HOLD-OUT SCORING FOR EFFICIENT GAUSSIAN DAG LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

High-dimensional Gaussian DAG learning faces a statistical–computational gap: methods with sharp sample complexity rely on computationally expensive subset search and a supplied indegree bound, whereas polynomial-time alternatives have less favorable sample complexity. We introduce HOST, an efficient DAG learning algorithm that replaces subset search with nodewise hold-out scoring and convex regression, without requiring a supplied indegree bound. Our key insight is that recovering a correct ordering does not require uniformly small estimation errors in ordering scores but only one-sided control of those errors. In the ordering step, HOST exploits the fact that score estimation using hold-out samples inflates ordering scores in expectation, which is the favorable direction for candidates that should not yet be selected. Given the ordering, HOST recovers parents by recursively removing indirect effects from total effects between two nodes. Under suitable conditions, HOST exactly recovers a p-node DAG of maximum indegree d with sample complexity of order $d \log p$ in polynomial time. Experiments show that HOST achieves competitive graph recovery while exhibiting favorable runtime scaling.

1 INTRODUCTION

Learning a directed acyclic graph (DAG) from observational data is a central problem in graphical modeling and causal discovery (Pearl, 1988; Spirtes et al., 2001). Classical approaches include constraint-based and score-based methods (Spirtes et al., 2001; Chickering, 2002). These approaches generally identify the graph only up to a Markov equivalence class. Additional assumptions on the data-generating process, such as linear non-Gaussian models or nonlinear additive-noise structure, can make causal directions identifiable within a Markov equivalence class (Shimizu et al., 2006; 2011; Hoyer et al., 2008; Peters et al., 2014). Other DAG-learning approaches include ordering-based methods that separate order search from edge selection (Bühlmann et al., 2014) and continuous optimization using a smooth acyclicity constraint (Zheng et al., 2018). More recent methods use derivatives of the log density for causal ordering and, in some cases, edge selection (Rolland et al., 2022; Montagna et al., 2023a;b; Xu et al., 2024).

We focus on Gaussian linear DAG models, in which each node is expressed as a linear function of its parents plus an independent Gaussian error. Suitable conditions involving direct effects and error variances make the DAG identifiable from the observational distribution (Peters & Bühlmann, 2014; Park & Kim, 2020; Park, 2020; Chen et al., 2019; Ghoshal & Honorio, 2018). With identifiability established, we ask whether Gaussian linear DAGs can be learned with both sample efficiency and computational tractability.

Existing methods exhibit a statistical–computational tradeoff. Forward methods, which iteratively search for nodes appearing early in an ordering, can achieve sharp sample complexity in terms of the maximum indegree d. For example, Gao et al. (2022) develop an algorithm that attains the information-theoretically optimal sample complexity $n \asymp d \log(p/d)$ under suitable conditions, where n is the sample size and p denotes the number of variables. However, several forward methods require a supplied indegree bound $q \geq d$ and costly subset search with complexity $p^{O(q)}$, which becomes prohibitive as q grows (Chen et al., 2019; Gao et al., 2022; Daskalakis et al., 2025).

Backward procedures recursively identify and eliminate nodes from the end of an ordering using sparse inverse-covariance or regression methods (Ghoshal & Honorio, 2018; Park et al., 2021). These

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of Gaussian DAG learning algorithms. The results for Chen et al. (2019) concern ordering recovery only. In this table, $\gtrsim$ suppresses constants, including model-dependent factors. $\widetilde{O}$ suppresses logarithms. Irrep. denotes irrepresentability. Input q indicates whether a supplied indegree bound $q \geq d$ is required.

Reference	Sample complexity	Computational complexity	Irrep.	Input q
Ghoshal & Honorio (2018)	$n \gtrsim d_{\mathrm{m}}^4 \log p$	$\widetilde{O}(p^3 + p d_{\mathrm{m}}^4)$	×	×
Chen et al. (2019)	$n \gtrsim d^2 \log p$	$p^{O(q)}$	×	✓
Park et al. (2021)	$n \gtrsim d_{\mathrm{m}}^2 \log p$	$O\{n(p^3 + p^2 d_{\mathrm{m}}^2)\}$	✓	×
Gao et al. (2022)	$n \gtrsim d \log(p/d)$	$p^{O(q)}$	×	✓
Daskalakis et al. (2025), Alg. 1	$n \gtrsim d \log(p/d)$	$p^{O(q)}$	×	✓
Daskalakis et al. (2025), Alg. 2/3	$n \gtrsim d^4 \log p$	$\mathrm{poly}(n, p)$	×	✓
HOST	$n \gtrsim d \log p$	$O(np^3 + p^4)$	×	×

methods are computationally tractable, as they require sequential convex optimization, but their statistical guarantees depend on the maximum moralized degree d_{m}. The maximum moralized degree can substantially exceed the maximum indegree d because it also accounts for children and co-parents. These results motivate a procedure that replaces subset search with convex optimization, removes the need for a supplied indegree bound, and achieves sample complexity with a mild dependence on d rather than the potentially much larger d_{m}.

The need for subset search with a supplied indegree bound in forward methods arises because forward ordering recovery repeatedly compares scores to determine which remaining node has a sufficiently small score to be selected next. When the same observations are used for fitting and scoring, in-sample overfitting can produce artificially small scores and cause a node with unselected parents to be selected too early. To control this error uniformly across candidates, existing subset-search methods restrict the candidate models using a supplied indegree bound.

Our key observation is that accurate estimation of every score is unnecessary since recovering a compatible ordering requires only one-sided score control. We control overestimation for nodes whose parents have all been selected and underestimation for nodes with unselected parents. On held-out data, fitting error inflates scores in expectation, which is the favorable direction for nodes that should not yet be selected. Accurate estimation of scores is therefore needed only for nodes whose parents have all been selected. We exploit this asymmetry through HOST (Hold-Out Ordering and Structural Thresholding), which combines nodewise hold-out scoring for ordering recovery with recursive direct-effect reconstruction for parent recovery. Our main contributions are as follows.

- **Ordering recovery without a supplied indegree bound.** HOST uses Lasso-based regressions with a two-fold sample split to control score underestimation without dependence on the regression dimension, while accurate estimation of scores is required only for nodes whose parents have all been selected. This guarantees ordering recovery with d-dependent sample complexity.
- **Parent recovery without irrepresentability.** Given the recovered ordering, HOST reconstructs direct effects from total effects by processing earlier nodes in reverse order and subtracting indirect effects. The ordering fits enter only through bounded corrections, allowing one threshold per node to recover parents without exact Lasso support recovery or an irrepresentability condition, even when the ordering fits omit true parents or direct and indirect effects cancel in the total effect.
- **Near-optimal sample complexity in polynomial time.** Under explicit model conditions, HOST recovers the DAG with sample complexity of order $d \log p$ and takes polynomial time without subset search. Under the fixed-parameter convention of Gao et al. (2022), this bound is within a factor of order $\log d$ of the minimax-optimal sample complexity $d \log(p/d)$ for $2 \leq d \leq p/2$.

2 Preliminaries

2.1 Notation

For a vector $v \in \mathbb{R}^m$, a matrix $M \in \mathbb{R}^{m \times n}$, and index sets $I \subseteq [m]$ and $J \subseteq [n]$, let v_I and M_{IJ} denote the corresponding subvector and submatrix with singleton indices suppressed. We write $\|\cdot\|_1$

and $\|\cdot\| = \|\cdot\|_2$ for the ℓ_1 and Euclidean norms, respectively. We use I_k for the $k \times k$ identity matrix and $\lambda_{\min}(M)$ for the smallest eigenvalue when M is symmetric. $a \gtrsim b$ and $a \asymp b$ denote inequality and equivalence up to positive constant factors, respectively. For a DAG $G = (V, E)$ with node set $V = [p] = \{1, \dots, p\}$ and directed edge set $E \subseteq V \times V$, write $k \to j$ if $(k, j) \in E$. A directed path from a to b is a sequence $\Gamma = (v_0 \to \cdots \to v_r)$ of distinct nodes with $v_0 = a$, $v_r = b$, and length $\ell(\Gamma) = r \geq 1$. Let $\mathcal{P}(a, b)$ denote the set of these paths. An ordering is a permutation $\pi = (\pi_1, \dots, \pi_p)$ of V, compatible with G if $k \to j$ implies $\pi^{-1}(k) < \pi^{-1}(j)$. For $j \in V$, define $\mathrm{Pa}(j) = \{k \in V : k \to j\}$ and $\mathrm{An}(j) = \{k \in V : \mathcal{P}(k, j) \neq \emptyset\}$. Let $d = \max_{j \in V} |\mathrm{Pa}(j)|$ be the maximum indegree, and let q denote a supplied indegree bound. A set $S \subseteq V$ is ancestral if $\mathrm{Pa}(j) \subseteq S$ for every $j \in S$. For ancestral S, let $R(S) = V \setminus S$, $\mathrm{Src}(S) = \{j \in R(S) : \mathrm{Pa}(j) \subseteq S\}$, and $\mathrm{Nsrc}(S) = R(S) \setminus \mathrm{Src}(S)$ denote the remaining nodes, sources, and nonsources, respectively.

2.2 Gaussian Linear DAG Models

Gaussian linear DAG model. The Gaussian linear DAG model is defined by the structural equations

$$X_j = \sum_{k \in \mathrm{Pa}(j)} B_{kj} X_k + \varepsilon_j, \qquad \varepsilon_j \overset{\text{ind}}{\sim} N(0, \sigma_j^2), \qquad j \in V, \tag{1}$$

where $\sigma_j > 0$ and B_{kj} is the direct effect from k to j. The nonzero entries of B define the edges of the DAG G. The resulting vector $X \sim N(0, \Sigma)$ has covariance and precision matrices

$$\Sigma = \mathrm{Cov}(X) = (I_p - B)^{-\top} \Sigma^{\varepsilon} (I_p - B)^{-1} \succ 0, \qquad \Omega = \Sigma^{-1},$$

where $\Sigma^{\varepsilon} = \mathrm{diag}(\sigma_1^2, \dots, \sigma_p^2)$ is the covariance matrix of the independent errors. The directed-path expansion of (1) expresses each variable in terms of these errors (Sullivant et al., 2010):

$$X_j = \varepsilon_j + \sum_{k \in \mathrm{An}(j)} B_{k \rightsquigarrow j} \varepsilon_k, \qquad B_{k \rightsquigarrow j} = \sum_{\Gamma \in \mathcal{P}(k,j)} \prod_{t=1}^{\ell(\Gamma)} B_{v_{t-1} v_t}. \tag{2}$$

For each directed path $\Gamma \in \mathcal{P}(k, j)$, the product of the direct-effect coefficients along its edges gives the effect transmitted along Γ, and summing over these paths yields the total effect $B_{k \rightsquigarrow j}$.

DAG identifiability. For $S \subseteq V$ and $j \notin S$, define the population ordering score of j given S as

$$\mathcal{S}_j(S) = \mathbb{E}[(X_j - X_S^{\top} \theta_j(S))^2], \quad \theta_j(S) = \operatorname*{argmin}_{\theta \in \mathbb{R}^{|S|}} \mathbb{E}[(X_j - X_S^{\top} \theta)^2].$$

Under Gaussianity, the least-squares residual is independent of X_S, so $\mathcal{S}_j(S) = \mathrm{Var}(X_j \mid X_S)$. When the selected set S is ancestral, the remaining sources $\mathrm{Src}(S)$ have all their parents in S and can appear next in a compatible ordering. The following lemma gives a score separation under which minimizing the population score selects a remaining source.

Lemma 2.1 (Forward Identifiability)**.** *Suppose every ancestral $S \subsetneq V$ satisfies*

$$\min_{k \in \mathrm{Nsrc}(S)} \mathcal{S}_k(S) > \min_{j \in \mathrm{Src}(S)} \mathcal{S}_j(S),$$

where a minimum over the empty set is $+\infty$. Then an ordering compatible with G can be identified from the observational distribution.

Starting from $S = \emptyset$, successively adding score minimizers to S keeps S ancestral, so the selection order is compatible with G. The proof follows an inductive argument based on conditional variances, similar to those in Chen et al. (2019) and Park (2020), so we omit the details.

For ancestral S and $j \notin S$, conditioning on X_S is equivalent to conditioning on ε_S, so (2) and independence of the errors give the explicit form of the population ordering score

$$\mathcal{S}_j(S) = \sigma_j^2 + \sum_{a \in \mathrm{An}(j) \setminus S} B_{a \rightsquigarrow j}^2 \sigma_a^2. \tag{3}$$

For a remaining source, every parent has been selected and the score is simply σ_j^2. For a remaining nonsource j, choose an unselected parent a with no descendant among the other unselected parents of j. Then $B_{a \rightsquigarrow j} = B_{aj} \neq 0$, so at least one additional variance term is positive. Equal error variances therefore suffice for the score separation in Lemma 2.1 (Peters & Bühlmann, 2014). With heterogeneous errors, the comparison instead depends jointly on direct effects and error variances.

3 One-Sided Score Control for Ordering Recovery

3.1 Score Estimation Along Forward Search

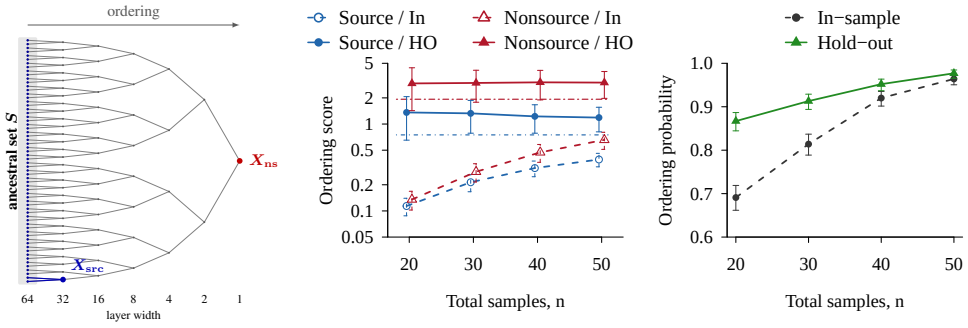
Implementing the ordering rule in Lemma 2.1 requires estimating the population scores $\mathcal{S}_j(S)$ at each selected ancestral set S. A natural approach uses the residual mean square from a regression of each remaining node on X_S as a score estimator. The following proposition describes the corresponding population regression coefficients.

Proposition 3.1 (Regression Coefficients at an Ancestral Set)**.** *In the Gaussian linear DAG model, let $S \subsetneq V$ be ancestral and $j \notin S$. The coefficient $\theta_{kj}(S)$ of X_k in the population regression of X_j on X_S satisfies*

$$\theta_{kj}(S) = B_{k \rightsquigarrow j} - \sum_{a \in S \cap \mathrm{An}(j)} B_{ka} B_{a \rightsquigarrow j} = B_{kj} + \sum_{a \in \mathrm{An}(j) \setminus S} B_{ka} B_{a \rightsquigarrow j}, \qquad k \in S.$$

The proof is given in Appendix A.1. For a remaining source j, all parents lie in S, so $\theta_j(S) = B_{Sj}$ and the regression coincides with its structural equation. For a remaining nonsource, the coefficients also contain indirect effects through unselected nodes. Thus the maximum indegree d does not bound the number of nonzero coefficients in the regression of a remaining nonsource on X_S.

Motivating example. In the layered DAG of Figure 1(a), every node after the first layer has two parents, yet the regression of the designated remaining nonsource X_{ns} on X_S includes indirect effects from all 64 variables in S. The regression of the designated remaining source X_{src} involves only its two parents.



(a) Layered DAG (b) Ordering scores (c) Ordering probability

Figure 1: Panel (a) shows a 127-node layered DAG with the ancestral set S shaded, the designated remaining source X_{src} in blue, and the designated remaining nonsource X_{ns} in red. Panel (b) shows mean Lasso ordering scores for in-sample (In) and hold-out (HO) scoring using X_S on a log scale. Error bars indicate one standard deviation across repetitions, computed on the original score scale. Horizontal lines mark population scores of 0.75 for X_{src} and 1.93 for X_{ns}. Panel (c) shows the fraction of repetitions with X_{src} scoring below X_{ns}, with 95% Wilson intervals. Appendix D.1 gives the settings.

Minimax lower bounds for residual variance estimation in high-dimensional Gaussian regression can depend on the number of nonzero regression coefficients (Verzelen & Gassiat, 2018). As Proposition 3.1 and the preceding example show, indirect effects can make this number much larger than the maximum indegree d for remaining nonsources, making accurate estimation of every score difficult even when d is small. However, preserving the population score comparison in Lemma 2.1 suffices for ordering recovery. In other words, it suffices to bound score overestimation for remaining sources and score underestimation for remaining nonsources so that the estimated scores preserve the comparison.

3.2 Hold-Out Scoring

Score underestimation can be controlled through hold-out scoring because fitting the regression on an independent sample prevents overfitting to the evaluation observations. The following lemma

provides a bound on score underestimation that does not depend on the true regression coefficients or the accuracy of their estimates.

Lemma 3.2 (Hold-Out Lower Tail)**.** *Let $\mathcal{F}$ be a sigma-field. For integers $n \geq 1$ and $s \geq 0$, suppose $\mathbf{X} \in \mathbb{R}^{n \times s}$ and both vectors $\theta, \widetilde{\theta} \in \mathbb{R}^s$ are $\mathcal{F}$-measurable with finite entries. Suppose $Y = \mathbf{X}\theta + \xi$, where $\xi \mid \mathcal{F} \sim N(0, vI_n)$ for a fixed $v > 0$. Then for $\widetilde{\mathcal{S}} = \|Y - \mathbf{X}\widetilde{\theta}\|_2^2/n$ and every $0 < t < v$,*

$$\mathbb{P}\{\widetilde{\mathcal{S}} < v - t \mid \mathcal{F}\} \leq \exp\{-nt^2/(4v^2)\} \qquad \textit{almost surely}.$$

The proof is given in Appendix A.2. For the regression on X_S at a fixed ancestral set S, apply the lemma with $v = \mathcal{S}_j(S)$ and $\theta = \theta_j(S)$, conditioning on the fitted coefficients and held-out values of X_S. With $\mathbf{X}$ denoting the design matrix of the n held-out observations, the conditional expected score is $\mathbb{E}[\widetilde{\mathcal{S}} \mid \mathcal{F}] = \mathcal{S}_j(S) + \|\mathbf{X}(\theta_j(S) - \widetilde{\theta})\|^2/n \geq \mathcal{S}_j(S)$. Thus coefficient estimation error contributes a nonnegative term, providing intuition for the lower-tail bound. The lemma controls score underestimation for remaining nonsources regardless of the true coefficients, their estimates, or the size of S. For remaining sources, score overestimation is controlled through the accuracy of the fitted regression. Proposition 3.1 gives $\theta_j(S) = B_{Sj}$, so the population regression involves at most d nonzero coefficients. Estimation bounds for penalized regression can therefore provide the complementary control of score overestimation with sample requirements depending on d.

Figure 1(b) and (c) illustrate the distinction between accurate score estimation and preservation of the score comparison. In-sample scoring uses all n observations for both fitting and evaluation, while hold-out scoring uses one half for fitting and the other for evaluation. In-sample scores substantially underestimate the population scores and reduce their separation in panel (b), allowing the remaining nonsource to be selected first. Hold-out scores exceed the population scores on average, yet panel (c) shows a higher probability of placing the remaining source below the remaining nonsource.

4 HOST Algorithm and Guarantees

HOST builds the graph one node at a time. It selects the next node by hold-out scoring, recovers its parents among the selected nodes, and retains a local ordinary least squares (OLS) fit and its residual for later iterations. The parent-recovery step uses these residuals to separate direct effects from indirect effects, with the ordering fits guiding a bounded correction before thresholding.

Initialization. Let $\mathbf{X}$ be the data matrix from n independent observations. Split the observations independently of their values into $I_1 \cup I_2 = [n]$, with sizes $n_\ell = |I_\ell|$ differing by at most one. For $\ell \in \{1, 2\}$, write $\bar{\ell}$ for the other fold index. Use the same positive ordering penalty λ_{ord} on both folds, positive parent thresholds τ_j, and a fixed clipping parameter $0 \leq \delta < 1$. Choose $\widehat{\pi}_1 = \mathrm{argmin}_{j \in V} \widehat{\Sigma}_{jj}$, where $\widehat{\Sigma}_{jj} = \|\mathbf{X}_j\|^2/n$, and initialize the estimated parent set $\widehat{T}_{\widehat{\pi}_1} = \emptyset$ and the residual $\widehat{r}_{\widehat{\pi}_1} = \mathbf{X}_{\widehat{\pi}_1}$.

Ordering by hold-out scoring. At iteration $r \geq 2$, the selected nodes form $\widehat{S}_r = \{\widehat{\pi}_1, \ldots, \widehat{\pi}_{r-1}\}$. For every remaining node j, fit a Lasso regression on each half of the sample,

$$\widehat{\theta}_j^{(\ell)}(r) \in \operatorname*{argmin}_{\theta} \Big\{ \frac{\|\mathbf{X}_{I_\ell j} - \mathbf{X}_{I_\ell \widehat{S}_r}\theta\|^2}{2n_\ell} + \lambda_{\mathrm{ord}}\|\theta\|_1 \Big\}, \qquad \ell = 1, 2. \tag{4}$$

Combine the two opposite-fold residual sums of squares to obtain the hold-out ordering score,

$$\widehat{\mathcal{S}}_j(r) = \frac{1}{n} \sum_{\ell=1}^{2} \|\mathbf{X}_{I_{\bar{\ell}} j} - \mathbf{X}_{I_{\bar{\ell}} \widehat{S}_r} \widehat{\theta}_j^{(\ell)}(r)\|^2. \tag{5}$$

The remaining node with the smallest score becomes $\widehat{\pi}_r$, with ties resolved by node label. For $j = \widehat{\pi}_r$, let $\bar{\theta}_{kj} = \{\widehat{\theta}_{kj}^{(1)}(r) + \widehat{\theta}_{kj}^{(2)}(r)\}/2$ be the averaged coefficient of X_k for each earlier node k. Define the updated ordering coefficients by refitting each nonzero coordinate as

$$\check{\theta}_{kj} = \mathrm{argmin}_{b \in \mathbb{R}} \sum_{i=1}^{n} \Big(\mathbf{X}_{ij} - \sum\nolimits_{a \in \widehat{S}_r \setminus \{k\}} \mathbf{X}_{ia}\bar{\theta}_{aj} - \mathbf{X}_{ik} b \Big)^2, \qquad \bar{\theta}_{kj} \neq 0, \tag{6}$$

and set $\check{\theta}_{kj} = 0$ for $\bar{\theta}_{kj} = 0$. This update aims to reduce Lasso shrinkage and guides correction of parent estimates from local fits below, without requiring the ordering Lasso to recover the parent set.

