
SUPPLEMENTARY MATERIAL

Differentiable Structure Learning for General Binary Data

A Preliminary Technical Results

In this appendix, we present several key technical results that are essential to our proofs.

Lemma 1. *Suppose $X \sim \text{MultiBernoulli}(\mathbf{p})$ with $\mathbf{p} > 0$. Then for every subset $S \subseteq [p]$, the marginal vector X_S satisfies*

$$X_S \sim \text{MultiBernoulli}(\mathbf{p}_S),$$

where $\mathbf{p}_S > 0$.

The marginal vector X_S remains multivariate Bernoulli with natural parameter \mathbf{p}_S , and positivity of \mathbf{p} indicates the positivity of \mathbf{p}_S .

Lemma 2. *Suppose $X \sim \text{MultiBernoulli}(\mathbf{p})$ in the natural parameterization, or equivalently $X \sim \text{MultiBernoulli}(\mathbf{f})$ in the general parameterization. Then*

$$\mathbf{p} > 0 \iff |\mathbf{f}| < \infty.$$

The general parameter vector \mathbf{p} is strictly positive if and only if its corresponding natural parameter \mathbf{f} is strictly finite.

Lemma 3. *Let $\mathbf{p} > 0$ and suppose $X \sim \text{MultiBernoulli}(\mathbf{p})$. Fix any topological order π and index $j \in [p]$. Define the population negative log-likelihood*

$$\ell(w) = \mathbb{E} \left[\log(1 + \exp(w^\top \Phi(X_{\pi(1)}, \dots, X_{\pi(j-1)}))) - X_{\pi(j)} w^\top \Phi(X_{\pi(1)}, \dots, X_{\pi(j-1)}) \right], \quad w \in \mathbb{R}^{2^{j-1}}.$$

Then $\ell(w)$ is strictly convex and therefore admits a unique minimizer $w_{\pi,j}^*$.

This optimization problem is equivalent to fitting, in the population limit, a logistic regression of $X_{\pi(j)}$ on $\Phi((X_{\pi(1)}, \dots, X_{\pi(j-1)}))$. Because $\mathbf{f}_{\pi,j}$, as defined in Section 4.2, is one solution to the optimization above, and we show optimal solution is unique, then $w_{\pi,j}^* = \mathbf{f}_{\pi,j}$. Consequently, logistic regression perfectly recovers the true natural parameter.

Corollary 2. *Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be differentiable and m -strongly convex, i.e.*

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{m}{2} \|y - x\|^2 \quad \forall x, y \in \mathbb{R}^d$$

Denote by x^* the unique minimizer f , so $f(x^*) = \min_x f(x)$. Then for any constant $c \geq f(x^*)$ the subset

$$L_c = \{x \in \mathbb{R}^p : f(x) \leq c\}$$

is bounded.

Corollary 2 asserts that any strongly convex function possesses bounded level sets, which is used in the proof of Theorem 3.

B Proofs

In this appendix, we provide detailed proofs of the main results.

B.1 Proof of Lemma 1

From the proof of Corollary 1, we know that $X_S \sim \text{MultiBernoulli}(\mathbf{p}_S)$. Moreover, because $\mathbf{p} > 0$, then

$$P(X_S = x_s) = \sum_{x_{[p] \setminus S} \in \{0,1\}^{p-|S|}} P(X_S = x_s, X_{[p] \setminus S} = x_{[p] \setminus S}) > 0 \quad (16)$$

Therefore, for any x_s , $P(X_S = x_s) > 0$. So, $\mathbf{p}_S > 0$.

806 B.2 Proof of Lemma 2

807 **Sufficiency** By (62), each natural-parameter probability satisfies

$$\exp(f^{j_1 j_2 \dots j_r}) \quad (17)$$

$$= \frac{\prod \text{p(even \# zeros among } j_1, j_2, \dots, j_r \text{ components and other components are all zero)}}{\prod \text{p(odd \# zeros among } j_1, j_2, \dots, j_r \text{ components and other components are all zero)}}, \quad (18)$$

808 and since $p > 0$, every term $\exp(f^{j_1 \dots j_r})$ is strictly positive and finite. Hence

$$0 < \exp(f^{j_1 \dots j_r}) < \infty \iff |f^{j_1 \dots j_r}| < \infty. \quad (19)$$

809 **Necessity** Note that

$$S^{j_1 j_2 \dots j_r} = \sum_{1 \leq s \leq r} f^{j_s} + \sum_{1 \leq s < t \leq r} f^{j_s j_t} + \dots + f^{j_1 j_2 \dots j_r} \quad (20)$$

810 If $|f| < \infty$, then $|S^{j_1 \dots j_r}| < \infty$, so

$$0 < \exp(S^{j_1 j_2 \dots j_r}) < \infty \quad (21)$$

811 The joint probability of observing ones in positions j_1, \dots, j_r is

$$\begin{aligned} & \text{p}(j_1, j_2, \dots, j_r \text{ positions are one, others are zero}) \\ &= \frac{\exp(S^{j_1 j_2 \dots j_r})}{\exp(b(f))}. \\ &= \frac{\exp(S^{j_1 j_2 \dots j_r})}{\sum_{r=1}^K \left[1 + \left(\sum_{1 \leq j_1 < j_2 < \dots < j_r \leq K} \exp[S^{j_1 j_2 \dots j_r}] \right) \right]} \end{aligned} \quad (22)$$

812 which lies in the interval $(0, 1)$:

$$0 < p(j_1, j_2, \dots, j_r \text{ positions are one, others are zero}) < 1 \quad (23)$$

813 B.3 Proof of Lemma 3

814 Fix a topological order π and an index $j \in [p]$. To simplify notation, set

$$Y = X_{\pi(j)} \in \{0, 1\}, \quad Z = \Phi((X_{\pi(1)}, \dots, X_{\pi(j-1)})) \in \{0, 1\}^{2^{j-1}}, \quad (24)$$

815 and let $w \in \mathbb{R}^{2^{j-1}}$ be the parameter vector. The population objective is

$$\ell(w) = \mathbb{E}[\log(1 + \exp(w^\top Z)) - Y(w^\top Z)]. \quad (25)$$

816 A straightforward calculation shows

$$\nabla^2 \ell(w) = \mathbb{E}[\sigma(w^\top Z)(1 - \sigma(w^\top Z)) Z Z^\top], \quad \sigma(c) = \frac{1}{1 + e^{-c}}. \quad (26)$$

817 Since $Z \in \{0, 1\}^{2^{j-1}}$, every inner product $w^\top Z$ is finite and thus $\sigma(w^\top Z)(1 - \sigma(w^\top Z)) > 0$.

818 Define

$$m(w) = \min_{z \in \{0, 1\}^{2^{j-1}}} \sigma(w^\top z)(1 - \sigma(w^\top z)) > 0. \quad (27)$$

819 The marginal distribution of $(X_{\pi(1)}, \dots, X_{\pi(j-1)})$ is still a multivariate Bernoulli distribution, with
820 natural parameters $\mathbf{p}_{\pi, j-1}$. i.e., $(X_{\pi(1)}, \dots, X_{\pi(j-1)}) \sim \text{MultiBernoulli}(\mathbf{p}_{\pi, j-1})$. By lemma 1,

821 $q_{\min, j-1} = \min_{i \in [2^{j-1}]} [\mathbf{p}_{\pi, j-1}]_i > 0$. Hence for any nonzero $\beta \in \mathbb{R}^{2^{j-1}}$,

$$\begin{aligned} \beta^\top \nabla^2 \ell(w) \beta &= \sum_{z \in \{0, 1\}^{2^{j-1}}} p(z) \sigma(w^\top z)(1 - \sigma(w^\top z)) (\beta^\top z)^2 \\ &\geq q_{\min, j-1} m(w) \sum_{z \in \{0, 1\}^{2^{j-1}}} (\beta^\top z)^2 > 0, \end{aligned} \quad (28)$$

822 because the sum is over all $z \in \{0, 1\}^{2^{j-1}}$. So $(\beta^\top z)^2 > 0$ for at least one z when $\beta \neq 0$.
823 Since the Hessian is everywhere positive definite, $\ell(w)$ is strictly convex and therefore has a unique
824 minimizer.

