A. Missing Proofs for the Learning Algorithm

Lemma 4.6. *Let $D_x(\mathbf{V})$ be the Bayes net as defined in Equation (4). Then Algorithm 1 takes $\tilde{O}\left(\frac{n|\Sigma|^{2kd}}{\alpha^k\varepsilon^2}\right)$ samples and $\tilde{O}\left(\frac{n^2|\Sigma|^{4kd}}{\alpha^k\varepsilon^2}\right)$ time and returns a Bayes net $\hat{D}_x(\mathbf{V})$ such that $d_{\mathrm{TV}}(D_x, \hat{D}_x) \leqslant \varepsilon$ with probability at least 3/4.*

Proof. We run Algorithm 2 with the parameters $m = 20n|\Sigma|^{kd+k+2}\log(n|\Sigma|^{kd+k})/(\alpha^k\varepsilon^2)$ and $t = 10\log(n|\Sigma|^{kd+k})$ to learn D_x as $\hat{D}_x$. We get from Theorem 4.3 for the distributions D_x and $\hat{D}_x$:

$$\mathrm{KL}(D_x, \hat{D}_x) \leqslant \sum_i \sum_{\mathbf{a}} D_x(\mathbf{Z}'_i = \mathbf{a})\ \mathrm{KL}(D_x(V_i \mid \mathbf{Z}'_i = \mathbf{a}), \hat{D}_x(V_i \mid \mathbf{Z}'_i = \mathbf{a})) \quad (5)$$

Our strategy is to learn $D_x(V_i \mid \mathbf{Z}'_i = \mathbf{a})$ by conditional sampling from either $P(\mathbf{V} \setminus \mathbf{Z}'_i \mid \mathbf{Z}'_i = \mathbf{a})$ or $P(\mathbf{V} \setminus (\mathbf{Z}'_i \cup \{X\}) \mid \mathbf{Z}'_i = \mathbf{a}, X = x)$ as appropriate.

First consider the summands of Equation (5) where $V_i \in \mathbf{S}_1$. In this case $D_x(V_i \mid \mathbf{Z}'_i) = P(V_i \mid \mathbf{Z}_i)$ and $\mathbf{Z}'_i = \mathbf{Z}_i \subseteq \mathbf{Pa}^+(\mathbf{S}_1)$ where P is α-strongly positive w.r.t. the later set. Hence at least $m\alpha/2$ samples turn up with $\mathbf{Z}_i = \mathbf{a}$ from Chernoff and union bounds except with 1/40 probability for a large enough n. Conditioned on this, Theorem 4.4 gives $\mathbf{E}[\mathrm{KL}(D_x(V_i \mid \mathbf{Z}'_i = \mathbf{a}), \hat{D}_x(V_i \mid \mathbf{Z}'_i = \mathbf{a}))] \leqslant 2(|\Sigma|-1)/(m\alpha) \leqslant \varepsilon^2/(10n|\Sigma|^{kd+k})$ where $\hat{D}_x$ is the add-1 estimator on the conditional samples with $\mathbf{Z}_i = \mathbf{a}$. Since $D_x(\mathbf{Z}'_i = \mathbf{a}) \leqslant 1$ each summand is also upper-bounded by $\varepsilon^2/(10n|\Sigma|^{kd+k})$.

Next we consider the summands $(i, \mathbf{a})$ with $V_i \notin \mathbf{S}_1$. For these summands, if $X \notin \mathbf{Z}_i$, we have $\mathbf{Z}'_i = \mathbf{Z}_i$, $D_x(V_i \mid \mathbf{Z}'_i) = P(V_i \mid \mathbf{Z}_i)$ and $P(\mathbf{Z}'_i = \mathbf{a}) \geqslant \alpha^k D_x(\mathbf{Z}'_i = \mathbf{a})/|\Sigma|$, the last inequality by marginalization of Lemma 4.5 Part 3. over $\mathbf{V} \setminus (\mathbf{Z}'_i \cup \{X\})$. If $X \in \mathbf{Z}_i$, we have $\mathbf{Z}'_i = \mathbf{Z}_i \setminus \{X\}$, $D_x(V_i \mid \mathbf{Z}'_i) = P(V_i \mid \mathbf{Z}_i \setminus \{X\}, X = x)$ and $P(\mathbf{Z}'_i = \mathbf{a}, X = x) \geqslant \alpha^k D_x(\mathbf{Z}'_i = \mathbf{a})/|\Sigma|$, the later inequality by marginalization of Lemma 4.5 Part 2. over $\mathbf{V} \setminus (\mathbf{Z}'_i \cup \{X\})$. Let $N_{i,\mathbf{a}}$ be the number of samples with with $\mathbf{Z}'_i = \mathbf{a}$ if $X \notin \mathbf{Z}_i$ and with $\mathbf{Z}'_i \circ X = \mathbf{a} \circ x$ if $X \in \mathbf{Z}_i$. Then $N_{i,\mathbf{a}} \sim \mathrm{Binomial}(m, p)$ where $p \geqslant \alpha^k D_x(\mathbf{Z}'_i = \mathbf{a})/|\Sigma|$.

