\!\!\perp \mathbf{S} \setminus \mathbf{S}' \mid \mathbf{X} \cup \mathbf{S}'} \leq \varepsilon$, $\Delta_{\mathbf{X} \perp\!\!\!\perp \mathbf{S}' \setminus \mathbf{S} \mid \mathbf{S}} \leq \varepsilon$, and $|\mathbf{\Sigma}_{\mathbf{S}'}| \leq |\mathbf{\Sigma}_{\mathbf{S}}|$ whenever $\mathbf{C}_k \neq \emptyset$.

For subsequent analytical purposes, let us define an intermediate term $Z_{\mathbf{x},\mathbf{y}}$ as follows:

$$Z_{\mathbf{x},\mathbf{y}} = \sum_{\mathbf{s} \cup \mathbf{s}'} \frac{1}{\mathbb{P}(\mathbf{x} \mid \mathbf{s})} \cdot \mathbb{P}(\mathbf{x}, \mathbf{s} \cup \mathbf{s}') \cdot \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s}') \tag{14}$$

By triangle inequality, we have $\left|T_{\mathbf{S},\mathbf{x},\mathbf{y}} - T_{\mathbf{S}',\mathbf{x},\mathbf{y}}\right| = \left|T_{\mathbf{S},\mathbf{x},\mathbf{y}} - Z_{\mathbf{x},\mathbf{y}} + Z_{\mathbf{x},\mathbf{y}} - T_{\mathbf{S}',\mathbf{x},\mathbf{y}}\right| \leq |T_{\mathbf{S},\mathbf{x},\mathbf{y}} - Z_{\mathbf{x},\mathbf{y}}| + \left|Z_{\mathbf{x},\mathbf{y}} - T_{\mathbf{S}',\mathbf{x},\mathbf{y}}\right|$. We will bound each of these terms separately.

1. Bounding $|T_{\mathbf{S},\mathbf{x},\mathbf{y}} - Z_{\mathbf{x},\mathbf{y}}|$.

$$|T_{\mathbf{S},\mathbf{x},\mathbf{y}} - Z_{\mathbf{x},\mathbf{y}}| = \left| \sum_{\mathbf{s} \cup \mathbf{s}'} \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s} \cup \mathbf{s}') \cdot \mathbb{P}(\mathbf{s}' \setminus \mathbf{s} \mid \mathbf{x}, \mathbf{s}) \cdot \mathbb{P}(\mathbf{s}) - \sum_{\mathbf{s} \cup \mathbf{s}'} \frac{1}{\mathbb{P}(\mathbf{x} \mid \mathbf{s})} \cdot \mathbb{P}(\mathbf{x}, \mathbf{s} \cup \mathbf{s}') \cdot \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s}') \right|$$

(By Eq. (4) and Eq. (14))

$$= \left| \sum_{\mathbf{s} \cup \mathbf{s}'} \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s} \cup \mathbf{s}') \cdot \frac{\mathbb{P}(\mathbf{x}, \mathbf{s} \cup \mathbf{s}')}{\mathbb{P}(\mathbf{x}, \mathbf{s})} \cdot \mathbb{P}(\mathbf{s}) - \sum_{\mathbf{s} \cup \mathbf{s}'} \frac{1}{\mathbb{P}(\mathbf{x} \mid \mathbf{s})} \cdot \mathbb{P}(\mathbf{x}, \mathbf{s} \cup \mathbf{s}') \cdot \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s}') \right|$$

$$= \left| \sum_{\mathbf{s} \cup \mathbf{s}'} \frac{1}{\mathbb{P}(\mathbf{x} \mid \mathbf{s})} \cdot \mathbb{P}(\mathbf{x}, \mathbf{s} \cup \mathbf{s}') \cdot \big(\mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s} \cup \mathbf{s}') - \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s}') \big) \right|$$

(Pull out common terms)

$$\leq \sum_{\mathbf{s} \cup \mathbf{s}'} \frac{1}{\mathbb{P}(\mathbf{x} \mid \mathbf{s})} \cdot \mathbb{P}(\mathbf{x}, \mathbf{s} \cup \mathbf{s}') \cdot \big| \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s} \cup \mathbf{s}') - \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s}') \big|$$

(Triangle inequality and non-negative of probabilities)

$$\leq \frac{1}{\alpha_{\mathbf{S}}} \cdot \sum_{\mathbf{s} \cup \mathbf{s}'} \mathbb{P}(\mathbf{x}, \mathbf{s} \cup \mathbf{s}') \cdot \big| \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s} \cup \mathbf{s}') - \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s}') \big|$$

(By definition of $\alpha_{\mathbf{S}}$ in Eq. (2))

$$\leq \frac{1}{\alpha_{\mathbf{S}}} \cdot \sum_{\mathbf{y}, \mathbf{x}, \mathbf{s} \cup \mathbf{s}'} \mathbb{P}(\mathbf{x}, \mathbf{s} \cup \mathbf{s}') \cdot \big| \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s} \cup \mathbf{s}') - \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s}') \big|$$

(Summing over more terms)

$$\leq \frac{\varepsilon}{\alpha_{\mathbf{S}}}$$

(Since $\Delta_{\mathbf{Y} \perp\!\!\!\perp \mathbf{S} \setminus \mathbf{S}' \mid \mathbf{X} \cup \mathbf{S}'} \leq \varepsilon$ and using Eq. (3))

2. Bounding $\left|Z_{\mathbf{x},\mathbf{y}} - T_{\mathbf{S}',\mathbf{x},\mathbf{y}}\right|$.

$$\left|T_{\mathbf{S}',\mathbf{x},\mathbf{y}} - Z_{\mathbf{x},\mathbf{y}}\right| = \left| \sum_{\mathbf{s} \cup \mathbf{s}'} \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s}') \cdot \mathbb{P}(\mathbf{s}') \cdot \mathbb{P}(\mathbf{s} \setminus \mathbf{s}' \mid \mathbf{s}') - \sum_{\mathbf{s} \cup \mathbf{s}'} \frac{1}{\mathbb{P}(\mathbf{x} \mid \mathbf{s})} \cdot \mathbb{P}(\mathbf{x}, \mathbf{s} \cup \mathbf{s}') \cdot \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s}') \right|$$

(By Eq. (4) and Eq. (14))

$$= \left| \sum_{\mathbf{s} \cup \mathbf{s}'} \frac{1}{\mathbb{P}(\mathbf{x} \mid \mathbf{s})} \cdot \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s}') \cdot \mathbb{P}(\mathbf{s} \cup \mathbf{s}') \cdot \big(\mathbb{P}(\mathbf{x} \mid \mathbf{s}) - \mathbb{P}(\mathbf{x} \mid \mathbf{s} \cup \mathbf{s}') \big) \right|$$

(Pull out common terms)

$$\leq \sum_{\mathbf{s}\cup\mathbf{s}'} \frac{1}{\mathbb{P}(\mathbf{x}\mid\mathbf{s})} \cdot \mathbb{P}(\mathbf{s}\cup\mathbf{s}') \cdot \left|\mathbb{P}(\mathbf{x}\mid\mathbf{s}) - \mathbb{P}(\mathbf{x}\mid\mathbf{s}\cup\mathbf{s}')\right|$$

(Triangle inequality, non-negativity of probabilities, and since $\mathbb{P}(\mathbf{y}\mid\mathbf{x},\mathbf{s}') \leq 1$)

$$\leq \frac{1}{\alpha_{\mathbf{S}}} \cdot \sum_{\mathbf{s}\cup\mathbf{s}'} \mathbb{P}(\mathbf{s}\cup\mathbf{s}') \cdot \left|\mathbb{P}(\mathbf{x}\mid\mathbf{s}) - \mathbb{P}(\mathbf{x}\mid\mathbf{s}\cup\mathbf{s}')\right|$$

(By definition of $\alpha_{\mathbf{S}}$ in Eq. (2))

$$\leq \frac{\varepsilon}{\alpha_{\mathbf{S}}}$$

(Since $\Delta_{\mathbf{X}\perp\mathbf{S}'\setminus\mathbf{S}|\mathbf{S}} \leq \varepsilon$ and using Eq. (3))

Putting together.
We see that

$$\left|T_{\mathbf{S},\mathbf{x},\mathbf{y}} - T_{\mathbf{S}',\mathbf{x},\mathbf{y}}\right| \leq \left|T_{\mathbf{S},\mathbf{x},\mathbf{y}} - Z_{\mathbf{x},\mathbf{y}}\right| + \left|Z_{\mathbf{x},\mathbf{y}} - T_{\mathbf{S}',\mathbf{x},\mathbf{y}}\right| \leq \frac{\varepsilon}{\alpha_{\mathbf{S}}} + \frac{\varepsilon}{\alpha_{\mathbf{S}}} = \frac{2\varepsilon}{\alpha_{\mathbf{S}}}$$