Estimating total effects. Each $k \in \widehat{S}_r$ has a previously estimated parent set $\widehat{T}_k$. The direct-effect estimates from $\widehat{T}_k$ to k and the residual from its retained local OLS fit are reused from earlier iterations, and are given by

$$\widehat{B}^{\mathrm{LS}}_{\widehat{T}_k,k} = \mathbf{X}^{\dagger}_{\widehat{T}_k}\mathbf{X}_k, \quad \widehat{B}^{\mathrm{LS}}_{V\setminus\widehat{T}_k,k} = 0, \quad \widehat{r}_k = \mathbf{X}_k - \mathbf{X}\widehat{B}^{\mathrm{LS}}_{\cdot k}, \tag{7}$$

where † denotes the Moore–Penrose inverse. When $\widehat{T}_k = \mathrm{Pa}(k)$, the population residual is ε_k, and (2) gives $\mathrm{Cov}(\varepsilon_k, X_j)/\sigma_k^2 = B_{k\rightsquigarrow j}$. We therefore use $\widehat{B}_{k\rightsquigarrow j} = \widehat{r}_k^{\top}\mathbf{X}_j/\|\widehat{r}_k\|^2$, defined as zero when $\widehat{r}_k = 0$, to estimate the total effect of X_k on X_j.

Structural thresholding. For $j = \widehat{\pi}_r$, let $\widehat{T}_j^{(m)}$ and $\widehat{Q}_j^{(m)}$ contain the selected parents and nodes retained for subtraction after m steps, both initially empty. At step $m = 1, \ldots, r-1$, set $k = \widehat{\pi}_{r-m}$. Proposition 3.1 gives the direct-effect identity and its sample reconstruction:

$$B_{kj} = B_{k\rightsquigarrow j} - \sum\nolimits_{a\in \mathrm{An}(j)} B_{ka}B_{a\rightsquigarrow j}, \qquad \widehat{B}^{\mathrm{res}}_{kj} = \widehat{B}_{k\rightsquigarrow j} - \sum\nolimits_{a\in\widehat{Q}_j^{(m-1)}} \widehat{B}^{\mathrm{LS}}_{ka}\widehat{B}_{a\rightsquigarrow j}. \tag{8}$$

With $\mathrm{clip}_t(x) = \max\{-t, \min(x,t)\}$, apply a bounded correction using $\check{\theta}_{kj}$ to obtain a corrected direct-effect estimate. Select k as a parent of j when the magnitude of this estimate exceeds τ_j:

$$\widetilde{B}_{kj} = \widehat{B}^{\mathrm{res}}_{kj} + \mathrm{clip}_{\delta\tau_j}\left(\check{\theta}_{kj} - \widehat{B}^{\mathrm{res}}_{kj}\right), \qquad \widehat{T}_j^{(m)} = \widehat{T}_j^{(m-1)} \cup \{u \in \{k\} : |\widetilde{B}_{kj}| > \tau_j\}. \tag{9}$$

The correction is bounded by $\delta\tau_j$, so parent recovery does not require $\check{\theta}_{kj}$ to be accurate. Retain k if it is selected as a parent of j or belongs to the estimated parent set of a previously retained node:

$$\widehat{Q}_j^{(m)} = \widehat{Q}_j^{(m-1)} \cup \left(\{k\} \cap \left(\widehat{T}_j^{(m)} \cup \bigcup\nolimits_{a\in\widehat{Q}_j^{(m-1)}} \widehat{T}_a\right)\right). \tag{10}$$

After $r-1$ steps, return $\widehat{T}_j = \widehat{T}_j^{(r-1)}$ and use $\widehat{T}_j$ to compute the local OLS fit $\widehat{B}^{\mathrm{LS}}_{\cdot j}$ and residual $\widehat{r}_j$ in (7) for subsequent iterations.

Algorithm 1: HOST Hold-Out Ordering and Structural Thresholding

Input Data $\mathbf{X}$, penalty $\lambda_{\mathrm{ord}} > 0$, parent thresholds $\tau_j > 0$, and $\delta \in [0, 1)$
Output Estimated DAG $\widehat{G}$
Split $[n]$ into balanced folds I_1, I_2 independently of $\mathbf{X}$
Set $\widehat{\pi}_1 = \mathrm{argmin}_{j\in V}\widehat{\Sigma}_{jj}$, $\widehat{T}_{\widehat{\pi}_1} = \emptyset$, and residual $\widehat{r}_{\widehat{\pi}_1} = \mathbf{X}_{\widehat{\pi}_1}$
for $r = 2, \ldots, p$ *with* $\widehat{S}_r = \{\widehat{\pi}_1, \ldots, \widehat{\pi}_{r-1}\}$ **do**
  Fit foldwise Lasso regressions and compute hold-out scores for $j \notin \widehat{S}_r$ by (4)–(5)
  Select $j = \widehat{\pi}_r = \mathrm{argmin}_{a\notin\widehat{S}_r}\widehat{\mathcal{S}}_a(r)$ and compute $\check{\theta}_j$ by (6)
  Set $\widehat{T}_j^{(0)} = \widehat{Q}_j^{(0)} = \emptyset$ and compute $\widehat{B}_{k\rightsquigarrow j} = \widehat{r}_k^{\top}\mathbf{X}_j/\|\widehat{r}_k\|^2$ for $k \in \widehat{S}_r$
  for $m = 1, \ldots, r-1$ **do**
    Set $k = \widehat{\pi}_{r-m}$ and subtract indirect effects to obtain $\widehat{B}^{\mathrm{res}}_{kj}$ by (8)
    Correct and threshold by (9) to update $\widehat{T}_j^{(m)}$, then update $\widehat{Q}_j^{(m)}$ by (10)
  Set $\widehat{T}_j = \widehat{T}_j^{(r-1)}$ and store its OLS coefficients and residual by (7)
Return $\widehat{G} = (V, \{k \to j : j \in V,\ k \in \widehat{T}_j\})$

4.1 Theoretical Guarantees

Throughout this section, we assume $p \geq 2$ and even $n \geq 4$ and use a data-independent split into two equal halves. Define $B_{\min} = \min_{k\to j}|B_{kj}|$, with $B_{\min} = +\infty$ for an empty graph. Define $\Sigma_{\max} = \max_j \Sigma_{jj}$, $\sigma_{\max} = \max_j \sigma_j$, and $\Omega_{\max} = \max_k \Omega_{kk}$.

Assumption 1 (Forward Identifiability Margin)**.** There is a margin $\tau_\star > 0$ such that, for every ancestral $S \subsetneq V$ with at least one remaining nonsource,

$$\min_{k\in\mathrm{Nsrc}(S)} \mathcal{S}_k(S) - \min_{j\in\mathrm{Src}(S)} \mathcal{S}_j(S) \geq \tau_\star.$$

For fixed p, the strict separation in Lemma 2.1 yields a positive common margin because there are finitely many ancestral sets. With common error variance σ^2 and $E \neq \emptyset$, one can take $\tau_\star = \sigma^2 B_{\min}^2$.

Theorem 4.1 (Ordering Recovery)**.** *Suppose the data follow the Gaussian linear DAG model in Section 2.2 and Assumption 1 holds. Choose $\lambda_{\mathrm{ord}} \asymp \sigma_{\max}\sqrt{\Sigma_{\max}\log p/n}$, with sufficiently large universal constant factors. If*

$$n \gtrsim \left[\frac{d\Sigma_{\max}}{\lambda_{\min}(\Sigma)}\left(1+\frac{\sigma_{\max}^2}{\tau_\star}\right)+\left(1+\frac{\Sigma_{\max}}{\tau_\star}\right)^2\right]\log p, \tag{11}$$

then $\mathbb{P}\{\widehat{\pi}$ is compatible with $G\} \geq 1-p^{-2}$.

The proof is given in Appendix B.2. It combines score estimation bounds for remaining sources with hold-out lower-tail bounds for remaining nonsources. The argument uses conditional bounds for selected sets to account for dependence induced by adaptive node selection across iterations.

Theorem 4.2 (Parent Recovery)**.** *Under the Gaussian linear DAG model, fix $0 \leq \delta < 1$ and use $\tau_j \asymp \sqrt{\Sigma_{\max}\Omega_{\max}\log p/n}$ in Algorithm 1, with a sufficiently large constant depending only on δ. If*

$$n \gtrsim d + \log p, \qquad B_{\min} \gtrsim \sqrt{\Sigma_{\max}\Omega_{\max}\log p/n}, \tag{12}$$

then, for any ordering $\widehat{\pi}$ computed from the same observations,

$$\mathbb{P}\{\widehat{G} \neq G\} \leq \mathbb{P}\{\widehat{\pi} \text{ is incompatible with } G\} + p^{-2}.$$

The proof is given in Appendix B.3. It uniformly controls errors in the reconstructed direct effects over compatible orderings. An induction argument then shows that thresholding recovers all parents, even when the ordering is learned from the same data, without requiring accurate ordering fits. Ordering guarantees concern the exact minimizers in (4) at the fixed penalty specified above.

Theorem 4.3 (Exact DAG Recovery)**.** *Suppose Assumption 1 holds in the Gaussian linear DAG model. Choose the fixed ordering penalty and parent thresholds in Theorems 4.1 and 4.2. If*

$$n \gtrsim \left[\frac{d\Sigma_{\max}}{\lambda_{\min}(\Sigma)}\left(1+\frac{\sigma_{\max}^2}{\tau_\star}\right)+\left(1+\frac{\Sigma_{\max}}{\tau_\star}\right)^2+\frac{\Sigma_{\max}\Omega_{\max}}{B_{\min}^2}\right]\log p, \tag{13}$$

then $\mathbb{P}\{\widehat{G} = G\} \geq 1-p^{-1}$.

The proof in Appendix B.4 combines the ordering and parent bounds. This high-dimensional bound gives recovery probability tending to one as $p = p_n \to \infty$ under (13).

For computational complexity, we use cyclic coordinate descent for Lasso and singular value decomposition (SVD) for minimum-norm local OLS fits. We report complexity for dense implementations, suppressing solver iteration counts and numerical tolerance factors.

Proposition 4.4 (Computational Complexity)**.** *Under the computational setup above, Algorithm 1 has arithmetic complexity $O(np^3+p^4)$ for a fixed ordering penalty and fixed parent thresholds.*

Appendix B.5 provides the proof. The complexity bound concerns the numerical implementation, whereas the statistical guarantees use exact Lasso minimizers.

4.2 Comparison with Existing Forward-Search Methods

For equal-error-variance models, let σ^2 be the common error variance. Under $M^{-1}I \preceq \Sigma \preceq MI$ for a constant $M > 1$, Gao et al. (2022) establish the sample complexity lower bound $n \gtrsim \max\{B_{\min}^{-2}\log p,\ (M^2-1)^{-1}d\log(p/d)\}$ for exact DAG recovery when $1 \leq d \leq p/2$. Under their convention of fixed M, $B_{\min}$, and σ^2, their subset-search algorithm attains minimax-optimal $d\log(p/d)$ sample complexity. For HOST, the covariance bounds and $\tau_\star = \sigma^2 B_{\min}^2$ reduce (13) to the sufficient condition

$$n \gtrsim \left[dM^2(1+B_{\min}^{-2})+(1+M\sigma^{-2}B_{\min}^{-2})^2+M^2B_{\min}^{-2}\right]\log p.$$

Under the same convention, the sample complexity of HOST is within a factor of order $\log d$ of theirs for $2 \leq d \leq p/2$. Gao et al. (2022) require a supplied indegree bound q and $p^{O(q)}$ subset search, whereas HOST runs in polynomial time.

For equal-error-variance models without this eigenvalue restriction, Daskalakis et al. (2025) establish a sample complexity lower bound $n \gtrsim \max\{B_{\min}^{-4}, TB_{\min}^{-2}\} \log p$, where $T = \sigma^2 \Omega_{\max} = 1 + \max_i \sum_{j:\, i \to j} B_{ij}^2$. For HOST, using the parameter relationships $\tau_\star = \sigma^2 B_{\min}^2$, $\sigma^2/\lambda_{\min}(\Sigma) \le (d+1)T$, and $V = \Sigma_{\max}/\sigma^2 \ge 1 + B_{\min}^2$ for a nonempty DAG reduces (13) to the sufficient condition

$$n \gtrsim \left[d^2 VT(1 + B_{\min}^{-2}) + V^2 B_{\min}^{-4}\right] \log p. \tag{14}$$

Their Algorithm 1 has sample complexity $n \gtrsim \max\{B_{\min}^{-4}, TB_{\min}^{-2}\}\, d \log(p/d)$ and is based on subset search. For bounded V, this sample complexity is within a factor d of the lower bound and has milder d dependence than HOST, but Algorithm 1 requires a supplied indegree bound q and $p^{O(q)}$ computation. Their Algorithms 2 and 3 provide polynomial-time alternatives with sample complexity $n \gtrsim V d^4 T^3 B_{\min}^{-4} \log p$. For bounded V, HOST improves the dependence on d and T from d^4T^3 to d^2T, although the dependence on V differs.

5 Simulation Studies

We evaluate HOST across sample sizes, dimensions, and maximum indegrees, comparing its recovery accuracy and runtime with existing methods. All experiments use Gaussian linear DAGs. In this section, each node has at most d randomly selected parents. Nonzero direct effects have magnitudes uniform on $[0.4, 0.6]$ and symmetric signs, and independent Gaussian errors have variance 0.75. Appendix D details the experimental settings and includes additional graph families (D.5), ordering comparisons (D.6), and parent-recovery comparisons (D.7).

Figure 2 shows improving graph recovery with increasing sample size, with qualitative trends consistent with Theorem 4.3. The experiment covers p = 30, 50, 100, 200, d = 2, 3, 4, 5, and n = 128, 256, 512, 1024, 2048, with 200 repetitions per setting. The graph recovery rate is the fraction of exactly recovered graphs. For each (d, p), the rate tends toward one as n grows. Panels (a)–(d) vary maximum indegree d at fixed dimension p and generally show higher graph recovery rates for smaller d at a given sample size. Panels (e)–(h) vary dimension p at fixed maximum indegree d and show that larger graphs generally require more observations to attain a comparable graph recovery rate.

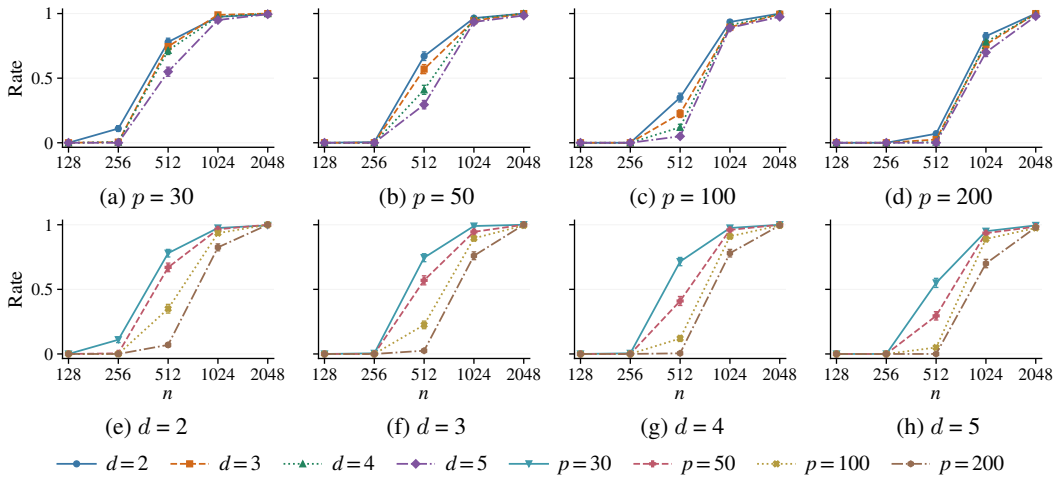
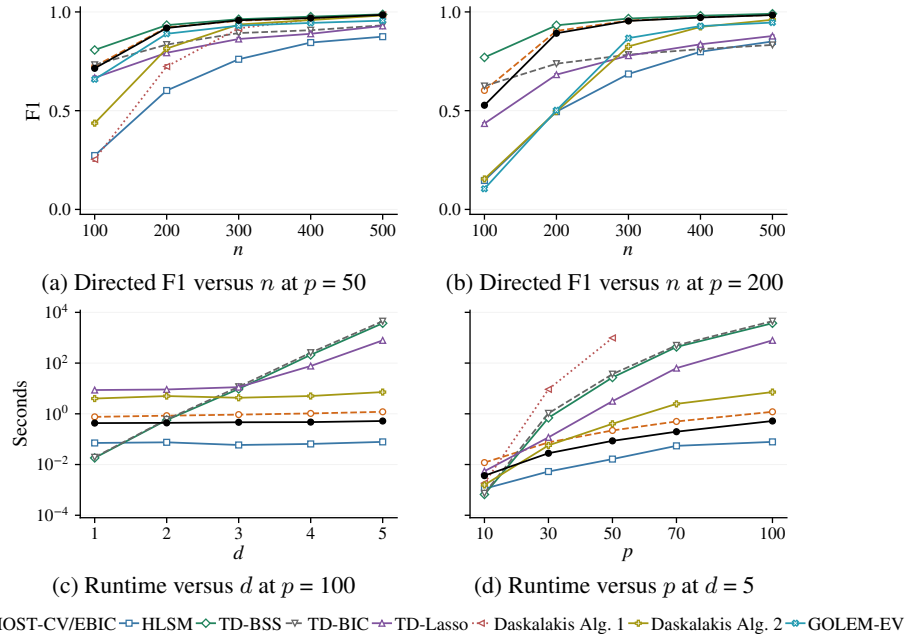


Figure 2: Graph recovery rate of HOST with ± 1 standard error over 200 repetitions. Panels (a)–(d) overlay curves for different d at fixed p, and (e)–(h) for different p at fixed d.

Figure 3 compares HOST with fixed tuning and HOST-CV/EBIC with data-selected tuning (Appendix C) against HLSM (Park et al., 2021), TD-BSS and TD-BIC (Gao et al., 2022), TD-Lasso (Chen et al., 2019), Daskalakis Algs. 1 and 2 (Daskalakis et al., 2025), and GOLEM-EV (Ng et al., 2020). Graph recovery is evaluated using the directed F1 score at p = 50, 200, d = 3, and n = 100, 200, 300, 400, 500, with 50 paired repetitions. Full-fit runtime, including parameter selection where applicable, is evaluated at n = 100, p = 10, 30, 50, 70, 100, and d = 1, 2, 3, 4, 5, with 20 paired repetitions. Daskalakis Alg. 1 is evaluated only for $p \le 50$, and GOLEM-EV is excluded from

timing. All recovery metrics and the full runtime comparison are reported in Appendices D.3 and D.4, respectively.



(a) Directed F1 versus n at $p = 50$

(b) Directed F1 versus n at $p = 200$

(c) Runtime versus d at $p = 100$

(d) Runtime versus p at $d = 5$

Figure 3: Recovery accuracy and runtime scaling. Recovery uses $d = 3$ and shows means ± 1 standard error over 50 paired repetitions. Runtime uses $n = 100$ and reports mean full-fit time in seconds (log scale), including tuning, over 20 paired repetitions.

Figure 3 shows the main empirical advantage of HOST: at sufficiently large sample sizes, it achieves recovery accuracy close to TD-BSS, while its runtime grows much more slowly with d and it requires no supplied indegree bound. In panels (a)–(b), HOST-CV/EBIC improves F1 over fixed-tuning HOST at $n = 100$ for both values of p. As the sample size increases, both HOST variants approach the F1 of TD-BSS. At $p = 200$, $d = 3$, and $n = 500$, HOST and HOST-CV/EBIC attain directed F1 0.984 and 0.985, respectively, compared with 0.991 for TD-BSS. Panel (c) shows that TD-BSS is fast for small d but becomes much slower as d increases. At $p = 100$, increasing d from 1 to 5 raises its runtime from 0.018 to 3716 seconds, whereas the runtime of HOST changes only from 0.43 to 0.52 seconds and that of HOST-CV/EBIC changes from 0.76 to 1.20 seconds. Thus, at $d = 5$, HOST is more than 7,000 times faster than TD-BSS in this setting. Panel (d) shows a similar runtime advantage as p increases at $d = 5$. HLSM is consistently faster than HOST, but its F1 is consistently lower.

6 Conclusion

We prove a new theoretical principle underlying HOST: accurate estimation of every candidate score is unnecessary, and it suffices to control only errors that can cause an incorrect next-node selection. Building on this result, HOST achieves exact Gaussian DAG recovery with sample complexity of order $d \log p$ and polynomial computation, without a supplied indegree bound or subset search. Its parent-recovery analysis also avoids exact Lasso support recovery and restrictive conditions such as irrepresentability.

Beyond HOST, our analysis isolates a broader principle for fit-then-score structure learning: when correct selection depends only on one-sided score errors, held-out evaluation can remove the need for uniformly accurate estimation and prevent overfitting-induced premature selection. Extending this principle to nonlinear and non-Gaussian DAG models and developing adaptive tuning with recovery guarantees are promising directions for future work.

REPRODUCIBILITY STATEMENT

Appendices A and B provide the proofs, Appendix C describes the tuning rules, and Appendices D and E specify the simulation and real-data protocols. The code and instructions for reproducing the experiments will be made publicly available after publication. An anonymized version can be provided to reviewers upon request during review.

AI USE STATEMENT

We used OpenAI Codex to help draft sections of the manuscript, translate author-provided content into English, improve readability, identify relevant literature, provide critical ingredients for proving the authors' mathematical claims and assist in writing the proofs, implement methods by writing and debugging method, simulation, and analysis code, and clean and reformat datasets. We did not use generative AI tools to generate synthetic data sets, formulate mathematical claims, develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, or interpret results, and qualitative and thematic data analysis is not applicable to this work. The authors manually checked all mathematical claims and proofs, code, experimental results, and citations, and take full responsibility for the final content.

REFERENCES

Afonso S. Bandeira, Amit Singer, and Thomas Strohmer. Mathematics of data science. arXiv preprint arXiv:2607.11938, 2026. URL https://arxiv.org/abs/2607.11938v1.

Konstantina Biza, Ioannis Tsamardinos, and Sofia Triantafillou. Tuning causal discovery algorithms. In Manfred Jaeger and Thomas Dyhre Nielsen (eds.), *Proceedings of the 10th International Conference on Probabilistic Graphical Models*, volume 138 of *Proceedings of Machine Learning Research*, pp. 17–28. PMLR, 2020. URL https://proceedings.mlr.press/v138/biza20a.html.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014. doi: 10.1214/14-AOS1260.

Jiahua Chen and Zehua Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008. doi: 10.1093/biomet/asn034.

Wenyu Chen, Mathias Drton, and Y. Samuel Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 2019. doi: 10.1093/biomet/asz049.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002. URL https://www.jmlr.org/papers/v3/chickering02b.html.

Constantinos Costis Daskalakis, Anthimos Vardis Kandiros, and Rui Yao. Learning Gaussian DAG models without condition number bounds. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 12554–12589. PMLR, 2025. URL https://proceedings.mlr.press/v267/daskalakis25a.html.

Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, fifth edition, 2019. URL https://sites.math.duke.edu/~rtd/PTE/PTE5_011119.pdf.

Rina Foygel and Mathias Drton. Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems*, volume 23, pp. 604–612, 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/hash/072b030ba126b2f4b2374f342be9ed44-Abstract.html.

Jerome H. Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007. doi: 10.1214/07-AOAS131.

Jerome H. Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. doi: 10.18637/jss.v033.i01.

Juan L. Gamella, Jonas Peters, and Peter Bühlmann. Causal chambers as a real-world physical testbed for AI methodology. *Nature Machine Intelligence*, 7:107–118, 2025. doi: 10.1038/s42256-024-00964-x.

Ming Gao, Wai Ming Tai, and Bryon Aragam. Optimal estimation of Gaussian DAG models. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 8738–8757. PMLR, 2022. URL https://proceedings.mlr.press/v151/gao22a.html.

Asish Ghoshal and Jean Honorio. Learning linear structural equation models in polynomial time and sample complexity. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1466–1475. PMLR, 2018. URL https://proceedings.mlr.press/v84/ghoshal18a.html.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, fourth edition, 2013. doi: 10.56021/9781421407944. URL https://www.press.jhu.edu/books/title/10678/matrix-computations.

Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, volume 21, pp. 689–696, 2008. URL https://papers.nips.cc/paper_files/paper/2008/hash/f7664060cc52bc6f3d620bcedc94a4b6-Abstract.html.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. doi: 10.1214/aos/1015957395.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Causal discovery with score matching on additive models with arbitrary noise. In *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pp. 726–751, 2023a. URL https://proceedings.mlr.press/v213/montagna23a.html.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Scalable causal discovery with score matching. In *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pp. 752–771, 2023b. URL https://proceedings.mlr.press/v213/montagna23b.html.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and DAG constraints for learning linear DAGs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17943–17954, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d04d42cdf14579cd294e5079e0745411-Abstract.html.

Gunwoong Park. Identifiability of additive noise models using conditional variances. *Journal of Machine Learning Research*, 21(75):1–34, 2020. URL https://www.jmlr.org/papers/v21/19-664.html.

Gunwoong Park and Youngwhan Kim. Identifiability of Gaussian linear structural equation models with homogeneous and heterogeneous error variances. *Journal of the Korean Statistical Society*, 49(1):276–292, 2020. doi: 10.1007/s42952-019-00019-7.

Gunwoong Park, Sang Jun Moon, Sion Park, and Jong-June Jeon. Learning a high-dimensional linear structural equation model via ℓ_1-regularized regression. *Journal of Machine Learning Research*, 22(102):1–41, 2021. URL https://www.jmlr.org/papers/v22/20-1005.html.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, California, 1988.

Jonas Peters and Peter Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014. doi: 10.1093/biomet/ast043.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014. URL https://www.jmlr.org/papers/v15/peters14a.html.

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18741–18753, 2022. URL https://proceedings.mlr.press/v162/rolland22a.html.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005. doi: 10.1126/science.1105809.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030, 2006. URL https://www.jmlr.org/papers/v7/shimizu06a.html.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12(33):1225–1248, 2011. URL https://www.jmlr.org/papers/v12/shimizu11a.html.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, Massachusetts, second edition, 2001. doi: 10.7551/mitpress/1754.001.0001.

Seth Sullivant, Kelli Talaska, and Jan Draisma. Trek separation for Gaussian graphical models. *The Annals of Statistics*, 38(3):1665–1685, 2010. doi: 10.1214/09-AOS760. URL https://arxiv.org/abs/0812.1938.

Ryan J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013. doi: 10.1214/13-EJS815.

Nicolas Verzelen and Elisabeth Gassiat. Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli*, 24(4B):3683–3710, 2018. doi: 10.3150/17-BEJ975.

Zhuopeng Xu, Yujie Li, Cheng Liu, and Ning Gui. Ordering-based causal discovery for linear and nonlinear relations. In *Advances in Neural Information Processing Systems*, volume 37, pp. 4315–4340, 2024. doi: 10.52202/079017-0141. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/081b08068e4733ae3e7ad019fe8d172f-Abstract-Conference.html.

Peng Zhao and Bin Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7(90):2541–2563, 2006. URL https://jmlr.org/papers/v7/zhao06a.html.

Xun Zheng, Bryon Aragam, Pradeep K. Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31, pp. 9492–9503, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/e347c51419ffb23ca3fd5050202f9c3d-Abstract.html.

A Proofs for Regression Coefficients and Hold-Out Bounds

A.1 Proof of Proposition 3.1

Proof. If $S = \emptyset$, there is no coefficient to identify. Suppose $S \neq \emptyset$. Since S is ancestral, every parent of a node in S also lies in S, so the structural equations restricted to S give

$$\varepsilon_S = (I_{|S|} - B_{SS}^\top) X_S.$$

Acyclicity makes $I_{|S|} - B_{SS}^\top$ invertible. Thus X_S and ε_S determine each other and are independent of $\varepsilon_{R(S)}$. Split the error expansion in (2) as

$$X_j = \sum_{a \in S} B_{a \rightsquigarrow j} \varepsilon_a + \varepsilon_j + \sum_{a \in R(S) \setminus \{j\}} B_{a \rightsquigarrow j} \varepsilon_a.$$

Taking conditional expectations given X_S removes the errors indexed by $R(S)$. Then substituting $\varepsilon_a = X_a - \sum_{k \in S} B_{ka} X_k$ for $a \in S$ gives

$$\begin{aligned}
\mathbb{E}(X_j \mid X_S) &= \sum_{a \in S} B_{a \rightsquigarrow j} \varepsilon_a \\
&= \sum_{a \in S} B_{a \rightsquigarrow j} X_a - \sum_{a \in S} \sum_{k \in S} B_{ka} B_{a \rightsquigarrow j} X_k \\
&= \sum_{k \in S} \left(B_{k \rightsquigarrow j} - \sum_{a \in S} B_{ka} B_{a \rightsquigarrow j} \right) X_k. \qquad (15)
\end{aligned}$$

Since $\Sigma_{SS} \succ 0$, the linear conditional mean in (15) identifies the unique population regression coefficient $\theta_j(S)$. Comparing coefficients and using $B_{a \rightsquigarrow j} = 0$ for $a \notin \mathrm{An}(j)$ yields

$$\theta_{kj}(S) = B_{k \rightsquigarrow j} - \sum_{a \in S \cap \mathrm{An}(j)} B_{ka} B_{a \rightsquigarrow j}, \qquad k \in S. \qquad (16)$$

For $k \neq j$, partitioning $\mathcal{P}(k, j)$ by the first edge and using the path-sum definition in (2) gives

$$\begin{aligned}
B_{k \rightsquigarrow j} &= B_{kj} + \sum_{a \in \mathrm{An}(j)} B_{ka} \left(\sum_{\Gamma \in \mathcal{P}(a,j)} \prod_{t=1}^{\ell(\Gamma)} B_{v_{t-1} v_t} \right) \\
&= B_{kj} + \sum_{a \in \mathrm{An}(j)} B_{ka} B_{a \rightsquigarrow j}. \qquad (17)
\end{aligned}$$

Substituting (17) into (16) cancels the terms indexed by $S \cap \mathrm{An}(j)$, yielding

$$\theta_{kj}(S) = B_{kj} + \sum_{a \in \mathrm{An}(j) \setminus S} B_{ka} B_{a \rightsquigarrow j}.$$

For $k \neq j$ with $j \notin \mathrm{An}(k)$, take $S = \mathrm{An}(j) \cup \mathrm{An}(k) \cup \{k\}$, which is ancestral and excludes j. Since $\mathrm{An}(j) \subseteq S$, the two expressions for $\theta_{kj}(S)$ give the population identity in (8). If $j \in \mathrm{An}(k)$, every term in that identity is zero by acyclicity. □

A.2 Proof of Lemma 3.2

Proof. The conditional Gaussian model and $\mathcal{F}$-measurability of $\mathbf{X}, \theta, \widetilde{\theta}$ give $Y - \mathbf{X}\widetilde{\theta} \mid \mathcal{F} \sim N(\mathbf{X}(\theta - \widetilde{\theta}), v I_n)$. Since $n\widetilde{\mathcal{S}} = \|Y - \mathbf{X}\widetilde{\theta}\|_2^2$, Gaussian integration gives, for $u > 0$,

$$\mathbb{E}[e^{-nu\widetilde{\mathcal{S}}} \mid \mathcal{F}] = (1 + 2vu)^{-n/2} \exp\left\{ -\frac{u \|\mathbf{X}(\theta - \widetilde{\theta})\|_2^2}{1 + 2vu} \right\} \leq (1 + 2vu)^{-n/2}. \qquad (18)$$

Applying the Markov inequality conditionally on $\mathcal{F}$ and using (18) gives

$$\begin{aligned}
\mathbb{P}\{\widetilde{\mathcal{S}} < v - t \mid \mathcal{F}\} &= \mathbb{P}\{e^{-nu\widetilde{\mathcal{S}}} > e^{-nu(v-t)} \mid \mathcal{F}\} \\
&\leq e^{nu(v-t)} \mathbb{E}[e^{-nu\widetilde{\mathcal{S}}} \mid \mathcal{F}] \\
&\leq \exp\{nu(v-t) - (n/2)\log(1 + 2vu)\}. \qquad (19)
\end{aligned}$$

Using $\log(1 + z) \geq z - z^2/2$ for $z \geq 0$ in (19) and taking $u = t/(2v^2)$, we obtain

$$\begin{aligned}
\mathbb{P}\{\widetilde{\mathcal{S}} < v - t \mid \mathcal{F}\} &\leq \exp\{-nut + nv^2u^2\} \\
&= \exp\{-nt^2/(4v^2)\} \qquad \text{almost surely.}
\end{aligned}$$

□

B Recovery and Computational Guarantees

B.1 Auxiliary lemmas

Lemmas B.1 and B.2 control squared residual norms and inner products between independent predictor and error columns. Lemmas B.3 and B.4 control the designs uniformly over the sparse-regression cones. Lemma B.5 controls the martingale maxima in the parent-recovery proof.

Lemma B.1 (Laurent & Massart (2000, Lemma 1))**.** *Let $g_1, \dots, g_r$ be independent standard Gaussian variables, and let $a \in [0,\infty)^r \setminus \{0\}$ be a deterministic vector of weights. For $Z = \sum_{i=1}^r a_i(g_i^2 - 1)$ and every $x > 0$,*

$$\mathbb{P}\{Z \geq 2\|a\|_2\sqrt{x} + 2\|a\|_\infty x\} \leq e^{-x}, \tag{20}$$

$$\mathbb{P}\{Z \leq -2\|a\|_2\sqrt{x}\} \leq e^{-x}. \tag{21}$$

Lemma B.2 (Gaussian squares and products)**.** *Let $r \geq 1$ be an integer, and let the $2r$ variables $g_1, \dots, g_r, h_1, \dots, h_r$ be mutually independent standard Gaussian variables. For every $x > 0$,*

$$\mathbb{P}\left\{\sum_{i=1}^r g_i^2 > r + 2\sqrt{rx} + 2x\right\} \leq e^{-x}, \tag{22}$$

$$\mathbb{P}\left\{\sum_{i=1}^r g_i^2 \leq r - 2\sqrt{rx}\right\} \leq e^{-x}, \tag{23}$$

$$\mathbb{P}\left\{\left|\sum_{i=1}^r g_i h_i\right| > 2(\sqrt{rx} + x)\right\} \leq 2e^{-x}. \tag{24}$$

Proof. Apply (20) and (21) with $a = \mathbf{1} = (1, \dots, 1)^\top$ to obtain (22) and (23), respectively.

For (24), set $Z = \sum_{i=1}^r g_i h_i$ and $g = (g_1, \dots, g_r)^\top$. Equation (22) gives

$$\mathbb{P}\{\|g\|_2 > \sqrt{r} + \sqrt{2x}\} \leq e^{-x}. \tag{25}$$

Since $Z \mid g \sim N(0, \|g\|_2^2)$, the Gaussian tail bound yields

$$\mathbb{P}\{|Z| > \sqrt{2x}\|g\|_2 \mid g\} \leq e^{-x}. \tag{26}$$

Whenever $\|g\|_2 \leq \sqrt{r} + \sqrt{2x}$,

$$\sqrt{2x}\|g\|_2 \leq \sqrt{2rx} + 2x \leq 2(\sqrt{rx} + x).$$

Combining (25) and (26) therefore gives

$$\begin{aligned}\mathbb{P}\{|Z| > 2(\sqrt{rx} + x)\} &\leq \mathbb{P}\{\|g\|_2 > \sqrt{r} + \sqrt{2x}\} + \mathbb{E}\big[\mathbb{P}\{|Z| > \sqrt{2x}\|g\|_2 \mid g\}\big] \\ &\leq 2e^{-x}.\end{aligned}$$

□

Lemma B.3 (Bandeira et al. (2026, Theorem 13.17))**.** *Let $r, m \geq 1$ be integers, let $M \in \mathbb{R}^{r\times m}$ have independent $N(0,1)$ entries, and let $\mathcal{T}$ be a nonempty closed subset of the unit sphere in $\mathbb{R}^m$. Define*

$$a_r = \mathbb{E}\|h\|_2, \quad h \sim N(0, I_r), \qquad \omega(\mathcal{T}) = \mathbb{E}\sup_{u\in\mathcal{T}}\langle g, u\rangle, \quad g \sim N(0, I_m).$$

For every $\varepsilon > \omega(\mathcal{T})/a_r$, with probability at least

$$1 - 2\exp\left\{-\frac{a_r^2}{2}\left(\varepsilon - \frac{\omega(\mathcal{T})}{a_r}\right)^2\right\},$$

the following bounds hold simultaneously for all $u \in \mathcal{T}$:

$$(1-\varepsilon)\|u\|_2 \leq \left\|\frac{1}{a_r}Mu\right\|_2 \leq (1+\varepsilon)\|u\|_2.$$

Here $r^2/(r+1) \leq a_r^2 \leq r$.

Lemma B.4 (Gaussian maximum)**.** *Let $Z_1, \dots, Z_N$ be centered Gaussian variables, not necessarily independent, with $N \geq 1$ and $\max_j \mathbb{E}Z_j^2 \leq \sigma^2$ for some $\sigma \geq 0$. Then*

$$\mathbb{E} \max_{1 \leq j \leq N} |Z_j| \leq \sigma \sqrt{2 \log(2N)}. \tag{27}$$

Proof. For $t > 0$, Jensen's inequality and $\mathbb{E}e^{\pm tZ_j} \leq e^{t^2\sigma^2/2}$ give

$$\begin{aligned} \exp\left(t\, \mathbb{E} \max_{1 \leq j \leq N} |Z_j| \right) &\leq \mathbb{E} \exp\left(t \max_{1 \leq j \leq N} |Z_j| \right) \\ &\leq \sum_{j=1}^{N} \mathbb{E}\big(e^{tZ_j} + e^{-tZ_j}\big) \leq 2Ne^{t^2\sigma^2/2}. \end{aligned}$$

Taking logarithms and dividing by t yields

$$\mathbb{E} \max_{1 \leq j \leq N} |Z_j| \leq \frac{\log(2N)}{t} + \frac{t\sigma^2}{2}.$$

Taking the infimum over $t > 0$ proves the claim. □

Lemma B.5 (Doob's inequality, Durrett (2019, Theorem 4.4.2))**.** *Let X_m be a submartingale, $\overline{X}_n = \max_{0 \leq m \leq n} X_m^+$, $\lambda > 0$, and $A = \{\overline{X}_n \geq \lambda\}$. Then*

$$\lambda \mathbb{P}(A) \leq \mathbb{E}[X_n \mathbf{1}_A] \leq \mathbb{E}X_n^+. \tag{28}$$

Here $x^+ = \max\{x, 0\}$ and $\mathbf{1}_A$ is the indicator of A.

B.2 Proof of Theorem 4.1

B.2.1 Setup

If $d = 0$, every ordering is compatible. Assume $d \geq 1$ and keep the penalty $\lambda_{\mathrm{ord}} > 0$ deterministic. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ have independent $N(0, \Sigma)$ rows, and let $\mathbf{W} \in \mathbb{R}^{n \times p}$ be the structural error matrix, with independent columns $\mathbf{W}_j \sim N(0, \sigma_j^2 I_n)$. Condition throughout on the data-independent split $I_1 \cup I_2 = [n]$. Without loss of generality, assume that n is even, so $|I_1| = |I_2| = n/2$. Odd n changes only numerical constants. Let $\bar{\ell}$ denote the other fold. Both fold designs have independent $N(0, \Sigma)$ rows, and the two folds are independent. Applying (1) and (2) to all observations gives

$$\mathbf{X}_j = \sum_{a \in \mathrm{Pa}(j)} B_{aj} \mathbf{X}_a + \mathbf{W}_j, \tag{29}$$

$$\mathbf{X}_j = \sum_{a \in \mathrm{An}(j)} B_{a \rightsquigarrow j} \mathbf{W}_a + \mathbf{W}_j \qquad (j \in V). \tag{30}$$

For an ancestral set $A \subseteq V$, restriction of (29) gives $\mathbf{X}_A(I_{|A|} - B_{AA}) = \mathbf{W}_A$. Acyclicity makes $I_{|A|} - B_{AA}$ invertible, so

$$\mathbf{X}_A = \mathbf{W}_A (I_{|A|} - B_{AA})^{-1}. \tag{31}$$

For each node $f \in V$, let $\mathcal{D}_f$ contain f and its descendants, and let U_f be its complement:

$$\mathcal{D}_f = \{f\} \cup \{k \in V : f \in \mathrm{An}(k)\}, \qquad U_f = V \setminus \mathcal{D}_f.$$

The set U_f is ancestral, since a parent outside U_f would make its child a descendant of f. Thus (31) shows that $\mathbf{X}_{U_f}$ is determined by $\mathbf{W}_{U_f}$.

For a fixed predictor set $S \subsetneq V$, response node $j \notin S$, and fold $\ell \in \{1, 2\}$, define the Lasso fit by

$$\widehat{\theta}_{Sj}^{(\ell)} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^{|S|}} \left\{ \frac{\|\mathbf{X}_{I_\ell j} - \mathbf{X}_{I_\ell S}\theta\|_2^2}{n} + \lambda_{\mathrm{ord}} \|\theta\|_1 \right\}. \tag{32}$$

For nonempty S, positive definiteness of Σ_{SS} gives the fitting design $\mathbf{X}_{I_\ell S}$ a nondegenerate joint Gaussian density, so its columns are in general position almost surely. For such designs, the minimizer in (32) is unique for every response (Tibshirani, 2013, Lemma 3). Since $\lambda_{\mathrm{ord}} > 0$, comparison with

the zero vector gives $\lambda_{\mathrm{ord}}\|\widehat{\theta}^{(\ell)}_{Sj}\|_1 \le \|\mathbf{X}_{I_\ell j}\|_2^2/n$, so minimizers are locally bounded as the fitting data vary. Joint continuity of the objective and uniqueness then imply continuity of the minimizer on the Borel set of general-position designs. Extending the fit by zero elsewhere gives a version measurable with respect to $\sigma(\mathbf{X}_{I_\ell S}, \mathbf{X}_{I_\ell j})$. Take the empty vector when $S = \emptyset$. A finite intersection over predictor sets, response nodes, and folds shows that these versions agree with all HOST fits on one event of probability one. Define its loss on the other fold and the combined score by

$$\widehat{L}^{(\ell)}_{j,S} = \frac{2\|\mathbf{X}_{I_{\bar{\ell}} j} - \mathbf{X}_{I_{\bar{\ell}} S}\widehat{\theta}^{(\ell)}_{Sj}\|_2^2}{n}, \qquad \widehat{\mathcal{S}}_{j,S} = \frac{1}{2}\sum_{\ell=1}^{2} \widehat{L}^{(\ell)}_{j,S}. \tag{33}$$

Write $\widehat{S}_r$ for the nodes selected before step r, with $\widehat{S}_1 = \emptyset$. For $r \ge 2$, equations (32)–(33) at $S = \widehat{S}_r$ give $\widehat{\theta}^{(\ell)}_{\widehat{S}_r, j} = \widehat{\theta}^{(\ell)}_j(r)$ and $\widehat{\mathcal{S}}_{j,\widehat{S}_r} = \widehat{\mathcal{S}}_j(r)$ from (4)–(5). At $S = \emptyset$, (33) gives $\widehat{\mathcal{S}}_{j,\emptyset} = \|\mathbf{X}_j\|_2^2/n = \widehat{\Sigma}_{jj}$, so its minimizer agrees with the initialization of Algorithm 1.