825 **B.4 Proof of Theorem 1**

826 Let π be any permutation of $\{1, \dots, p\}$. Then the joint distribution of X can be written as

$$P(X) = \prod_{j=1}^p P(X_{\pi(j)} \mid X_{\pi(1), \dots, \pi(j-1)}) \quad (29)$$

827 In the population limit ($n \rightarrow \infty$), each conditional law has the Bernoulli form

$$P(X_{\pi(j)} \mid X_{\pi(1)}, \dots, X_{\pi(j-1)}) = q^{X_{\pi(j)}} (1 - q)^{1 - X_{\pi(j)}} \quad (30)$$

828 where

$$q = \text{logistic} \left(\mathbf{f}_{\pi, j}^\top \Phi \left((X_{\pi(1)}, \dots, X_{\pi(j-1)}) \right) \right) \quad (31)$$

829 Lemma 3 guarantees that logistic regression uniquely recovers $\mathbf{f}_{\pi, j}$. Moreover, the structural equation
830 model

$$Y_{\pi(j)} \sim \text{Bernoulli} \left(\text{logistic} \left(\mathbf{f}_{\pi, j}^\top \Phi \left((Y_{\pi(1)}, \dots, Y_{\pi(j-1)}) \right) \right) \right) \quad (32)$$

831 It induces exactly the same conditionals as X . i.e.,

$$Y_{\pi(j)} \mid Y_{\pi(1)}, \dots, Y_{\pi(j-1)} \stackrel{d}{=} X_{\pi(j)} \mid X_{\pi(1)}, \dots, X_{\pi(j-1)}, \quad (33)$$

832 Hence

$$Y \sim \text{MultiBernoulli}(\mathbf{p}), \quad (34)$$

833 establishing the desired result.

834 **B.5 Proof of Theorem 2**

835 By Theorem 1, for every $(\mathbf{f}_\pi, G_\pi) \in \mathcal{E}_{\min}(\mathbf{p})$, the vector \mathbf{f}_π defines, via the structural equation
836 model (7), a distribution

$$X \sim \text{MultiBernoulli}(\mathbf{p}) \quad (35)$$

837 that is Markov with respect to G_π . By definition of $\mathcal{E}_{\min}(\mathbf{p})$, each G_π is a sparsest graph in the
838 equivalence class $\mathcal{E}(\mathbf{p})$. Since all sparsest Markov representations lie in the same Markov equivalence
839 class, it follows that for any two pairs

$$(\mathbf{f}_{\pi_1}, G_{\pi_1}), (\mathbf{f}_{\pi_2}, G_{\pi_2}) \in \mathcal{E}_{\min}(\mathbf{p}), \quad (36)$$

840 their Markov classes coincide:

$$\mathcal{M}(G_{\pi_1}) = \mathcal{M}(G_{\pi_2}). \quad (37)$$

841 **B.6 Proof of Theorem 3**

842 The proof relies on Theorems 4 and 5 of Deng et al. [15]. We verify Assumptions A and B from that
843 work.

844 **Assumption A (1)** This requires the equivalence class to be finite. Since

$$|\mathcal{E}_{\min}(\mathbf{p})| \leq p! \quad (38)$$

845 the condition holds.

846 **Assumption A (2)** This requires the weighted adjacency matrix $W(\mathbf{H})$ to be L -Lipschitz. Recall
 847 that in (10),

$$[W(\mathbf{H})]_{ij} = \sum_{S \subseteq [p], i \in S} (h^{j,S})^2. \quad (39)$$

848 For simplicity we instead use the equivalent form

$$[W(\mathbf{H})]_{ij} = \sum_{S \subseteq [p], i \in S} |h^{j,S}|. \quad (40)$$

849 Let \mathbf{H}_1 and \mathbf{H}_2 be two parameter values. We show there exists L such that

$$\|W(\mathbf{H}_1) - W(\mathbf{H}_2)\|_2 \leq L \|\mathbf{H}_1 - \mathbf{H}_2\|_2. \quad (41)$$

850 First,

$$\|W(\mathbf{H}_1) - W(\mathbf{H}_2)\|_2 = \sqrt{\sum_j \sum_i \left(\sum_{S \subseteq [p], i \in S} |h_1^{j,S}| - \sum_{S \subseteq [p], i \in S} |h_2^{j,S}| \right)^2}. \quad (42)$$

851 Meanwhile,

$$\|\mathbf{H}_1 - \mathbf{H}_2\|_2 = \sqrt{\sum_j \sum_{S \subseteq [p]} (h_1^{j,S} - h_2^{j,S})^2}. \quad (43)$$

852 By Cauchy–Schwarz,

$$\begin{aligned} \left(\sum_{S \subseteq [p], i \in S} |h_1^{j,S}| - \sum_{S \subseteq [p], i \in S} |h_2^{j,S}| \right)^2 &= \left(\sum_{S \subseteq [p], i \in S} (|h_1^{j,S}| - |h_2^{j,S}|) \right)^2 \\ &\leq |S \subseteq [p], i \in S| \sum_{S \subseteq [p], i \in S} (|h_1^{j,S}| - |h_2^{j,S}|)^2 \\ &\leq 2^{p-1} \sum_{S \subseteq [p], i \in S} (h_1^{j,S} - h_2^{j,S})^2 \\ &\leq 2^{p-1} \sum_{S \subseteq [p]} (h_1^{j,S} - h_2^{j,S})^2 \end{aligned} \quad (44)$$

853 Hence

$$\begin{aligned} \|W(\mathbf{H}_1) - W(\mathbf{H}_2)\|_2 &= \sqrt{\sum_j \sum_i \left(\sum_{S \subseteq [p], i \in S} |h_1^{j,S}| - \sum_{S \subseteq [p], i \in S} |h_2^{j,S}| \right)^2} \\ &\leq \sqrt{\sum_j \sum_i 2^{p-1} \sum_{S \subseteq [p]} (h_1^{j,S} - h_2^{j,S})^2} \\ &\leq \sqrt{2^{p-1} p \sum_j \sum_{S \subseteq [p]} (h_1^{j,S} - h_2^{j,S})^2} \\ &= \sqrt{p} 2^{(p-1)/2} \sqrt{\sum_j \sum_{S \subseteq [p]} (h_1^{j,S} - h_2^{j,S})^2} \\ &= \sqrt{p} 2^{(p-1)/2} \|\mathbf{H}_1 - \mathbf{H}_2\|_2 \end{aligned} \quad (45)$$

854 Thus one may take $L = \sqrt{p} 2^{(p-1)/2}$.

855 **Assumption B** This requires $\mathbb{E}[s(\mathbf{H}; \mathbf{X})]$ to have bounded level sets. From Lemma 3 we know that
 856 under $\mathbf{p} > 0$, each population logistic loss is strongly convex. Since $\mathbb{E}[s(\mathbf{H}; \mathbf{X})]$ is a sum of p such
 857 strongly convex terms, it is itself strongly convex. By Corollary 2, any strongly convex function has
 858 bounded sublevel sets.

859 B.7 Proof of Theorem 4

860 Let π^* be the topological ordering of G guaranteed by Assumption A. Then

$$P(X) = \prod_{j=1}^p P(X_{\pi^*(j)} \mid X_{\pi^*(1), \dots, \pi^*(j-1)}) \quad (46)$$

861 Write $S_{\pi, j} = \{\pi^*(1), \dots, \pi^*(j-1)\}$. Under the linear SEM assumption,

$$X_{\pi^*(j)} \mid X_{S_{\pi, j}} = x_{S_{\pi, j}} \sim \text{Bernoulli}(\text{logistic}([w_{\pi^*(j)}]_{S_{\pi, j}}^\top X_{S_{\pi, j}} + c_{\pi^*(j)})) \quad (47)$$

862 where $w_{\pi^*(j)}$ and $c_{\pi^*(j)}$ are finite. Hence for any $x_{S_{\pi, j}}$,

$$0 < \text{logistic}(w_{\pi^*(j), S_{\pi, j}}^\top x_{S_{\pi, j}} + c_{\pi^*(j)}) < 1, \quad (48)$$

863 so both

$$P(X_{\pi^*(j)} = 1 \mid X_{S_{\pi, j}} = x_{S_{\pi, j}}) \quad \text{and} \quad P(X_{\pi^*(j)} = 0 \mid X_{S_{\pi, j}} = x_{S_{\pi, j}}) \quad (49)$$

864 lie in $(0, 1)$, implying $\mathbf{p} > 0$.