Failure rate. Note that there are at most $\binom{|\mathbf{A}|}{k}$ possible candidate sets in $\mathbf{C}_k$ for each $k \in \{0, 1, \ldots, |\mathbf{A}|\}$. Since we invoked two calls to APPROXCONDIND in iteration k, each with failure parameter $\delta w_k/2$, union bound tells us that the probability of *any* call failing across all calls is at most

$$\sum_{k=0}^{|\mathbf{S}'|} 2 \cdot \frac{\delta w_k}{2} \cdot \binom{|\mathbf{A}|}{k} = \sum_{k=0}^{|\mathbf{S}'|} \delta \cdot \frac{1}{|\mathbf{A}| \cdot \binom{|\mathbf{A}|}{k}} \cdot \binom{|\mathbf{A}|}{k} = \sum_{k=0}^{|\mathbf{S}'|} \frac{\delta}{|\mathbf{A}|} \leq \frac{\delta \cdot |\mathbf{S}|}{|\mathbf{A}|} \leq \delta$$

Sample complexity. Since we are using union bound to bound our overall failure probability, we can reuse samples in all our calls to APPROXCONDIND. Thus, the total sample complexity is attributed to the final call when $k = |\mathbf{S}'|$. Such an invocation of APPROXCONDIND uses $\widetilde{\mathcal{O}}\left(\frac{1}{\varepsilon^2} \cdot \sqrt{|\mathbf{\Sigma}_{\mathbf{X}}| \cdot |\mathbf{\Sigma}_{\mathbf{Y}}| \cdot |\mathbf{\Sigma}_{\mathbf{A}\setminus\mathbf{S}'}| \cdot |\mathbf{\Sigma}_{\mathbf{S}'}|} \cdot \log \frac{1}{\delta w_k}\right)$ samples according to Lemma 18 and $w_k = \left(|\mathbf{A}| \cdot \binom{|\mathbf{A}|}{k}\right)^{-1}$, so the total number of samples used is at most

$$\widetilde{\mathcal{O}}\left(\frac{1}{\varepsilon^2} \cdot \sqrt{|\mathbf{\Sigma}_{\mathbf{X}}| \cdot |\mathbf{\Sigma}_{\mathbf{Y}}| \cdot |\mathbf{\Sigma}_{\mathbf{A}\setminus\mathbf{S}'}| \cdot |\mathbf{\Sigma}_{\mathbf{S}'}|} \cdot \log \frac{1}{\delta w_k}\right) \subseteq \widetilde{\mathcal{O}}\left(\frac{|\mathbf{S}'|}{\varepsilon^2} \cdot \sqrt{|\mathbf{\Sigma}_{\mathbf{X}}| \cdot |\mathbf{\Sigma}_{\mathbf{Y}}| \cdot |\mathbf{\Sigma}_{\mathbf{A}}|} \cdot \log \frac{1}{\delta}\right)$$

We omit $\log |\mathbf{A}|$ within $\widetilde{\mathcal{O}}(\cdot)$ because $|\mathbf{A}| \leq |\mathbf{\Sigma}_{\mathbf{A}}|$. ■

C.5. Deferred probabilistic manipulations

In the proof of Lemma 10 and Theorem 6, we skipped the full derivation of

$$\sum_{\mathbf{s}} \mathbb{P}(\mathbf{s}) \cdot \mathbb{P}(\mathbf{y}\mid\mathbf{x},\mathbf{s}) \cdot \sum_{\mathbf{s}'\setminus\mathbf{s}} \mathbb{P}(\mathbf{s}'\setminus\mathbf{s}\mid\mathbf{y},\mathbf{x},\mathbf{s}) = \sum_{\mathbf{s}\cup\mathbf{s}'} \mathbb{P}(\mathbf{s}\cup\mathbf{s}') \cdot \mathbb{P}(\mathbf{y}\mid\mathbf{x},\mathbf{s}\cup\mathbf{s}') \cdot \frac{\mathbb{P}(\mathbf{x}\mid\mathbf{s}\cup\mathbf{s}')}{\mathbb{P}(\mathbf{x}\mid\mathbf{s})}$$

It is obtained via a series of standard probabilistic manipulations:

$$\sum_{\mathbf{s}} \mathbb{P}(\mathbf{s}) \cdot \mathbb{P}(\mathbf{y}\mid\mathbf{x},\mathbf{s}) \cdot \sum_{\mathbf{s}'\setminus\mathbf{s}} \mathbb{P}(\mathbf{s}'\setminus\mathbf{s}\mid\mathbf{y},\mathbf{x},\mathbf{s})$$

$$= \sum_{\mathbf{s}\cup\mathbf{s}'} \mathbb{P}(\mathbf{s}) \cdot \mathbb{P}(\mathbf{y},\mathbf{s}'\setminus\mathbf{s}\mid\mathbf{x},\mathbf{s})$$

(Since $\mathbb{P}(\mathbf{y}\mid\mathbf{x},\mathbf{s}) \cdot \mathbb{P}(\mathbf{s}'\setminus\mathbf{s}\mid\mathbf{y},\mathbf{x},\mathbf{s}) = \mathbb{P}(\mathbf{y},\mathbf{s}'\setminus\mathbf{s}\mid\mathbf{x},\mathbf{s})$)

$$
\begin{aligned}
&= \sum_{\mathbf{s}\cup\mathbf{s}'} \frac{\mathbb{P}(\mathbf{s})}{\mathbb{P}(\mathbf{s} \mid \mathbf{x})} \cdot \mathbb{P}(\mathbf{y}, \mathbf{s} \cup \mathbf{s}' \mid \mathbf{x}) && \text{(Since } \mathbb{P}(\mathbf{y}, \mathbf{s} \cup \mathbf{s}' \mid \mathbf{x}) = \mathbb{P}(\mathbf{s} \mid \mathbf{x}) \cdot \mathbb{P}(\mathbf{y}, \mathbf{s}' \setminus \mathbf{s} \mid \mathbf{x}, \mathbf{s})\text{)} \\
&= \sum_{\mathbf{s}\cup\mathbf{s}'} \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s} \cup \mathbf{s}') \cdot \mathbb{P}(\mathbf{s} \cup \mathbf{s}' \mid \mathbf{x}) \cdot \frac{\mathbb{P}(\mathbf{s})}{\mathbb{P}(\mathbf{s} \mid \mathbf{x})} \\
& && \text{(Since } \mathbb{P}(\mathbf{y}, \mathbf{s} \cup \mathbf{s}' \mid \mathbf{x}) = \mathbb{P}(\mathbf{s} \cup \mathbf{s}' \mid \mathbf{x}) \cdot \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s} \cup \mathbf{s}')\text{)} \\
&= \sum_{\mathbf{s}\cup\mathbf{s}'} \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s} \cup \mathbf{s}') \cdot \mathbb{P}(\mathbf{s} \cup \mathbf{s}') \cdot \frac{\mathbb{P}(\mathbf{s}) \cdot \mathbb{P}(\mathbf{x} \mid \mathbf{s} \cup \mathbf{s}')}{\mathbb{P}(\mathbf{x}) \cdot \mathbb{P}(\mathbf{s} \mid \mathbf{x})} \\
& && \text{(Since } \mathbb{P}(\mathbf{s} \cup \mathbf{s}' \mid \mathbf{x}) = \tfrac{\mathbb{P}(\mathbf{x}, \mathbf{s}\cup\mathbf{s}')}{\mathbb{P}(\mathbf{x})} = \tfrac{\mathbb{P}(\mathbf{s}\cup\mathbf{s}') \cdot \mathbb{P}(\mathbf{x} \mid \mathbf{s}\cup\mathbf{s}')}{\mathbb{P}(\mathbf{x})}\text{)} \\
&= \sum_{\mathbf{s}\cup\mathbf{s}'} \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s} \cup \mathbf{s}') \cdot \mathbb{P}(\mathbf{s} \cup \mathbf{s}') \cdot \frac{\mathbb{P}(\mathbf{x} \mid \mathbf{s} \cup \mathbf{s}')}{\mathbb{P}(\mathbf{x} \mid \mathbf{s})} && \text{(Since } \mathbb{P}(\mathbf{x} \mid \mathbf{s}) = \tfrac{\mathbb{P}(\mathbf{x}) \cdot \mathbb{P}(\mathbf{s} \mid \mathbf{x})}{\mathbb{P}(\mathbf{s})}\text{)} \\
&= \sum_{\mathbf{s}\cup\mathbf{s}'} \mathbb{P}(\mathbf{s} \cup \mathbf{s}') \cdot \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s} \cup \mathbf{s}') \cdot \frac{\mathbb{P}(\mathbf{x} \mid \mathbf{s} \cup \mathbf{s}')}{\mathbb{P}(\mathbf{x} \mid \mathbf{s})} \\
& && \text{(Swap positions of } \mathbb{P}(\mathbf{s} \cup \mathbf{s}') \text{ and } \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s} \cup \mathbf{s}')\text{)}
\end{aligned}
$$