We partition the summands $(i, \mathbf{a})$ with $V_i \notin \mathbf{S}_1$ into two sets: *heavy* if $D_x[\mathbf{Z}'_i = \mathbf{a}] \geqslant \varepsilon^2/(10n|\Sigma|^{kd+k+1})$ and *light* otherwise. Consider the event "all heavy $(i, \mathbf{a})$s satisfy $N_{i,\mathbf{a}} \geqslant m\alpha^k D_x(\mathbf{Z}'_i = \mathbf{a})/(2|\Sigma|)$". It is easy to see from from our definition of m and heaviness that this event holds except with 1/40 probability from Chernoff and union bounds for a large enough n. Hence for the rest of the argument, we condition on this event. In this case, all heavy items satisfy $N_{i,\mathbf{a}} \geqslant t$ from our definition of m and t.

For the summands $(i, \mathbf{a})$ with $V_i \notin \mathbf{S}_1$, we get the following.

- If $(i, \mathbf{a})$ is heavy then from Theorem 4.4 $\mathbf{E}[D_x(\mathbf{Z}'_i = \mathbf{a})\ \mathrm{KL}(D_x(V_i \mid \mathbf{Z}'_i = \mathbf{a}), \hat{D}_x(V_i \mid \mathbf{Z}'_i = \mathbf{a}))] \leqslant \frac{D_x(\mathbf{Z}'_i = \mathbf{a})(|\Sigma|-1)}{N_{i,a}} \leqslant \varepsilon^2/(10n|\Sigma|^{kd+k})$, using the lower bound of $N_{i,\mathbf{a}}$ from the previous paragraph.
- If a light $(i, \mathbf{a})$ satisfy $N_{i,\mathbf{a}} \geqslant t$, we get $\mathbf{E}[D_x(\mathbf{Z}'_i = \mathbf{a})\ \mathrm{KL}(D_x(V_i \mid \mathbf{Z}'_i = \mathbf{a}), \hat{D}_x(V_i \mid \mathbf{Z}'_i = \mathbf{a}))] \leqslant \frac{\varepsilon^2}{10n|\Sigma|^{kd+k+1}}\frac{|\Sigma|-1}{t} \leqslant \varepsilon^2/(10n|\Sigma|^{kd+k})$ from Theorem 4.4.
- $(i, \mathbf{a})$s which do not satisfy $N_{i,\mathbf{a}} \geqslant t$ must be light for which we define the conditional distribution to be uniform. In this case, $\mathrm{KL}(D_x(V_i \mid \mathbf{Z}'_i = \mathbf{a}), \hat{D}_x(V_i \mid \mathbf{Z}'_i = \mathbf{a})) = \sum_{\sigma\in\Sigma} D_x(V_i = \sigma \mid \mathbf{Z}'_i = \mathbf{a}) \ln(|\Sigma| D_x(V_i \mid \mathbf{Z}'_i = \mathbf{a})) = \ln|\Sigma| - H(D_x(V_i \mid \mathbf{Z}'_i = \mathbf{a})) \leqslant \ln|\Sigma|$, where $0 \leqslant H(\cdot) \leqslant \ln|\Sigma|$ is the Shannon entropy function. Hence in this case also $\mathbf{E}[D_x(\mathbf{Z}'_i = \mathbf{a})\ \mathrm{KL}(D_x(V_i \mid \mathbf{Z}'_i = \mathbf{a}), \hat{D}_x(V_i \mid \mathbf{Z}'_i = \mathbf{a}))] \leqslant \varepsilon^2/(10n|\Sigma|^{kd+k})$.

Thus each of the $n|\Sigma|^{kd+k}$ summands in the r.h.s. of Equation (5) is at most $\varepsilon^2/(10n|\Sigma|^{kd+k})$ in expectation. We get $\mathbf{E}[\mathrm{KL}(D_x, \hat{D}_x)] \leqslant \varepsilon^2/10$. From Markov's and Pinsker's inequalities, $d_{\mathrm{TV}}(D_x, \hat{D}_x) \leqslant \varepsilon$ except $1/5$ probability.

The total error probability is at most $1/4$ so far. □

Next we improve the success probability of the above learning algorithm. We repeat Algorithm 2 independently $O(\log\frac{1}{\delta})$ times and use the following result to achieve $(1-\delta)$ success probability.

Theorem A.1 (Theorem 2.9 in (Bhattacharyya et al., 2020) restated)**.** *Fix any $0 < \varepsilon, \delta < 1$. Suppose we are given an algorithm that learns an unknown Bayes net P over Σ^N on a graph G with indegree $\leqslant \Delta$ as a Bayes net $\hat{P}$ on G such that $d_{\mathrm{TV}}(P, \hat{P}) \leqslant \varepsilon$ with probability at least 3/4 using $m(\varepsilon)$ samples and $t(\varepsilon)$ time. Then we can output a distribution P' on G such that $d_{\mathrm{TV}}(P, P') \leqslant \varepsilon$ with probability at least $(1-\delta)$ using $O(m(\varepsilon/4)\log\frac{1}{\delta})$ samples and $O(t(\varepsilon/4)\log\frac{1}{\delta} + |\Sigma|^{2\Delta}N^2\varepsilon^{-2}\log^3\frac{1}{\delta})$ time.*

We get the following final theorem for learning P_x.