865 Although every $(\mathbf{f}_\pi, G_\pi) \in \mathcal{E}(\mathbf{p})$ generates $X \sim \text{MultiBernoulli}(\mathbf{p})$ via (7), the linear optimization
 866 (94) only admits first-order models. Thus any (\mathbf{f}_π, G_π) with higher-order terms is misspecified.
 867 Define

$$\mathcal{E}^{\text{linear}}(\mathbf{p}) = \{(\mathbf{f}_\pi, G_\pi) : (\mathbf{f}_\pi, G_\pi) \text{ where } \mathbf{f}_\pi \text{ only has first order term, } \forall \pi\} \quad (50)$$

868 By Assumption A, $(W^0, G^0) \in \mathcal{E}^{\text{linear}}(\mathbf{p})$. Let

$$\mathcal{E}_{\min}^{\text{linear}}(\mathbf{p}) = \{(\mathbf{f}_\pi, G_\pi) : (\mathbf{f}_\pi, G_\pi) \text{ is the minimal element, } (\mathbf{f}_\pi, G_\pi) \in \mathcal{E}^{\text{linear}}(\mathbf{p})\} \quad (51)$$

869 If (W^0, G^0) is not already minimal, we simply replace it by the sparsest element in $\mathcal{E}^{\text{linear}}(\mathbf{p})$, so
 870 that $(W^0, G^0) \in \mathcal{E}_{\min}^{\text{linear}}(\mathbf{p})$.

871 In the proof of Theorem 3, we verified that our model meets Assumptions A and B of Deng et al.
 872 [15]. Therefore, by Theorem 4 of Deng et al. [15], we have

$$\mathcal{O}_{\infty, \lambda, \delta}^{\text{linear}} = \mathcal{E}_{\min}^{\text{linear}}(\mathbf{p}), \quad \text{and hence} \quad (W^0, G^0) \in \mathcal{O}_{\infty, \lambda, \delta}^{\text{linear}}. \quad (52)$$

873 B.8 Proof of Corollary 1

874 Since $X \sim \text{MultiBernoulli}(\mathbf{p})$ and $X = (X_1, \dots, X_p)$. Then, $\forall S \subseteq [p]$, the range of $X_S \in$
 875 $\{0, 1\}^{|S|}$. The corresponding density mass

$$P(X_S = x_S) = \sum_{x_{[p] \setminus S} \in \{0, 1\}^{p-|S|}} P(X_S = x_S, X_{[p] \setminus S} = x_{[p] \setminus S}) \quad (53)$$

876 In such way, we could enumerate all the value for the $x_S \in \{0, 1\}^{|S|}$, and put them together to get the
 877 natural parameter for X_S , i.e., \mathbf{p}_S . As a consequence, $X_S \sim \text{MultiBernoulli}(\mathbf{p}_S)$.

878 For any $j \in [p]$, and any $S \in [p] \setminus j$, since $X_j \in \{0, 1\}$, so the conditional distribution $P(X_j \mid X_S)$ is
 879 Bernoulli distribution, and the probability

$$\begin{aligned} P(X_j = x_j \mid X_S = x_S) &= \frac{P(X_j = x_j, X_S = x_S)}{P(X_S = x_S)} \\ &= \frac{P(X_j = x_j, X_S = x_S)}{P(X_j = 1, X_S = x_S) + P(X_j = 0, X_S = x_S)} \end{aligned} \quad (54)$$

880 B.9 Proof of Corollary 2

881 Since f is strongly convex,

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{m}{2} \|y - x\|^2 \quad \forall x, y \in \mathbb{R}^d \quad (55)$$

882 Take $x = x^*$, then

$$f(y) \geq f(x^*) + \frac{m}{2} \|y - x^*\|^2 \quad \forall x, y \in \mathbb{R}^d \quad (56)$$

883 Rearranging,

$$f(y) \leq c \Rightarrow \frac{m}{2} \|y - x^*\|^2 \leq c - f(x^*) \Rightarrow \|y - x^*\| \leq \sqrt{\frac{2(c - f(x^*))}{m}} \quad (57)$$

884 Thus, $L_c \subseteq \left\{ y : \|y - x^*\| \leq \sqrt{\frac{2(c - f(x^*))}{m}} \right\}$, a bounded ball.

885 C Supplementary Technical Details and Examples

886 In this appendix, we collect additional technical derivations, algorithms, and definitions that support
887 our main theorems.

- 888 • In C.1, we give the explicit parameter transformation between the general parameter \mathbf{p} and
889 the natural parameter \mathbf{f} of the multivariate Bernoulli model [11];
- 890 • In C.2, we derive the logistic form of the conditional distributions in equation (3)
- 891 • In C.3, we formalize the graded-lexicographic ordering used to index the interaction features.
- 892 • In C.4, we present the population-level recovery algorithms for each topological order π .
- 893 • In C.5, we define the structural equation model framework underlying Theorem 1.
- 894 • In C.6, we review faithfulness, Markov equivalence, and the Sparsest Markov Representation
895 necessary for Theorems 2 and 3
- 896 • In C.7, we provide the derivation of our score function $s(\mathbf{H}; \mathbf{X})$ (negative log-likelihood
897 function).
- 898 • In C.8, we provide theoretical justification for the previous work [12, 4] with our general
899 framework.

900 C.1 Parameter Transformation in the Multivariate Bernoulli Distribution

901 All the material in this subsection can be found in [11]. We include here for the completeness.

902 The multivariate Bernoulli distribution has two different parameterization, one is using general
903 parameter \mathbf{p} and another one is using natural parameter \mathbf{f} .

904 The density is expressed by general parameter \mathbf{p} .

$$\begin{aligned} P(Y_1 = y_1, Y_2 = y_2, \dots, Y_K = y_K) &= p(y_1, y_2, \dots, y_K) \\ &= p(0, 0, \dots, 0)^{\prod_{j=1}^K (1 - y_j)} \\ &\quad \times p(1, 0, \dots, 0)^{[y_1 \prod_{j=2}^K (1 - y_j)]} \\ &\quad \times p(0, 1, \dots, 0)^{[(1 - y_1) y_2 \prod_{j=3}^K (1 - y_j)]} \dots \\ &\quad \times p(1, 1, \dots, 1)^{\prod_{j=1}^K y_j}, \end{aligned} \quad (58)$$

905 The density is expressed by natural parameter \mathbf{f} .

$$P(Y_1 = y_1, Y_2 = y_2, \dots, Y_K = y_K) = \exp \left(f^0 + \sum_{r=1}^p \left(\sum_{1 \leq j_1 < j_2 < \dots < j_r \leq p} f^{j_1 j_2 \dots j_r} B^{j_1 j_2 \dots j_r}(y) \right) \right) \quad (59)$$

906 To simplify the notation, we could define the quantity S to be

$$S^{j_1 j_2 \dots j_r} = \sum_{1 \leq s \leq r} f^{j_s} + \sum_{1 \leq s < t \leq r} f^{j_s j_t} + \dots + f^{j_1 j_2 \dots j_r} \quad (60)$$

907 and also define the interaction function B

$$B^{j_1 j_2 \dots j_r}(y) = y_{j_1} y_{j_2} \dots y_{j_r} \quad (61)$$

908 The following lemma shows the one-to-one mapping between general parameter \mathbf{p} and natural
909 parameter \mathbf{f} .