In the proof of Theorem 6, we also skipped the derivations of

$$
\sum_{\mathbf{s}\cup\mathbf{s}'} \mathbb{P}(\mathbf{x} \mid \mathbf{s}\cup\mathbf{s}') \cdot \mathbb{P}(\mathbf{s}\cup\mathbf{s}') \cdot \left| \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s}') - \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s} \cup \mathbf{s}') \right| \leq \varepsilon \quad \text{and} \quad \sum_{\mathbf{s}\cup\mathbf{s}'} \mathbb{P}(\mathbf{s}\cup\mathbf{s}') \cdot \left| \mathbb{P}(\mathbf{x} \mid \mathbf{s}) - \mathbb{P}(\mathbf{x} \mid \mathbf{s} \cup \mathbf{s}') \right| \leq \varepsilon
$$

under the assumptions of $\Delta_{\mathbf{Y} \perp\!\!\!\perp \mathbf{S} \setminus \mathbf{S}' \mid \mathbf{X} \cup \mathbf{S}'} \leq \varepsilon$ and $\Delta_{\mathbf{X} \perp\!\!\!\perp \mathbf{S}' \setminus \mathbf{S} \mid \mathbf{S}} \leq \varepsilon$ respectively. We derive them below:

$$
\begin{aligned}
&\sum_{\mathbf{s}\cup\mathbf{s}'} \mathbb{P}(\mathbf{x} \mid \mathbf{s} \cup \mathbf{s}') \cdot \mathbb{P}(\mathbf{s} \cup \mathbf{s}') \cdot \left| \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s}') - \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s} \cup \mathbf{s}') \right| \\
&= \sum_{\mathbf{s}\cup\mathbf{s}'} \mathbb{P}(\mathbf{x}, \mathbf{s} \cup \mathbf{s}') \cdot \left| \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s}') - \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s} \cup \mathbf{s}') \right| && \text{(Since } \mathbf{X} \cap (\mathbf{S} \cup \mathbf{S}') = \emptyset\text{)} \\
&= \sum_{\mathbf{s}\cup\mathbf{s}'} \mathbb{P}(\mathbf{x}, \mathbf{s}') \cdot \mathbb{P}(\mathbf{s} \setminus \mathbf{s}' \mid \mathbf{x}, \mathbf{s}') \cdot \left| \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s}') - \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s} \cup \mathbf{s}') \right| \\
&= \sum_{\mathbf{s}\cup\mathbf{s}'} \mathbb{P}(\mathbf{x}, \mathbf{s}') \cdot \left| \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s}') \cdot \mathbb{P}(\mathbf{s} \setminus \mathbf{s}' \mid \mathbf{x}, \mathbf{s}') - \mathbb{P}(\mathbf{y} \cup (\mathbf{s} \setminus \mathbf{s}') \mid \mathbf{x}, \mathbf{s}') \right| \\
&\leq \sum_{\mathbf{y}, \mathbf{x}, \mathbf{s}\cup\mathbf{s}'} \mathbb{P}(\mathbf{x}, \mathbf{s}') \cdot \left| \mathbb{P}(\mathbf{y} \mid \mathbf{x}, \mathbf{s}') \cdot \mathbb{P}(\mathbf{s} \setminus \mathbf{s}' \mid \mathbf{x}, \mathbf{s}') - \mathbb{P}(\mathbf{y} \cup (\mathbf{s} \setminus \mathbf{s}') \mid \mathbf{x}, \mathbf{s}') \right| \\
& && \text{(Since we sum over all values of } \mathbf{\Sigma}_{\mathbf{X}} \text{ and } \mathbf{\Sigma}_{\mathbf{Y}}\text{)} \\
&\leq \varepsilon && \text{(when } \Delta_{\mathbf{Y} \perp\!\!\!\perp \mathbf{S} \setminus \mathbf{S}' \mid \mathbf{X} \cup \mathbf{S}'} \leq \varepsilon\text{)}
\end{aligned}
$$

$$
\begin{aligned}
&\sum_{\mathbf{s}\cup\mathbf{s}'} \mathbb{P}(\mathbf{s} \cup \mathbf{s}') \cdot \left| \mathbb{P}(\mathbf{x} \mid \mathbf{s}) - \mathbb{P}(\mathbf{x} \mid \mathbf{s} \cup \mathbf{s}') \right| \\
&= \sum_{\mathbf{s}\cup\mathbf{s}'} \mathbb{P}(\mathbf{s}) \cdot \mathbb{P}(\mathbf{s}' \setminus \mathbf{s} \mid \mathbf{s}) \cdot \left| \mathbb{P}(\mathbf{x} \mid \mathbf{s}) - \mathbb{P}(\mathbf{x} \cup (\mathbf{s}' \setminus \mathbf{s}) \mid \mathbf{s}) \right| \\
&= \sum_{\mathbf{s}\cup\mathbf{s}'} \mathbb{P}(\mathbf{s}) \cdot \left| \mathbb{P}(\mathbf{x} \mid \mathbf{s}) \cdot \mathbb{P}(\mathbf{s}' \setminus \mathbf{s} \mid \mathbf{s}) - \mathbb{P}(\mathbf{x} \cup (\mathbf{s}' \setminus \mathbf{s}) \mid \mathbf{s}) \right|
\end{aligned}
$$

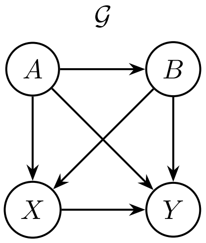
$$\begin{aligned}
&\le \sum_{\mathbf{x},\mathbf{s}\cup\mathbf{s}'} \mathbb{P}(\mathbf{s}) \cdot \left|\mathbb{P}(\mathbf{x}\mid\mathbf{s})\cdot\mathbb{P}(\mathbf{s}'\setminus\mathbf{s}\mid\mathbf{s}) - \mathbb{P}(\mathbf{x}\cup(\mathbf{s}'\setminus\mathbf{s})\mid\mathbf{s})\right| \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{(Since we sum over all values of } \mathbf{\Sigma_X}) \\
&\le \varepsilon \qquad\qquad\qquad\qquad\qquad\qquad \text{(when } \Delta_{\mathbf{X}\perp\mathbf{S}'\setminus\mathbf{S}\mid\mathbf{S}} \le \varepsilon)
\end{aligned}$$

C.6. Deferred derivations for hardness proof

In the proof of Lemma 9, we argued that the distribution $\mathbb{P}$ described in Fig. 4 has the following well-defined conditional probabilities:

a	b	$\mathbb{P}(b \mid a)$	$\mathbb{P}(X=0 \mid a,b)$	$\mathbb{P}(X=0 \mid a)$	$\sum_x \lvert\mathbb{P}(x \mid a,b) - \mathbb{P}(x \mid a)\rvert$
0	0	$\sqrt{\varepsilon}/2$	$1-\alpha+\sqrt{\varepsilon}/2$	$1-\alpha+\varepsilon/4$	$\sqrt{\varepsilon}-\varepsilon/2$
0	1	$1-\sqrt{\varepsilon}/2$	$1-\alpha$	$1-\alpha+\varepsilon/4$	$\varepsilon/2$
1	0	$1-\sqrt{\varepsilon}/2$	α	$\alpha-\varepsilon/4$	$\varepsilon/2$
1	1	$\sqrt{\varepsilon}/2$	$\alpha-\sqrt{\varepsilon}/2$	$\alpha-\varepsilon/4$	$\sqrt{\varepsilon}-\varepsilon/2$

For convenience, we produce Fig. 4 below.



$$A = \begin{cases} 1 & \text{w.p. } \frac{\varepsilon}{4\alpha}\cdot\frac{\alpha-\varepsilon/4}{1-\sqrt{\varepsilon}/2} \\ 0 & \text{else} \end{cases} \qquad X = \begin{cases} A & \text{w.p. } 1-\alpha \\ 1-A & \text{w.p. } \alpha-\sqrt{\varepsilon}/2 \\ B & \text{w.p. } \sqrt{\varepsilon}/2 \end{cases}$$

$$B = \begin{cases} 1-A & \text{w.p. } 1-\sqrt{\varepsilon} \\ 0 & \text{w.p. } \sqrt{\varepsilon}/2 \\ 1 & \text{w.p. } \sqrt{\varepsilon}/2 \end{cases} \qquad Y = \begin{cases} 1 & \text{if } X=0, A=1, B=0 \\ 0 & \text{else} \end{cases}$$

Figure 5: Reproduced: Probability distribution $\mathbb{P}$ defined over 4 binary variables $\{A,B,X,Y\}$ in a topological ordering of $A \prec B \prec X \prec Y$ with parameters ε and α, where $0 < \sqrt{\varepsilon} \le \alpha \le 1/2$.

We first check that all the (conditional) probabilities of $\mathbb{P}$ are well-defined. Since $0 < \sqrt{\varepsilon} \le \alpha \le 1/2$, the only non-straightforward term to verify is $\mathbb{P}(A=1)$. Observe that

$$\frac{\varepsilon}{4\alpha}\cdot\frac{\alpha-\varepsilon/4}{1-\sqrt{\varepsilon}/2} \le 1 \iff \varepsilon\cdot(\alpha-\varepsilon/4) \le 4\alpha\cdot(1-\sqrt{\varepsilon}/2) \iff 2\alpha\sqrt{\varepsilon}+\alpha\varepsilon-\varepsilon^2/4 \le 4\alpha$$

which is true as $0<\varepsilon<\sqrt{\varepsilon}<\alpha\le 1$ implies $2\alpha\sqrt{\varepsilon}+\alpha\varepsilon-\varepsilon^2/4 \le 3\alpha\sqrt{\varepsilon} \le 3\alpha \le 4\alpha$. Therefore, $0 \le \mathbb{P}(A=1) \le 1$.