Theorem 2.3. *[Evaluation and Generation] For any intervention x to X and parameter $\varepsilon \in (0, 1)$, there is an algorithm that takes $m = \tilde{O}\left(\frac{|\Sigma|^{2kd}n}{\alpha^k\varepsilon^2}\right)$ samples from P, and in $O\left(mn|\Sigma|^{2kd}\right)$ time, learns a distribution $\hat{P}$ satisfying $d_{\mathrm{TV}}(P_x, \hat{P}) \leqslant \varepsilon$ such that*

- *Evaluation: Given an assignment $\mathbf{w}$ to $\mathbf{V} \setminus \{X\}$ computing $\hat{P}(\mathbf{w})$ takes $O(n|\Sigma|(kd+k))$ time*
- *Generation: Obtaining an independent sample from $\hat{P}$ takes $O(n|\Sigma|(kd+k))$ time .*

Algorithm 2 Learning D_x

Input: Samples from P, parameters m, t
Output: A Bayes net $\hat{D}_x$ according to the factorization of Equation (4)
Get m samples from P
for every vertex $V_i \in \mathbf{S}_1$ **do**
  for every fixing $\mathbf{Z}_i = \mathbf{a}$, where $\mathbf{Z}_i$ are the effective parents of V_i **do**
    $\hat{D}_x(V_i \mid \mathbf{Z}'_i = \mathbf{a}) \leftarrow$ the add-1 empirical distribution (see Theorem 4.4) at node i in the subset of samples where $\mathbf{Z}_i = \mathbf{a}$
  end for
end for
for every vertex $V_i \in \mathbf{V} \setminus \mathbf{S}_1$ **do**
  for every fixing $\mathbf{Z}_i \setminus \{X\} = \mathbf{a}$, where $\mathbf{Z}_i$ are the effective parents of V_i **do**
    if $X \in \mathbf{Z}_i$ **then**
      $N_{i,\mathbf{a}} \leftarrow$ the number of samples with $\mathbf{Z}_i \setminus \{X\} = \mathbf{a}$ and $X = x$
      if $N_{i,\mathbf{a}} \geqslant t$ **then**
        $\hat{D}_x(V_i \mid \mathbf{Z}_i \setminus \{X\} = \mathbf{a}) \leftarrow$ the add-1 empirical distribution at node i in the subset of samples where $\mathbf{Z}_i \setminus X = \mathbf{a}$ and $X = x$
      else
        $\hat{D}_x(V_i \mid \mathbf{Z}_i \setminus \{X\} = \mathbf{a}) \leftarrow$ the uniform distribution over Σ
      end if
    else
      $N_{i,\mathbf{a}} \leftarrow$ the number of samples with $\mathbf{Z}_i = \mathbf{a}$
      if $N_{i,\mathbf{a}} \geqslant t$ **then**
        $\hat{D}_x(V_i \mid \mathbf{Z}_i = \mathbf{a}) \leftarrow$ the add-1 empirical distribution at node i in the subset of samples where $\mathbf{Z}_i = \mathbf{a}$
      else
        $\hat{D}_x(V_i \mid \mathbf{Z}_i = \mathbf{a}) \leftarrow$ the uniform distribution over Σ
      end if
    end if
  end for
end for

Proof. We use Algorithm 2 with $m = 20n|\Sigma|^{kd+k+2}\log(n|\Sigma|^{kd+k})/(\alpha^k\varepsilon^2)$ and $t = 10\log(n|\Sigma|^{kd+k})$, which from Lemma 4.6, guarantees 3/4 success probability for learning D_x within total variation distance at most ε. Then we use Theorem A.1 to improve the success probability to $1-\delta$. The final time and sample complixities follow from Theorem A.1.

This gives us a distribution $\hat{D}_x$ over $\mathbf{V}$, whose marginal distribution on all variables but X, we use for evaluation and sampling. Once we have learnt D_x, sampling and evaluation takes $O(n|\Sigma|(kd+k))$ time. □

B. Lower Bound

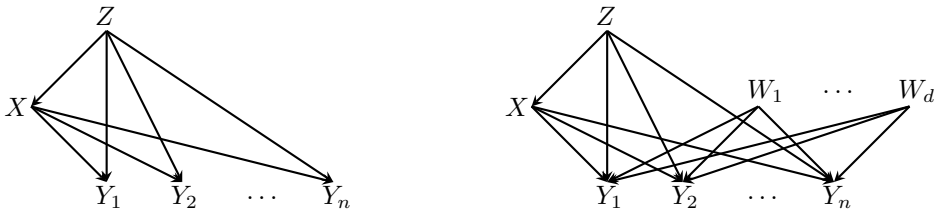
For the lower bound we use a well-known packing argument based on Fano's inequality which says if there is a class of 2^K distributions with pairwise KL distance at most β then $\Omega(K/\beta)$ samples are needed to identify a uniformly random distribution from the class. The KL distance is known to satisfy certain chain rule which we use in the following proof (see eg. Lemma 6 in (Przytycki, 2011)). We first recall Theorem 2.6.

Theorem 2.6. *Fix integers $d, k \geqslant 1$ and a set Σ of size $\geqslant 2$. For all sufficiently large n, there exists an ADMG G with n nodes and in-degree d so that the following hold. G contains a node X such that $|\mathbf{Pa}(X)| = d$ and $|\mathbf{S}_1| = k$ (where $\mathbf{S}_1$ is the c-component containing X). For any $Z \in \mathbf{Pa}(X) \cup \mathbf{S}_1$, there exists a causal Bayes net $\mathcal{P}$ on G over Σ-valued variables such that:*

(i) For the observational distribution P, the marginal $P|_{(\mathbf{Pa}(X)\cup\mathbf{S}_1)\setminus\{Z\}}$ is uniform but the marginal $P|_{\mathbf{Pa}(X)\cup\mathbf{S}_1}$ has mass at most α at some assignment.
(ii) There exists an intervention x on X such that learning the distribution P_x upto d_{TV}-distance ε with probability $9/10$ requires $\Omega(n|\Sigma|^d/\alpha\varepsilon^2)$ samples from P.