910 **Lemma 4** (Parameter transformation). *For the multivariate Bernoulli model, the general parameters
911 and natural parameters have the following relationship.*

$$\exp(f^{j_1 j_2 \dots j_r}) \quad (62)$$

$$= \frac{\prod p(\text{even \# zeros among } j_1, j_2, \dots, j_r \text{ components and other components are all zero})}{\prod p(\text{odd \# zeros among } j_1, j_2, \dots, j_r \text{ components and other components are all zero})}, \quad (63)$$

912 where $\#$ refers to the number of zeros among the superscript $y_{j_1} \dots y_{j_r}$ of f . In addition,

$$\exp(S^{j_1 j_2 \dots j_r}) = \frac{p(j_1, j_2, \dots, j_r \text{ positions are one, others are zero})}{p(0, 0, \dots, 0)} \quad (64)$$

913 and conversely the general parameters can be represented by the natural parameters

$$p(j_1, j_2, \dots, j_r \text{ positions are one, others are zero}) = \frac{\exp(S^{j_1 j_2 \dots j_r})}{\exp(b(\mathbf{f}))}. \quad (65)$$

914 where

$$b(\mathbf{f}) = \log \sum_{r=1}^K \left[1 + \left(\sum_{1 \leq j_1 < j_2 < \dots < j_r \leq K} \exp[S^{j_1 j_2 \dots j_r}] \right) \right] \quad (66)$$

915 C.2 Conditional distribution of multivariate Bernoulli distribution

916 In this part, we derive the conditional distribution of multivariate Bernoulli distribution. Especially,

$$\begin{aligned} & P(X_p = 1 \mid X_1 = x_1, \dots, X_{p-1} = x_{p-1}) \\ &= \text{logistic} \left(\sum_{r=1}^p \left(\sum_{1 \leq j_1 < j_2 < \dots < j_r = p \leq p} f^{j_1 \dots j_{r-1} p} x_{j_1} \dots x_{j_{r-1}} x_p \right) \right) \\ &= \text{logistic} (f^p + f^{1p} x_1 + f^{2p} x_2 \dots f^{p-1, p} f_{p-1} + f^{12p} x_1 x_2 + \dots + f^{1 \dots p} x_1 \dots x_{p-1}) \end{aligned} \quad (67)$$

917 It is known the multivariate Bernoulli distribution in exponential form can be written as

$$P(X_1 = x_1, \dots, X_p = x_p) = \exp \left(f^0 + \sum_{r=1}^p \left(\sum_{1 \leq j_1 < j_2 < \dots < j_r \leq p} f^{j_1 j_2 \dots j_r} B^{j_1 j_2 \dots j_r}(x) \right) \right) \quad (68)$$

918 Then,

$$\begin{aligned} P(X_p = 1 \mid X_{-p} = x_{-p}) &= \frac{P(X_p = 1, X_{-p} = x_{-p})}{P(X_{-p} = x_{-p})} \\ &= \frac{P(X_p = 1, X_{-p} = x_{-p})}{P(X_p = 1, X_{-p} = x_{-p}) + P(X_p = 0, X_{-p} = x_{-p})} \\ &= \frac{1}{1 + \frac{P(X_p = 0, X_{-p} = x_{-p})}{P(X_p = 1, X_{-p} = x_{-p})}} \end{aligned} \quad (69)$$

919 where $X_{-p} = (X_1, \dots, X_{p-1})$.

$$\begin{aligned}
& P(X_1 = x_1, \dots, X_p = x_p) \\
&= \exp \left(f^0 + \sum_{r=1}^p \left(\sum_{1 \leq j_1 < j_2 < \dots < j_r \leq p} f^{j_1 j_2 \dots j_r} B^{j_1 j_2 \dots j_r}(x) \right) \right) \\
&= \exp \left(f^0 + \sum_{r=1}^p \left(\sum_{\substack{1 \leq j_1 < j_2 < \dots < j_r \leq p \\ p \in \{j_1, \dots, j_r\}}} f^{j_1 j_2 \dots j_r} B^{j_1 j_2 \dots j_r}(x) + \sum_{\substack{1 \leq j_1 < j_2 < \dots < j_r \leq p \\ p \notin \{j_1, \dots, j_r\}}} f^{j_1 j_2 \dots j_r} B^{j_1 j_2 \dots j_r}(x) \right) \right) \tag{70}
\end{aligned}$$

920 Then,

$$\begin{aligned}
& P(X_1 = x_1, \dots, X_p = 1) \\
&= \exp \left(f^0 + \sum_{r=1}^p \left(\sum_{\substack{1 \leq j_1 < j_2 < \dots < j_r \leq p \\ p \in \{j_1, \dots, j_r\}}} f^{j_1 j_2 \dots j_r} B^{j_1 j_2 \dots j_r}((x_{-p}, 1)) + \sum_{\substack{1 \leq j_1 < j_2 < \dots < j_r \leq p \\ p \notin \{j_1, \dots, j_r\}}} f^{j_1 j_2 \dots j_r} B^{j_1 j_2 \dots j_r}(x) \right) \right) \tag{71}
\end{aligned}$$

921

$$P(X_1 = x_1, \dots, X_p = 0) = \exp \left(f^0 + \sum_{r=1}^p \left(\sum_{\substack{1 \leq j_1 < j_2 < \dots < j_r \leq p \\ p \notin \{j_1, \dots, j_r\}}} f^{j_1 j_2 \dots j_r} B^{j_1 j_2 \dots j_r}(x) \right) \right) \tag{72}$$

922 Finally, put them together,

$$\begin{aligned}
& P(X_p = 1 \mid X_{-p} = x_{-p}) \\
&= \frac{1}{1 + \frac{P(X_p=0, X_{-p}=x_{-p})}{P(X_p=1, X_{-p}=x_{-p})}} \\
&= \frac{1}{1 + \exp \left(- \sum_{r=1}^p \left(\sum_{\substack{1 \leq j_1 < j_2 < \dots < j_r \leq p \\ p \in \{j_1, \dots, j_r\}}} f^{j_1 j_2 \dots j_r} B^{j_1 j_2 \dots j_r}((x_{-p}, 1)) \right) \right)} \\
&= \sigma \left(f^p + f^{1p} x_1 + f^{2p} x_2 \dots f^{p-1,p} f_{p-1} + f^{12p} x_1 x_2 + \dots + f^{1 \dots p} x_1 \dots x_{p-1} \right) \tag{73}
\end{aligned}$$

923 where $\sigma(x) = \frac{1}{1 + \exp(-x)}$

924 C.3 Graded-lexicographic order

925 To index the 2^p interaction features and corresponding parameters in a consistent way, we use the
926 *graded-lexicographic* order on subsets of $[p]$, where for any finite set S , $|S|$ denotes its cardinality
927 (the number of elements in S).

928 **Definition 4** (Graded-lexicographic order). *Let π be a permutation of $[p]$. For any two subsets*
929 *$S, T \subseteq [p]$, we say $S \prec_{\text{grlex}} T$ if either*

- 930 1. $|S| < |T|$, or
- 931 2. $|S| = |T|$ and, when listing the elements of S and T in ascending order under π , the first
932 index at which they differ belongs to S .

933 Under this rule, all subsets are grouped by increasing cardinality, and ties are broken by the usual lex
934 order induced by π .

935 **Example.** Take $p = 3$ and the identity order $\pi = (1, 2, 3)$. Then the graded-lexicographic sequence
936 of subsets is

$$\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}. \tag{74}$$

Algorithm 1: RECOVERPARENTS(\mathbf{p}, π, j)

```
1:  $\text{PA}(\pi(j)) \leftarrow \{\}$ , and  $X_{\pi,j} \leftarrow (X_{\pi(1)}, \dots, X_{\pi(j)})$ 
2: Compute the general parameters  $\mathbf{p}_{\pi,j}$  of  $X_{\pi,j}$  // Compute the marginal distribution of  $X_{\pi,j}$ 
3: Convert  $\mathbf{p}_{\pi,j}$  to natural parameters  $\mathbf{f}_{\pi,j}$  utilizing Lemma 4
4: for  $i = 1, \dots, j - 1$  do
    | if  $\sum_{S \subseteq [\pi(1), \dots, \pi(j-1), \pi(j)] \setminus [\pi(i), \pi(j)]} (f_{\pi,j}^{\pi(i), \pi(j), S})^2 > 0$  then
    | |  $\text{PA}(\pi(j)) \leftarrow \text{PA} \cup \{\pi(i)\}$ 
5: return  $(\text{PA}(\pi(j)), \mathbf{f}_{\pi,j})$ 
```

Algorithm 2: RECOVERDAG(\mathbf{p}, π)

Input: Probability vector \mathbf{p} (or empirical count) and topological sort π

Output: DAG G_π , and natural parameters \mathbf{f}_π

```
1  $G_\pi \leftarrow$  empty graph and  $\mathbf{f}_\pi \leftarrow \{\}$ 
2 for  $j = 1, 2, \dots, p$  do
3    $(\text{PA}(\pi(j)), \mathbf{f}_{\pi,j}) \leftarrow \text{RECOVERPARENTS}(\mathbf{p}, \pi, j)$  //  $\text{PA}(\pi(j))$ : the parents of node  $\pi(j)$ 
4   for  $i \in \text{PA}(\pi(j))$  do
5     | Add edge  $X_i \rightarrow X_j$  to  $G_\pi$ 
6    $\mathbf{f}_\pi \leftarrow \mathbf{f}_\pi \cup \{\mathbf{f}_{\pi,j}\}$ 
```

937 Accordingly, the extended feature map $\Phi(X) = [B^S(X)]_{S \subseteq [3]}$ becomes

$$\Phi(X) = [1, X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3, X_1X_2X_3]. \quad (75)$$

938 Similarly, for any j and order π , the parameter block $\mathbf{f}_{\pi,j} \in \mathbb{R}^{2^{j-1}}$ is arranged so that its entries
939 align one-to-one with $\Phi(X_{\pi(1)}, \dots, X_{\pi(j-1)})$ in graded-lexicographic order.