We now proceed to verify the conditional probabilities shown in the table above. For instance,

$$\begin{aligned}
&\mathbb{P}(X=0 \mid A=0) \\
&= \mathbb{P}(B=0 \mid A=0)\cdot\mathbb{P}(X=0 \mid A=0,B=0) + \mathbb{P}(B=1 \mid A=0)\cdot\mathbb{P}(X=0 \mid A=0,B=1) \\
&= (\sqrt{\varepsilon}/2)\cdot(1-\alpha+\sqrt{\varepsilon}/2) + (1-\sqrt{\varepsilon}/2)\cdot(1-\alpha) \\
&= 1-\alpha+\varepsilon/4
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}(X = 0 \mid A = 1) \\
&= \mathbb{P}(B = 0 \mid A = 1) \cdot \mathbb{P}(X = 0 \mid A = 1, B = 0) + \mathbb{P}(B = 1 \mid A = 1) \cdot \mathbb{P}(X = 0 \mid A = 1, B = 1) \\
&= (1 - \sqrt{\varepsilon}/2) \cdot \alpha + (\sqrt{\varepsilon}/2) \cdot (\alpha - \sqrt{\varepsilon}/2) \\
&= \alpha - \varepsilon/4 \\
&= 1 - \mathbb{P}(X = 0 \mid A = 0)
\end{aligned}$$

The detailed workings for $\sum_x |\mathbb{P}(x \mid a, b) - \mathbb{P}(x \mid a)|$ for different values of $a, b \in \{0, 1\}$ are as follows:

When $A = 0$ and $B = 0$:

$$\begin{aligned}
&\sum_x |\mathbb{P}(x \mid a, b) - \mathbb{P}(x \mid a)| \\
=\ & |\mathbb{P}(X = 0 \mid A = 0, B = 0) - \mathbb{P}(X = 0 \mid A = 0)| + |\mathbb{P}(X = 1 \mid A = 0, B = 0) - \mathbb{P}(X = 1 \mid A = 0)| \\
=\ & \left|(1 - \alpha + \sqrt{\varepsilon}/2) - (1 - \alpha + \varepsilon/4)\right| + \left|(\alpha - \sqrt{\varepsilon}/2) - (\alpha - \varepsilon/4)\right| \\
=\ & 2\left(\sqrt{\varepsilon}/2 - \varepsilon/4\right) \\
=\ & \sqrt{\varepsilon} - \varepsilon/2
\end{aligned}$$

When $A = 0$ and $B = 1$:

$$\begin{aligned}
&\sum_x |\mathbb{P}(x \mid a, b) - \mathbb{P}(x \mid a)| \\
=\ & |\mathbb{P}(X = 0 \mid A = 0, B = 1) - \mathbb{P}(X = 0 \mid A = 0)| + |\mathbb{P}(X = 1 \mid A = 0, B = 1) - \mathbb{P}(X = 1 \mid A = 0)| \\
=\ & |(1 - \alpha) - (1 - \alpha + \varepsilon/4)| + |(\alpha) - (\alpha - \varepsilon/4)| \\
=\ & 2\,(\varepsilon/4) \\
=\ & \varepsilon/2
\end{aligned}$$

When $A = 1$ and $B = 0$:

$$\begin{aligned}
&\sum_x |\mathbb{P}(x \mid a, b) - \mathbb{P}(x \mid a)| \\
=\ & |\mathbb{P}(X = 0 \mid A = 1, B = 0) - \mathbb{P}(X = 0 \mid A = 1)| + |\mathbb{P}(X = 1 \mid A = 1, B = 0) - \mathbb{P}(X = 1 \mid A = 1)| \\
=\ & |(\alpha) - (\alpha - \varepsilon/4)| + |(1 - \alpha) - (1 - \alpha + \varepsilon/4)| \\
=\ & 2\,(\varepsilon/4) \\
=\ & \varepsilon/2
\end{aligned}$$

When $A = 1$ and $B = 1$:

$$\sum_x |\mathbb{P}(x \mid a, b) - \mathbb{P}(x \mid a)|$$

$$
\begin{aligned}
&= |\mathbb{P}(X = 0 \mid A = 1, B = 1) - \mathbb{P}(X = 0 \mid A = 1)| + |\mathbb{P}(X = 1 \mid A = 1, B = 1) - \mathbb{P}(X = 1 \mid A = 1)| \\
&= \left|(\alpha - \sqrt{\varepsilon}/2) - (\alpha - \varepsilon/4)\right| + \left|(1 - \alpha + \sqrt{\varepsilon}/2) - (1 - \alpha + \varepsilon/4)\right| \\
&= 2\left(\sqrt{\varepsilon}/2 - \varepsilon/4\right) \\
&= \sqrt{\varepsilon} - \varepsilon/2
\end{aligned}
$$

C.7. Estimating causal effects using AMBA and BAMBA

If we re-express the results of Theorem 1, Theorem 4 and Theorem 6 in terms of an upper bound on error for a fixed number of samples n, we get the following three corollaries.

Corollary 31 (Estimation corollary) *Suppose we are given (1) failure tolerance $\delta > 0$, (2) n i.i.d. samples from distribution $\mathbb{P}(\mathbf{V})$, (3) a subset $\mathbf{A} \subseteq \mathbf{V}$ with $\alpha_{\mathbf{A}} = \max_{\mathbf{a} \in \mathbf{\Sigma_A}} \mathbb{P}(\mathbf{x} \mid \mathbf{a})$. Then, there is an algorithm that produces an estimate $\widehat{T}_{\mathbf{A},\mathbf{x},\mathbf{y}}$ such that $\Pr(|\widehat{T}_{\mathbf{A},\mathbf{x},\mathbf{y}} - T_{\mathbf{A},\mathbf{x},\mathbf{y}}| \leq \varepsilon) \geq 1 - \delta$ for some error term*

$$
\varepsilon \in \widetilde{\mathcal{O}}\left(\frac{|\mathbf{\Sigma_A}|}{n\alpha_{\mathbf{A}}} + \frac{1}{\sqrt{n\alpha_{\mathbf{A}}}} + \sqrt{\frac{|\mathbf{\Sigma_A}|}{n}} \right)
$$

Proof From Theorem 1, we know that $\widetilde{\mathcal{O}}\left(\left(\frac{|\mathbf{\Sigma_A}|}{\varepsilon\alpha_{\mathbf{A}}} + \frac{1}{\varepsilon^2\alpha_{\mathbf{A}}} + \frac{|\mathbf{\Sigma_A}|}{\varepsilon^2}\right) \cdot \log\left(\frac{1}{\delta}\right)\right)$ samples suffice to produce an estimate $\widehat{T}_{\mathbf{A},\mathbf{x},\mathbf{y}}$ such that $\Pr(|\widehat{T}_{\mathbf{A},\mathbf{x},\mathbf{y}} - T_{\mathbf{A},\mathbf{x},\mathbf{y}}| \leq \varepsilon) \geq 1 - \delta$. Ignoring the logarithmic terms and constant factors, the result follows by re-expressing $n \leq \frac{|\mathbf{\Sigma_A}|}{\varepsilon\alpha_{\mathbf{A}}} + \frac{1}{\varepsilon^2\alpha_{\mathbf{A}}} + \frac{|\mathbf{\Sigma_A}|}{\varepsilon^2}$ in terms of ε. ■

Corollary 32 (AMBA corollary) *Suppose we are given (1) failure tolerance $\delta > 0$, (2) n i.i.d. samples from distribution $\mathbb{P}(\mathbf{V})$, and (3) an arbitrary subset $\mathbf{A} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$. Then, there is an algorithm that produces a subset $\mathbf{S} \subseteq \mathbf{A}$ such that $\Pr\left(\Delta_{\mathbf{X} \perp\!\!\!\perp \mathbf{A} \setminus \mathbf{S} \mid \mathbf{S}} > \varepsilon\right) \geq 1 - \delta$ and $\Pr\left(|T_{\mathbf{S},\mathbf{x},\mathbf{y}} - T_{\mathbf{A},\mathbf{x},\mathbf{y}}| \leq \varepsilon\right) \geq 1 - \delta$ for some error term*

$$
\varepsilon \in \widetilde{\mathcal{O}}\left(\frac{1}{\alpha_{\mathbf{S}}} \cdot \sqrt{\frac{|\mathbf{S}|}{n}} \cdot (|\mathbf{\Sigma_X}| \cdot |\mathbf{\Sigma_A}|)^{\frac{1}{4}} \right)
$$