Proof. We first show the lower bound where Z is a parent of X, and $d = 2$. Later we show how to prove the full theorem.

Our ADMG on $n+2$ variables: $Z, X, Y_1, Y_2, \ldots, Y_n$ consists of n triangles with Z, X, Y_j for every j where Z is



(a) With no control variables

(b) With control variables

Figure 4. ADMGs for Theorem 2.6

the source and Y_j is the sink. Let $\mathbf{Y} = \langle Y_1, Y_2, \ldots, Y_n \rangle$. Please refer to Figure 4a.

Z is uniform over $\{0,1\}$. $X = \bar{Z}$ with probability α and $X = Z$ with probability $1-\alpha$. Thus X, Z jointly satisfy α-strong positivity. Each $Y_j \mid X, Z$ is one among the following two conditional distributions:

$$\begin{aligned} D_1 : Y_j &= \mathrm{Bern}(1/2 + \varepsilon/\sqrt{n}) \text{ if } X \neq Z, \\ Y_j &= \mathrm{Bern}(1/2) \text{ if } X = Z \\ D_2 : Y_j &= \mathrm{Bern}(1/2 - \varepsilon/\sqrt{n}) \text{ if } X \neq Z \\ Y_j &= \mathrm{Bern}(1/2) \text{ if } X = Z \end{aligned}$$

We create a class $\mathcal{C}_\varepsilon$ of causal models using a code $\mathcal{C} \subset \{0,1\}^n$. This code has size $2^{\Omega(n)}$, and any two of them $c, d \in \mathcal{C}$ satisfy the following: there are $\Theta(n)$ positions where c is 1 and d is 0. Showing existence of such a code is standard. Given a code as above, corresponding to every $c \in \mathcal{C}$, we create a product distribution $\mathbf{Y} \mid X, Z$: the 1 positions of c use the distribution D_1 and the 0 positions of c use the distribution D_2. Together with the distributions of X, Z this defines a causal Bayes net $\mathcal{P}^c$.

We first lower-bound the distance between the interventional distributions for any two members $\mathcal{P}^c, \mathcal{P}^d \in \mathcal{C}_\varepsilon$. Let $\mathbf{S}$ be the subset of indices from $[n]$ of size $\Theta(n)$ where c is 1 and d is 0.

$$d_{\mathrm{TV}}(P^c_{X=1}(\mathbf{Y}), P^d_{X=1}(\mathbf{Y})) \geqslant d_{\mathrm{TV}}(P^c_{X=1}(\mathbf{S}), P^d_{X=1}(\mathbf{S}))$$

With 1/2 probability, when $Z = 1$ every dimension of both the distributions are $\mathrm{Bern}(1/2)$ and therefore have $d_{\mathrm{TV}} = 0$. We focus on the other case when every dimension of $P^c_{X=1}(\mathbf{S})$ follows D_1 and $P^d_{X=1}(\mathbf{S})$ follows D_2.

$$\begin{aligned} &d_{\mathrm{TV}}(P^c_{X=1}(\mathbf{S}), P^d_{X=1}(\mathbf{S})) \\ &= 1/2 \cdot d_{\mathrm{TV}}(\mathrm{Bern}(1/2 + \varepsilon/\sqrt{n})^{|\mathbf{S}|}, \mathrm{Bern}(1/2 - \varepsilon/\sqrt{n})^{|\mathbf{S}|}) \\ &\geqslant 1/2 \cdot d_{\mathrm{TV}}(\mathrm{Bern}(1/2 + \varepsilon/\sqrt{n})^{|\mathbf{S}|}, \mathrm{Bern}(1/2)^{|\mathbf{S}|}) \end{aligned}$$

Claim B.1. *$d_{\mathrm{TV}}(\mathrm{Bern}(1/2 + \varepsilon/\sqrt{n})^l, \mathrm{Bern}(1/2)^l) \geqslant \Theta(\varepsilon)$ for $l = \Theta(n)$, $l \leqslant n$, $\varepsilon \leqslant 1/4$.*

Proof.

$$\begin{aligned} &d_{\mathrm{TV}}(\mathrm{Bern}(1/2 + \varepsilon/\sqrt{n})^l, \mathrm{Bern}(1/2)^l) \\ &= \sum_{i=0}^{l} \binom{l}{i} |(1/2 + \varepsilon/\sqrt{n})^i (1/2 - \varepsilon/\sqrt{n})^{l-i} - 1/2^l| \\ &\geqslant \sum_{i=0}^{l/2} \binom{l}{i} 2^{-l} (1 - (1 + 2\varepsilon/\sqrt{n})^i (1 - 2\varepsilon/\sqrt{n})^{l-i}) \\ &\geqslant \sum_{i=0}^{l/2} \binom{l}{i} 2^{-l} (1 - \exp(2\varepsilon i/\sqrt{n}) \exp(-2\varepsilon(l-i)/\sqrt{n})) \\ &\geqslant \sum_{i=l/2-\sqrt{l}}^{l/2} \binom{l}{i} 2^{-l} (1 - \exp(-2\varepsilon(l-2i)/\sqrt{n})) \\ &\geqslant \sum_{i=l/2-\sqrt{l}}^{l/2} \binom{l}{i} 2^{-l} 2\varepsilon(l/2 - i)/\sqrt{n} \\ &= \sum_{j=0}^{\sqrt{l}} \binom{l}{l/2 - j} 2^{-l} 2\varepsilon j/\sqrt{n} \end{aligned}$$