B.2.2 Upper bounds for remaining-source scores

For $1 \le d \le p$, define the union of Lasso cones

$$\mathcal{C}_d = \bigcup_{\substack{T \subseteq [p] \\ |T| \le d}} \{v \in \mathbb{R}^p : \|v_{T^c}\|_1 \le 3\|v_T\|_1\}.$$

Lemma B.6 (Covariance bounds for the fold designs)**.** *Fix a sufficiently large universal constant C_{cov}. Under the setup in Section B.2.1, for $p \ge 2$ and $1 \le d \le p$, the condition*

$$n \ge C_{\mathrm{cov}} \frac{d\Sigma_{\max}}{\lambda_{\min}(\Sigma)} \log(2p) \tag{34}$$

implies that, with probability at least $1 - 4p^{-8}$,

$$\frac{1}{4} v^\top \Sigma v \le \frac{2\|\mathbf{X}_{I_\ell} v\|_2^2}{n} \le \frac{9}{4} v^\top \Sigma v \quad (v \in \mathcal{C}_d,\ \ell = 1, 2). \tag{35}$$

Proof. Write $\kappa = \Sigma_{\max}/\lambda_{\min}(\Sigma) \ge 1$ and $\mathcal{T} = \{v \in \mathcal{C}_d : v^\top \Sigma v = 1\}$. Every nonzero $v \in \mathcal{C}_d$ satisfies $v/\sqrt{v^\top \Sigma v} \in \mathcal{T}$, so it suffices to prove (35) on $\mathcal{T}$. For each such v, choose $J \subseteq [p]$ with $|J| \le d$ and $\|v_{J^c}\|_1 \le 3\|v_J\|_1$. The cone condition and Cauchy–Schwarz give

$$\|v\|_1 \le 4\|v_J\|_1 \le 4\sqrt{d}\,\|v\|_2 \le 4\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}},$$

where the last inequality follows from $1 = v^\top \Sigma v \ge \lambda_{\min}(\Sigma)\|v\|_2^2$. For $g \sim N(0, I_p)$, each coordinate of $\Sigma^{1/2} g$ has variance at most $\Sigma_{\max}$. The Hölder inequality and Lemma B.4 therefore yield

$$\omega(\Sigma^{1/2}\mathcal{T}) \le 4\sqrt{\frac{d}{\lambda_{\min}(\Sigma)}}\,\mathbb{E}\|\Sigma^{1/2} g\|_\infty \le 4\sqrt{2d\kappa \log(2p)}.$$

The matrix $\mathbf{X}_{I_\ell}\Sigma^{-1/2}$ has independent standard Gaussian entries, and $\Sigma^{1/2}\mathcal{T}$ is a closed subset of the unit sphere. Apply Lemma B.3 with

$$\varepsilon = \frac{\omega(\Sigma^{1/2}\mathcal{T}) + 4\sqrt{\log p}}{a_{n/2}}.$$

Since $r^2/(r+1) \ge r - 1$, the lemma gives $\sqrt{n/2 - 1} \le a_{n/2} \le \sqrt{n/2}$. Taking a union bound over the two folds then gives, with probability at least $1 - 4p^{-8}$, simultaneously for $v \in \mathcal{T}$ and $\ell = 1, 2$,

$$\|\mathbf{X}_{I_\ell} v\|_2 \ge \sqrt{n/2 - 1} - 4\sqrt{2d\kappa \log(2p)} - 4\sqrt{\log p},$$
$$\|\mathbf{X}_{I_\ell} v\|_2 \le \sqrt{n/2} + 4\sqrt{2d\kappa \log(2p)} + 4\sqrt{\log p}.$$

For sufficiently large C_{cov}, condition (34) implies

$$\left(4\sqrt{2d\kappa\log(2p)}+4\sqrt{\log p}\right)^2 \le 96d\kappa\log(2p) \le \frac{n/2-4}{4}.$$

Consequently,

$$\|\mathbf{X}_{I_\ell}v\|_2 \ge \sqrt{n/2-1}-\frac{1}{2}\sqrt{n/2-4} \ge \frac{1}{2}\sqrt{n/2},$$
$$\|\mathbf{X}_{I_\ell}v\|_2 \le \sqrt{n/2}+\frac{1}{2}\sqrt{n/2-4} \le \frac{3}{2}\sqrt{n/2}.$$

Squaring and rescaling from $\mathcal{T}$ to $\mathcal{C}_d$ proves (35). □

Lemma B.7 (Uniform upper bounds for remaining-source scores)**.** *Fix sufficiently large universal constants C_{ord} and C_{L}. Suppose $d \ge 1$, condition* (34) *holds, and*

$$\lambda_{\mathrm{ord}} \ge C_{\mathrm{ord}}\sigma_{\max}\sqrt{\frac{\Sigma_{\max}\log p}{n}}. \tag{36}$$

With probability at least $1-6p^{-6}$, every ancestral $S \subsetneq V$, every $j \in \mathrm{Src}(S)$, and both $\ell \in \{1,2\}$ satisfy

$$\widehat{L}^{(\ell)}_{j,S} \le \sigma_j^2\left(1+8\sqrt{\frac{\log p}{n}}+\frac{32\log p}{n}\right)+\frac{C_{\mathrm{L}}d\lambda_{\mathrm{ord}}^2}{\lambda_{\min}(\Sigma)}. \tag{37}$$

Inequality (37) *also holds with $\widehat{L}^{(\ell)}_{j,S}$ replaced by $\widehat{\mathcal{S}}_{j,S}$.*

Proof. Lemma B.6 gives (35) with failure probability $4p^{-8}$. If $a \notin \mathcal{D}_j$, then $j \notin \{a\}\cup \mathrm{An}(a)$. The error representation (30) and the independence of the columns of $\mathbf{W}$ therefore imply that $\mathbf{X}_a$ is independent of $\mathbf{W}_j$.

For independent $g, h \sim N(0, I_{n/2})$, the Gaussian fold distributions give

$$\frac{\mathbf{X}_{I_\ell a}^\top \mathbf{W}_{I_\ell j}}{\sigma_j\sqrt{\Sigma_{aa}}} \overset{d}{=} g^\top h = \sum_{i=1}^{n/2} g_i h_i \qquad (a \notin \mathcal{D}_j),$$
$$\frac{\|\mathbf{W}_{I_\ell j}\|_2^2}{\sigma_j^2} \overset{d}{=} \|h\|_2^2 = \sum_{i=1}^{n/2} h_i^2 \qquad (j \in V).$$

Apply (24) and (22) with $r = n/2$ and $x = 8\log p$. Using $n/2 \ge 8\log p$ from (34) and the penalty condition (36) gives, for fixed indices, with respective failure probabilities $2p^{-8}$ and p^{-8},

$$\frac{2|\mathbf{X}_{I_\ell a}^\top \mathbf{W}_{I_\ell j}|}{n} \le 2\sigma_j\sqrt{\Sigma_{aa}}\left(4\sqrt{\frac{\log p}{n}}+\frac{16\log p}{n}\right)$$
$$\le 16\sigma_{\max}\sqrt{\frac{\Sigma_{\max}\log p}{n}} \le \lambda_{\mathrm{ord}}/2 \quad (a \notin \mathcal{D}_j), \tag{38}$$
$$\frac{2\|\mathbf{W}_{I_\ell j}\|_2^2}{n} \le \sigma_j^2\left(1+8\sqrt{\frac{\log p}{n}}+\frac{32\log p}{n}\right) \quad (j \in [p]). \tag{39}$$

A union bound over both folds and at most p^2 pairs (a, j) in (38), and p nodes in (39), gives respective failure probabilities $4p^{-6}$ and $2p^{-7}$.

Work on the event where (35), (38), and (39) all hold. These three events do not depend on the predictor set S. Fix an ancestral S, a remaining source $j \in \mathrm{Src}(S)$, and a fitting fold ℓ. Since S is ancestral and $j \notin S$, it contains no descendant of j. Thus $T = \mathrm{Pa}(j) \subseteq S \subseteq U_j$ and $|T| \le d$. Define the estimation error $\Delta \in \mathbb{R}^p$ by

$$\Delta_S = \widehat{\theta}^{(\ell)}_{Sj} - B_{Sj}, \qquad \Delta_{S^c} = 0.$$

The true coefficient vector $B_{\cdot j}$ is supported on $T \subseteq S$. By (29), optimality of the Lasso fit gives

$$\begin{aligned}
&\frac{\|\mathbf{W}_{I_\ell j} - \mathbf{X}_{I_\ell}\Delta\|_2^2}{n} + \lambda_{\mathrm{ord}}\|B_{\cdot j} + \Delta\|_1 \\
&= \frac{\|\mathbf{W}_{I_\ell j}\|_2^2}{n} - \frac{2\mathbf{W}_{I_\ell j}^\top \mathbf{X}_{I_\ell}\Delta}{n} + \frac{\|\mathbf{X}_{I_\ell}\Delta\|_2^2}{n} + \lambda_{\mathrm{ord}}\|B_{\cdot j} + \Delta\|_1 \\
&\le \frac{\|\mathbf{W}_{I_\ell j}\|_2^2}{n} + \lambda_{\mathrm{ord}}\|B_{\cdot j}\|_1.
\end{aligned}$$

Since Δ is supported on $S \subseteq U_j$ and $B_{T^c,j} = 0$, equation (38) and the triangle inequality give

$$\begin{aligned}
\frac{\|\mathbf{X}_{I_\ell}\Delta\|_2^2}{n} &\le \frac{2\mathbf{W}_{I_\ell j}^\top \mathbf{X}_{I_\ell}\Delta}{n} + \lambda_{\mathrm{ord}}\big(\|B_{\cdot j}\|_1 - \|B_{\cdot j} + \Delta\|_1\big) \\
&\le \frac{\lambda_{\mathrm{ord}}}{2}\|\Delta\|_1 + \lambda_{\mathrm{ord}}\big(\|\Delta_T\|_1 - \|\Delta_{T^c}\|_1\big).
\end{aligned}$$

Rearranging yields the basic inequality

$$\frac{2\|\mathbf{X}_{I_\ell}\Delta\|_2^2}{n} + \lambda_{\mathrm{ord}}\|\Delta_{T^c}\|_1 \le 3\lambda_{\mathrm{ord}}\|\Delta_T\|_1. \tag{40}$$

Hence $\Delta \in \mathcal{C}_d$. The lower bound in (35) and Cauchy–Schwarz give

$$\frac{\|\Sigma^{1/2}\Delta\|_2^2}{4} \le \frac{2\|\mathbf{X}_{I_\ell}\Delta\|_2^2}{n} \le 3\lambda_{\mathrm{ord}}\|\Delta_T\|_1 \le 3\lambda_{\mathrm{ord}}\sqrt{\frac{|T|}{\lambda_{\min}(\Sigma)}}\,\|\Sigma^{1/2}\Delta\|_2.$$

The cone inequality $\|\Delta_{T^c}\|_1 \le 3\|\Delta_T\|_1$ from (40) also gives

$$\|\Delta\|_1 \le 4\|\Delta_T\|_1 \le 4\sqrt{\frac{|T|}{\lambda_{\min}(\Sigma)}}\,\|\Sigma^{1/2}\Delta\|_2.$$

Consequently,

$$\|\Sigma^{1/2}\Delta\|_2^2 \le \frac{12^2|T|\lambda_{\mathrm{ord}}^2}{\lambda_{\min}(\Sigma)}, \qquad \lambda_{\mathrm{ord}}\|\Delta\|_1 \le \frac{4\cdot 12|T|\lambda_{\mathrm{ord}}^2}{\lambda_{\min}(\Sigma)}. \tag{41}$$

On the other fold, expanding the residual norm and applying (38), the upper bound in (35), and (41) gives

$$\begin{aligned}
\widehat{L}_{j,S}^{(\ell)} &= \frac{2\|\mathbf{W}_{I_{\bar\ell} j} - \mathbf{X}_{I_{\bar\ell}}\Delta\|_2^2}{n} \\
&\le \frac{2\|\mathbf{W}_{I_{\bar\ell} j}\|_2^2}{n} + \frac{9}{4}\|\Sigma^{1/2}\Delta\|_2^2 + \lambda_{\mathrm{ord}}\|\Delta\|_1 \\
&\le \frac{2\|\mathbf{W}_{I_{\bar\ell} j}\|_2^2}{n} + \frac{C_{\mathrm{L}}|T|\lambda_{\mathrm{ord}}^2}{\lambda_{\min}(\Sigma)}.
\end{aligned} \tag{42}$$

Using (39) and $|T| \le d$ in (42) proves (37). If $S = \emptyset$, then $T = \emptyset$ and $\Delta = 0$, so (42) still holds. Averaging the two fold bounds according to (33) proves (37) for $\widehat{\mathcal{S}}_{j,S}$. The events in (35), (38), and (39) establish (37) for every ancestral S and remaining source j, so no union over predictor sets is needed. Adding their failure probabilities gives

$$4p^{-8} + 4p^{-6} + 2p^{-7} \le 6p^{-6}. \qquad \square$$

B.2.3 Lower bounds for remaining-nonsource scores

Lemma B.8 (Conditional lower tail for a selected predictor set)**.** *Fix a node $f \in V$ and a proper descendant k of f. Let $S \subseteq U_f$ be a random predictor set measurable with respect to $\mathbf{W}_{U_f}$. For each fold $\ell \in \{1,2\}$, define*

$$\mathcal{F}_{f,\ell} = \sigma(\mathbf{W}_{U_f}, \mathbf{W}_{I_\ell \mathcal{D}_f}).$$

For every $0 < t < \mathcal{S}_k(U_f)$, almost surely,

$$\mathbb{P}\left\{\widehat{L}_{k,S}^{(\ell)} < \mathcal{S}_k(U_f) - t \mid \mathcal{F}_{f,\ell}\right\} \le \exp\left\{-\frac{nt^2}{8\mathcal{S}_k(U_f)^2}\right\}. \tag{43}$$

Proof. Since S is $\mathbf{W}_{U_f}$-measurable, (31) makes both fold designs on S measurable with respect to $\mathbf{W}_{U_f}$. The fitting-fold errors $\mathbf{W}_{I_\ell V}$ are also $\mathcal{F}_{f,\ell}$-measurable and determine $\mathbf{X}_{I_\ell k}$ through (30). Thus the minimizer rule fixed in Section B.2.1 makes $\widehat{\theta}^{(\ell)}_{Sk}$ and $\mathbf{X}_{I_{\bar{\ell}}S}\widehat{\theta}^{(\ell)}_{Sk}$ $\mathcal{F}_{f,\ell}$-measurable.

By (30) and Proposition 3.1, the held-out response satisfies

$$\mathbf{X}_{I_{\bar{\ell}}k} = \mathbf{X}_{I_{\bar{\ell}}U_f}\theta_k(U_f) + \mathbf{W}_{I_{\bar{\ell}}k} + \sum_{a\in\mathcal{D}_f\setminus\{k\}} B_{a\rightsquigarrow k}\mathbf{W}_{I_{\bar{\ell}}a}. \tag{44}$$

The errors $\mathbf{W}_{I_{\bar{\ell}}a}$ for $a \in \mathcal{D}_f$ are mutually independent and independent of $\mathcal{F}_{f,\ell}$, because observations in the two folds and errors at different nodes are independent. The error term in (44) has conditional distribution

$$\mathbf{W}_{I_{\bar{\ell}}k} + \sum_{a\in\mathcal{D}_f\setminus\{k\}} B_{a\rightsquigarrow k}\mathbf{W}_{I_{\bar{\ell}}a} \;\Big|\; \mathcal{F}_{f,\ell} \sim N\big(0, \mathcal{S}_k(U_f)I_{n/2}\big)$$

by (3). Apply Lemma 3.2 to (44), with design $\mathbf{X}_{I_{\bar{\ell}}U_f}$, coefficient $\theta_k(U_f)$, and fitted coefficient equal to $\widehat{\theta}^{(\ell)}_{Sk}$ on S and zero on $U_f \setminus S$. This design and these coefficients are $\mathcal{F}_{f,\ell}$-measurable. By (33), the resulting loss in dimension $n/2$ is $\widehat{L}^{(\ell)}_{k,S}$, so the conclusion of Lemma 3.2 is (43). □

B.2.4 Completion of the proof

We prove recovery by induction on the number of selected nodes. Let

$$\mathcal{A}_r = \bigcap_{s=1}^{r}\{\mathrm{Pa}(\widehat{\pi}_s) \subseteq \widehat{S}_s\}, \qquad r = 0, \dots, p,$$

be the event that the first r choices are all remaining sources, with $\mathcal{A}_0$ the sure event. On $\mathcal{A}_{r-1}$, the set $\widehat{S}_r$ is ancestral. Define the source upper-bound event by

$$\mathcal{E}_{\mathrm{src}} = \bigcap_{\substack{S\subsetneq V\\ S \text{ ancestral}}} \bigcap_{a\in\mathrm{Src}(S)} \Big\{\widehat{\mathcal{S}}_{a,S} \le \sigma_a^2 + \tau_\star/4\Big\}. \tag{45}$$

First we bound $\mathbb{P}(\mathcal{E}^c_{\mathrm{src}})$. Use the fixed multiplier C_{ord} in Lemma B.7 and choose $\lambda_{\mathrm{ord}} = C_{\mathrm{ord}}\sigma_{\max}\sqrt{\Sigma_{\max}\log p/n}$, as prescribed in Theorem 4.1. The first term of (11), with a sufficiently large fixed factor, ensures the design condition (34) and

$$\frac{C_{\mathrm{L}} d\lambda^2_{\mathrm{ord}}}{\lambda_{\min}(\Sigma)} = C_{\mathrm{L}}C^2_{\mathrm{ord}}\frac{d\Sigma_{\max}}{\lambda_{\min}(\Sigma)}\frac{\sigma^2_{\max}\log p}{n} \le \frac{\tau_\star}{8}.$$

The second term of (11) controls the fluctuation of the error norms. Since $\sigma_j^2 \le \Sigma_{jj} \le \Sigma_{\max}$, it ensures

$$\sigma_j^2\left(8\sqrt{\frac{\log p}{n}} + \frac{32\log p}{n}\right) \le \Sigma_{\max}\left(8\sqrt{\frac{\log p}{n}} + \frac{32\log p}{n}\right) \le \frac{\tau_\star}{8}.$$

Lemma B.7 therefore gives

$$\mathbb{P}(\mathcal{E}^c_{\mathrm{src}}) \le 6p^{-6}. \tag{46}$$

We next bound the probability that a correct prefix gains a remaining nonsource, then iterate over r.

Lemma B.9 (One-step ordering error)**.** Under the sample-size and penalty conditions of Theorem 4.1, every $r \in [p]$ satisfies

$$\mathbb{P}(\mathcal{E}_{\mathrm{src}} \cap \mathcal{A}_{r-1} \cap \mathcal{A}^c_r) \le 2p^{-6}. \tag{47}$$

Proof. Suppose the first $r-1$ choices are remaining sources, so the selected set $\widehat{S}_r$ is ancestral. For any remaining nonsource k, following unselected parents backwards from k reaches a remaining source f. Since f is unselected and $\widehat{S}_r$ is ancestral, none of its descendants has been selected either. Thus all previous choices lie in U_f, giving $\widehat{S}_r \subseteq U_f$ and $|U_f| \ge r-1$.

Fix a node f with $|U_f| \ge r-1$. We now determine the value of $\widehat{S}_r$ whenever $\widehat{S}_r \subseteq U_f$, using only the data $\mathbf{X}_{U_f}$. At a specified selected set $T \subseteq U_f$, every candidate $a \in U_f \setminus T$ uses a response and

predictors from $\mathbf{X}_{U_f}$, so its fits and score depend only on these data. If the next HOST choice is in U_f, it must be the smallest-labeled node in

$$\operatorname*{argmin}_{a \in U_f \setminus T} \widehat{\mathcal{S}}_{a,T}. \tag{48}$$

Starting from the empty set, successive minimization in (48) determines each choice as long as HOST selects within U_f. The scores of candidates in $V \setminus U_f$ can determine when HOST first leaves U_f, but cannot change the choices preceding that exit.

By the measurability properties in Section B.2.1, applying (48) $r-1$ times defines a $\mathbf{W}_{U_f}$-measurable set

$$S_{r,f} = S_{r,f}(\mathbf{W}_{U_f}) \subseteq U_f, \qquad |S_{r,f}| = r-1.$$

Thus fixing r, f, and $\mathbf{W}_{U_f}$ fixes $S_{r,f}$. This set agrees with the actual selected set $\widehat{S}_r$ whenever $\widehat{S}_r \subseteq U_f$:

$$\widehat{S}_r \subseteq U_f \quad \Longrightarrow \quad \widehat{S}_r = S_{r,f} \qquad \text{almost surely.} \tag{49}$$

Now fix $k \in \mathcal{D}_f \setminus \{f\}$. On $\mathcal{E}_{\mathrm{src}} \cap \mathcal{A}_{r-1}$, if HOST selects k while f is a remaining source, then (49) gives $\widehat{S}_r = S_{r,f}$, and score minimization gives

$$\widehat{\mathcal{S}}_{k,S_{r,f}} \le \widehat{\mathcal{S}}_{f,S_{r,f}} \le \sigma_f^2 + \tau_\star/4 < \sigma_f^2 + 3\tau_\star/4. \tag{50}$$

Equation (50) shows that selecting k under these conditions requires $\widehat{\mathcal{S}}_{k,S_{r,f}} < \sigma_f^2 + 3\tau_\star/4$. We now bound the unconditional probability of this score inequality.

With U_f as the predictor set, all parents of f are included, whereas a parent of k on a path from f lies outside it. Thus f is the unique remaining source at U_f, and every proper descendant of f, including k, is a remaining nonsource. Equation (3) gives $\mathcal{S}_f(U_f) = \sigma_f^2$, and choosing the zero coefficient vector gives $\mathcal{S}_k(U_f) \le \Sigma_{kk}$. Assumption 1 therefore implies

$$\sigma_f^2 + \tau_\star \le \mathcal{S}_k(U_f) \le \Sigma_{kk} \le \Sigma_{\max}. \tag{51}$$

Since $S_{r,f}$ is $\mathbf{W}_{U_f}$-measurable, (43) applies with $S = S_{r,f}$ and $t = \tau_\star/4$. The combined score $\widehat{\mathcal{S}}_{k,S_{r,f}}$ averages the two fold losses by (33), and (51) gives $\sigma_f^2 + 3\tau_\star/4 \le \mathcal{S}_k(U_f) - \tau_\star/4$. Taking expectations in (43) therefore yields

$$\begin{aligned}
&\mathbb{P}\{\widehat{\mathcal{S}}_{k,S_{r,f}} < \sigma_f^2 + 3\tau_\star/4\} \\
&\quad \le \sum_{\ell=1}^{2} \mathbb{E}\Big[\mathbb{P}\Big\{\widehat{L}^{(\ell)}_{k,S_{r,f}} < \mathcal{S}_k(U_f) - \tau_\star/4 \,\Big|\, \mathcal{F}_{f,\ell}\Big\}\Big] \\
&\quad \le 2\exp\left\{-\frac{n\tau_\star^2}{128\Sigma_{\max}^2}\right\} \le 2p^{-8}.
\end{aligned} \tag{52}$$

The bound by $2p^{-8}$ in (52) follows from the second term of (11), with its fixed factor sufficiently large. Every error on $\mathcal{E}_{\mathrm{src}} \cap \mathcal{A}_{r-1}$ yields a remaining source f and a selected descendant k satisfying (50). Summing (52) over at most p^2 such pairs gives $p^2 \cdot 2p^{-8} = 2p^{-6}$, proving (47). □

Finally, $\mathcal{A}_r \subseteq \mathcal{A}_{r-1}$, so Lemma B.9 gives, for $r = 1, \dots, p$,

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_{\mathrm{src}} \cap \mathcal{A}_r^c) &\le \mathbb{P}(\mathcal{E}_{\mathrm{src}} \cap \mathcal{A}_{r-1}^c) + 2p^{-6} \\
&\le 2rp^{-6},
\end{aligned} \tag{53}$$

where the last inequality follows by induction from $\mathbb{P}(\mathcal{E}_{\mathrm{src}} \cap \mathcal{A}_0^c) = 0$. Since $\mathcal{A}_p$ is exactly the event that $\widehat{\pi}$ is compatible with G, combining (46) and (53) at $r = p$ gives

$$\begin{aligned}
\mathbb{P}\{\widehat{\pi} \text{ is incompatible with } G\} &\le \mathbb{P}(\mathcal{E}_{\mathrm{src}}^c) + \mathbb{P}(\mathcal{E}_{\mathrm{src}} \cap \mathcal{A}_p^c) \\
&\le 6p^{-6} + 2p^{-5} \le p^{-2}.
\end{aligned}$$

B.3 Proof of Theorem 4.2

We first control the fixed true-parent fits. A direct expansion of the reconstruction error then gives martingale increments, whose exponential bounds yield a uniform reconstruction event. On this event, thresholding and induction recover all parents along every compatible ordering.