Proof From Theorem 4, we know that $\widetilde{\mathcal{O}}\left(\frac{|\mathbf{S}|}{\varepsilon^2} \cdot \sqrt{|\mathbf{\Sigma_X}| \cdot |\mathbf{\Sigma_A}|} \cdot \log\frac{1}{\delta}\right)$ samples suffice to produce a subset $\mathbf{S} \subseteq \mathbf{A}$ such that $\Pr\left(\Delta_{\mathbf{X} \perp\!\!\!\perp \mathbf{A} \setminus \mathbf{S} \mid \mathbf{S}} > \varepsilon\right) \geq 1 - \delta$ and $\Pr\left(|T_{\mathbf{S},\mathbf{x},\mathbf{y}} - T_{\mathbf{A},\mathbf{x},\mathbf{y}}| \leq \frac{\varepsilon}{\alpha_{\mathbf{S}}}\right) \geq 1 - \delta$. Ignoring the logarithmic terms and constant factors, the result follows by re-expressing $n = \frac{|\mathbf{S}|}{(\varepsilon')^2} \cdot \sqrt{|\mathbf{\Sigma_X}| \cdot |\mathbf{\Sigma_A}|}$ in terms of $\varepsilon' = \varepsilon\alpha_{\mathbf{S}} \leq \varepsilon$. ■

Corollary 33 (BAMBA corollary) *Suppose we are given (1) failure tolerance $\delta > 0$, (2) n i.i.d. samples from distribution $\mathbb{P}(\mathbf{V})$, (3) an arbitrary subset $\mathbf{A} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$, and (4) an ε-Markov blanket $\mathbf{S} \subseteq \mathbf{A}$. Then, there is an algorithm that produces a subset $\mathbf{S}' \subseteq \mathbf{A}$ such that $|\mathbf{\Sigma_{S'}}| \leq |\mathbf{\Sigma_S}|$ and $\Pr\left(|T_{\mathbf{S}',\mathbf{x},\mathbf{y}} - T_{\mathbf{A},\mathbf{x},\mathbf{y}}| \leq \varepsilon\right) \geq 1 - \delta$ for some error term*

$$
\varepsilon \in \widetilde{\mathcal{O}}\left(\frac{1}{\alpha_{\mathbf{S}}} \cdot \sqrt{\frac{|\mathbf{S}'|}{n}} \cdot (|\mathbf{\Sigma_X}| \cdot |\mathbf{\Sigma_Y}| \cdot |\mathbf{\Sigma_A}|)^{\frac{1}{4}} \right)
$$

Proof From Theorem 6, we know that $\widetilde{\mathcal{O}}\left(\frac{|\mathbf{S}'|}{\varepsilon^2} \cdot \sqrt{|\mathbf{\Sigma_X}| \cdot |\mathbf{\Sigma_Y}| \cdot |\mathbf{\Sigma_A}|} \cdot \log \frac{1}{\delta}\right)$ samples suffice to produce a subset $\mathbf{S}' \subseteq \mathbf{A}$ such that $|\mathbf{\Sigma}_{\mathbf{S}'}| \leq |\mathbf{\Sigma_S}|$ and $\Pr\left(|T_{\mathbf{S}',\mathbf{x},\mathbf{y}} - T_{\mathbf{A},\mathbf{x},\mathbf{y}}| \leq \frac{\varepsilon}{\alpha_\mathbf{S}}\right) \geq 1 - \delta$. Ignoring the logarithmic terms and constant factors, the result follows by re-expressing $n = \frac{|\mathbf{S}'|}{(\varepsilon')^2} \cdot \sqrt{|\mathbf{\Sigma_X}| \cdot |\mathbf{\Sigma_Y}| \cdot |\mathbf{\Sigma_A}|}$ in terms of $\varepsilon' = \varepsilon \alpha_\mathbf{S} \leq \varepsilon$. ■

In light of Corollary 31, Corollary 32, and Corollary 33, there are a couple of ways one could attempt to estimate $\mathbb{P}_\mathbf{x}(\mathbf{y})$ when given a valid adjustment set $\mathbf{Z} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$:

1. Directly estimate using $\mathbf{Z}$. By Corollary 31, this yields an error of

$$|\mathbb{P}_\mathbf{x}(\mathbf{y}) - \widehat{\mathbb{P}}_\mathbf{x}(\mathbf{y})| = |T_{\mathbf{Z},\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{Z},\mathbf{x},\mathbf{y}}| \in \widetilde{\mathcal{O}}\left(\frac{|\mathbf{\Sigma_Z}|}{n\alpha_\mathbf{Z}} + \frac{1}{\sqrt{n\alpha_\mathbf{Z}}} + \sqrt{\frac{|\mathbf{\Sigma_Z}|}{n}}\right)$$

2. Use AMBA on $\mathbf{Z}$ to produce a subset $\mathbf{S} \subseteq \mathbf{Z}$ and estimate using $\mathbf{S}$. By Corollary 31 and Corollary 32, this yields an error of

$$\begin{aligned}|\mathbb{P}_\mathbf{x}(\mathbf{y}) - \widehat{\mathbb{P}}_\mathbf{x}(\mathbf{y})| &= |T_{\mathbf{Z},\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{S},\mathbf{x},\mathbf{y}}| \leq |T_{\mathbf{Z},\mathbf{x},\mathbf{y}} - T_{\mathbf{S},\mathbf{x},\mathbf{y}}| + |T_{\mathbf{S},\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{S},\mathbf{x},\mathbf{y}}| \\ &\in \widetilde{\mathcal{O}}\left(\frac{1}{\alpha_\mathbf{S}} \cdot \sqrt{\frac{|\mathbf{S}|}{n}} \cdot (|\mathbf{\Sigma_X}| \cdot |\mathbf{\Sigma_Z}|)^{\frac{1}{4}} + \frac{|\mathbf{\Sigma_S}|}{n\alpha_\mathbf{S}} + \frac{1}{\sqrt{n\alpha_\mathbf{S}}} + \sqrt{\frac{|\mathbf{\Sigma_S}|}{n}}\right)\end{aligned}$$

3. Use AMBA on $\mathbf{Z}$ to produce a subset $\mathbf{S} \subseteq \mathbf{Z}$, then use BAMBA to further produce subset $\mathbf{S}'$, and then estimate using $\mathbf{S}'$. By Corollary 31, Corollary 32, and Corollary 33, this yields an error of

$$|\mathbb{P}_\mathbf{x}(\mathbf{y}) - \widehat{\mathbb{P}}_\mathbf{x}(\mathbf{y})| = |T_{\mathbf{Z},\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{S},\mathbf{x},\mathbf{y}}| \leq |T_{\mathbf{Z},\mathbf{x},\mathbf{y}} - T_{\mathbf{S}',\mathbf{x},\mathbf{y}}| + |T_{\mathbf{S}',\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{S}',\mathbf{x},\mathbf{y}}|$$

$$\in \widetilde{\mathcal{O}}\left(\frac{1}{\alpha_\mathbf{S}} \cdot \sqrt{\frac{|\mathbf{S}|}{n}} \cdot (|\mathbf{\Sigma_X}| \cdot |\mathbf{\Sigma_Z}|)^{\frac{1}{4}} + \frac{1}{\alpha_\mathbf{S}} \cdot \sqrt{\frac{|\mathbf{S}'|}{n}} \cdot (|\mathbf{\Sigma_X}| \cdot |\mathbf{\Sigma_Y}| \cdot |\mathbf{\Sigma_Z}|)^{\frac{1}{4}} + \frac{|\mathbf{\Sigma}_{\mathbf{S}'}|}{n\alpha_{\mathbf{S}'}} + \frac{1}{\sqrt{n\alpha_{\mathbf{S}'}}} + \sqrt{\frac{|\mathbf{\Sigma}_{\mathbf{S}'}|}{n}}\right)$$

In any cases 2 and 3, with appropriate constant factors, we see that

$$|\mathbb{P}_\mathbf{x}(\mathbf{y}) - \widehat{\mathbb{P}}_\mathbf{x}(\mathbf{y})| = |T_{\mathbf{Z},\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{S},\mathbf{x},\mathbf{y}}| \leq |T_{\mathbf{Z},\mathbf{x},\mathbf{y}} - T_{\mathbf{S},\mathbf{x},\mathbf{y}}| + |T_{\mathbf{S},\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{S},\mathbf{x},\mathbf{y}}| \leq \varepsilon + \varepsilon = 2\varepsilon$$

The following lemma tells us that $\alpha_\mathbf{Z} \leq \alpha_\mathbf{S}$ and $\alpha_\mathbf{Z} \leq \alpha_{\mathbf{S}'}$, i.e. $\frac{1}{\alpha_\mathbf{S}} \leq \frac{1}{\alpha_\mathbf{Z}}$ and $\frac{1}{\alpha_{\mathbf{S}'}} \leq \frac{1}{\alpha_\mathbf{Z}}$, so it is always beneficial to use a smaller subset with respect to the error incurred by estimation in Corollary 31.