The third line in the above uses the fact that in the range $0, 1, \ldots, l/2$; $1/2^l$ is larger than the other term. The fourth line uses $e^x \geqslant 1 + x$ and $e^{-x} \geqslant 1 - x$. The sixth line uses $1 - e^{-x} \geqslant x/2$ whenever $x \leqslant 1$. The ratio $\binom{l}{l/2} / \binom{l}{l/2-j}$ can be upper-bounded by $\exp(j^2/(l/2 - j + 1)) = O(1)$ for $0 \leqslant j \leqslant \sqrt{l}$ and $\binom{l}{l/2} \simeq 2^l/\sqrt{l}$, which gives $\Theta(\varepsilon)$ in the last summation. □

Next we upper bound the KL distance of any two observational distributions from $\mathcal{C}_\varepsilon$. Considering the extreme case, we only upper bound the pairs P^c and P^d whose all the coordinates of $\mathbf{Y}$ are different: one is D_1, other is D_2. Note that the distributions $P^c|_{X \cup Z} = P^d|_{X \cup Z} := P(X, Z)$ (say), which gives (by the chain rule):

$$\begin{aligned}\mathrm{KL}(P^c, P^d) &= \sum_{x,z} P(x,z)\mathrm{KL}(P^c|_{\mathbf{Y}}, P^d|_{\mathbf{Y}}) \\ &= \sum_{x\neq z} P(x,z)\mathrm{KL}(P^c|_{\mathbf{Y}}, P^d|_{\mathbf{Y}}) \\ &\quad \text{(as when } x = z \text{ both are } \mathrm{Bern}(1/2)^n\text{)} \\ &= \alpha\mathrm{KL}(P^c|_{\mathbf{Y}}, P^d|_{\mathbf{Y}})\end{aligned}$$

$P^c|_{\mathbf{Y}}$ and $P^d|_{\mathbf{Y}}$ are product distributions (in the extreme case) whose each component pairs are distributed as either $\mathrm{Bern}(1/2+\varepsilon/\sqrt{n}), \mathrm{Bern}(1/2-\varepsilon/\sqrt{n})$ or $\mathrm{Bern}(1/2-\varepsilon/\sqrt{n}), \mathrm{Bern}(1/2+\varepsilon/\sqrt{n})$. Using additive property of KL we get $\mathrm{KL}(P^c, P^d) = \Theta(\alpha\varepsilon^2)$.

Therefore from Fano's inequality, learning each interventional distribution up to $\Theta(\varepsilon)$ distance with probability 2/3 requires $\Omega(n/\alpha\varepsilon^2)$ samples.

We next improve the lower bound by $|\Sigma|^d$ factor where d is the indegree. Please refer to Figure 4b. We pick $|\Sigma|^d$ random models from $\mathcal{C}_\varepsilon$ and use them as the conditional distributions for $\mathbf{Y} \mid X, Z$. Then we create d more control variables $\mathbf{W} = \langle W_1, \ldots, W_d\rangle$ which are uniformly distributed over Σ^d and indicates which of the hard distributions is followed by $\mathbf{Y}$. Any algorithm that want to learn the interventional distribution $X = 1$ for this model in d_{TV} distance ε have to learn a constant fraction of $|\Sigma|^d$ many hard interventional distributions from $\mathcal{C}_\varepsilon$ in d_{TV} distance $O(\varepsilon)$ over the choices of $\mathbf{W}$. We have already established that for every fixing of $\mathbf{W}$ learning the interventional distributions with 2/3 probability requires $\Omega(n/\alpha\varepsilon^2)$ samples. From Chernoff's bound, learning $9\cdot|\Sigma|^d/10$ many distributions with 9/10 probability would require $\Omega(n|\Sigma|^d/\alpha\varepsilon^2)$ samples.

We next show how to add hidden variables to the graph. Instead of Z being a parent of X in Figure 4b, suppose that Z is confounded with X. That is, there is a hidden variable U that is a parent of X as well as Z. Now, we can define the same causal models that we analyzed earlier, with U taking the place of the old Z, and the new Z copying the value of U. The analysis remains unchanged, as Z is not affected by an intervention on X. Also, the degree of X and the size of the c-component can be made arbitrarily large by adding dummy variables. □

C. Evaluation of Marginal Interventions

Here we discuss the problem of estimating $P_x|_{\mathbf{F}}$, i.e., the marginal interventional distribution of the intervention x to X on a subset of the observables $\mathbf{F} \subseteq \mathbf{V}$. For ease of exposition, we can assume that the vertices of G are $\mathbf{An}^+(\mathbf{F})$ as other variables do not play a role in $P_x|_{\mathbf{F}}$ and hence can be pruned out from the model; here $\mathbf{An}^+(\mathbf{F})$ denotes the set of all observable ancestors of $\mathbf{F}$, including $\mathbf{F}$. Tian and Pearl (Tian & Pearl, 2002b) provided an algorithm for this identification question when the ADMG satisfies Assumption 2.1 (See Theorem 4 of (Tian & Pearl, 2002b)), a sufficient condition for identifiability[6]. Later works (Shpitser & Pearl, 2006; Huang & Valtorta, 2008) generalized this result of Tian and Pearl for more general interventions, thus exhibiting a sufficient and necessary identifiability graphical condition for this problem.