940 C.4 Procedure for recovering causal graph and parameters

941 To formalize the recovery procedure from Section 4.2, we present Algorithms 1 and 2.

942 **Algorithm 1: Parent and parameter recovery.** For a fixed topological order π and node index
943 $j \in [p]$, this algorithm

- 944 1. computes the marginal probabilities $\mathbf{p}_{\pi,j}$ of $(X_{\pi(1)}, \dots, X_{\pi(j)})$,
- 945 2. converts $\mathbf{p}_{\pi,j}$ to the natural-parameter block $\mathbf{f}_{\pi,j}$ via Lemma 4, and
- 946 3. selects the parent set $\text{PA}_\pi(j)$ using the nonzero-coefficient criterion in (4).

947 For estimating the natural-parameter block $\mathbf{f}_{\pi,j}$, Section 4.2 uses a logistic regression approach [21].
948 In Algorithm 1, we instead compute $\mathbf{f}_{\pi,j}$ by applying the mapping of Lemma 4 to the marginal
949 probabilities. Under the positivity assumption $\mathbf{p} > 0$, these two methods are equivalent and yield
950 identical estimates for $\mathbf{f}_{\pi,j}$.

951 **Algorithm 2: Equivalence-class enumeration.** This algorithm iterates over all $p!$ permutations
952 $\pi \in \mathfrak{S}_p$ and, for each, calls Algorithm 1 for every $j = 1, \dots, p$. It assembles the corresponding DAG
953 G_π and parameter collection $\mathbf{f}_\pi = \{\mathbf{f}_{\pi,j}\}_{j=1}^p$. The output is the full equivalence class

$$\mathcal{E}(\mathbf{p}) = \{(\mathbf{f}_\pi, G_\pi) : (\mathbf{f}_\pi, G_\pi) \text{ is returned for some } \pi\}.$$

954 Although we state the algorithm in terms of the population vector \mathbf{p} , in practice one can simply input
955 the data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, since empirical frequencies convert \mathbf{X} into \mathbf{p} .

956 Building on Algorithms 1 and 2, Algorithm 3 then enumerates all graph-parameter pairs in $\mathcal{E}(\mathbf{p})$ and
957 retains only those with the fewest edges, thereby recovering the minimal equivalence class $\mathcal{E}_{\min}(\mathbf{p})$
958 as defined in (8).

Algorithm 3: RECOVERSPARESTDAG(p)

Input: Probability vector p (empirical count)

```

1  $\mathcal{S} \leftarrow \{\}$ 
2 for each  $\pi \in \mathfrak{G}_p$  do
3    $(G_\pi, \mathbf{f}_\pi) \leftarrow \text{RECOVERGRAPH}(p, \pi) // \mathfrak{G}_p$ : set of all the permutation on  $p$  variables
4    $\mathcal{S} \leftarrow \mathcal{S} \cup \{(G_\pi, \mathbf{f}_\pi)\}$ 
5 return  $((G_\pi, \mathbf{f}_\pi) \in \mathcal{S} : s_{G_\pi} \leq s_{G_{\tilde{\pi}}}, \forall (G_{\tilde{\pi}}, \mathbf{f}_{\tilde{\pi}}) \in \mathcal{S})$ 

```

C.5 Structural equation model

An structural equation model (SEM) [32] $(X, f, P(N))$ over the random vector $X = (X_1, \dots, X_p)$ is a collection of p structural equations of the form:

$$X_j = f_j(X, N_j), \quad \partial_k f_j = 0 \text{ if } k \notin \text{PA}(j), \quad (76)$$

where $f = (f_j)_{j=1}^p$ is a collection of functions $f_j : \mathbb{R}^{p+1} \rightarrow \mathbb{R}$, here $N = (N_1, \dots, N_p)$ is a vector of independent noises with distribution $P(N)$, and $\text{PA}(j)$ denotes the set of parents of node j . Here, $\partial_k f_j$ denotes the partial derivative of f_j w.r.t. X_k , which is identically zero when f_j is independent of X_k , i.e. $f_j(X, N_j) = f_j(X_{\text{PA}(j)}, N_j)$. The graphical structure induced by the SEM, assumed to be a DAG, will be represented by the following $p \times p$ weighted adjacency matrix B :

$$B = B(f), \quad B_{ij} = \|\partial_i f_j\|_2, \quad (77)$$

and we use $G(B)$ to denote the corresponding binary adjacency matrix.

The structural-equation framework in (76) provides a fully generative foundation for causal discovery by specifying, for each variable, an explicit functional dependence on its direct causes. Its generality encompasses a wide range of models—linear SEMs, additive-noise models, post-nonlinear models, generalized linear models, and more expressive non-linear SEMs. A major virtue of SEMs is their universality: *any* joint distribution can be represented in the form (76) (Proposition 7.1 of 32). In practice, however, researchers impose strong parametric restrictions—such as linearity, additivity, or low-order interactions—that may fail to capture the full complexity of real-world data.

C.6 Faithfulness, sparsest Markov representation, Markov equivalence class

We formally define the concepts mentioned in Section 5.

Definition 5 (Faithfulness [32]). *A pair (G, P) is said to be faithful if:*

$$X_i \perp\!\!\!\perp X_j \mid X_K \iff X_i \text{ and } X_j \text{ are } d\text{-separated by } X_K \text{ in } G, \quad (78)$$

for all disjoint subsets $\{i, j\}, K \subseteq V$. That is, every conditional independence in P corresponds exactly to a d -separation in G , and vice versa.

Definition 6 (Markov Equivalence Class [38]). *Two DAGs G_1 and G_2 on the same vertex set V are Markov equivalent if they encode the same set of conditional independence relations—equivalently, they have the same skeleton (undirected edges) and the same set of v-structures (induced subgraphs of the form $i \rightarrow k \leftarrow j$ with i and j not adjacent). The Markov equivalence class of a DAG G is*

$$\mathcal{M}(G) = \{G' : G' \text{ is a DAG and } G' \text{ is Markov equivalent to } G\}. \quad (79)$$

Definition 7 (Sparsest Markov Representation [35]). *A pair (G^0, P) satisfies the Sparsest Markov Representation (SMR) assumption if:*

1. (G^0, P) satisfies the Markov property, i.e. every d -separation in G^0 implies the corresponding conditional independence in P .
2. For any other DAG $G \notin \mathcal{M}(G^0)$ satisfying the Markov property with respect to P , we have

$$|E(G)| > |E(G^0)|.$$

Equivalently, G^0 is the (unique up to Markov equivalence) sparsest DAG compatible with P .

990 C.7 Derivation of the Logistic Loss

991 First, consider a single example (X, y) with feature vector $X \in \mathbb{R}^m$, binary label $y \in \{0, 1\}$, and
 992 parameter vector $w \in \mathbb{R}^m$. Let

$$q = \text{logistic}(w^\top X) = \frac{1}{1 + \exp(-w^\top X)}. \quad (80)$$

993 The (negative) log-likelihood is

$$\begin{aligned} \ell(w; X, y) &= -\log(q^y (1 - q)^{1-y}) \\ &= -y \log q - (1 - y) \log(1 - q). \end{aligned} \quad (81)$$

994 Noting that

$$\log \frac{q}{1 - q} = w^\top X, \quad \log(1 - q) = -\log(1 + \exp(w^\top X)), \quad (82)$$

995 we obtain the familiar logistic-loss form:

$$\ell(w; X, y) = \log(1 + \exp(w^\top X)) - y(w^\top X). \quad (83)$$

996 In our setting, each “feature” is replaced by the *extended* feature matrix $\Phi(\mathbf{X}) \in \mathbb{R}^{n \times 2^p}$, and each
 997 “label” is one column of the data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$. Stacking over all p columns and averaging over
 998 n samples yields

$$\begin{aligned} \ell(\mathbf{H}; \mathbf{X}) &= \frac{1}{n} \sum_{j=1}^p \sum_{i=1}^n \left[\log(1 + \exp[\Phi(\mathbf{X})\mathbf{H}]_{ij}) - \mathbf{X}_{ij} [\Phi(\mathbf{X})\mathbf{H}]_{ij} \right] \\ &= \frac{1}{n} \sum_{j=1}^p \mathbf{1}_n^\top \left(\log(\mathbf{1}_n + \exp(\Phi(\mathbf{X})\mathbf{H})) - \mathbf{X}_j \circ (\Phi(\mathbf{X})\mathbf{H}) \right), \end{aligned} \quad (84)$$

999 where \circ denotes the Hadamard product and $\mathbf{1}_n \in \mathbb{R}^n$ is the all-ones vector.