B.3.1 SETUP

Use the data matrix $\mathbf{X}$ and structural errors $\mathbf{W}$ from Section B.2.1. Fix one deterministic ordering compatible with G and label $V = [p]$ accordingly throughout this section. Write $\mathcal{F}_k = \sigma(\mathbf{W}_1, \dots, \mathbf{W}_k)$, with $\mathcal{F}_0$ trivial. This ordering fixes the order in which we expose the structural errors; the final recovery argument applies to any compatible ordering supplied to HOST. Until that argument, all local fits use the true parent sets:

$$\begin{aligned} D_k &= \mathbf{X}_{\mathrm{Pa}(k)}, \qquad P_k = D_k D_k^{\dagger}, \\ \widehat{B}^{\mathrm{LS}}_{\mathrm{Pa}(k),k} &= D_k^{\dagger}\mathbf{X}_k, \qquad \widehat{B}^{\mathrm{LS}}_{V\setminus\mathrm{Pa}(k),k} = 0, \\ \widehat{r}_k &= (I_n - P_k)\mathbf{X}_k = (I_n - P_k)\mathbf{W}_k. \end{aligned} \tag{54}$$

For an empty design, $P_k = 0$. The sample size assumptions below ensure $n > d$, so every nonempty D_k has full column rank almost surely. For $i \in \mathrm{Pa}(k)$, put $h_{ik} = D_k[(D_k^{\top} D_k)^{-1}]_{\cdot i}$, with columns indexed by $\mathrm{Pa}(k)$, and put $h_{ik} = 0$ otherwise. Every parent precedes its child, so D_k, P_k, h_{ik} are $\mathcal{F}_{k-1}$-measurable and $\mathbf{W}_k$ is independent of $\mathcal{F}_{k-1}$. The OLS error identity and total-effect estimates are as follows, with any zero-denominator ratio defined as zero:

$$\widehat{B}^{\mathrm{LS}}_{ik} - B_{ik} = h_{ik}^{\top}\mathbf{W}_k, \qquad P_k h_{ik} = h_{ik}, \tag{55}$$

$$\widehat{B}_{k\rightsquigarrow j} = \frac{\widehat{r}_k^{\top}\mathbf{X}_j}{\|\widehat{r}_k\|_2^2} \qquad (k \neq j). \tag{56}$$

We use the following covariance relations. For $i \in T \subseteq V$, additional Gaussian conditioning can only reduce conditional variance, so

$$(\Sigma_{TT}^{-1})_{ii} = \frac{1}{\mathrm{Var}(X_i \mid X_{T\setminus\{i\}})} \leq \frac{1}{\mathrm{Var}(X_i \mid X_{V\setminus\{i\}})} = \Omega_{ii}. \tag{57}$$

The SEM also gives $\Omega = (I - B)\,\mathrm{diag}(\sigma_k^{-2} : k \in V)(I - B)^{\top}$, hence

$$\Omega_{ii} = \sum_{k\in V} \frac{(\mathbf{1}\{k = i\} - B_{ik})^2}{\sigma_k^2}, \tag{58}$$

where $\mathbf{1}\{\cdot\}$ denotes an indicator.

B.3.2 EVENTS FOR THE FIXED-PARENT ESTIMATES

Fix a sufficiently large universal constant C_{par}. We use two events for the OLS coefficients and total-effect estimates in (54)–(56). The first event controls the local fits. For each node k, let

$$\mathcal{E}_k = \{D_k^{\top} D_k \succeq n\Sigma_{\mathrm{Pa}(k),\mathrm{Pa}(k)}/4\}, \tag{59}$$

with $\mathcal{E}_k$ the whole space for an empty design. Since D_k is $\mathcal{F}_{k-1}$-measurable, so is $\mathcal{E}_k$. Define

$$\mathcal{E}_{\mathrm{LS}} = \bigcap_{k\in V} \big(\mathcal{E}_k \cap \{\|\widehat{r}_k\|_2^2 \geq n\sigma_k^2/2\}\big) \cap \bigcap_{(i,k)\in E} \{|\widehat{B}^{\mathrm{LS}}_{ik} - B_{ik}| \leq |B_{ik}|\}. \tag{60}$$

On $\mathcal{E}_{\mathrm{LS}}$, the local coefficient errors are bounded by the corresponding edge magnitudes, and the designs and residual denominators are bounded away from degeneracy. The second event controls the combination of OLS and total-effect estimates used for structural subtraction in HOST:

$$\mathcal{E}_{\mathrm{rec}} = \left\{ \max_{\substack{i\neq j \\ j\notin \mathrm{An}(i)}} \left| \widehat{B}_{i\rightsquigarrow j} - \sum_{a\in\mathrm{An}(j)} \widehat{B}^{\mathrm{LS}}_{ia}\widehat{B}_{a\rightsquigarrow j} - B_{ij} \right| \leq C_{\mathrm{par}}\sqrt{\frac{\Sigma_{\max}\Omega_{\max}\log p}{n}} \right\}. \tag{61}$$

The pairs in (61) include each target and every node earlier than that target in a compatible ordering. We first bound the failure probabilities of these two events and then show how they yield recovery.

Lemma B.10 (Local-fit event)**.** Use the true-parent fits from Section B.3.1, and suppose

$$n \geq C_{\mathrm{par}}(d + \log p), \qquad nB_{ik}^2 \geq C_{\mathrm{par}}\sigma_k^2(\Sigma^{-1}_{\mathrm{Pa}(k),\mathrm{Pa}(k)})_{ii}\log p \quad (i \to k). \tag{62}$$

Then $\mathbb{P}(\mathcal{E}_{\mathrm{LS}}^c) \leq 5p^{-6}$.

Proof. On $\mathcal{E}_k$, inverse monotonicity and (57) give

$$\|h_{ik}\|_2^2 = [(D_k^\top D_k)^{-1}]_{ii} \le \frac{4}{n}(\Sigma^{-1}_{\mathrm{Pa}(k),\mathrm{Pa}(k)})_{ii} \le \frac{4\Omega_{ii}}{n} \qquad (i \in \mathrm{Pa}(k)). \tag{63}$$

The coefficient error satisfies $(\widehat{B}^{\mathrm{LS}}_{ik} - B_{ik}) \mid D_k \sim N(0, \sigma_k^2\|h_{ik}\|_2^2)$. On $\mathcal{E}_k$, the Gaussian tail bound, (63), and the signal condition in (62) give

$$\mathbb{P}\{|\widehat{B}^{\mathrm{LS}}_{ik} - B_{ik}| > |B_{ik}| \mid D_k\} \le 2\exp\left\{-\frac{nB_{ik}^2}{8\sigma_k^2(\Sigma^{-1}_{\mathrm{Pa}(k),\mathrm{Pa}(k)})_{ii}}\right\} \le 2p^{-8}. \tag{64}$$

Also $\|\widehat{r}_k\|_2^2/\sigma_k^2 \mid D_k \sim \chi^2_{r_k}$ with $r_k = n - |\mathrm{Pa}(k)|$. For sufficiently large C_{par}, the sample condition gives $r_k - 2\sqrt{8r_k \log p} \ge n/2$, so (23) yields

$$\mathbb{P}\{\|\widehat{r}_k\|_2^2 < n\sigma_k^2/2 \mid D_k\} \le p^{-8}. \tag{65}$$

For a nonempty parent set, put $m = |\mathrm{Pa}(k)|$. The matrix $D_k\Sigma^{-1/2}_{\mathrm{Pa}(k),\mathrm{Pa}(k)}$ is standard Gaussian, and the unit sphere $\mathcal{T} = \{z \in \mathbb{R}^m : \|z\|_2 = 1\}$ has Gaussian width $\omega(\mathcal{T}) = \mathbb{E}\|g\|_2 = a_m \le \sqrt{m} \le \sqrt{d}$, where $g \sim N(0, I_m)$. Apply Lemma B.3 with $r = n$, $M = D_k\Sigma^{-1/2}_{\mathrm{Pa}(k),\mathrm{Pa}(k)}$, and $\varepsilon = (a_m + 4\sqrt{\log p})/a_n$. Using $a_n^2 \ge n^2/(n+1) \ge n - 1$ gives

$$\inf_{\|z\|_2=1} \|D_k\Sigma^{-1/2}_{\mathrm{Pa}(k),\mathrm{Pa}(k)}z\|_2 \ge \sqrt{n-1} - \sqrt{d} - 4\sqrt{\log p} \ge \sqrt{n}/2$$

with failure probability at most $2\exp\{-a_n^2(\varepsilon - a_m/a_n)^2/2\} = 2p^{-8}$. Hence $\mathbb{P}(\mathcal{E}_k^c) \le 2p^{-8}$; this also holds for an empty parent set. Taking expectations in the conditional bounds and summing over nodes and edges proves

$$\mathbb{P}(\mathcal{E}_{\mathrm{LS}}^c) \le (3p + 2|E|)p^{-8} \le 5p^{-6}. \tag{66}$$

□

Lemma B.11 (Martingale bounds for reconstruction)**.** Suppose $n > d$, and fix $i \ne j$ with $j \notin \mathrm{An}(i)$. Put $K = \mathrm{An}(j) \cup \{i\}$ and define

$$\begin{aligned} \widetilde{B}^{\mathrm{LS}}_{ik} &= \mathrm{clip}_{2|B_{ik}|}(\widehat{B}^{\mathrm{LS}}_{ik}), \qquad R_k = \max\{\|\widehat{r}_k\|_2^2, n\sigma_k^2/2\}, \\ b_k &= \mathbf{1}\{k = i\} - B_{ik}, \qquad u_k = \frac{(\mathbf{1}\{k = i\} - \widetilde{B}^{\mathrm{LS}}_{ik})\widehat{r}_k}{R_k} \quad (k \in K), \\ V_m &= \sum_{\substack{k \in K \\ k \le m}} u_k, \qquad v_{ij} = V_p, \qquad M_m = \mathbb{E}[\mathbf{X}_j \mid \mathcal{F}_m]. \end{aligned} \tag{67}$$

Here V_m and M_m are defined for every $m = 0, \dots, p$, whether or not $m \in K$; the empty sum gives $V_0 = 0$. Then $(V_m)_{m=0}^p$ is a martingale, and for every $\mathcal{F}_{k-1}$-measurable vector $w \in \mathbb{R}^n$,

$$\mathbb{E}[\exp\{w^\top u_k\} \mid \mathcal{F}_{k-1}] \le \exp\left\{\frac{8b_k^2\|w\|_2^2}{n^2\sigma_k^2}\right\} \qquad (k \in K). \tag{68}$$

The event

$$\mathcal{G}_{ij} = \left\{\max_{0 \le m \le p} \|V_m\|_2^2 \le \frac{64\Omega_{ii}}{n}, \quad \max_{0 \le m \le p} \|M_m\|_2^2 \le 4n\Sigma_{jj}\right\} \tag{69}$$

satisfies $\mathbb{P}(\mathcal{G}_{ij}^c) \le 2\exp\{-n/2\}$.

Proof. Conditional on $\mathcal{F}_{k-1}$, the Gaussian projections $P_k\mathbf{W}_k$ and $\widehat{r}_k = (I_n - P_k)\mathbf{W}_k$ are independent. By (55), the clipped coefficient depends only on the first projection, whereas $\widehat{r}_k/R_k$ is an odd function of the second. Hence

$$\mathbb{E}[u_k \mid \mathcal{F}_{k-1}] = 0 \qquad (k \in K). \tag{70}$$

We first verify the martingale claim. Each V_m is $\mathcal{F}_m$-measurable and integrable: its summands are measurable functions of the errors exposed by step m, and the clipped coefficients and denominator

floors give $\|u_k\|_2 \le 4|b_k|\|\widehat{r}_k\|_2/(n\sigma_k^2)$, which has finite expectation. If $m \in K$, then $V_m = V_{m-1} + u_m$, so (70) gives

$$\mathbb{E}[V_m \mid \mathcal{F}_{m-1}] = V_{m-1} + \mathbb{E}[u_m \mid \mathcal{F}_{m-1}] = V_{m-1}.$$

If $m \notin K$, then $V_m = V_{m-1}$ and the same conditional-expectation identity holds. Thus $(V_m)_{m=0}^p$ is a martingale. Since $v_{ij} = V_p$, repeated conditioning also gives $V_m = \mathbb{E}[v_{ij} \mid \mathcal{F}_m]$.

We next prove the conditional exponential-moment bound (68). Iterating this bound will control the exponential moment of $\|v_{ij}\|_2^2$; together with the corresponding Gaussian bound for $\mathbf{X}_j$, this will let us apply Doob's inequality to control the maxima of V_m and M_m in (69). For $\mathcal{F}_{k-1}$-measurable w and a standard Gaussian variable $Z \sim N(0,1)$, we obtain

$$\begin{aligned}
&\mathbb{E}[\exp\{w^\top u_k\} \mid \mathcal{F}_{k-1}] \\
&= \mathbb{E}[\cosh(w^\top u_k) \mid \mathcal{F}_{k-1}] && \Big(\because\ u_k \mid \mathcal{F}_{k-1} \overset{d}{=} -u_k \mid \mathcal{F}_{k-1}\Big) \\
&\le \mathbb{E}\left[\cosh\left(\frac{2b_k w^\top \widehat{r}_k}{R_k}\right)\middle|\mathcal{F}_{k-1}\right] && \Big(\because\ |\mathbf{1}\{k=i\} - \widetilde{B}_{ik}^{\mathrm{LS}}| \le 2|b_k|\Big) \\
&\le \mathbb{E}\left[\cosh\left(\frac{4b_k w^\top \widehat{r}_k}{n\sigma_k^2}\right)\middle|\mathcal{F}_{k-1}\right] && \big(\because\ R_k \ge n\sigma_k^2/2\big) \\
&= \mathbb{E}_Z \cosh\left(\frac{4b_k\|(I_n - P_k)w\|_2 Z}{n\sigma_k}\right) && \big(\because\ \widehat{r}_k \mid \mathcal{F}_{k-1} \sim N(0, \sigma_k^2(I_n - P_k))\big) \\
&= \exp\left\{\frac{8b_k^2\|(I_n - P_k)w\|_2^2}{n^2\sigma_k^2}\right\} && \Big(\because\ \mathbb{E}\cosh(tZ) = e^{t^2/2}\Big) \\
&\le \exp\left\{\frac{8b_k^2\|w\|_2^2}{n^2\sigma_k^2}\right\} && \big(\because\ \|(I_n - P_k)w\|_2 \le \|w\|_2\big)\,.
\end{aligned}$$

The path expansion and (58) yield

$$\sum_{k\in K} \sigma_k^2 B_{k\rightsquigarrow j}^2 = \Sigma_{jj} - \sigma_j^2, \qquad \sum_{k\in K} \frac{b_k^2}{\sigma_k^2} \le \Omega_{ii}. \tag{71}$$

Fix deterministic $w \in \mathbb{R}^n$. For $m \in K$, conditioning on $\mathcal{F}_{m-1}$ and applying (68) give

$$\begin{aligned}
&\mathbb{E}\exp\{w^\top V_m\} \\
&= \mathbb{E}\big[\exp\{w^\top V_{m-1}\}\mathbb{E}[\exp\{w^\top u_m\} \mid \mathcal{F}_{m-1}]\big] && (\because\ V_m = V_{m-1} + u_m) \\
&\le \exp\left\{\frac{8b_m^2\|w\|_2^2}{n^2\sigma_m^2}\right\}\mathbb{E}\exp\{w^\top V_{m-1}\} && (\because\ (68)).
\end{aligned}$$

Write $K = \{k_1 < \cdots < k_s\}$ and set $k_0 = 0$. Since $V_m = V_{m-1}$ for $m \notin K$, we have $V_p = V_{k_s}$ and $V_{k_r - 1} = V_{k_{r-1}}$ for $1 \le r \le s$. The successive applications are

$$\begin{aligned}
&\mathbb{E}\exp\{w^\top v_{ij}\} \\
&= \mathbb{E}\exp\{w^\top V_{k_s}\} && (\because\ v_{ij} = V_p = V_{k_s}) \\
&\le \exp\left\{\frac{8b_{k_s}^2\|w\|_2^2}{n^2\sigma_{k_s}^2}\right\}\mathbb{E}\exp\{w^\top V_{k_s-1}\} && (\because\ (68)) \\
&= \exp\left\{\frac{8b_{k_s}^2\|w\|_2^2}{n^2\sigma_{k_s}^2}\right\}\mathbb{E}\exp\{w^\top V_{k_{s-1}}\} && (\because\ V_{k_s-1} = V_{k_{s-1}}) \\
&\le \left[\prod_{r=1}^{s}\exp\left\{\frac{8b_{k_r}^2\|w\|_2^2}{n^2\sigma_{k_r}^2}\right\}\right]\mathbb{E}\exp\{w^\top V_{k_0}\} \\
&= \prod_{k\in K}\exp\left\{\frac{8b_k^2\|w\|_2^2}{n^2\sigma_k^2}\right\} && (\because\ V_{k_0} = V_0 = 0) \\
&= \exp\left\{\frac{8\|w\|_2^2}{n^2}\sum_{k\in K}\frac{b_k^2}{\sigma_k^2}\right\}
\end{aligned}$$

$$\leq \exp\left\{\frac{8\Omega_{ii}\|w\|_2^2}{n^2}\right\} \qquad \left(\because \sum_{k\in K}\frac{b_k^2}{\sigma_k^2} \leq \Omega_{ii}\right). \tag{72}$$

For an independent $g \sim N(0, I_n)$, Gaussian integration and (72) give

$$\begin{aligned}
&\mathbb{E}\exp\left\{\frac{n^2\|v_{ij}\|_2^2}{64\Omega_{ii}}\right\} \\
&= \mathbb{E}_g\mathbb{E}\exp\left\{\frac{ng^\top v_{ij}}{\sqrt{32\Omega_{ii}}}\right\} && \left(\because\ \mathbb{E}_g\exp\{g^\top x\} = \exp\{\|x\|_2^2/2\}\right) \\
&\leq \mathbb{E}_g\exp\{\|g\|_2^2/4\} && \left(\because\ (72),\ w = ng/\sqrt{32\Omega_{ii}}\right) \\
&= 2^{n/2} && \left(\because\ \mathbb{E}\exp\{\chi_n^2/4\} = (1-2/4)^{-n/2}\right).
\end{aligned} \tag{73}$$

The expectation interchange in the first equality follows from the nonnegative integrand. Similarly, the response has the exponential norm bound

$$\begin{aligned}
\mathbb{E}\exp\{\|\mathbf{X}_j\|_2^2/(4\Sigma_{jj})\} &= (1-2/4)^{-n/2} && \left(\because\ \|\mathbf{X}_j\|_2^2/\Sigma_{jj} \sim \chi_n^2\right) \\
&= 2^{n/2}.
\end{aligned} \tag{74}$$

We now apply the maximal inequality directly to V_m and M_m. For $a > 0$ with $\mathbb{E}\exp\{a\|v_{ij}\|_2^2\} < \infty$, the identity $V_m = \mathbb{E}[v_{ij} \mid \mathcal{F}_m]$ and conditional Jensen first ensure integrability:

$$\begin{aligned}
\mathbb{E}\exp\{a\|V_m\|_2^2\} &= \mathbb{E}\exp\{a\|\mathbb{E}[v_{ij} \mid \mathcal{F}_m]\|_2^2\} && \left(\because\ V_m = \mathbb{E}[v_{ij} \mid \mathcal{F}_m]\right) \\
&\leq \mathbb{E}\exp\{a\|v_{ij}\|_2^2\} && \left(\because\ \nabla^2\exp\{a\|x\|_2^2\} \succeq 0\right).
\end{aligned}$$

The submartingale property follows from the same convexity and the martingale property of V_m:

$$\begin{aligned}
&\mathbb{E}[\exp\{a\|V_{m+1}\|_2^2\} \mid \mathcal{F}_m] \\
&\geq \exp\{a\|\mathbb{E}[V_{m+1} \mid \mathcal{F}_m]\|_2^2\} && \left(\because\ \nabla^2\exp\{a\|x\|_2^2\} \succeq 0\right) \\
&= \exp\{a\|V_m\|_2^2\} && (\because\ \mathbb{E}[V_{m+1} \mid \mathcal{F}_m] = V_m).
\end{aligned}$$

Thus $\left(\exp\{a\|V_m\|_2^2\}\right)_{m=0}^p$ is a nonnegative submartingale. Taking $a = n^2/(64\Omega_{ii})$, whose integrability follows from (73), and applying Lemma B.5 with $\lambda = \exp\{n\}$ gives

$$\begin{aligned}
&\mathbb{P}\left\{\max_{0\leq m\leq p}\|V_m\|_2^2 > \frac{64\Omega_{ii}}{n}\right\} \\
&= \mathbb{P}\left\{\max_{0\leq m\leq p}\exp\left\{\frac{n^2\|V_m\|_2^2}{64\Omega_{ii}}\right\} > \exp\{n\}\right\} && \left(\because\ \frac{n^2}{64\Omega_{ii}}\frac{64\Omega_{ii}}{n} = n\right) \\
&\leq \exp\{-n\}\mathbb{E}\exp\left\{\frac{n^2\|V_p\|_2^2}{64\Omega_{ii}}\right\} && (\because\ \text{Lemma B.5},\ \lambda = \exp\{n\}) \\
&= \exp\{-n\}\mathbb{E}\exp\left\{\frac{n^2\|v_{ij}\|_2^2}{64\Omega_{ii}}\right\} && (\because\ V_p = v_{ij}) \\
&\leq \exp\{-n/2\} && (\because\ (73)).
\end{aligned} \tag{75}$$

For $M_m = \mathbb{E}[\mathbf{X}_j \mid \mathcal{F}_m]$, the tower property gives $\mathbb{E}[M_{m+1} \mid \mathcal{F}_m] = M_m$ for every $0 \leq m < p$, so $(M_m)_{m=0}^p$ is a martingale. Moreover, (30) and the topological labeling make $\mathbf{X}_j$ measurable with respect to $\mathcal{F}_j$. Hence $M_m = \mathbf{X}_j$ for $m \geq j$, in particular $M_p = \mathbf{X}_j$. The same Jensen argument, now with $a = 1/(4\Sigma_{jj})$ and (74), makes $\left(\exp\{\|M_m\|_2^2/(4\Sigma_{jj})\}\right)_{m=0}^p$ a nonnegative submartingale.