We consider the following setting: Suppose $\mathcal{P}$ is an unknown causal Bayes net over a known ADMG G on n observable variables $\mathbf{V}$ that satisfies (Assumption 2.1) and α-strong positivity with respect to a variable $X \in \mathbf{V}$ (Assumption 2.2) and let $\mathbf{F} \subseteq \mathbf{V}$. Let d denote the maximum in-degree of the graph G, k denote the size of its largest c-component, and $f = |\mathbf{F}|$. When the graph being referred to is unclear, we will subscript notation (eg: $\mathbf{Pa}_H(V)$ indicates the observable parents of V in graph H) to indicate the graph on which the operator is defined on.

We show finite sample bounds for estimating $P_x|_{\mathbf{F}}$ when the underlying ADMG satisfies Assumption 2.1, thus making results of (Tian & Pearl, 2002b) quantitative. Estimating such causal effects under the necessary and sufficient graphical conditions of (Shpitser & Pearl, 2006; Tian & Pearl, 2002b) in the finite sample regime is an important and an interesting open question which we leave for future work. As mentioned in Section 2, the required marginal distribution $P_x|_{\mathbf{F}}$ can be estimated by taking $O(|\Sigma|^f/\varepsilon^2)$ samples from the generator $\hat{P}_x$, and we can use Theorem 2.3 to obtain the generator distribution $\hat{P}_x$ where we require $O(|\Sigma|^{5kd}n/\alpha^k\varepsilon^2)$ many samples from the observational distribution P. Hence we get Corollary 2.4.

The time complexity of the algorithm (of Corollary 2.4) described above is exponential in f. To handle problems that arise in practice for small $\mathbf{F}$'s, it is of interest to develop efficient algorithms for estimating $P_x|_{\mathbf{F}}$. In such cases the approach discussed above is superfluous, as the sample complexity depends linearly on n, the total number of variables in the model, which could be unnecessarily large. Theorem 2.5, restated below, shows a sample and time-efficient algorithm when f is very small (e.g. constant).

Theorem 2.5. *For any subset $\mathbf{F} \subseteq \mathbf{V}$ with $|\mathbf{F}| = f$, intervention x to X and parameter $\varepsilon \in (0,1)$, there is an algorithm that takes $m = \tilde{O}\left(\frac{|\Sigma|^{5(f+k(d+1))^2}}{\alpha^k\varepsilon^2}\right)$ samples from P and runs in $O(m(f+k(d+1))|\Sigma|^{2(f+k(d+1))^2})$ time and returns an evaluator for a distribution $\tilde{P}_{\mathbf{F}}$ on $\mathbf{F}$ such that*

[6] Recall that (Tian & Pearl, 2002b) proved Assumption 2.1 is necessary and sufficient for identifiability of P_x. However to identify $P_x|_{\mathbf{F}}$, Assumption 2.1 was known to be only sufficient for identifiability.

$d_{\mathrm{TV}}(P_x|_{\mathbf{F}}, \tilde{P}_{\mathbf{F}}) \leqslant \varepsilon$.

The rest of this section is dedicated towards proving Theorem 2.5.

First let us discuss a high level idea of our algorithm for Theorem 2.5. The idea to handle cases with small f is to restrict our attention to the marginal distribution $P(\mathbf{W})$ over a small set of vertices $\mathbf{W}$ and then apply Theorem 3 of (Tian & Pearl, 2002b) (Theorem 3.6 here) over $\mathbf{W}$. However restriction to $\mathbf{W}$ could potentially modify the parent/child or c-component relationships (or both) across the vertices of $\mathbf{W}$. Hence, to apply Theorem 3.6, the underlying causal graph over the vertices of $\mathbf{W}$ should be obtained via a formal approach in such a way that the topological ordering and the conditional independence relations across vertices of $\mathbf{W}$ are preserved. To do that we make use of a well-known latent projection algorithm (Verma & Pearl, 1990) to reduce the given ADMG G (over observables $\mathbf{V}$) to a different ADMG H (over observables $\mathbf{W}$). A similar reduction using the latent projection (Verma & Pearl, 1990) has also appeared in a slightly different context (Tikka & Karvanen, 2018) – where the objective is to improve the efficiency of the algorithm.

We carefully choose the set $\mathbf{W}$ and prune all the other variables $\mathbf{V} \setminus \mathbf{W}$ from the graph G, and then apply the latent projection to obtain H such that:

(A) The required causal effect is identifiable in H;
(B) Theorem 2.3 can be applied by maintaining bounds on in-degree and c-component size of this new graph H.

We will show that for $\mathbf{W} = \mathbf{F} \cup \mathbf{Pa}^{+}_{G}(\mathbf{S}_1)$ – both (A) and (B) hold. We will prove (A) while we prove (B); although (A) can easily be verified by pruning the vertices of $\mathbf{V} \setminus \mathbf{W}$, one by one, by using Corollary 16 of (Tikka & Karvanen, 2018). Before we prove (B) we will first describe the reduction procedure so that the essense of the argument becomes clearer. Our reduction procedure is discussed next.

The reduction consists of two steps: The first step is to simplify the given graph G to a much smaller graph G' (defined over observables $\mathbf{W}$) by ignoring all the other variables of G (i.e., ignoring $\mathbf{V} \setminus \mathbf{W}$). By ignoring a certain variable we mean that the variable is considered to be hidden. Although the observable vertices of G' is $\mathbf{W}$, as desired, G' is not an ADMG[7]. Since ADMGs are, in general, easy to analyze and parse through we then convert this general causal graph G' to an ADMG H using a known reduction technique – and this is the second step. This reduction procedure, which we call Reduction$(G, \mathbf{W})$ is formally discussed next.