1000 C.8 Theoretical justification for previous works

1001 Under Assumption A, there exists a topological ordering π consistent with G such that each

$X_{\pi(j)}$ is generated by a linear combination of $(X_{\pi(1)}, \dots, X_{\pi(j-1)}) \in \mathbb{R}^{j-1}$ via the logistic link, (85)

1002 rather than via the logistic link on $\Phi((X_{\pi(1)}, \dots, X_{\pi(j-1)})) \in \mathbb{R}^{2^{j-1}}$. Importantly, Algorithms 1
 1003 and 2 remain valid under this assumption. For any other topological sort $\tilde{\pi} \neq \pi$, the output $(G_{\tilde{\pi}}, f_{\tilde{\pi}})$
 1004 from Algorithm 2 may include higher-order interaction terms; nevertheless, by the structural equation
 1005 model (7), it still recovers the exact distribution

$$X \sim \text{MultiBernoulli}(\mathbf{p}), \quad (86)$$

1006 so Theorem 1 continues to hold. Moreover, under Assumption A we can reduce the dimensionality of
 1007 the optimization (13) from $\mathbf{H} \in \mathbb{R}^{2^p \times p}$ to $\mathbf{H} \in \mathbb{R}^{(p+1) \times p}$. We formalize this reduction below.

1008 Define the parameter matrix

$$\mathbf{H}_j = (\underbrace{h^{j,0}}_{\text{constant}}, \underbrace{h^{j,1}, \dots, h^{j,p}}_{\text{first order}})^\top \in \mathbb{R}^{p+1} \quad \mathbf{H} = (\mathbf{H}_1, \dots, \mathbf{H}_p) \in \mathbb{R}^{(p+1) \times p} \quad (87)$$

1009 where $h^{j,0}$ is the intercept and $h^{j,1}, \dots, h^{j,p}$ are the first-order coefficients.

1010 The induced adjacency matrix is

$$[W(\mathbf{H})]_{ij} = |h^{j,i}| \quad (88)$$

1011 Self-loops are forbidden, so we impose

$$h^{j,j} = 0 \quad \forall j \in [p] \quad (89)$$

1012 Redefine the feature map row-wise as

$$\Phi(X) = [1, X_1, \dots, X_p], \quad (90)$$

1013 so that for a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\Phi(\mathbf{X})$ applies Φ to each row.

1014 The score (negative log-likelihood) remains

$$\ell(\mathbf{H}; \mathbf{X}) = \frac{1}{n} \sum_{i=1}^p \mathbf{1}_n^\top (\log(\mathbf{1}_n + \exp(\Phi(\mathbf{X})\mathbf{H})) - \mathbf{X}_i \circ (\Phi(\mathbf{X})\mathbf{H})) \quad (91)$$

1015 We continue to use the quasi-MCP penalty [15], defined by

$$\text{quasi-MCP: } p_{\lambda, \delta}(t) = \lambda \left[\left(|t| - \frac{t^2}{2\delta} \right) \mathbb{1}(|t| < \delta) + \frac{\delta}{2} \mathbb{1}(|t| > \delta) \right] \quad (92)$$

1016 Our final score function is as below

$$s(\mathbf{H}; \lambda, \delta, \mathbf{X}) = s(\mathbf{H}; \mathbf{X}) + p_{\lambda, \delta}(W(\mathbf{H})) \quad (93)$$

1017 We formulate this task as the single continuous optimization problem

$$\begin{aligned} \min_{\mathbf{H}} \quad & s(\mathbf{H}; \lambda, \delta, \mathbf{X}) \\ \text{subject to} \quad & h(W(\mathbf{H})) = 0 \\ & h^{j,j} = 0 \quad \forall j \in [p] \end{aligned} \quad (94)$$

1018 Define the global optimal solution of (94) as

$$\mathcal{O}_{n, \lambda, \delta}^{\text{linear}} = \{(\mathbf{H}^*, G(W(\mathbf{H}^*))) : \mathbf{H}^* \text{ is a minimizer of (94)}\} \quad (95)$$

1019 Let $\mathcal{O}_{\infty, \lambda, \delta}^{\text{linear}}$ denote the set of minimizers of (94) when the empirical loss $s(\mathbf{H}; \lambda, \delta, \mathbf{X})$ is replaced by
1020 its population counterpart $\mathbb{E}[s(\mathbf{H}; \lambda, \delta, \mathbf{X})]$. Let us collect all the parameters in assumption A.

$$H^0 = \begin{bmatrix} c_1 & \dots & c_p \\ w_1 & \dots & w_p \end{bmatrix} \in \mathbb{R}^{p+1 \times p} \quad (96)$$

1021 **Theorem 4.** Suppose Assumption A holds, then $X \sim \text{MultiBernoulli}(\mathbf{p})$ where $\mathbf{p} > 0$. Moreover,
1022 there exist $\lambda, \delta > 0$ sufficiently small such that $(H^0, G^0) \in \mathcal{O}_{\infty, \lambda, \delta}^{\text{linear}}$ where G^0 is the ground truth
1023 graph in Assumption A.

1024 It is important to note that under this assumption

$$\mathcal{O}_{\infty, \lambda, \delta}^{\text{linear}} \neq \mathcal{E}_{\min}(\mathbf{p}), \quad (97)$$

1025 because there may exist a topological sort π for which $(\mathbf{f}_\pi, G_\pi) \in \mathcal{E}_{\min}(\mathbf{p})$ involves higher-order
1026 terms, whereas every solution in $\mathcal{O}_{\infty, \lambda, \delta}^{\text{linear}}$ contains only first-order terms.

1027 D Experiments

1028 In this section, we present comprehensive experimental details, including the graph types evaluated,
1029 the data-generation process, the baseline methods for comparison, the steps required to reproduce our
1030 implementation, and the evaluation metrics employed.

1031 D.1 Experimental Setting

1032 In this section, we outline the process for generating graphs and data. For each model, a random graph
1033 G is generated using one of two types of random graph models: Erdős-Rényi(ER) or Scale-Free (SF).
1034 The models are specified to have, on average, kp edges, where $k \in \{1, 2, 4\}$. These configurations
1035 are denoted as ER k or SF k , respectively.

- 1036 • *Erdős-Rényi(ER)*, Random graphs whose edges are add independently with equal probability.
1037 We simulated models with $p, 2p$ and $4p$ edges (in expectation) each, denoted by ER1, ER2,
1038 and ER4 respectively.
- 1039 • *Scale-free network(SF)*. Network simulated according to the preferential attachment process.
1040 We simulated scale-free network with $p, 2p$ and $4p$ edges and $\beta = 1$, where β is the exponent
1041 used in the preferential attachment process.