Lemma 34 *For any value* $\mathbf{x}$ *for* $\mathbf{X}$ *and subsets* $\mathbf{A} \subseteq \mathbf{B} \subseteq \mathbf{V} \setminus \mathbf{X}$, *we have*

$$\alpha_\mathbf{A} = \min_\mathbf{a} \mathbb{P}(\mathbf{x} \mid \mathbf{a}) \geq \min_\mathbf{b} \mathbb{P}(\mathbf{x} \mid \mathbf{b}) = \alpha_\mathbf{B}$$

Proof Fix an arbitrary values of $\mathbf{x}$ for $\mathbf{X}$ and $\mathbf{a}$ for $\mathbf{A}$, we see that

$$\mathbb{P}(\mathbf{x} \mid \mathbf{a}) = \sum_{\mathbf{b} \setminus \mathbf{a}} \mathbb{P}(\mathbf{x}, \mathbf{b} \setminus \mathbf{a} \mid \mathbf{a}) \geq \min_\mathbf{b} \mathbb{P}(\mathbf{x} \mid \mathbf{b}) \cdot \sum_{\mathbf{b} \setminus \mathbf{a}} \mathbb{P}(\mathbf{b} \setminus \mathbf{a} \mid \mathbf{a}) = \min_\mathbf{b} \mathbb{P}(\mathbf{x} \mid \mathbf{b})$$

Therefore, $\min_{\mathbf{a}} \mathbb{P}(\mathbf{x} \mid \mathbf{a}) \geq \min_{\mathbf{b}} \mathbb{P}(\mathbf{x} \mid \mathbf{b})$. ■

Observe that $\frac{1}{\alpha_{\mathbf{S}}} \leq \frac{1}{\alpha_{\mathbf{Z}}}$ from Lemma 34 and $|\mathbf{\Sigma_S}| \leq |\mathbf{\Sigma_Z}|$ since $\mathbf{S} \subseteq \mathbf{Z}$. So, the second approach of estimating $\mathbb{P}_{\mathbf{x}}(\mathbf{y})$ using the subset $\mathbf{S} \subseteq \mathbf{Z}$ produced by AMBA would yield an asymptotically smaller error than directly using $\mathbf{Z}$ whenever $\frac{1}{\alpha_{\mathbf{S}}} \cdot \sqrt{\frac{|\mathbf{S}|}{n}} \cdot (|\mathbf{\Sigma_X}| \cdot |\mathbf{\Sigma_Z}|)^{\frac{1}{4}} \leq \frac{|\mathbf{\Sigma_Z}|}{n\alpha_{\mathbf{S}}} + \frac{1}{\sqrt{n\alpha_{\mathbf{S}}}} + \sqrt{\frac{|\mathbf{\Sigma_Z}|}{n}}$. This happens when

$$|\mathbf{S}| \cdot \sqrt{\frac{|\mathbf{\Sigma_X}|}{|\mathbf{\Sigma_Z}|}} < \max\left\{\frac{|\mathbf{\Sigma_Z}|}{n}, \frac{\alpha_{\mathbf{S}}}{|\mathbf{\Sigma_Z}|}, \alpha_{\mathbf{S}}^2\right\} \tag{15}$$

Observe that we know all terms in Eq. (15) except for $\alpha_{\mathbf{S}}$. For small n, say when $n \ll |\mathbf{\Sigma_Z}|$, the first term justifies estimating using the subset $\mathbf{S}$ produced by AMBA instead of directly estimating using $\mathbf{Z}$. However, for large n, one would need to make the decision based on $\alpha_{\mathbf{S}}$. A similar kind of decision has to be made whether the third approach, of running AMBA to produce $\mathbf{S} \subseteq \mathbf{Z}$ then BAMBA to produce $\mathbf{S}' \subseteq \mathbf{Z}$, would yield a smaller estimation error. Note that $|\mathbf{\Sigma_{S'}}| \leq |\mathbf{\Sigma_S}|$ would imply $|\mathbf{S}'| \leq |\mathbf{S}|$ when all variables have the same domain size.

Theorem 7 (PAC causal effect estimation with positivity) *Suppose we are given (1) $\varepsilon > 0$, (2) $\delta > 0$, (3) n i.i.d. samples from $\mathbb{P}(\mathbf{V})$, (4) an interventional query $\mathbb{P}_{\mathbf{x}}(\mathbf{y})$, (5) a valid adjustment set $\mathbf{Z} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$, and (6) guaranteed that $\alpha_{\mathbf{S}} \geq \alpha \in (0, 1)$ for any $\mathbf{S} \subseteq \mathbf{Z}$. Then, there is an algorithm that outputs a subset $\mathbf{S}^* \subseteq \mathbf{Z}$ and an estimate $\widehat{\mathbb{P}}_{\mathbf{x}}(\mathbf{y}) = \widehat{T}_{\mathbf{S}^*,\mathbf{x},\mathbf{y}}$ such that $\Pr\left(\left|\widehat{\mathbb{P}}_{\mathbf{x}}(\mathbf{y}) - \mathbb{P}_{\mathbf{x}}(\mathbf{y})\right| \leq \varepsilon\right) \geq 1 - \delta$ for some error term*

$$\varepsilon \in \widetilde{\mathcal{O}}\left(\frac{1}{n} \cdot \frac{|\mathbf{\Sigma_{S^*}}|}{\alpha} + \frac{1}{\sqrt{n}} \cdot \left(\frac{\sqrt{|\mathbf{Z}|} \cdot (|\mathbf{\Sigma_X}| \cdot |\mathbf{\Sigma_Y}| \cdot |\mathbf{\Sigma_Z}|)^{\frac{1}{4}}}{\alpha} + \frac{1}{\sqrt{\alpha}} + \sqrt{|\mathbf{\Sigma_{S^*}}|}\right)\right).$$

Moreover, if there exists a Markov blanket $\mathbf{S}$ of $\mathbf{X}$ such that $|\mathbf{S}| \cdot \sqrt{\frac{|\mathbf{\Sigma_X}|}{|\mathbf{\Sigma_Z}|}} < \max\left\{\frac{|\mathbf{\Sigma_Z}|}{n}, \frac{\alpha_{\mathbf{S}}}{|\mathbf{\Sigma_Z}|}, \alpha_{\mathbf{S}}^2\right\}$, then $|\mathbf{S}^*| \leq |\mathbf{S}|$. Note that we hide a dependency on $\log(1/\delta)$ using the $\widetilde{\mathcal{O}}(\cdot)$ notation for readability.

Proof Consider the following algorithm:

1. Run AMBA to obtain $\mathbf{S} \subseteq \mathbf{Z}$
2. Check if $|\mathbf{S}| \cdot \sqrt{\frac{|\mathbf{\Sigma_X}|}{|\mathbf{\Sigma_Z}|}} < \max\left\{\frac{|\mathbf{\Sigma_Z}|}{n}, \frac{\alpha_{\mathbf{S}}}{|\mathbf{\Sigma_Z}|}, \alpha_{\mathbf{S}}^2\right\}$ according to Eq. (15)
3. If so, run BAMBA to obtain $\mathbf{S}' \subseteq \mathbf{Z}$ and produce estimate $\widehat{\mathbb{P}}_{\mathbf{x}}(\mathbf{y}) = \widehat{T}_{\mathbf{S}',\mathbf{x},\mathbf{y}}$
4. Otherwise, produce estimate $\widehat{\mathbb{P}}_{\mathbf{x}}(\mathbf{y}) = \widehat{T}_{\mathbf{Z},\mathbf{x},\mathbf{y}}$

That is, depending on Eq. (15), we decide to perform estimation based on $\mathbf{S}^* = \mathbf{S}'$ or $\mathbf{S}^* = \mathbf{Z}$. It remains to show that the bound holds for each case separately while noting that $\alpha_{\mathbf{S}}, \alpha_{\mathbf{S}'}, \alpha_{\mathbf{Z}} \geq \alpha$.