[7]Recall that an ADMG is a graph where the unobservables are root nodes and have exactly two observable children – denoted by bidirected edges.

C.1. Reduction: Pruning G to a simpler graph H

Reduction$(G, \mathbf{W})$

1. Let G' be the graph obtained from G by considering $\mathbf{V} \setminus \mathbf{W}$ as hidden variables.
2. **Projection Algorithm (G' to H) (Tian & Pearl, 2002a; Verma & Pearl, 1990)**. The projection algorithm reduces the causal graph G' to an ADMG H by the following procedure:
   (a) For each observable variable $V_i \in \mathbf{V}$ of G', add an observable variable V_i in H.
   (b) For each pair of observable variables $V_i, V_j \in \mathbf{V}$, if there exists a directed edge from V_i to V_j in G', or if there exists a *directed* path from V_i to V_j that contains only unobservable variables in G', then add a directed edge from V_i to V_j in H.
   (c) For each pair of observable variables $V_i, V_j \in \mathbf{V}$, if there exists an unobservable variable U such that there exist two *directed* paths in G' from U to V_i and from U to V_j such that both the paths contain only unobservable variables, then add a bidirected edge between V_i and V_j in H.
3. **Return** H

C.2. Properties of Reduction$(G, \mathbf{W})$

It is well-known that the projection algorithm (G' to H) (Tian & Pearl, 2002a; Verma & Pearl, 1990) preserves some of the important properties such as topological ordering and conditional independence relations. Before we discuss those, let us revisit the equivalent definitions of parents and c-components for general causal graphs with hidden variables.

Definition C.1 (Effective Parents for general causal graphs)**.** *Given a general causal graph G' and a vertex $V_j \in \mathbf{V}$,* the effective parents of V_j *is the set of all observable vertices V_i such that either V_i is a parent of V_j or there exists a directed path from V_i to V_j that contains only unobservable variables in G'.*

Definition C.2 (c-component for general causal graphs)**.** *For a given general causal graph G', two observable vertices V_i and V_j are related by the* c-component relation *if (i) there exists an unobservable variable U such that G' contains two paths (a) from U to V_i; and (b) from U to V_j, where both the paths use only unobservable variables, or (ii) there exists another vertex $V_z \in \mathbf{V}$ such that V_i and V_z (and) V_j and V_z are related by the c-component relation.*

The below lemma illustrates: "c-component that contains X remains the same in G and H."

Lemma C.3. *Let $\mathbf{S}_1$ denotes the c-component that contains X in G, $\mathbf{W} = \mathbf{Y} \cup \mathbf{Pa}_G^+(\mathbf{S}_1)$ and $H = Reduction(G, \mathbf{W})$. Then, (i) the c-component that contains X in H is also $\mathbf{S}_1$; (ii) H satisfies Assumption 2.1; (iii) $\mathbf{Pa}_G^+(\mathbf{S}_1)$ and $\mathbf{Pa}_H^+(\mathbf{S}_1)$ are the same; (iv) H satisfies Assumption 2.2.*

Proof. Let $\mathbf{C}$ denote the c-component of H that contains X. Note that $\mathbf{S}_1 \subseteq \mathbf{C}$ since all those bidirected edges in G that forms $\mathbf{S}_1$ are retained in H (because $\mathbf{S}_1 \subseteq \mathbf{W}$). Next we will prove that no other vertex of H share a bidirected edge with $\mathbf{S}_1$ in H. Suppose for contradiction there exists a vertex $W_i \in \mathbf{W}$ that share a bidirected edge with some node $W_j \in \mathbf{S}_1$ in H. This implies, during the reduction, there exists two paths in G' (U to W_i and U to W_j) such that all the variables included in these two paths, other than W_i and W_j, are unobservables in G' which means all those vertices belong to $\mathbf{V} \setminus \mathbf{W}$, a contradiction to the fact that $\mathbf{Pa}_G^+(\mathbf{S}_1)$ is contained in $\mathbf{W}$. This proves (i). Since $\mathbf{Pa}_G^+(\mathbf{S}_1) \subseteq \mathbf{W}$ there can not exist a bidirected edge between X and a child of X in H which proves (ii).

Note that $\mathbf{Pa}_G^+(\mathbf{S}_1) \subseteq \mathbf{Pa}_H^+(\mathbf{S}_1)$ – because $\mathbf{W}$ contains $\mathbf{Pa}_G^+(\mathbf{S}_1)$. Now suppose, for contradiction, $\mathbf{Pa}_G^+(\mathbf{S}_1) \subset \mathbf{Pa}_H^+(\mathbf{S}_1)$. Then during the reduction step there must have been an edge from an unobservable to $\mathbf{S}_1$ in G', which can not be true as $\mathbf{W}$ contains $\mathbf{Pa}_G^+(\mathbf{S}_1)$ and none of those variables are treated as hidden variables in the reduction. This proves (iii). Since $\mathbf{S}_1$ and $\mathbf{Pa}^+(\mathbf{S}_1)$ remains unchanged in both G and H, Assumption 2.2 still holds in H which proves (iii). □

The projection algorithm (G' to H) is known to preserve the following set of properties.

- The c-components of H and G' are identical and the c-component factorization formula (Equation (20) in Lemma 2 of (Tian & Pearl, 2002a)) holds even for the general causal graph (See Section 5 of (Tian & Pearl, 2002a) for more details). They show this based on a known previously known reduction from G' to H (Verma & Pearl, 1990). The proof is based on the fact that for any subset $\mathbf{S} \subseteq \mathbf{V}$ of observable variables, the induced subgraphs $G'[\mathbf{S}]$ and $H[\mathbf{S}]$ require the same set of conditional independence constraints.
- The effective parents (see Definition C.1) of every observable node in G' is the same as the (observable) parent set of the corresponding node in H.
- The observable vertices of G' and H are the same.
- Also, the topological ordering of the observable nodes of G' and H are the same.