Algorithm 4: Generate data matrix \mathbf{X}

Input: DAG G , sample size n , interaction type $\tau \in \{1\text{st}+2\text{nd}, 1\text{st}+\text{pth}, 1\text{st}+2\text{nd}+\text{pth}, 2\text{nd}, \text{pth}\}$

Output: $\mathbf{X} \in \{0, 1\}^{n \times p}$

```

1 Compute a topological ordering  $\pi$  of  $G$ 
2 for  $j \leftarrow 1$  to  $p$  do
3   Sample  $w_{\text{PA}(\pi(j))} = (w_k)_{k=1}^{2^{|\text{PA}(\pi(j))|}} \in \mathbb{R}^{2^{|\text{PA}(\pi(j))|}}$ ,  $w_k \stackrel{\text{iid}}{\sim} \text{Unif}([-2, -1] \cup [1, 2])$ .
4   if  $\tau = 1\text{st}+2\text{nd}$  then
5      $q \leftarrow \text{logistic}(w_{\text{PA}(\pi(j))}^\top \Phi^{1\text{st}+2\text{nd}}(\mathbf{X}_{\text{PA}(\pi(j))}))$ 
6   else if  $\tau = 1\text{st}+\text{pth}$  then
7      $q \leftarrow \text{logistic}(w_{\text{PA}(\pi(j))}^\top \Phi^{1\text{st}+\text{pth}}(\mathbf{X}_{\text{PA}(\pi(j))}))$ 
8   else if  $\tau = 2\text{nd}$  then
9      $q \leftarrow \text{logistic}(w_{\text{PA}(\pi(j))}^\top \Phi^{2\text{nd}}(\mathbf{X}_{\text{PA}(\pi(j))}))$ 
10  else
11     $q \leftarrow \text{logistic}(w_{\text{PA}(\pi(j))}^\top \Phi^{\text{pth}}(\mathbf{X}_{\text{PA}(\pi(j))}))$ 
12   $\mathbf{X}_{\pi(j)} \sim \text{Bernoulli}(q)$ 

```

1042 **General binary data** Since we wish to study structure learning for general binary data, Theorem 1
 1043 implies that for any $p > 0$,

$$X \sim \text{MultiBernoulli}(\mathbf{p}) \quad (98)$$

1044 can be generated via the SEM (7). To allow different interaction orders, define the following extended
 1045 feature maps for $X = (X_1, \dots, X_p)$:

$$\begin{aligned}
 \Phi^{1\text{st}+2\text{nd}}(X) &= (\underbrace{1}_{\text{constant}}, \underbrace{X_1, \dots, X_p}_{\text{first order}}, \underbrace{X_1 X_2, \dots, X_{p-1} X_p}_{\text{second order}}, \underbrace{0}_{\text{third order}}, \underbrace{0}_{\text{forth to } (p-1)\text{-th order}}, \underbrace{0}_{p\text{-th order}})^\top \in \mathbb{R}^{2^p} \\
 \Phi^{1\text{st}+2\text{nd}+\text{pth}}(X) &= (\underbrace{1}_{\text{constant}}, \underbrace{X_1, \dots, X_p}_{\text{first order}}, \underbrace{X_1 X_2, \dots, X_{p-1} X_p}_{\text{second order}}, \underbrace{0}_{\text{third order}}, \underbrace{0}_{\text{forth to } (p-1)\text{-th order}}, \underbrace{X_1 \dots X_p}_{p\text{-th order}})^\top \in \mathbb{R}^{2^p} \\
 \Phi^{1\text{st}+\text{pth}}(X) &= (\underbrace{1}_{\text{constant}}, \underbrace{X_1, \dots, X_p}_{\text{first order}}, \underbrace{0}_{\text{second order}}, \underbrace{0}_{\text{third order}}, \underbrace{0}_{\text{forth to } (p-1)\text{-th order}}, \underbrace{X_1 X_2 \dots X_p}_{p\text{-th order}})^\top \in \mathbb{R}^{2^p} \\
 \Phi^{2\text{nd}}(X) &= (\underbrace{1}_{\text{constant}}, \underbrace{0}_{\text{first order}}, \underbrace{X_1 X_2, \dots, X_{p-1} X_p}_{\text{second order}}, \underbrace{0}_{\text{third order}}, \underbrace{0}_{\text{forth to } (p-1)\text{-th order}}, \underbrace{0}_{p\text{-th order}})^\top \in \mathbb{R}^{2^p} \\
 \Phi^{\text{pth}}(X) &= (\underbrace{1}_{\text{constant}}, \underbrace{0}_{\text{first order}}, \underbrace{0}_{\text{second order}}, \underbrace{0}_{\text{third order}}, \underbrace{0}_{\text{forth to } (p-1)\text{-th order}}, \underbrace{X_1 X_2 \dots X_p}_{p\text{-th order}})^\top \in \mathbb{R}^{2^p}
 \end{aligned} \quad (99)$$

1046 where in each vector the nonzero blocks correspond respectively to the constant term, first-order
 1047 terms, second-order terms, and highest-order term. By convention, if $X = \emptyset$, then all four maps
 1048 reduce to the scalar $(1) \in \mathbb{R}$. When applied to a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, each $\Phi(\mathbf{X})$ operates
 1049 row-wise.

1050 Finally, given a random DAG $B \in \{0, 1\}^{p \times p}$ sampled from one of our graph models, we generate \mathbf{X}
 1051 using Algorithm 4, choosing the desired interaction map according to whether we study first+second,
 1052 first+highest, second, or highest-order interactions.

1053 **Simulation** We generate random dataset $\mathbf{X} \in \mathbb{R}^{n \times p}$ by sampling i.i.d from the models described
 1054 above. For each simulation, we produce datasets with n sample cross graphs with p nodes.

- 1055 • **(Small Graph)** $p = \{5, 6, 7, 8, 9\}$, $k = \{1, 2\}$, $n = 10000$ and graph types: {ER,SF}
- 1056 • **(Large Graph)** $p = \{10, 20, 30, 40\}$, $k = \{1, 2, 4\}$, $n = 1000$ and graph types: {ER,SF}

1057 **Implementation** For each dataset, we applied several structural learning algorithms, including
 1058 fast greedy equivalence search (FGES [34]), constraint-based methods (PC [38]), DAGMA [4],

NOTEARS [47]. The implementation details are provided in the following paragraph. After running the algorithms, a post-processing threshold of 0.3 was applied to the estimated B_{est} to prune small values, following the same procedure in [46, 4].

- Fast Greedy Equivalence Search (FGES [34]) is based on greedy search and assumes linear dependency between variables. The implementation is based on the py-tetrad package, available at <https://github.com/cmu-phil/py-tetrad>. We use `search.use_bdeu(sample_prior=10, structure_prior=0)`.
- PC [38] is constraint-based method and based on uses conditional independence induced by causal relationships to learn those causal relationships. The implementation is based on the py-tetrad package, available at <https://github.com/cmu-phil/py-tetrad>. We use `search.use_chi_square(alpha=0.1)`
- NOTEARS-MLP [47] is a continuous DAG-learning method that employs a least-squares loss with ℓ_1 regularization. Its Python implementation is available at <https://github.com/xunzheng/notears>. In our variant, we insert a sigmoid activation $\sigma(x) = 1/(1 + \exp(-x))$ on the final layer and replace the original loss with the cross-entropy (logistic) loss to accommodate binary data. After estimating the weighted adjacency matrix W_{est} via NOTEARS-MLP, we prune all entries below a threshold of 0.3, compute a topological ordering of the resulting graph, and then apply Algorithm 2 with first and second order to obtain the final structure. Finally, we remove any remaining edges whose weight does not exceed 1.0 to eliminate spurious connections. We name this method as NOTEARS-MLP-REG. These heuristic methods are applied to larger graphs with $d \in \{10, 20, 30, 40\}$.
- DAGMA [4] is a continuous DAG-learning algorithm that achieves improved accuracy and faster computation, with barrier methods. Its implementation can be found at <https://github.com/kevinsbello/dagma>. To highlight that the original DAGMA only models first-order interactions, we refer to it as DAGMA-1ST. By solving (13) while incorporating all higher-order interactions, we arrive at our extended method, denoted DAGMA-HO. For small graphs, we implement the full formulation (13) including all higher-order interactions; hence, DAGMA-HO is applied for $d \in \{5, 6, 7, 8, 9\}$.

Hyperparameter tuning Theorem 3 indicates that one should ideally choose small values of λ and δ for the quasi-MCP penalty. In practice, however, achieving the global optimum of (13) is infeasible, and if λ and δ are too small the penalty becomes ineffective and the algorithm may fail to recover the sparsest solution. To mitigate this, we adopt the continuation strategy of Deng et al. [15]: start with relatively large λ and δ , solve (13) via DAGMA-HO to obtain an initial estimate \mathbf{H}_{est} , then iteratively shrink λ and δ by a factor $\gamma < 1$, using the previous estimate as the warm start for the next run of DAGMA-HO. We terminate when the negative log-likelihood $s(\mathbf{H}_{\text{est}}; \mathbf{X})$ ceases to decrease. Empirically, $\gamma = 0.5$, $\lambda = 0.05$, and $\delta = 0.2$ perform well in our experiments.