Case 1: $|\mathbf{S}| \cdot \sqrt{\frac{|\mathbf{\Sigma_X}|}{|\mathbf{\Sigma_Z}|}} < \max\left\{\frac{|\mathbf{\Sigma_Z}|}{n}, \frac{\alpha_{\mathbf{S}}}{|\mathbf{\Sigma_Z}|}, \alpha_{\mathbf{S}}^2\right\}$, so we estimate using $\mathbf{S}^* = \mathbf{S}'$ produced from BAMBA

This incurs an error of

$$|\mathbb{P}_{\mathbf{x}}(\mathbf{y}) - \widehat{\mathbb{P}}_{\mathbf{x}}(\mathbf{y})| = |T_{\mathbf{Z},\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{S},\mathbf{x},\mathbf{y}}| \leq |T_{\mathbf{Z},\mathbf{x},\mathbf{y}} - T_{\mathbf{S}',\mathbf{x},\mathbf{y}}| + |T_{\mathbf{S}',\mathbf{x},\mathbf{y}} - \widehat{T}_{\mathbf{S}',\mathbf{x},\mathbf{y}}|$$

$$\in \widetilde{\mathcal{O}}\left(\frac{1}{\alpha}\cdot\sqrt{\frac{|\mathbf{S}|}{n}}\cdot(|\mathbf{\Sigma_X}|\cdot|\mathbf{\Sigma_Z}|)^{\frac{1}{4}}+\frac{1}{\alpha}\cdot\sqrt{\frac{|\mathbf{S}'|}{n}}\cdot(|\mathbf{\Sigma_X}|\cdot|\mathbf{\Sigma_Y}|\cdot|\mathbf{\Sigma_Z}|)^{\frac{1}{4}}+\frac{|\mathbf{\Sigma_{S'}}|}{n\alpha}+\frac{1}{\sqrt{n\alpha}}+\sqrt{\frac{|\mathbf{\Sigma_{S'}}|}{n}}\right)$$

(From Corollary 31, Corollary 32, and Corollary 33)

$$\subseteq \widetilde{\mathcal{O}}\left(\frac{1}{\alpha}\cdot\sqrt{\frac{|\mathbf{Z}|}{n}}\cdot(|\mathbf{\Sigma_X}|\cdot|\mathbf{\Sigma_Y}|\cdot|\mathbf{\Sigma_Z}|)^{\frac{1}{4}}+\frac{|\mathbf{\Sigma_{S'}}|}{n\alpha}+\frac{1}{\sqrt{n\alpha}}+\sqrt{\frac{|\mathbf{\Sigma_{S'}}|}{n}}\right)$$

(Since $\max\{|\mathbf{S}|,|\mathbf{S}'|\}\leq|\mathbf{Z}|$)

$$\subseteq \widetilde{\mathcal{O}}\left(\frac{1}{n}\cdot\frac{|\mathbf{\Sigma_{S^*}}|}{\alpha}+\frac{1}{\sqrt{n}}\cdot\left(\frac{\sqrt{|\mathbf{Z}|}\cdot(|\mathbf{\Sigma_X}|\cdot|\mathbf{\Sigma_Y}|\cdot|\mathbf{\Sigma_Z}|)^{\frac{1}{4}}}{\alpha}+\frac{1}{\sqrt{\alpha}}+\sqrt{|\mathbf{\Sigma_{S^*}}|}\right)\right)$$

(Since $\mathbf{S}^*=\mathbf{S}'$)

Case 2: $|\mathbf{S}|\cdot\sqrt{\frac{|\mathbf{\Sigma_X}|}{|\mathbf{\Sigma_Z}|}}\geq\max\left\{\frac{|\mathbf{\Sigma_Z}|}{n},\frac{\alpha_{\mathbf{S}}}{|\mathbf{\Sigma_Z}|},\alpha_{\mathbf{S}}^2\right\}$, so we estimate using $\mathbf{S}^*=\mathbf{Z}$
This incurs an error of

$$|\mathbb{P}_{\mathbf{x}}(\mathbf{y})-\widehat{\mathbb{P}}_{\mathbf{x}}(\mathbf{y})|=|T_{\mathbf{Z},\mathbf{x},\mathbf{y}}-\widehat{T}_{\mathbf{Z},\mathbf{x},\mathbf{y}}|$$

$$\in \widetilde{\mathcal{O}}\left(\frac{|\mathbf{\Sigma_Z}|}{n\alpha}+\frac{1}{\sqrt{n\alpha}}+\sqrt{\frac{|\mathbf{\Sigma_Z}|}{n}}\right)$$

(From Corollary 31)

$$\subseteq \widetilde{\mathcal{O}}\left(\frac{|\mathbf{\Sigma_{S^*}}|}{n\alpha}+\frac{1}{\sqrt{n\alpha}}+\sqrt{\frac{|\mathbf{\Sigma_{S^*}}|}{n}}\right)$$

(Since $\mathbf{S}^*=\mathbf{S}'$)

$$\subseteq \widetilde{\mathcal{O}}\left(\frac{1}{n}\cdot\frac{|\mathbf{\Sigma_{S^*}}|}{\alpha}+\frac{1}{\sqrt{n}}\cdot\left(\frac{\sqrt{|\mathbf{Z}|}\cdot(|\mathbf{\Sigma_X}|\cdot|\mathbf{\Sigma_Y}|\cdot|\mathbf{\Sigma_Z}|)^{\frac{1}{4}}}{\alpha}+\frac{1}{\sqrt{\alpha}}+\sqrt{|\mathbf{\Sigma_{S^*}}|}\right)\right)$$

(Adding more terms)

Therefore, we see that the error upper bound holds for either case. ■

Appendix D. Experimental results

For the purposes of empirically demonstrating and evaluating our algorithm, we use a simple setting with a single treatment variable and a single outcome variable, i.e., $\mathbf{X}=\{X\}$ and $\mathbf{Y}=\{Y\}$; thus, we use unbolded fonts in this section. We run our experiments in the standard statistical setting, where each algorithm is provided with a dataset of n samples, rather than being given sample access to the distribution $\mathbb{P}(\mathbf{V})$. The code for replicating our experiments can be found at `https://github.com/csquires/amba-bamba-clear2025`.

In Section D.1, we describe the baseline algorithm against which we compare, along with some details about how these algorithms are implemented. In Section D.2, we describe our procedure for synthetic data generation, and in Section D.3, we describe our evaluation metrics and our approach for hyperparameter selection. We conclude in Section D.4 by discussing our results.

D.1. Comparisons and Implementation Details

We compare our AMBA and BAMBA methods to three other methods:

- Z-ADJUST: Adjust by the given adjustment set $\mathbf{Z}$

- MB-ADJUST: Adjust by the Markov blanket of X in the (unknown) data-generating graph $\mathcal{G}$
- MIN-ADJUST: Adjust by the minimum-sized adjustment set for $\mathbb{P}_x(y)$ in the (unknown) data-generating graph $\mathcal{G}$

We note that MB-ADJUST and MIN-ADJUST require *oracle* knowledge of the graph and are not possible to run in practical (non-synthetic) scenarios; we have included them here for the sake of comparison.

Implementation details One of our focuses in this empirical demonstration is to slightly adapt the methods to better reflect how they are used in practice. Our methods and others use two key subroutines: *conditional independence testing* and *estimation by covariate adjustment*, which both require certain design choices.

First, let us describe our estimators of marginal and conditional probabilities used for both subroutines. Let $\mathbf{A}$, $\mathbf{B}$ be disjoint sets. For each $\mathbf{a}$ and $\mathbf{b}$, let $n_{\mathbf{a}}$ denote the number of samples where $\mathbf{A} = \mathbf{a}$, and let $n_{\mathbf{a},\mathbf{b}}$ denote the number of samples where $\mathbf{A} = \mathbf{a}$ and $\mathbf{B} = \mathbf{b}$. Then, given $\beta \geq 0$, for all $\mathbf{a}$ and $\mathbf{b}$, we define

$$\widehat{P}_\beta(\mathbf{a}) := \frac{n_{\mathbf{a}} + \beta}{n + \beta \cdot |\mathbf{\Sigma}_{\mathbf{A}}|} \tag{16}$$

and

$$\widehat{P}_\beta(\mathbf{a} \mid \mathbf{b}) := \frac{n_{\mathbf{a},\mathbf{b}} + \beta}{n_{\mathbf{b}} + \beta \cdot |\mathbf{\Sigma}_{\mathbf{A}}|}, \tag{17}$$

where $\frac{0}{0} := 0$. This technique, which can be interpreted as adding "pseudocounts" for each value of $\mathbf{A}$, is known as *additive* or *Laplace* smoothing, and is well-known to have favorable statistical properties when β is sufficiently small (Kamath et al., 2015). Hence, it is likely that our theory extends to such an estimator. For the sake of estimation by covariate adjustment, given the set $\mathbf{A}$, we use the following estimator of $T_{\mathbf{A},x,y}$:

$$\widehat{T}_{\mathbf{A},x,y} := \sum_{\mathbf{a}} \widehat{P}_\beta(\mathbf{a}) \cdot \widehat{P}_\beta(\mathbf{y} \mid \mathbf{a}, x).$$

For the sake of conditional independence testing, we find that a simple *plug-in* estimator of $\Delta_{\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}}$ worked better than the more complicated conditional independence testers cited in our theoretical analysis. In particular, we use the following estimator:

$$\widehat{\Delta}_{\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}} := \sum_{\mathbf{c}} \widehat{P}_0(\mathbf{c}) \cdot \left| \widehat{P}_0(\mathbf{a}, \mathbf{b} \mid \mathbf{c}) - \widehat{P}_0(\mathbf{a} \mid \mathbf{c}) \cdot \widehat{P}_0(\mathbf{b} \mid \mathbf{c}) \right|.$$

Then, we have a hyperparameter τ controlling the threshold for conditional independence; i.e., $\widehat{\Delta}_{\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}} > \tau$, we reject conditional independence. For a given choice of $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$, this estimator requires computing the empirical counts $n_{\mathbf{a},\mathbf{b},\mathbf{c}}$. In our algorithms, such computations can often be re-used, e.g. $n_{\mathbf{a},\mathbf{b},\mathbf{c}} = n_{\mathbf{a}',\mathbf{b}',\mathbf{c}'}$ whenever $\mathbf{A} \cup \mathbf{B} \cup \mathbf{C} = \mathbf{A}' \cup \mathbf{B}' \cup \mathbf{C}'$. Thus, our implementation performs *memoization* of such empirical counts to reduce computational complexity.