Applying Lemma B.5 again yields

$$\begin{aligned}
&\mathbb{P}\Big\{\max_{0\le m\le p}\|M_m\|_2^2 > 4n\Sigma_{jj}\Big\}\\
&= \mathbb{P}\Big\{\max_{0\le m\le p}\exp\Big\{\frac{\|M_m\|_2^2}{4\Sigma_{jj}}\Big\} > \exp\{n\}\Big\} \qquad \Big(\because\ \frac{4n\Sigma_{jj}}{4\Sigma_{jj}} = n\Big)\\
&\le \exp\{-n\}\mathbb{E}\exp\Big\{\frac{\|M_p\|_2^2}{4\Sigma_{jj}}\Big\} \qquad (\because\ \text{Lemma B.5},\ \lambda = \exp\{n\})\\
&= \exp\{-n\}\mathbb{E}\exp\Big\{\frac{\|\mathbf{X}_j\|_2^2}{4\Sigma_{jj}}\Big\} \qquad (\because\ M_p = \mathbf{X}_j)\\
&\le \exp\{-n/2\} \qquad (\because\ (74)).
\end{aligned}$$

Finally, the definition of $\mathcal{G}_{ij}$ in (69) and a union bound give

$$\begin{aligned}
\mathbb{P}(\mathcal{G}_{ij}^c) &\le \mathbb{P}\Big\{\max_m\|V_m\|_2^2 > 64\Omega_{ii}/n\Big\} + \mathbb{P}\Big\{\max_m\|M_m\|_2^2 > 4n\Sigma_{jj}\Big\}\\
&\le 2\exp\{-n/2\}.
\end{aligned}$$

□

Lemma B.12 (Uniform reconstruction)**.** Use the true-parent fits from Section B.3.1. If $n \ge C_{\mathrm{par}}(d + \log p)$, then

$$\mathbb{P}(\mathcal{E}_{\mathrm{LS}} \cap \mathcal{E}_{\mathrm{rec}}^c) \le 2p^{-6}.$$

Proof. Fix $i \ne j$ with $j \notin \mathrm{An}(i)$ and use the notation of Lemma B.11. Recall that $K = \mathrm{An}(j) \cup \{i\}$. Define

$$H = \widehat{B}_{i\rightsquigarrow j} - \sum_{a\in\mathrm{An}(j)} \widehat{B}_{ia}^{\mathrm{LS}}\widehat{B}_{a\rightsquigarrow j} - B_{ij}. \tag{76}$$

On $\mathcal{E}_{\mathrm{LS}}$, clipping and denominator floors are inactive:

$$\widetilde{B}_{ik}^{\mathrm{LS}} = \widehat{B}_{ik}^{\mathrm{LS}}, \qquad R_k = \|\widehat{r}_k\|_2^2 \qquad (k \in K).$$

For each node $\ell \in V = [p]$, set $\gamma_\ell = \mathbf{1}\{\ell = j\} + B_{\ell\rightsquigarrow j}$. Then (30) gives $\mathbf{X}_j = \sum_{\ell\in V}\gamma_\ell\mathbf{W}_\ell$: the term $\ell = j$ is $\mathbf{W}_j$, the ancestor terms are their transmitted errors, and all other terms are zero. Substituting (56) into the definition of H gives, on $\mathcal{E}_{\mathrm{LS}}$,

$$\begin{aligned}
H &= \frac{\widehat{r}_i^\top\mathbf{X}_j}{R_i} - \sum_{a\in\mathrm{An}(j)}\widetilde{B}_{ia}^{\mathrm{LS}}\frac{\widehat{r}_a^\top\mathbf{X}_j}{R_a} - B_{ij}\\
&= \left(\sum_{k\in K}u_k\right)^\top\mathbf{X}_j - B_{ij}\\
&= \sum_{k\in K}\sum_{\ell\in V}\gamma_\ell u_k^\top\mathbf{W}_\ell - B_{ij}\\
&= \underbrace{\sum_{k\in K}\gamma_k u_k^\top\mathbf{W}_k}_{\mathrm{I}} + \underbrace{\sum_{k\in K}\sum_{\substack{\ell\in V\\ \ell\ne k}}\gamma_\ell u_k^\top\mathbf{W}_\ell}_{\mathrm{II}} - B_{ij}.
\end{aligned}$$

For I, we obtain on $\mathcal{E}_{\mathrm{LS}}$

$$\begin{aligned}
\mathrm{I} &= \sum_{k\in K} \gamma_k(\mathbf{1}\{k=i\} - \widehat{B}^{\mathrm{LS}}_{ik}) \frac{\widehat{r}_k^\top \mathbf{W}_k}{R_k} && (\because\ \widetilde{B}^{\mathrm{LS}}_{ik} = \widehat{B}^{\mathrm{LS}}_{ik}) \\
&= \sum_{k\in K} \gamma_k(\mathbf{1}\{k=i\} - \widehat{B}^{\mathrm{LS}}_{ik}) && (\because\ \widehat{r}_k^\top \mathbf{W}_k = R_k) \\
&= \sum_{k\in K} B_{k\rightsquigarrow j}(\mathbf{1}\{k=i\} - \widehat{B}^{\mathrm{LS}}_{ik}) && (\because\ j \notin K \Rightarrow \gamma_k = B_{k\rightsquigarrow j}) \\
&= \sum_{k\in K} B_{k\rightsquigarrow j}(\mathbf{1}\{k=i\} - B_{ik} - h_{ik}^\top \mathbf{W}_k) && (\because\ (55)) \\
&= B_{ij} - \sum_{k\in K} B_{k\rightsquigarrow j} h_{ik}^\top \mathbf{W}_k && (\because\ (8)).
\end{aligned} \tag{77}$$

For II, partition the pairs (k,ℓ) with $k \neq \ell$ according to which index is larger:

$$\begin{aligned}
\mathrm{II} &= \sum_{k\in K} \sum_{\substack{\ell\in V \\ k<\ell}} \gamma_\ell u_k^\top \mathbf{W}_\ell + \sum_{k\in K} \sum_{\substack{\ell\in V \\ \ell<k}} \gamma_\ell u_k^\top \mathbf{W}_\ell \\
&= \sum_{\ell\in V} \gamma_\ell \left(\sum_{\substack{k\in K \\ k<\ell}} u_k \right)^\top \mathbf{W}_\ell + \sum_{k\in K} u_k^\top \left(\sum_{\substack{\ell\in V \\ \ell<k}} \gamma_\ell \mathbf{W}_\ell \right).
\end{aligned}$$

In the first term, fix ℓ and collect all earlier increments u_k. Their sum is

$$\sum_{\substack{k\in K \\ k<\ell}} u_k = V_{\ell-1} \qquad \left(\because\ V_m = \sum_{\substack{k\in K \\ k\le m}} u_k \right).$$

In the second term, fix k and collect the errors revealed before stage k. Since $\mathcal{F}_{k-1} = \sigma(\mathbf{W}_1, \ldots, \mathbf{W}_{k-1})$, independence and centering give

$$\mathbb{E}[\mathbf{W}_\ell \mid \mathcal{F}_{k-1}] = \begin{cases} \mathbf{W}_\ell, & \ell < k, \\ 0, & \ell \ge k. \end{cases}$$

Consequently,

$$\begin{aligned}
\sum_{\substack{\ell\in V \\ \ell<k}} \gamma_\ell \mathbf{W}_\ell &= \mathbb{E}\left[\sum_{\ell\in V} \gamma_\ell \mathbf{W}_\ell \,\middle|\, \mathcal{F}_{k-1} \right] \\
&= \mathbb{E}[\mathbf{X}_j \mid \mathcal{F}_{k-1}] && \left(\because\ \mathbf{X}_j = \sum_{\ell\in V} \gamma_\ell \mathbf{W}_\ell \right) \\
&= M_{k-1}.
\end{aligned}$$

Substituting these two sums back into II gives

$$\begin{aligned}
\mathrm{II} &= \sum_{\ell\in V} \gamma_\ell V_{\ell-1}^\top \mathbf{W}_\ell + \sum_{k\in K} u_k^\top M_{k-1} \\
&= V_{j-1}^\top \mathbf{W}_j + \sum_{\ell\in \mathrm{An}(j)} B_{\ell\rightsquigarrow j} V_{\ell-1}^\top \mathbf{W}_\ell + \sum_{k\in K} u_k^\top M_{k-1} && (\because\ \gamma_\ell = \mathbf{1}\{\ell=j\} + B_{\ell\rightsquigarrow j}) \\
&= V_{j-1}^\top \mathbf{W}_j + \sum_{k\in K} B_{k\rightsquigarrow j} V_{k-1}^\top \mathbf{W}_k + \sum_{k\in K} u_k^\top M_{k-1} && (\because\ B_{k\rightsquigarrow j} = 0 \text{ for } k \notin \mathrm{An}(j)).
\end{aligned}$$

Combining $H = \mathrm{I} + \mathrm{II} - B_{ij}$ with (77) yields, on $\mathcal{E}_{\mathrm{LS}}$,

$$H = V_{j-1}^\top \mathbf{W}_j + \sum_{k\in K} \left\{ B_{k\rightsquigarrow j} \mathbf{W}_k^\top (V_{k-1} - h_{ik}) + u_k^\top M_{k-1} \right\}. \tag{78}$$

We next bound the reconstruction error H by an exponential-moment argument. To use conditional expectations without conditioning on the full event $\mathcal{E}_{\mathrm{LS}}$, put

$$h_{ik}^{\circ} = \mathbf{1}_{\mathcal{E}_k} h_{ik}.$$

Both $\mathcal{E}_k$ and h_{ik} depend only on the parent design, so h_{ik}° is $\mathcal{F}_{k-1}$-measurable. Moreover, (63) gives $\|h_{ik}^{\circ}\|_2 \le 2\sqrt{\Omega_{ii}/n}$, and $h_{ik}^{\circ} = h_{ik}$ on $\mathcal{E}_{\mathrm{LS}}$. Define, for $k = 1, \dots, p$,

$$Z_k = \begin{cases} V_{j-1}^{\top} \mathbf{W}_j, & k = j, \\ B_{k \rightsquigarrow j} \mathbf{W}_k^{\top} (V_{k-1} - h_{ik}^{\circ}) + u_k^{\top} M_{k-1}, & k \in K, \\ 0, & k \notin K \cup \{j\}. \end{cases}$$

These cases are disjoint because $j \notin K$. For $k \in K$, $\mathcal{F}_{k-1}$-measurability of $V_{k-1}, h_{ik}^{\circ}, M_{k-1}$ gives

$$\begin{aligned} \mathbb{E}[Z_k \mid \mathcal{F}_{k-1}] &= B_{k \rightsquigarrow j} (V_{k-1} - h_{ik}^{\circ})^{\top} \mathbb{E}[\mathbf{W}_k \mid \mathcal{F}_{k-1}] + M_{k-1}^{\top} \mathbb{E}[u_k \mid \mathcal{F}_{k-1}] \\ &= 0. \end{aligned}$$

The same conditional mean is zero for $k = j$, since $\mathbb{E}[\mathbf{W}_j \mid \mathcal{F}_{j-1}] = 0$. Thus $(\sum_{k=1}^{m} Z_k)_{m=0}^{p}$ is a martingale. On $\mathcal{E}_{\mathrm{LS}}$, (78) gives $\sum_{k=1}^{p} Z_k = H$.

For $t \in \mathbb{R}$, the two terms of Z_k have the following conditional exponential bounds:

$$\begin{aligned} &\mathbb{E}\big[\exp\{2t B_{k \rightsquigarrow j} \mathbf{W}_k^{\top} (V_{k-1} - h_{ik}^{\circ})\} \mid \mathcal{F}_{k-1}\big] \\ &\quad = \exp\{2t^2 \sigma_k^2 B_{k \rightsquigarrow j}^2 \|V_{k-1} - h_{ik}^{\circ}\|_2^2\} \qquad (\because\ \mathbf{W}_k \mid \mathcal{F}_{k-1} \sim N(0, \sigma_k^2 I_n)), \\ &\mathbb{E}[\exp\{2t u_k^{\top} M_{k-1}\} \mid \mathcal{F}_{k-1}] \\ &\quad \le \exp\left\{ \frac{32 t^2 b_k^2 \|M_{k-1}\|_2^2}{n^2 \sigma_k^2} \right\} \qquad (\because\ (68),\ w = 2tM_{k-1}). \end{aligned}$$

Conditional Cauchy–Schwarz therefore yields, for $k \in K$,

$$\begin{aligned} &\mathbb{E}[\exp\{tZ_k\} \mid \mathcal{F}_{k-1}] \\ &\le \Big(\mathbb{E}[\exp\{2t B_{k \rightsquigarrow j} \mathbf{W}_k^{\top} (V_{k-1} - h_{ik}^{\circ})\} \mid \mathcal{F}_{k-1}] \Big)^{1/2} \Big(\mathbb{E}[\exp\{2t u_k^{\top} M_{k-1}\} \mid \mathcal{F}_{k-1}] \Big)^{1/2} \\ &\le \exp\left\{ t^2 \left(\sigma_k^2 B_{k \rightsquigarrow j}^2 \|V_{k-1} - h_{ik}^{\circ}\|_2^2 + \frac{16 b_k^2 \|M_{k-1}\|_2^2}{n^2 \sigma_k^2} \right) \right\}. \end{aligned}$$

Accordingly, define the $\mathcal{F}_{k-1}$-measurable quantities

$$v_k = \begin{cases} \sigma_j^2 \|V_{j-1}\|_2^2, & k = j, \\ \sigma_k^2 B_{k \rightsquigarrow j}^2 \|V_{k-1} - h_{ik}^{\circ}\|_2^2 + \dfrac{16 b_k^2 \|M_{k-1}\|_2^2}{n^2 \sigma_k^2}, & k \in K, \\ 0, & k \notin K \cup \{j\}, \end{cases} \qquad Q = \sum_{k=1}^{p} v_k.$$

For each fixed $t \in \mathbb{R}$, we will show that

$$\left(\exp\left\{ t \sum_{k=1}^{m} Z_k - t^2 \sum_{k=1}^{m} v_k \right\} \right)_{m=0}^{p}$$

is a nonnegative supermartingale with respect to $(\mathcal{F}_m)_{m=0}^{p}$, starting at one. This will bound its terminal expectation by one.

At node j, V_{j-1} is $\mathcal{F}_{j-1}$-measurable and $\mathbf{W}_j \sim N(0, \sigma_j^2 I_n)$ is independent of $\mathcal{F}_{j-1}$. Hence

$$Z_j \mid \mathcal{F}_{j-1} \sim N\big(0, \sigma_j^2 \|V_{j-1}\|_2^2\big) = N(0, v_j).$$

The Gaussian moment-generating function therefore gives

$$\mathbb{E}[\exp\{tZ_j\} \mid \mathcal{F}_{j-1}] = \exp\{t^2 v_j / 2\} \le \exp\{t^2 v_j\}.$$

Thus $\mathbb{E}[\exp\{tZ_k - t^2 v_k\} \mid \mathcal{F}_{k-1}] \le 1$ for every k. For $m = 1, \dots, p$, it follows that

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{t\sum_{k=1}^{m} Z_k - t^2 \sum_{k=1}^{m} v_k\right\}\middle| \mathcal{F}_{m-1}\right] \\
&= \exp\left\{t\sum_{k=1}^{m-1} Z_k - t^2 \sum_{k=1}^{m-1} v_k\right\} \mathbb{E}[\exp\{tZ_m - t^2 v_m\} \mid \mathcal{F}_{m-1}] \\
&\le \exp\left\{t\sum_{k=1}^{m-1} Z_k - t^2 \sum_{k=1}^{m-1} v_k\right\}.
\end{aligned}$$

This proves the nonnegative supermartingale property. Taking expectations in the preceding inequality removes the last increment from both sums. Repeating this step for $m = p, p-1, \dots, 1$, with both sums empty at $m = 0$, gives

$$\begin{aligned}
\mathbb{E}[\mathbf{1}_{\mathcal{E}_{\mathrm{LS}}} \exp\{tH - t^2 Q\}] &= \mathbb{E}\left[\mathbf{1}_{\mathcal{E}_{\mathrm{LS}}} \exp\left\{t\sum_{k=1}^{p} Z_k - t^2 Q\right\}\right] \\
&\le \mathbb{E}\exp\left\{t\sum_{k=1}^{p} Z_k - t^2 \sum_{k=1}^{p} v_k\right\} \\
&\le \mathbb{E}\exp\left\{t\sum_{k=1}^{p-1} Z_k - t^2 \sum_{k=1}^{p-1} v_k\right\} \\
&\le \mathbb{E}\exp\left\{t\sum_{k=1}^{0} Z_k - t^2 \sum_{k=1}^{0} v_k\right\} \\
&= 1 \qquad (t \in \mathbb{R}).
\end{aligned} \tag{79}$$

We now bound Q on $\mathcal{G}_{ij}$. In particular, (69) gives

$$\|V_{j-1}\|_2^2 \le \max_{0 \le m \le p} \|V_m\|_2^2 \le \frac{64\Omega_{ii}}{n}.$$

By (69) and (63),

$$\|V_{k-1} - h_{ik}^{\circ}\|_2 \le \|V_{k-1}\|_2 + \|h_{ik}^{\circ}\|_2 \le 8\sqrt{\Omega_{ii}/n} + 2\sqrt{\Omega_{ii}/n} = 10\sqrt{\Omega_{ii}/n}.$$

Using also $\|M_{k-1}\|_2^2 \le 4n\Sigma_{jj}$ from (69), we obtain

$$\begin{aligned}
Q = \sum_{k=1}^{p} v_k &= \sigma_j^2 \|V_{j-1}\|_2^2 + \sum_{k \in K} \left\{\sigma_k^2 B_{k \rightsquigarrow j}^2 \|V_{k-1} - h_{ik}^{\circ}\|_2^2 + \frac{16 b_k^2 \|M_{k-1}\|_2^2}{n^2 \sigma_k^2}\right\} \\
&\le \frac{64\Omega_{ii}\sigma_j^2}{n} + \frac{100\Omega_{ii}}{n} \sum_{k \in K} \sigma_k^2 B_{k \rightsquigarrow j}^2 + \frac{64\Sigma_{jj}}{n} \sum_{k \in K} \frac{b_k^2}{\sigma_k^2} \\
&\le \frac{\Omega_{ii}}{n}\{64\sigma_j^2 + 100(\Sigma_{jj} - \sigma_j^2) + 64\Sigma_{jj}\} \qquad (\because \text{(71)}) \\
&= \frac{\Omega_{ii}}{n}(164\Sigma_{jj} - 36\sigma_j^2) \\
&\le \frac{164\Omega_{ii}\Sigma_{jj}}{n}.
\end{aligned} \tag{80}$$

For $u > 0$ and $t > 0$, on $\mathcal{G}_{ij} \cap \{H > u\}$ we have

$$tH - t^2 Q > tu - \frac{164 t^2 \Omega_{ii}\Sigma_{jj}}{n}.$$

Exponential Markov and (79) therefore give

$$\begin{aligned}
&\mathbb{P}(\mathcal{E}_{\mathrm{LS}} \cap \mathcal{G}_{ij} \cap \{H > u\}) \\
&\le \mathbb{P}\left\{ \mathbf{1}_{\mathcal{E}_{\mathrm{LS}}} \exp\{tH - t^2 Q\} > \exp\left\{ tu - \frac{164 t^2 \Omega_{ii}\Sigma_{jj}}{n} \right\} \right\} \\
&\le \exp\left\{ -tu + \frac{164 t^2 \Omega_{ii}\Sigma_{jj}}{n} \right\} \mathbb{E}[\mathbf{1}_{\mathcal{E}_{\mathrm{LS}}} \exp\{tH - t^2 Q\}] \\
&\le \exp\left\{ -tu + \frac{164 t^2 \Omega_{ii}\Sigma_{jj}}{n} \right\} \qquad (\because \text{ (79)}) \\
&= \exp\left\{ -\frac{nu^2}{656 \Omega_{ii}\Sigma_{jj}} \right\} \qquad \left(\text{substituting } t = \frac{nu}{328 \Omega_{ii}\Sigma_{jj}} \right).
\end{aligned}$$

Applying the same argument to $-H$ uses (79) with $-t$ and gives the same bound. Adding the two tails and the probability of $\mathcal{G}_{ij}^c$ yields

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_{\mathrm{LS}} \cap \{|H| > u\}) &\le \mathbb{P}(\mathcal{G}_{ij}^c) + \mathbb{P}(\mathcal{E}_{\mathrm{LS}} \cap \mathcal{G}_{ij} \cap \{|H| > u\}) \\
&\le 2\exp\{-n/2\} + 2\exp\left\{ -\frac{nu^2}{656 \Omega_{ii}\Sigma_{jj}} \right\}. \qquad (81)
\end{aligned}$$

Finally, take $u = C_{\mathrm{par}} \sqrt{\Omega_{ii}\Sigma_{jj} \log p / n}$. Enlarge C_{par} so that $n \ge 18 \log p$ and $C_{\mathrm{par}}^2 \ge 5904 = 9 \cdot 656$. For $p \ge 2$, the two terms in (81) satisfy

$$2\exp\{-n/2\} + 2\exp\{-C_{\mathrm{par}}^2 \log p / 656\} \le 4p^{-9} \le 2p^{-8}.$$

The preceding bound holds for each fixed pair $i \ne j$ with $j \notin \mathrm{An}(i)$. Write H_{ij} for the error H in (76) to make its dependence on the pair explicit. By (61), $\mathcal{E}_{\mathrm{rec}}$ requires every such error to be below the common threshold. Its complement is therefore

$$\mathcal{E}_{\mathrm{rec}}^c = \bigcup_{\substack{i \ne j \\ j \notin \mathrm{An}(i)}} \left\{ |H_{ij}| > C_{\mathrm{par}} \sqrt{\frac{\Sigma_{\max}\Omega_{\max} \log p}{n}} \right\}.$$

For each pair, this threshold is at least the value of u used above, so

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{E}_{\mathrm{LS}} \cap \left\{ |H_{ij}| > C_{\mathrm{par}} \sqrt{\frac{\Sigma_{\max}\Omega_{\max} \log p}{n}} \right\} \right) \\
&\le \mathbb{P}\left(\mathcal{E}_{\mathrm{LS}} \cap \left\{ |H_{ij}| > C_{\mathrm{par}} \sqrt{\frac{\Sigma_{jj}\Omega_{ii} \log p}{n}} \right\} \right) \qquad (\because \ \Sigma_{jj}\Omega_{ii} \le \Sigma_{\max}\Omega_{\max}) \\
&\le 2p^{-8}.
\end{aligned}$$

A union bound over at most p^2 possible pairs (i, j) now gives

$$\mathbb{P}(\mathcal{E}_{\mathrm{LS}} \cap \mathcal{E}_{\mathrm{rec}}^c) \le \sum_{\substack{i \ne j \\ j \notin \mathrm{An}(i)}} 2p^{-8} \le p^2 \cdot 2p^{-8} = 2p^{-6}. \qquad (82)$$

□

B.3.3 Completion of the proof

We first obtain a single event on which all oracle reconstructions are accurate. We then show that, on this event, the algorithm uses exactly these reconstructions and makes every parent decision correctly.