Equipment The experiments are conducted in the following CPU architectures

- Intel Broadwell—28 cores @ 2.4 GHz with 64 GB memory per node
- Intel Skylake—40 cores @ 2.4 GHz with 96 GB memory per node

D.2 Metrics

- **Structural Hamming distance (SHD)**: A standard benchmark in the structure learning literature that counts the total number of edges additions, deletions, and reversals needed to convert the estimated graph into the true graph. Since our data specified in (1) is nonidentifiable, the Structural Hamming Distance (SHD) is calculated with respect to the completed partially directed acyclic graph (CPDAG) of the ground truth and B_{est} .

1105 E Additional Figures

1106 E.1 Small graphs

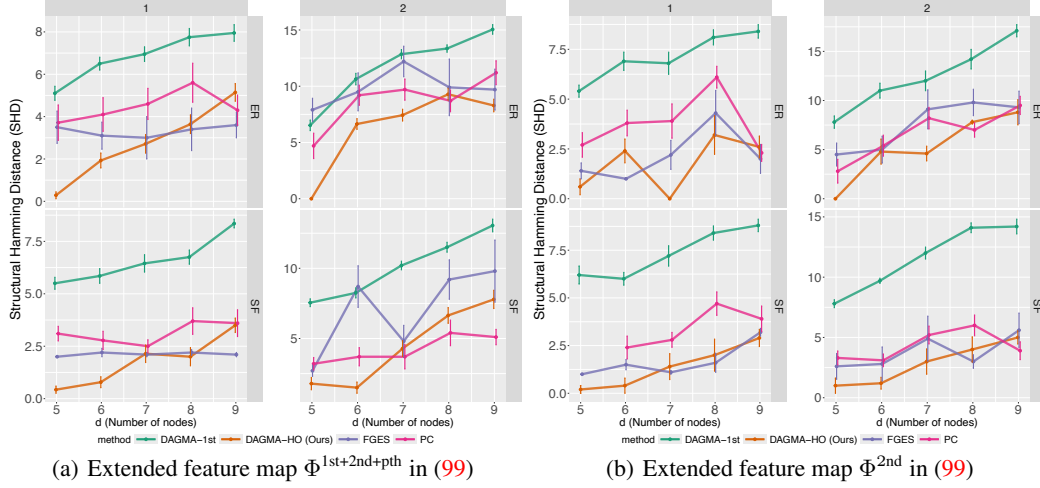


Figure 3: Results in terms of SHD between MECs of estimated graph and ground truth. Lower is better. Column: $k = \{1, 2\}$. Row: random graph types. $\{ER, SF\}$ - $k = \{\text{Scale-Free, Erdős-Rényi}\}$ graphs with kd expected edges. Here $p = \{5, 6, 7, 8, 9\}$. DAGMA [4] is renamed as “DAGMA-1st”, to emphasize only linear term is used. Error bars denote the standard error computed over 10 replications.

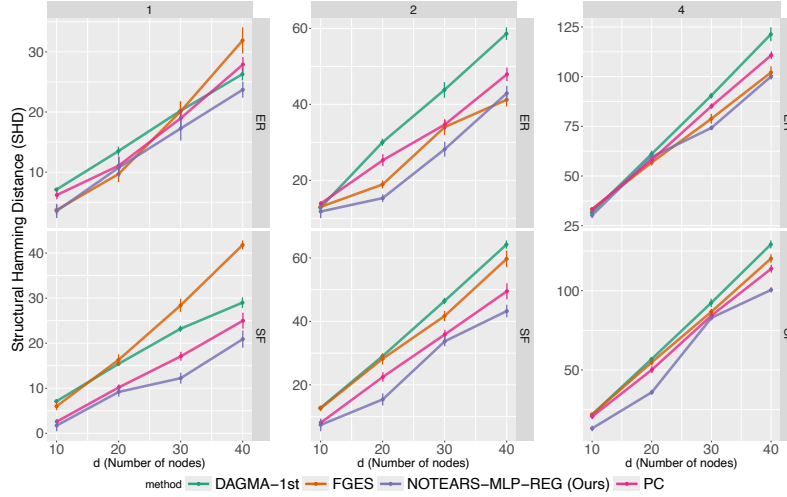


Figure 4: Results in terms of SHD between MECs of estimated graph and ground truth. Lower is better. Data are generated using extended feature map $\Phi^{1st+pth}$ in (99). Column: $k = \{1, 2, 4\}$. Row: random graph types. $\{ER, SF\}-k = \{\text{Scale-Free, Erdős-Rényi}\}$ graphs with kd expected edges. Here $p = \{10, 20, 30, 40\}$. NOTEARS-MLP-REG is our two stage approach. DAGMA [4] is renamed as “DAGMA-1st”, to emphasize only linear term is used.

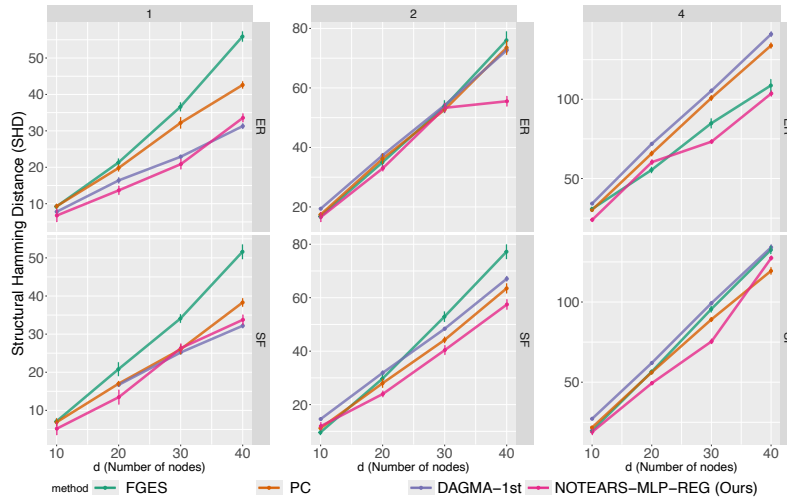


Figure 5: Results in terms of SHD between MECs of estimated graph and ground truth. Lower is better. Data are generated using extended feature map Φ^{2nd} in (99). Column: $k = \{1, 2, 4\}$. Row: random graph types. $\{ER, SF\}-k = \{\text{Scale-Free, Erdős-Rényi}\}$ graphs with kd expected edges. Here $p = \{10, 20, 30, 40\}$. NOTEARS-MLP-REG is our two stage approach. DAGMA [4] is renamed as “DAGMA-1st”, to emphasize only linear term is used.

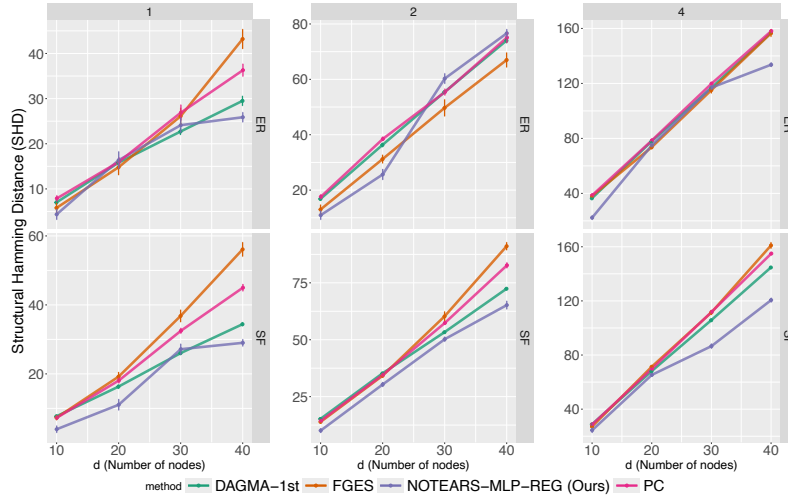


Figure 6: Results in terms of SHD between MECs of estimated graph and ground truth. Lower is better. Data are generated using extended feature map Φ^{pth} in (99). Column: $k = \{1, 2, 4\}$. Row: random graph types. $\{\text{ER}, \text{SF}\}-k = \{\text{Scale-Free}, \text{Erdős-Rényi}\}$ graphs with kd expected edges. Here $p = \{10, 20, 30, 40\}$. NOTEARS-MLP-REG is our two stage approach. DAGMA [4] is renamed as “DAGMA-1st”, to emphasize only linear term is used.