Finally, MIN-ADJUST requires an algorithm for finding the minimum-sized adjustment set. For this step, we use the `optimaladj` package, which implements the methods described in Smucler and Rotnitzky (2022).

D.2. Synthetic data generation

We use a single 10-node DAG structure $\mathcal{G}$ inspired by Fig. 1, with $M = 50$ different distributions $(P_m(\mathbf{V}))_{m=1}^{50}$ on this structure. For each $m = 1, 2, \ldots, 50$, we sample $D = 10$ datasets, each of size $n = 500$. In more detail:

> *Structure.* We use a fixed 10-node DAG $\mathcal{G}$ with $\mathbf{X} = \{X\}$, $\mathbf{Y} = \{Y\}$, $|\mathbf{Z}| = 8$, $|\text{pa}(x)| = 6$, and $|\mathbf{S}_{\text{min}}| = 2$, where $\mathbf{S}_{\text{min}}$ is the unique minimum-sized valid adjustment set for $P_x(y)$. In particular, the DAG has the following edges and no others: $X \to Y$, $V \to X$ for each $V \in \text{pa}(X)$, $W \to V$ for each $W \in \mathbf{S}_{\text{min}}$ and $V \in \text{pa}(X)$, and $W \to Y$ for each $W \in \mathbf{S}_{\text{min}}$.
>
> *Distribution, Step 1: Sampling from a beta prior.* For each $m = 1, 2, \ldots, 50$, we construct a distribution $\mathbb{P}^{(m)}(\mathbf{V})$ as follows. We take all variables in $\mathbf{V}$ to be binary. For each $V_i \in \mathbf{V}$ and each value $\mathbf{w} \in \{0,1\}^{|\text{pa}(V_i)|}$, we sample $p_{\mathbf{w}} \sim \text{Beta}(1,1)$, and assign $\mathbb{P}^{(m)}(V_i \mid \text{pa}(V_i) = \mathbf{w}) = p_{\mathbf{w}}$.
>
> *Distribution, Step 2: Tilting the distribution of Y.* By symmetries of the beta distribution, the true causal effect $\mathbb{P}_x^{(m)}(y)$ is likely to be close to 0.5; this creates and artificial bias towards estimator with (implicit) shrinkage toward 0.5. Thus, we *tilt* the distribution of Y toward higher values by taking $\widetilde{\mathbb{P}}^{(m)}(Y \mid \text{pa}(Y) = \mathbf{w}) \propto \mathbb{P}^{(m)}(Y \mid \text{pa}(Y) = \mathbf{w}) \cdot e^{2Y}$ for all $\mathbf{w} \in \{0,1\}^{|\text{pa}(Y)|}$.
>
> *Samples.* For each $m = 1, 2, \ldots, 50$, we draw $D = 10$ datasets of $n = 500$ samples, giving us an indexed family $\{\mathcal{D}^{(m,d)}\}_{m=1,d=1}^{50,10}$ of datasets.

D.3. Evaluation and hyperparameter selection

Empirical MSE Let $\text{ALG} \in \{\text{Z-ADJUST}, \text{MB-ADJUST}, \ldots\}$. For each $m \in [50]$ and $d \in [10]$, let $\widehat{P}^{(m,d)}(\text{ALG})$ denote the causal effect estimated by ALG when given dataset $\mathcal{D}^{(m,d)}$ as input. Then, for each $m \in [50]$, we compute the *empirical mean squared error* of ALG as follows:

$$\widehat{\text{MSE}}_m(\text{ALG}) = \frac{1}{50} \sum_{m=1}^{50} \left(\widehat{P}^{(m,d)}(\text{ALG}) - P_x^{(m)}(y) \right)^2$$

Note that, for each instance $m = 1, 2, \ldots, 50$, the difficulty of the estimation problem may be different, i.e., $\widehat{\text{MSE}}_m(\text{ALG})$ may depend as much on m as on ALG. This observation informs how we perform hyperparameter selection and report our final evaluation metric.

Hyperparameter selection To fairly select the best hyperparameters for each algorithm, we perform a simple grid search on a set of "hyperparameter selection" datasets $\{\mathcal{D}^{(m',d)}\}_{m'=51,d=1}^{100,10}$ that are only used for this step, and not for our final evaluation.[4]

Fix an algorithm ALG, and let ALG_η denote the algorithm run with hyperparameters η. For each η in the grid and for each $m' = 51, 52, \ldots, 100$, we compute $\widehat{\text{MSE}}_{m'}(\text{ALG}_\eta)$. Then, for each η in

[4] This procedure ensures a good *upper bound* on the best possible performance of each algorithm, and reflects real-world scenarios in which either (i) the practitioner has domain knowledge about which hyperparameters are likely to perform well, or (ii) the practitioner uses an effective method for hyperparameter selection. In general, data-driven hyperparameter selection is an interesting and difficult problem for *all* of these approaches, not just AMBA and BAMBA.

Algorithm	Smoothing coefficient β	CI threshold τ
Z-ADJUST	0.1	-
MB-ADJUST	0.1	-
MIN-ADJUST	0.1	-
AMBA	0.1	0.2
BAMBA	0.1	0.1

Table 1: **Selected hyperparameters for the tested algorithms.**

the grid, we compute the number of instances where that η performed better than all other values η' in the grid, and we select the η with the largest fraction of "best performances".

The possible hyperparameters for our algorithms are the smoothing coefficient β (for Eq. (16) and Eq. (17)) and the conditional independence threshold τ. We search over possible values $\beta \in \{0.1, 0.25, 0.5, 1.5, 2\}$ and $\tau \in \{0.05, 0.075, 0.1, 0.2\}$. The selected hyperparameters for all algorithms are given in Table 1.

Final evaluation metric Using the best hyperparameter values, we normalize the MSEs of each algorithm to account for the difficulty of the instance. In particular, we define

$$\widehat{\text{N-MSE}}_m(\textsc{Alg}) = \frac{\widehat{\text{MSE}}_m(\textsc{Alg})}{\widehat{\text{MSE}}_m(\textsc{Baseline})},$$

where we select BASELINE $=$ Z-ADJUST. In particular, for all $m = 1, \ldots, 50$, we always have $\widehat{\text{N-MSE}}_m(\textsc{Z-Adjust}) = 1$.

D.4. Results

In Figure 6, we demonstrate the performance of our algorithms and the baselines on our synthetic datasets. The performance levels match our theoretical predictions: AMBA has lower error that Z-ADJUST, since it typically uses a smaller adjustment set; similarly, BAMBA has lower error than AMBA.

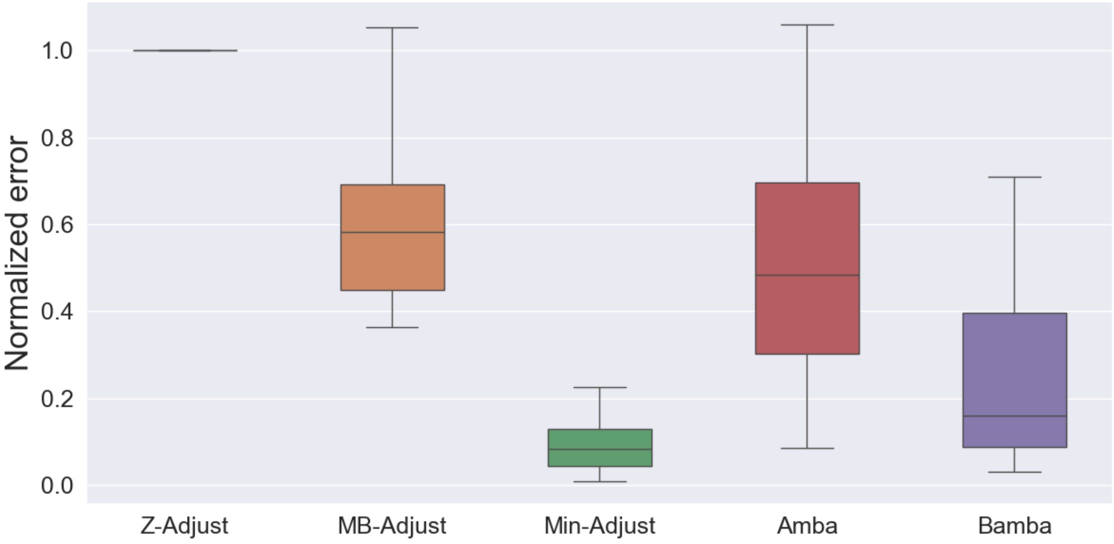


Figure 6: BAMBA **performs the best of any non-oracle algorithm on our synthetic evaluation.** We compared AMBA and BAMBA to four other approaches, including two approaches with additional oracle information (MB-ADJUST and MIN-ADJUST). The middle denotes the median, the lower and upper ends of the box denote the 25% and 75% quantiles, and the whiskers denote the range of the rest of the distribution (except for outliers). See text for more details.