Since $\sigma_k^2 \le \Sigma_{kk} \le \Sigma_{\max}$ and $(\Sigma^{-1}_{\mathrm{Pa}(k),\mathrm{Pa}(k)})_{ii} \le \Omega_{ii} \le \Omega_{\max}$, the sample and signal conditions in (12) imply (62). Splitting according to whether $\mathcal{E}_{\mathrm{LS}}$ holds, Lemmas B.10 and B.12 give

$$\mathbb{P}(\mathcal{E}_{\mathrm{rec}}^c) \le \mathbb{P}(\mathcal{E}_{\mathrm{LS}}^c) + \mathbb{P}(\mathcal{E}_{\mathrm{LS}} \cap \mathcal{E}_{\mathrm{rec}}^c) \le 5p^{-6} + 2p^{-6} = 7p^{-6} \le p^{-2}. \qquad (83)$$

Choose the threshold multiplier and then the signal factor in (12) so that

$$C_{\mathrm{par}}\sqrt{\frac{\Sigma_{\max}\Omega_{\max}\log p}{n}} \le (1-\delta)\min_j \tau_j, \qquad 2\max_j \tau_j < B_{\min}. \tag{84}$$

Fix a sample in $\mathcal{E}_{\mathrm{rec}}$ and any ordering $\widehat{\pi}$ compatible with G. We use induction over the targets $j = \widehat{\pi}_r$. Compatibility implies $\mathrm{Pa}(\widehat{\pi}_1) = \emptyset$, so the initialization $\widehat{T}_{\widehat{\pi}_1} = \emptyset$ is correct.

Suppose now that every target a preceding j has been recovered correctly. Then

$$\widehat{T}_a = \mathrm{Pa}(a) \quad \Longrightarrow \quad \mathbf{X}_{\widehat{T}_a}^{\dagger}\mathbf{X}_a = \mathbf{X}_{\mathrm{Pa}(a)}^{\dagger}\mathbf{X}_a.$$

Thus the OLS coefficients and residuals in (7) coincide with the true-parent fits in (54) used to define $\mathcal{E}_{\mathrm{rec}}$.

For this target j, the algorithm processes earlier nodes in reverse order. After m steps, we claim that

$$\widehat{T}_j^{(m)} = \mathrm{Pa}(j) \cap \{\widehat{\pi}_{r-m}, \dots, \widehat{\pi}_{r-1}\},$$
$$\widehat{Q}_j^{(m)} = \mathrm{An}(j) \cap \{\widehat{\pi}_{r-m}, \dots, \widehat{\pi}_{r-1}\}.$$

For $m = 0$, the processed set is empty and both identities hold by initialization. Suppose they hold after $m-1$ steps, and let $k = \widehat{\pi}_{r-m}$ be the next earlier node to be processed.

We first show that replacing $\widehat{Q}_j^{(m-1)}$ by $\mathrm{An}(j)$ does not change the sum in (8). By the induction hypothesis,

$$\widehat{Q}_j^{(m-1)} = \mathrm{An}(j) \cap \{\widehat{\pi}_{r-m+1}, \dots, \widehat{\pi}_{r-1}\} \subseteq \mathrm{An}(j).$$

Thus every node included in the reconstruction sum is an ancestor of j. Any ancestor missing from this set has not yet been processed. Since the current node is $k = \widehat{\pi}_{r-m}$, this gives

$$\begin{aligned} a \in \mathrm{An}(j) \setminus \widehat{Q}_j^{(m-1)} &\implies \widehat{\pi}^{-1}(a) \le r - m = \widehat{\pi}^{-1}(k) \\ &\implies k \notin \mathrm{Pa}(a) \qquad (\because\ k \in \mathrm{Pa}(a) \Rightarrow \widehat{\pi}^{-1}(k) < \widehat{\pi}^{-1}(a)) \\ &\implies \widehat{B}_{ka}^{\mathrm{LS}} = 0 \qquad (\because\ (54)). \end{aligned}$$

Consequently, every additional term obtained by extending the sum to all ancestors is zero:

$$\begin{aligned} \sum_{a\in\mathrm{An}(j)} \widehat{B}_{ka}^{\mathrm{LS}}\widehat{B}_{a\rightsquigarrow j} &= \sum_{a\in\widehat{Q}_j^{(m-1)}} \widehat{B}_{ka}^{\mathrm{LS}}\widehat{B}_{a\rightsquigarrow j} + \sum_{a\in\mathrm{An}(j)\setminus\widehat{Q}_j^{(m-1)}} \widehat{B}_{ka}^{\mathrm{LS}}\widehat{B}_{a\rightsquigarrow j} \\ &= \sum_{a\in\widehat{Q}_j^{(m-1)}} \widehat{B}_{ka}^{\mathrm{LS}}\widehat{B}_{a\rightsquigarrow j}. \end{aligned}$$

Substituting this equality into (8) gives

$$\widehat{B}_{kj}^{\mathrm{res}} = \widehat{B}_{k\rightsquigarrow j} - \sum_{a\in\widehat{Q}_j^{(m-1)}} \widehat{B}_{ka}^{\mathrm{LS}}\widehat{B}_{a\rightsquigarrow j} = \widehat{B}_{k\rightsquigarrow j} - \sum_{a\in\mathrm{An}(j)} \widehat{B}_{ka}^{\mathrm{LS}}\widehat{B}_{a\rightsquigarrow j}. \tag{85}$$

Since k precedes j in a compatible ordering, $j \notin \mathrm{An}(k)$. This pair is therefore covered by (61). The first inequality in (84) leaves room for the correction of size at most $\delta\tau_j$ in (9). Combining the reconstruction error bound with this correction gives

$$\begin{aligned} |\widetilde{B}_{kj} - B_{kj}| &\le |\widehat{B}_{kj}^{\mathrm{res}} - B_{kj}| + \delta\tau_j && (\because\ (9)) \\ &\le C_{\mathrm{par}}\sqrt{\frac{\Sigma_{\max}\Omega_{\max}\log p}{n}} + \delta\tau_j && (\because\ \mathcal{E}_{\mathrm{rec}}) \\ &\le (1-\delta)\tau_j + \delta\tau_j && (\because\ (84)) \\ &= \tau_j. \end{aligned} \tag{86}$$

The second inequality in (84), $B_{\min} > 2\max_j \tau_j$, now separates nonzero coefficients from zero. Together with (86), it gives both directions of the threshold decision:

$$\begin{aligned} B_{kj} = 0 &\implies |\widetilde{B}_{kj}| \le \tau_j, \\ B_{kj} \ne 0 &\implies |\widetilde{B}_{kj}| \ge |B_{kj}| - \tau_j \ge B_{\min} - \tau_j > \tau_j. \end{aligned}$$

Hence k is selected exactly when it is a parent of j, preserving the claimed identity for $\widehat{T}_j^{(m)}$.

It remains to verify the induction identity for $\widehat{Q}_j^{(m)}$. Only the current node $k = \widehat{\pi}_{r-m}$ can be added at this step, so it suffices to show that the update adds k exactly when $k \in \mathrm{An}(j)$.

If $k \in \mathrm{Pa}(j)$, the correctness of the parent decision gives $k \in \widehat{T}_j^{(m)}$, and (10) adds k. If $k \in \mathrm{An}(j) \setminus \mathrm{Pa}(j)$, choose a directed path $k \to a \to \cdots \to j$ of length at least two. Compatibility of the ordering gives

$$r - m = \widehat{\pi}^{-1}(k) < \widehat{\pi}^{-1}(a) < \widehat{\pi}^{-1}(j) = r.$$

Thus a is one of the earlier nodes already processed in reverse order. Since $a \in \mathrm{An}(j)$, the induction hypothesis implies

$$a \in \mathrm{An}(j) \cap \{\widehat{\pi}_{r-m+1}, \ldots, \widehat{\pi}_{r-1}\} = \widehat{Q}_j^{(m-1)}.$$

Moreover, $k \to a$ means $k \in \mathrm{Pa}(a) = \widehat{T}_a$, where the equality follows from correctness of the earlier target a. Hence the second condition in (10) also adds k.

Conversely, if the update adds k, then either $k \in \widehat{T}_j^{(m)}$ or $k \in \widehat{T}_a$ for some $a \in \widehat{Q}_j^{(m-1)}$. The first case implies $k \in \mathrm{Pa}(j) \subseteq \mathrm{An}(j)$ by the correctness of the parent decision. In the second case, the induction hypotheses give $k \in \mathrm{Pa}(a)$ and $a \in \mathrm{An}(j)$, so concatenating $k \to a$ with a path from a to j gives $k \in \mathrm{An}(j)$. We have therefore established

$$k \in \mathrm{An}(j) \quad \Longleftrightarrow \quad k \in \widehat{T}_j^{(m)} \text{ or } k \in \bigcup_{a \in \widehat{Q}_j^{(m-1)}} \widehat{T}_a. \tag{87}$$

Equations (10) and (87) give

$$\widehat{Q}_j^{(m)} = \widehat{Q}_j^{(m-1)} \cup \big(\{k\} \cap \mathrm{An}(j)\big) = \mathrm{An}(j) \cap \{\widehat{\pi}_{r-m}, \ldots, \widehat{\pi}_{r-1}\},$$

completing the induction over m. At $m = r - 1$, the parent-set identity gives $\widehat{T}_j = \mathrm{Pa}(j)$, so the OLS fit for j uses its true parents. Induction over $r = 1, \ldots, p$ therefore recovers all parent sets and hence G.

The event $\mathcal{E}_{\mathrm{rec}}$ covers all pairs in (61) and does not depend on the chosen compatible ordering. The preceding deterministic argument therefore also applies to a data-dependent ordering whenever it is compatible. In particular,

$$\{\widehat{G} \neq G\} \subseteq \{\widehat{\pi} \text{ is incompatible with } G\} \cup \mathcal{E}_{\mathrm{rec}}^c,$$

and hence

$$\begin{aligned} \mathbb{P}\{\widehat{G} \neq G\} &\leq \mathbb{P}\{\widehat{\pi} \text{ is incompatible with } G\} + \mathbb{P}(\mathcal{E}_{\mathrm{rec}}^c) \\ &\leq \mathbb{P}\{\widehat{\pi} \text{ is incompatible with } G\} + p^{-2} \qquad (\because \text{ (83)}). \end{aligned}$$

B.4 Proof of Theorem 4.3

Proof. The first two terms in (13) give (11). They also imply $n \gtrsim d + \log p$, because $\Sigma_{\max}/\lambda_{\min}(\Sigma) \geq 1$ and $p \geq 2$. For a nonempty graph, the last term of (13) is equivalent, up to a fixed factor, to

$$B_{\min}^2 \gtrsim \frac{\Sigma_{\max}\Omega_{\max}\log p}{n},$$

which is the signal condition in (12). For an empty graph this condition is vacuous. Theorems 4.1 and 4.2, with their stated penalties and thresholds, therefore give

$$\mathbb{P}\{\widehat{G} \neq G\} \leq \mathbb{P}\{\widehat{\pi} \text{ is incompatible with } G\} + p^{-2} \leq 2p^{-2} \leq p^{-1}.$$

□

B.5 Proof of Proposition 4.4

Proof. Fix the ordering penalty and parent thresholds. We use dense cyclic coordinate descent for Lasso (Friedman et al., 2010; 2007) and singular value decomposition (SVD) for local OLS, suppressing solver iteration counts and numerical tolerance factors as in Section 4.1. First form the full-sample and two fitting-fold cross-product matrices in $O(np^2)$ work.

Ordering fits and coefficient updates. At a prefix of size s, a coordinate update uses the stored cross-products and costs $O(s)$. A sweep over the s coordinates costs $O(s^2)$, giving the same cost per fit under the stated convention. Evaluating a fit on the other fold also costs $O(s^2)$, since the residual sum of squares is a quadratic form computed from the held-out cross-products. There are $p - s$ candidate nodes and two fitting folds, so summing over prefixes gives

$$O\left(\sum_{s=1}^{p-1} 2(p-s)s^2\right) = O(p^4).$$

For the selected node, the coordinate refits use the same stored cross-products and the original averaged coefficients. Together they require $O(s^2)$ work, or $O(p^3)$ over all nodes. Scanning the scores and resolving ties adds $O(p^2)$.

Local regressions and total effects. A local design with $s \le p - 1$ columns has a minimum-norm OLS solution computable by SVD in $O(ns \min\{n, s\})$ work (Golub & Van Loan, 2013). Forming the residual adds $O(ns + n)$, also accounting for empty designs. Each node requires only one retained local fit, so the total is at most $O(np^3)$. For node k with a nonzero residual, the row vector $\widehat{r}_k^\top \mathbf{X}/\|\widehat{r}_k\|_2^2$ supplies its total-effect estimates to all other nodes. This multiplication costs $O(np)$, hence $O(np^2)$ over all nodes. A zero residual gives a zero row of estimates.

Processing earlier nodes in reverse order. For each target j, store the current reconstruction for every unprocessed earlier node and a flag recording whether it is a parent of a retained node. Start with the total-effect estimates as the reconstructions and set all flags to zero. Processing an earlier node k takes constant work for clipping and thresholding. If k is retained, subtract $\widehat{B}^{\mathrm{LS}}_{uk}\widehat{B}_{k\rightsquigarrow j}$ from the reconstruction for each unprocessed $u \in \widehat{T}_k$ and set its flag. Local coefficients outside $\widehat{T}_k$ are zero, so these updates implement (8) and (10). With s earlier nodes, at most s retained nodes each update at most s entries. Thus the cost is $O(s^2)$ per target and $O(p^3)$ over all target nodes.

Combining the stages gives $O(np^3 + p^4)$ under the stated convention. □

C Practical Data-Driven Tuning

HOST-CV/EBIC selects a common ordering penalty by pooled nodewise two-fold cross-validation (CV), then selects a parent threshold for each node by the extended Bayesian information criterion (EBIC). The candidate grids $\mathcal{C}_{\mathrm{ord}}$ and $\mathcal{C}_{\mathrm{par}}$ are given in Appendices D.3 and E.

C.1 Ordering penalty selection

For each $c \in \mathcal{C}_{\mathrm{ord}}$, use the same penalty $\lambda_{\mathrm{ord}}(c) = c\sqrt{\log p/\lfloor n/2 \rfloor}$ in both folds. Let $\mathbf{Z}^{(\ell)}$ denote the column-centered data in fold I_ℓ, and write $\bar{\ell}$ for the other fold index. Fit a Lasso regression for each node using all other nodes in that fold.

Let $\mathrm{RSS}_j^{(\ell)}(c)$ denote the residual sum of squares on the other fold $I_{\bar{\ell}}$, evaluated using the coefficients and column means estimated from I_ℓ. Normalize each held-out RSS by the centered response sum of squares, then average over nodes and both folds to obtain

$$\mathrm{CV}^{(\ell)}(c) = \frac{1}{p}\sum_{j=1}^{p} \frac{\mathrm{RSS}_j^{(\ell)}(c)}{\|\mathbf{Z}_j^{(\bar{\ell})}\|_2^2}, \qquad \mathrm{CV}(c) = \frac{\mathrm{CV}^{(1)}(c) + \mathrm{CV}^{(2)}(c)}{2}.$$

Select the multiplier minimizing the pooled score to obtain one common ordering penalty.

C.2 Parent threshold selection

Recover parent sets sequentially along the estimated ordering. For node j, let ν_j be the number of earlier nodes. All full-sample covariance and EBIC calculations use the sample size n. Each multiplier $c \in \mathcal{C}_{\mathrm{par}}$ gives the threshold $\tau_j(c) = c\sqrt{\log p/n}$. Apply the reverse processing step of Algorithm 1 at this threshold to obtain a candidate parent set $\widehat{T}_j(c)$. The fitted coefficients and

residuals of earlier nodes are shared across candidates, while the retained sets $\widehat{Q}_j^{(m)}$ are recomputed for each threshold. When $\nu_j = 0$, the candidate parent set is empty.

Evaluate each distinct candidate set T by full-sample OLS after centering the columns of $\mathbf{X}$, with residual sum of squares

$$\mathrm{RSS}_j(T) = \min_{b \in \mathbb{R}^{|T|}} \|\mathbf{X}_j - \mathbf{X}_T b\|_2^2.$$

The implementation uses the nodewise score

$$\mathrm{EBIC}_j(T) = n \log\{\mathrm{RSS}_j(T)/n\} + |T| \log n + 4|T| \log \max\{\nu_j, 1\}. \tag{88}$$

The first two terms measure regression fit and model size. The final term adapts the graphical EBIC penalty with $\gamma = 1$ (Foygel & Drton, 2010) to the ν_j earlier nodes, following the EBIC construction of Chen & Chen (2008). EBIC is evaluated only for candidate sets T with a full-column-rank centered design, $\mathrm{rank}(\mathbf{X}_T) = |T|$, positive residual degrees of freedom, $n - |T| - 1 > 0$, and a positive residual sum of squares, $\mathrm{RSS}_j(T) > 0$. Otherwise, the EBIC score is set to $+\infty$.

For each node j, select the multiplier $\widehat{c}_j$ with the smallest finite EBIC score, breaking ties by fewer parents and then a larger multiplier.

D Simulation Protocol and Results

This section gives the data-generation procedures, algorithm settings, and evaluation rules for the main experiments, followed by comparisons across graph families and separate evaluations of ordering and parent recovery. The score diagnostic in Figure 1 is described separately below.

D.1 Score diagnostic in Figure 1

The graph has layer widths $64, 32, 16, 8, 4, 2, 1$, with each pair of nodes feeding one node in the next layer. All nodes have independent Gaussian errors with variance 0.75. The direct effects across the six successive layer transitions are $(0.4, 0.44, 0.48, 0.52, 0.56, 0.6)$, with all effects positive. The designated remaining source is the first node in the second layer, and the remaining nonsource is the single node in the final layer. Their regressions on the 64 first-layer nodes have two and 64 nonzero coefficients, respectively. Each regression coefficient for the node in the final layer is the product of the six direct effects, approximately 0.015. The population ordering scores are 0.75 for the designated node in the second layer and approximately 1.929 for the node in the final layer.

We use 1,000 repetitions for each n = 20, 30, 40, 50, fitting zero-intercept Lasso to raw observations with standardized predictors and fixed $\lambda = 0.1$. In-sample fitting and scoring use all observations. One-way hold-out fitting and scoring use a random equal split. Scores are mean squared residuals. We summarize the scores by means plus or minus one sample standard deviation across repetitions, computed before the logarithmic axis transformation, and estimate the pairwise ordering probability with 95% Wilson intervals.

D.2 Data generation

The graph-recovery and runtime experiments in Figures 2–3 use the Gaussian linear SEM in (1) and a common random DAG generator. Start with an auxiliary ordering $1, \ldots, p$. For each node j, draw its parent count uniformly from $\{0, \ldots, \min(d, j-1)\}$, then sample that many parents uniformly without replacement from $\{1, \ldots, j-1\}$. The parameter d specifies the maximum indegree allowed by the generator. Nonzero direct effects have independent magnitudes uniform on $[0.4, 0.6]$ and independent symmetric signs. Structural errors are independent Gaussian variables with variance 0.75. An independent uniform permutation of the node labels shuffles the variable ordering. The additional graph families in Appendix D.5 use the same coefficient and error distributions with different graph-generation procedures. Each generated dataset is column-centered once before fitting.

D.3 Graph Recovery Experiments

Figure 2 evaluates HOST over all combinations of p = 30, 50, 100, 200, d = 2, 3, 4, 5, and n = 128, 256, 512, 1024, 2048, with 200 repetitions per setting. It reports graph recovery rate, the fraction of exactly recovered graphs.

Figure 3(a)–(b) compares methods at p = 50, 200, d = 3, and n = 100, 200, 300, 400, 500, with 50 paired repetitions per setting. Methods receive the same dataset in each repetition. Daskalakis Alg. 1 is evaluated only at p = 50 due to the memory required to cache intermediate regression coefficients at larger dimensions. All graph-recovery experiments use the CPU implementations and tuning settings below. The implementations used for runtime measurement are specified in Appendix D.4.

HOST and HOST-CV/EBIC. Both variants use our Cython/C++ implementation with $\delta = 0.8$ and the same two-fold split for each dataset. HOST uses $\lambda_{\mathrm{ord}} = 0.5\sqrt{\log p/\lfloor n/2 \rfloor}$ and $\tau_j = c_{\mathrm{par}}\sqrt{\log p/n}$. The parent-threshold multiplier is $c_{\mathrm{par}} = 3$ in Figure 2 and $c_{\mathrm{par}} = 2$ in the recovery and runtime comparisons, including the additional graph families. HOST-CV/EBIC selects both tuning parameters using the rules in Appendix C. The pooled two-fold CV ordering grid is $\mathcal{C}_{\mathrm{ord}} = \{10^{-1+0.1k} : k = 0, \ldots, 20\}$. Parent recovery uses the full sample and the grid $\mathcal{C}_{\mathrm{par}} = \{\frac{1}{4}32^{k/32} : k = 0, \ldots, 32\}$, which contains 33 multipliers from 0.25 to 8.

HLSM. Our Cython/C++ implementation of Park et al. (2021) constructs the ordering backwards using OLS residual variances on Lasso-selected supports and selects parents from these supports, with standardized predictors and penalty $\lambda = 3\sqrt{\log p/n}$.

TD-BSS. Our Cython/C++ implementation of the subset-based procedure of Gao et al. (2022) uses the supplied indegree bound $q = d$ in ordering, with d given by an oracle. Parent selection compares the residual-variance increase on removing a candidate parent against $\eta = 3 \log p/n$.

TD-BIC. Our Cython/C++ implementation of TD-BIC proposed by Gao et al. (2022) uses the same ordering search and supplied indegree bound $q = d$ as TD-BSS. For each node, it selects a subset of at most q earlier nodes minimizing $n \log(\mathrm{RSS}/n) + |S| \log n$, including the empty subset.

Daskalakis Alg. 1. Our C++ implementation follows the published ordering and parent-selection procedure (Daskalakis et al., 2025). It uses the supplied indegree bound $q = d$ given by an oracle and coefficient threshold $b_{\min}/2 = 0.2$, using the generating signal lower bound $b_{\min} = 0.4$.

TD-Lasso. We use the official R implementation of the top-down procedure (Chen et al., 2019), with subset-search parameter $J = q$, supplied indegree bound $q = d$ given by an oracle, and cross-validated Lasso parent selection. The Lasso penalty is selected by ten-fold cross-validation using the one-standard-error rule.

Daskalakis Alg. 2. We use the published Python implementation of Algorithm 2 (Daskalakis et al., 2025), with Lasso penalty 0.01, supplied indegree bound $q = d$ given by an oracle, and coefficient threshold $b_{\min}/2 = 0.2$, using the generating lower bound $b_{\min} = 0.4$.

GOLEM-EV. We use the official Python/TensorFlow equal-variance implementation on CPU (Ng et al., 2020), with $\lambda_1 = 0.02$, $\lambda_2 = 5$, learning rate 0.001, and 100,000 optimizer updates. The official postprocessing routine thresholds edge weights at 0.3 and removes edges until the result is acyclic.

Metrics. True positives, false positives, and false negatives are counted on directed edges. Exact recovery means equality of the entire directed adjacency matrices. F1 is $2\mathrm{TP}/(2\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$, with value one when both edge sets are empty. Directed SHD is $\mathrm{FP} + \mathrm{FN}$, so reversing an edge contributes two errors. Metrics are computed for each repetition and then averaged. Standard errors are the sample standard deviations divided by the square root of the number of repetitions. The graph-recovery figures report means with error bars of one standard error across repetitions.