As is common in the causality literature we do not use any other property of G' besides the above in our analysis, and hence it is sufficient to derive conclusions from this modified graph which contains only a small number of vertices.

C.2.1. Proof of Theorem 2.5

We know from Lemma C.3 that whenever G satisfy Assumption 2.1, $H = \text{Reduction}(G, \mathbf{W})$ with $\mathbf{W} = \mathbf{Y} \cup \mathbf{Pa}_G^+(\mathbf{S}_1)$ satisfy Assumption 2.1 as well. For such graphs G and H, while both satisfy Assumption 2.1, it is well-known that an equivalent statement of Theorem 3.6 directly follows from the proof of Theorem 3 of (Tian & Pearl, 2002b) since their proof uses only the above mentioned properties. We will also extensively use (i) of Lemma C.3. This results in the following theorem.

Theorem C.4 (Theorem 3 of (Tian & Pearl, 2002b) with respect to ADMG $H = \text{Reduction}(G, \mathbf{W})$)**.** *Let P be a CBN over a causal graph $G = (\mathbf{V}, E^{\rightarrow} \cup E^{\leftrightarrow})$, $X \in \mathbf{V}$ be a designated variable and $\mathbf{Y} \subseteq \mathbf{V} \setminus \mathbf{X}$. Let $\mathbf{S}_1, \ldots, \mathbf{S}_\ell$ are the c-components of G and without loss of generality assume $X \in \mathbf{S}_1$. Suppose that G satisfies Assumption 2.1. Let $\mathbf{W} = \mathbf{Y} \cup \mathbf{Pa}_G^+(\mathbf{S}_1)$ for $\mathbf{Y} \subseteq \mathbf{V} \setminus \{X\}$. Let $H = Reduction(G, \mathbf{W})$ and let $\mathbf{S}'_1, \ldots, \mathbf{S}'_{\ell'}$ be the c-components of H where without loss of generality let $\mathbf{S}_1 = \mathbf{S}'_1$. Then for any setting x to X and any assignment $\mathbf{t}$ to $\mathbf{W} \setminus \{X\}$ the interventional distribution $P_x(\mathbf{t})$ is given by:*

$$P_x(\mathbf{t}) = P_{\mathbf{t}_{\mathbf{W}\setminus\mathbf{S}'_1}}(\mathbf{t}_{\mathbf{S}'_1\setminus\{X\}}) \cdot \prod_{j=2}^{\ell'} P_{\mathbf{t}_{\mathbf{W}\setminus(\mathbf{S}'_j\cup\{X\})}\circ x}(\mathbf{t}_{\mathbf{S}'_j})$$

$$= \sum_{\tilde{x}\in\Sigma} Q_{\mathbf{S}'_1}(\mathbf{t}\circ\tilde{x}) \cdot \prod_{j=2}^{\ell'} Q_{\mathbf{S}'_j}(\mathbf{t}\circ x)$$

This proves part (A) discussed before. Next we prove part (B): where we provide bounds on the in-degree and the cardinality of c-components of H.

Lemma C.5. *Let $\mathbf{S}_1$ be the c-component of G that contains X. Let $\mathbf{W} = \mathbf{Y} \cup \mathbf{Pa}_G^+(\mathbf{S}_1)$, $H = Reduction(G, \mathbf{W})$ and let the c-components of H are denoted by $\mathbf{S}'_1, \ldots, \mathbf{S}'_{\ell'}$ without loss of generality let $\mathbf{S}_1 = \mathbf{S}'_1$. Then:*

1. *The in-degree of H is at most $f + k(d+1)$.*
2. *$|\mathbf{S}'_i| \leqslant f + kd$, for every i.*

Proof. The fact that H contains at most $f + k(d+1)$ vertices provides the bound on the in-degree. We know from Lemma C.3 that $\mathbf{S}'_1$ is a c-component of H and the remaining vertices of H is $(\mathbf{Y} \cup \mathbf{Pa}_H^+(\mathbf{S}'_1)) \setminus \mathbf{S}'_1$ which is of size at most $f + kd$ (since $\mathbf{S}_1 = \mathbf{S}'_1$ and $\mathbf{Pa}_G^+(\mathbf{S}_1) = \mathbf{Pa}_H(\mathbf{S}_1)$). □

We have now gathered the tools required to prove Theorem 2.5.

Proof of Theorem 2.5. Let $\mathbf{W} = \mathbf{Y} \cup \mathbf{Pa}_G^+(\mathbf{S}_1)$ and let $H = \text{Reduction}(G, \mathbf{W})$. The reduction $H = \text{Reduction}(G, \mathbf{W})$ can be performed using breadth first search/depth first search which can be done in time linear in the size of the input graph. We obtained an equivalent statement of Theorem 3.6 in Theorem C.4. Also from Lemma C.3 we know that the model over the causal graph H satisfies both Assumption 2.1 and Assumption 2.2. Hence by substituting n by $f + k(d+1)$ – the cardinality of observables of H; k by $f + kd$ – the size of the largest c-component of H; and d by $f + k(d+1)$ – the in-degree of H, into Theorem 2.3 we obtain the desired bounds on sample and time complexities. □