Figure 4 reports exact recovery, F1, and SHD for the method comparison in Figure 3(a)–(b). At p = 200, n = 500, HOST and HOST-CV/EBIC attain F1 0.984 and 0.985, respectively, close to 0.991 for TD-BSS, which has higher exact recovery and lower SHD. At n = 100, HOST-CV/EBIC has higher F1 and lower SHD than HOST at both dimensions, while the two variants become similar as sample size increases. The recovery trends are similar to those in Figures 2 and 3(a)–(b).

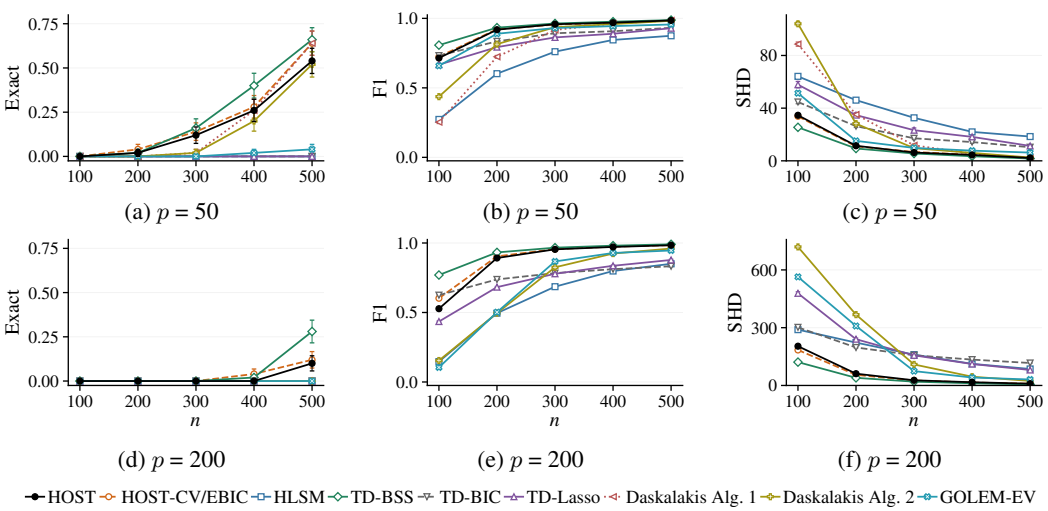


Figure 4: Graph recovery for $d = 3$. Panels (a)–(c) use $p = 50$, and (d)–(f) use $p = 200$. Columns show exact recovery, F1, and SHD. Points and error bars show means ± 1 standard error over 50 paired repetitions.

D.4 Runtime Experiments

The runtime experiments in Figures 3(c)–(d) and 5 fix $n = 100$ and use all combinations of $p = 10$, 30, 50, 70, 100 and $d = 1$, 2, 3, 4, 5, with 20 paired repetitions per setting. Each repetition uses the same input across methods. Daskalakis Alg. 1 is evaluated only at $p = 10$, 30, 50. GOLEM-EV is excluded from this single-core CPU runtime comparison because its published scalability experiments use GPU acceleration (Ng et al., 2020). HOST, HOST-CV/EBIC, HLSM, TD-BSS, TD-BIC, and Daskalakis Alg. 1 use our compiled Cython/C++ implementations for both graph recovery and runtime measurement.

For TD-Lasso, graph recovery uses the official R code, whereas runtime is measured with our C++ implementation. For Daskalakis Alg. 2, graph recovery uses the published Python code, whereas runtime is measured with our Cython/C++ implementation. The timing implementations preserve the algorithms, supplied information, constants, and per-dataset tuning rules specified in Appendix D.3. Their numerical agreement with the graph-recovery implementations is checked before timing.

The runtime experiments use Intel Xeon Platinum 8592+ processors. Each fit uses a single physical CPU core and one numerical-library thread. Full-fit time is the wall-clock time for the full fitting procedure on each prepared input, including ordering, parent selection, final graph construction, method-internal transformations, and tuning. This includes CV and EBIC parameter selection for HOST-CV/EBIC and ten-fold cross-validation for TD-Lasso. HOST uses ordering and parent-threshold multipliers 0.5 and 2 without parameter search.

Figure 5 extends the two runtime sweeps in Figure 3(c)–(d) to all settings, varying d at each fixed p and p at each fixed d. The full comparison shows runtime trends similar to those in Figure 3(c)–(d).

D.5 Additional Graph Families

Figure 6 compares recovery on ER, hub, and bipartite DAGs with $p = 30$, $n = 100$, 200, 300, 400, 500, and 50 paired repetitions. It includes the same methods and recovery implementations as Figure 3(a)–(b), including Daskalakis Alg. 1 and both HOST variants. Coefficient and error distributions, node permutations, and tuning settings follow the main method comparison. Subset-search methods receive the supplied indegree bound $q = d$, where d is the realized maximum indegree of each graph.

For ordered Erdős–Rényi DAGs, each potential edge from an earlier to a later node is sampled independently with probability $2/(p-1)$, giving expected edge count p. For hub DAGs, three nodes with at least six earlier nodes are selected as hubs, each receiving six uniformly sampled incoming

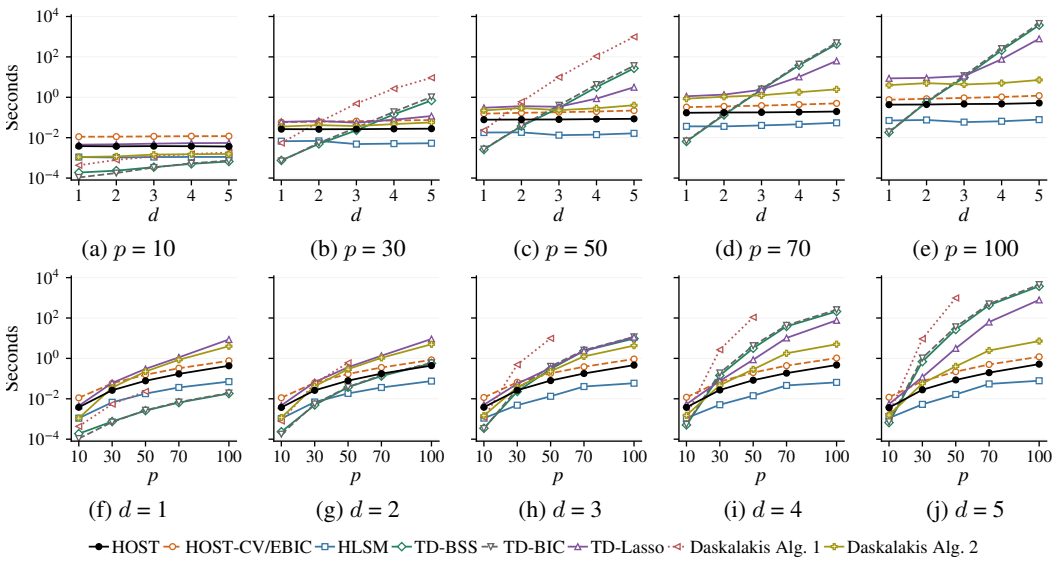


Figure 5: Mean full-fit time (seconds, log scale) at $n = 100$, including tuning, over 20 paired repetitions. Panels (a)–(e) vary d at fixed p, and (f)–(j) vary p at fixed d.

edges. Other nodes receive independently sampled incoming edges with probability 0.02, conditioned on having at most five parents. The maximum indegree is therefore exactly six. Bipartite DAGs have two groups of 15 nodes. Edges are sampled independently from the first group to the second with probability $2/15$, giving expected edge count 30, with no edges within either group.

Figure 6 shows that relative performance varies across graph families and recovery metrics. At $n = 500$, HOST achieves the highest exact recovery on ER graphs, whereas Daskalakis Alg. 1 achieves higher exact recovery on hub and bipartite graphs. The two HOST variants become similar in F1 and SHD as sample size increases, and both provide competitive graph recovery across all three graph families.

D.6 Ordering Recovery Comparison with Subset Search Methods

We compare the ordering steps of HOST and TD-BSS (Gao et al., 2022) at $p = 50$, $d = 2, 3, 4$, and $n = 100, 200, 300, 400, 500$, with 50 paired repetitions. Each graph contains the chain $1 \to 2 \to \cdots \to p$, which makes its compatible ordering unique. For $j \geq 2$, node j has exactly $\min(d, j-1)$ parents, including $j-1$, with any additional parents sampled uniformly from $\{1, \ldots, j-2\}$. Coefficients, error variances, and node permutations follow Appendix D. HOST uses the fixed ordering penalty in Appendix D.3, while TD-BSS receives supplied indegree bounds $q = 2, 3, 4, 5, 6$. We report the normalized Kendall distance $K/\binom{p}{2}$, where K counts node pairs ordered differently in the estimated and true orderings. We also compare ordering runtime.

Figure 7 shows that HOST achieves decreasing Kendall distance as sample size increases without receiving a supplied indegree bound, while TD-BSS generally achieves smaller Kendall distance when $q \geq d$. Supplying $q < d$ can substantially degrade TD-BSS recovery. When d is unknown, choosing a larger q reduces the risk of under-specification but sharply increases the cost of subset search. This comparison illustrates the computational advantage of ordering recovery without a supplied indegree bound.

D.7 Parent Recovery Comparison with Lasso

We isolate parent recovery by supplying the true ordering to HOST and Lasso. The graph construction follows Appendix D.6, with $p = 50$, $d = 2, 3, 4$, $n = 100, 200, 300, 400, 500$, and 50 paired repetitions. Lasso fits each node on all earlier variables using penalty $\lambda = c\sqrt{\log p/n}$ and selects variables with nonzero coefficients as parents. HOST uses parent threshold $\tau_j = c\sqrt{\log p/n}$ and $\delta = 0.8$. The

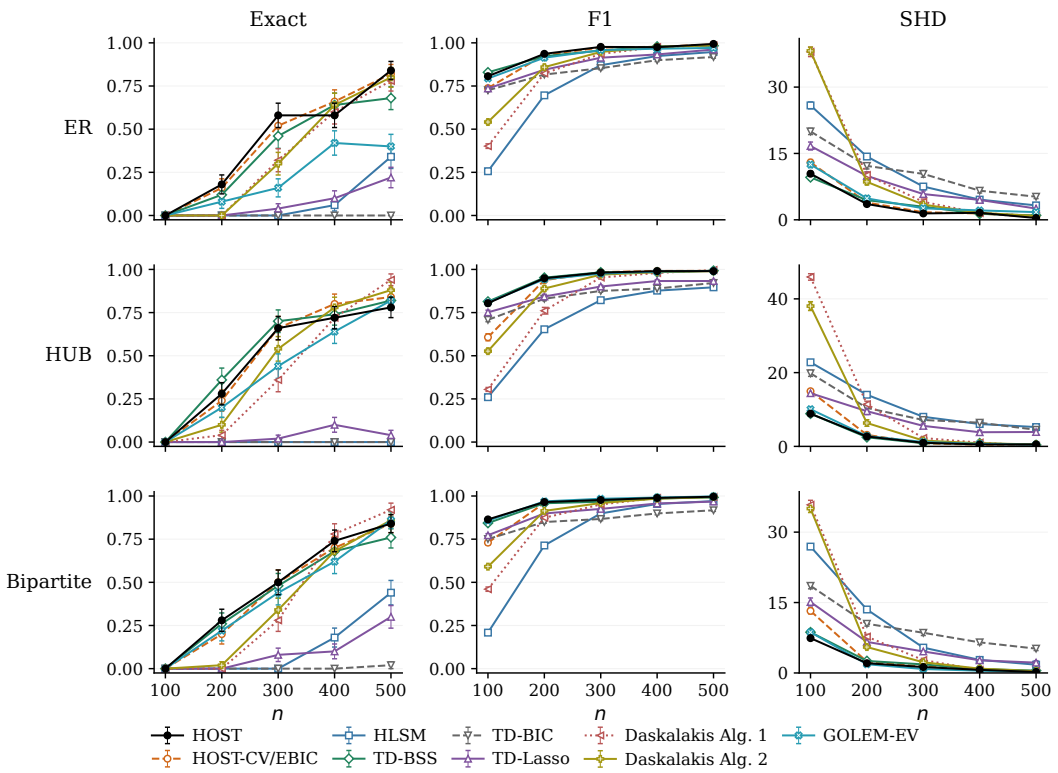


Figure 6: Graph recovery for ER, hub, and bipartite DAGs in the first, second, and third rows, respectively, with $p = 30$. Columns show exact recovery, F1, and SHD. Points and error bars show means ± 1 standard error across 50 paired repetitions.

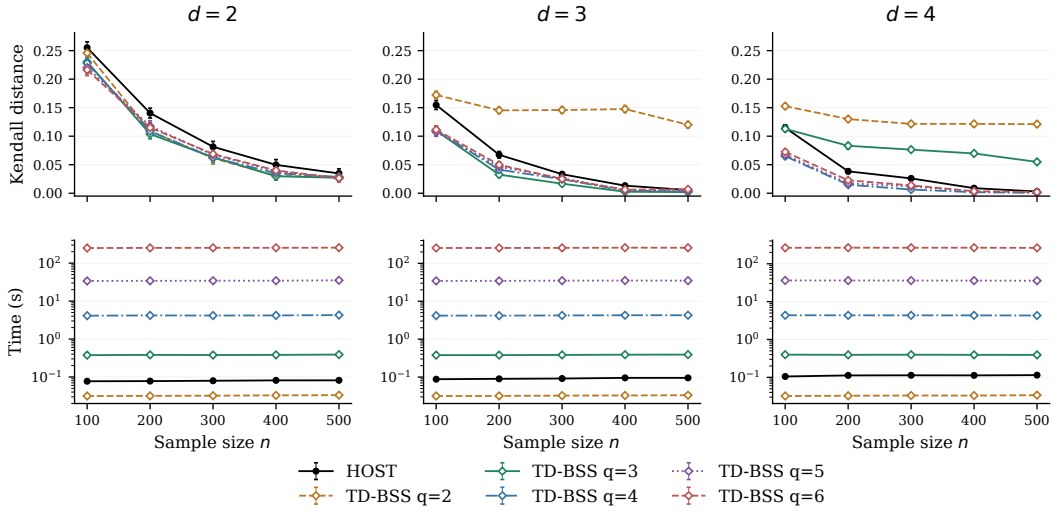


Figure 7: Ordering recovery and computational cost at $p = 50$. Rows show Kendall distance and runtime, and columns show $d = 2, 3, 4$. Points and error bars show means ± 1 standard error across 50 paired repetitions.

two-fold ordering fits used to construct the parent-recovery guide follow the true ordering and use a fixed penalty multiplier of 0.5. HOST reuses its main-comparison implementation, and Lasso uses the same coordinate-descent solver.

To remove differences due to tuning parameter selection, each method uses oracle tuning over $c \in \{0.5, 1, 2, 3, 4, 5\}$. In each repetition, one multiplier for the entire graph is chosen to minimize SHD against the true graph, with ties resolved in favor of the larger multiplier. SHD, precision, and recall are all computed from that same selected graph. Oracle tuning provides a reference for recovery achievable within this candidate pool.

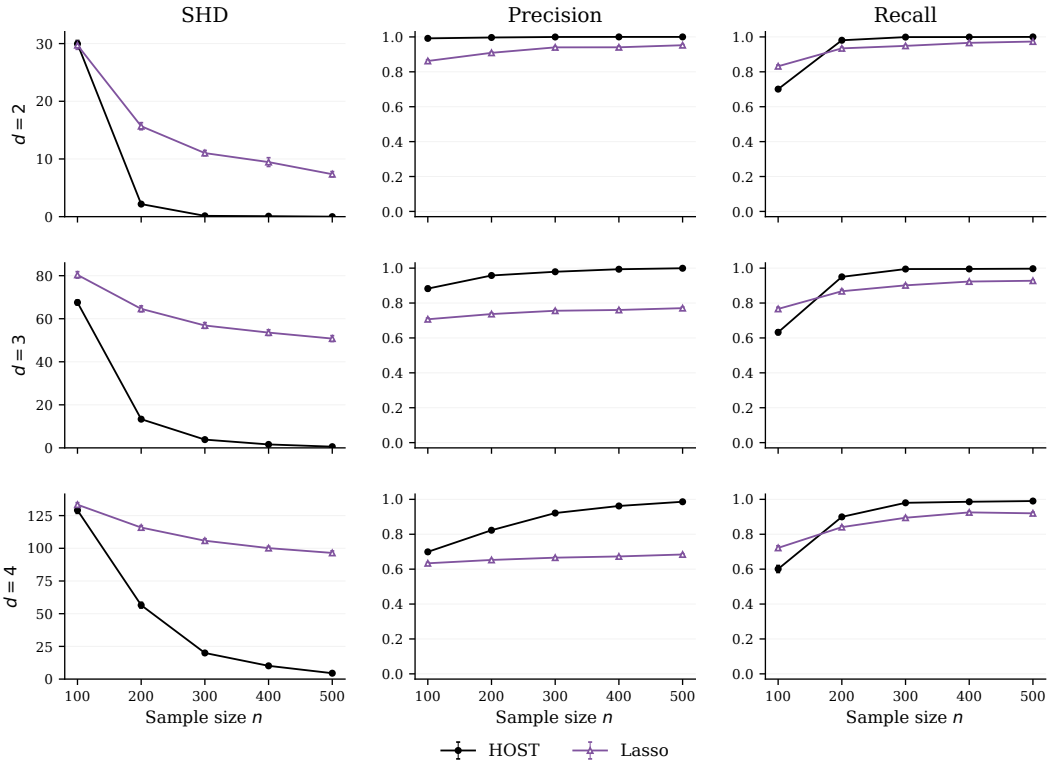


Figure 8: Parent recovery with the true ordering and oracle tuning at $p = 50$. Rows correspond to $d = 2, 3, 4$, and columns show SHD, precision, and recall. Each method selects $c \in \{0.5, 1, 2, 3, 4, 5\}$ by minimizing SHD separately in each repetition. Points and error bars show means ± 1 standard error across 50 paired repetitions.

Figure 8 shows that HOST generally achieves lower SHD than Lasso as sample size increases, with a larger gap at larger maximum indegrees, although Lasso performs better in some small-sample settings. The difference remains after each method receives its best multiplier in the candidate pool. Increasing d introduces more direct and indirect paths and can make covariance conditions for Lasso support recovery, including the irrepresentable condition (Zhao & Yu, 2006), harder to satisfy. The observed pattern is consistent with this limitation of Lasso. HOST instead reconstructs direct effects without requiring exact Lasso support recovery or irrepresentability, as established in Theorem 4.2.

E Real-Data Experiments

We compare HOST with six methods on the Sachs sample available from the `bnlearn` website (Sachs et al., 2005) ($n = 853$, $p = 11$) and the uniform reference experiment from the Light Tunnel dataset (Gamella et al., 2025) ($n = 10{,}000$, $p = 20$). Each dataset is column-centered using all observations before fitting. The supplied reference DAGs contain 17 edges for Sachs and 39 for Light Tunnel and are used only to evaluate directed F1 and SHD.

We use HOST, HLSM, TD-BSS, TD-BIC, TD-Lasso, and Daskalakis Algs. 1 and 2, following the implementations in Appendix D.3. GOLEM-EV is excluded from this comparison. All tuning multipliers range over $\mathcal{C} = \{2^{-7}, 2^{-6}, \ldots, 2^{7}\}$, and supplied indegree bounds range over $Q = \{1, 2, 4, 6, 8\}$.

For HOST, pooled two-fold CV selects $c_{\mathrm{ord}} \in \mathcal{C}$ and the common ordering penalty $\lambda_{\mathrm{ord}} = c_{\mathrm{ord}}\sqrt{\log p/\lfloor n/2 \rfloor}$. After recovering the ordering with this penalty, HOST processes nodes in that ordering and selects $c_j \in \mathcal{C}$ separately for each node by EBIC, giving parent threshold $\tau_j = c_j\sqrt{\log p/n}$. The CV and EBIC criteria follow Appendix C.

For each comparison method, we construct a candidate DAG G for every parameter combination in Table 2 and select the graph and parameters minimizing Gaussian BIC (Biza et al., 2020),

$$\mathrm{BIC}(G) = n\sum_{j=1}^{p}\log\left(\frac{\mathrm{RSS}_j(G)}{n}\right) + |E(G)|\log n.$$

Here $E(G)$ is the edge set of G, and $\mathrm{RSS}_j(G)$ is the residual sum of squares from OLS regression of node j on its parents in G, without an intercept on the centered data.

Table 2: Real-data tuning parameters and candidate construction.

Method	Parameters and selection procedure
HOST	Select $c_{\mathrm{ord}} \in \mathcal{C}$ by pooled two-fold CV, then $c_j \in \mathcal{C}$ for each node by EBIC.
HLSM	For each $c \in \mathcal{C}$, fit with Lasso penalty $\lambda = c\sqrt{\log p/n}$.
TD-BSS	For every $(q, c) \in Q \times \mathcal{C}$, fit with supplied indegree bound q and parent-selection threshold $\eta = c\log p/n$.
TD-BIC	For each $q \in Q$, select each parent set by minimizing nodewise BIC over subsets of at most q earlier nodes.
TD-Lasso	For each $q \in Q$, recover the ordering and select nodewise Lasso penalties by ten-fold cross-validation using the one-standard-error rule.
Daskalakis Alg. 1	For every $(q, c_b) \in Q \times \mathcal{C}$, fit with supplied indegree bound q and $b_{\min} = c_b\sqrt{\log p/n}$.
Daskalakis Alg. 2	For every $(q, c_b, c_\lambda) \in Q \times \mathcal{C}^2$, fit with supplied indegree bound q, $b_{\min} = c_b\sqrt{\log p/n}$, and Lasso penalty $\lambda = c_\lambda\sqrt{\log p/n}$.

Table 3: Directed F1 and SHD on the full Sachs and Light Tunnel datasets. Bold entries indicate the highest F1 or lowest SHD in each column.

	Sachs		Light Tunnel	
Method	F1	SHD	F1	SHD
HOST	**0.385**	**16**	**0.518**	**41**
HLSM	0.276	21	0.477	57
TD-BSS	0.303	23	0.344	80
TD-BIC	**0.385**	**16**	0.506	**41**
TD-Lasso	0.200	**16**	0.517	43
Daskalakis Alg. 1	0.357	18	0.411	66
Daskalakis Alg. 2	0.370	17	0.472	47

Table 3 shows that HOST and TD-BIC tie for the highest Sachs F1 (0.385) and, together with TD-Lasso, attain the lowest SHD (16). On Light Tunnel, HOST attains F1 0.518, close to 0.517 for TD-Lasso, and ties TD-BIC for the lowest SHD (41). These results indicate competitive recovery by HOST on both datasets among methods developed under similar modeling assumptions.