The Causal Description Gap: Information-Theoretic Separations Across Pearl's Hierarchy

Seyed Morteza Emadi
Kenan-Flagler Business School
University of North Carolina at Chapel Hill
seyed_emadi@kenan-flagler.unc.edu

Abstract

Pearl's causal hierarchy shows that observational, interventional, and counterfactual queries are qualitatively distinct. We ask a quantitative version of this question: how many additional bits are needed to specify higher-rung causal answers once lower-rung answers are known? We formalize this via *query-class description length*, the Kolmogorov complexity of the answers a structural causal model (SCM) gives to a class of queries, with system size given. Our main construction gives acyclic SCMs on n binary variables whose common observational distribution has constant description length, while single-variable intervention answers can require $\Theta(n^2)$ additional bits. For SCMs with independent noises, fixed finite gate-schema and noise libraries, and at most d parents per variable, a degree-sensitive upper bound gives an observational–interventional gap of at most $O(nd\log(en/d) + n\log n)$. This makes the quadratic construction order-optimal in the dense regime and a rooted-tree construction order-optimal for bounded indegree. The quadratic separation persists under ε-accurate total-variation descriptions for every fixed $\varepsilon < 1/4$. At the next rung, even knowing every hard-intervention distribution can leave a $\Theta(n)$ counterfactual description gap; a masked-parity construction extends this to $\Theta(n^2)$. These separation rates hold in the worst case and with high probability under uniform hidden parameters. A general ambiguity-to-bits theorem and Shannon analogue, together with explicit encodings, show that the gaps in our constructions equal the logarithm of residual higher-rung ambiguity up to lower-order terms.

1 Introduction

Predictive accuracy identifies what a system does under passive observation. *Causal* questions go further: *interventional* queries ("what if I set X?") and *counterfactual* queries ("what would have happened if X had been different?"). Pearl's causal hierarchy [11] formalizes these three levels, or rungs (association, intervention, counterfactual), and the Causal Hierarchy Theorem of Bareinboim et al. [1] shows that the hierarchy is generically strict: higher-rung queries cannot in general be answered from lower-rung information. But *how large* is the gap? Is the additional information bounded by a constant, or can it scale with system size?

We give a quantitative answer in bits. We study acyclic structural causal models (SCMs) on n binary variables, so each observation is an n-bit vector. An SCM specifies each variable as a function of its direct causes, called parents, and noise. Our baseline models use independent, finite-valued noises with rational probabilities. Hard interventions set selected variables to fixed values. We define *query-class description length* as the Kolmogorov complexity of a model's answers to a class

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

of queries, with n given. A *causal description gap* measures the additional bits needed to specify higher-rung answers when lower-rung answers are already known.

Residual information after non-identification. Identification asks whether lower-rung information determines a higher-rung answer. We construct families whose models share the same lower-rung answers but have exponentially many distinct higher-rung answers. Counting these possibilities and matching them with explicit encodings shows that the missing information can grow quadratically with system size.

Contributions.

1. **Framework and ambiguity principle.** For finite SCM families with finitely encoded query answers, a general counting theorem lower-bounds the worst-case additional bits by the logarithm of the number of compatible higher-rung answers, up to an additive constant. A Shannon analogue and explicit encodings give matching characterizations for our constructions.
2. **Observational–interventional separations and optimality.** Rooted-tree and bipartite families have $O(1)$-bit observational descriptions but interventional gaps of $\Theta(n \log n)$ and $\Theta(n^2)$. Let d denote the maximum number of parents of any variable. For fixed finite libraries of noise laws and gate schemas (mechanism rules specified uniformly across input sizes), an upper bound of $O(nd \log(en/d) + n \log n)$ covers all hard-intervention and counterfactual answers. It matches the tree and bipartite gaps for $d = O(1)$ and $d = \Theta(n)$, respectively.
3. **Counterfactual separations.** Our modular-XOR and masked-parity families have gaps of $\Theta(n)$ and $\Theta(nd \log(en/d))$, respectively, even given every distribution under hard interventions. The latter gap reaches $\Theta(n^2)$ when $d = \Theta(n)$. Both families assign positive probability to every observational state; no number of hard interventions resolves their counterfactual ambiguity.
4. **Robustness, learning, and experiments.** The bipartite family has two layers of m variables and one root, so $n = 2m + 1$. Its gap survives total-variation error $\varepsilon < 1/4$. No observational learner recovers the intervention answers with probability exceeding 2^{-m^2} under a uniform hidden graph, regardless of sample size. When each experiment supplies its exact post-intervention distribution, m intervention settings are necessary in the worst case and sufficient for identification.
5. **Assumptions and graph knowledge.** With independent noises and positivity (positive probability for every state), a known graph reduces the interventional gap to $O(1)$. Without graph knowledge, adding faithfulness gives a tight worst-case $\Theta(n \log n)$ bound. Faithfulness requires every observational conditional independence to be implied by the graph. Our family with shared latent causes (confounding) retains $\Theta(n)$ bits even given the observed graph.

The separation rates hold in the worst case and with high probability under uniform hidden parameters within the stated families. Table 1 summarizes the constructions; Sections 2 and 3 provide related work and the formal framework.

Family	Lower rung	Higher rung	Ambiguity	Gap	Where
Rooted trees	Obs	Int_1	n^{n-1}	$\Theta(n \log n)$	§4.1
Bipartite graphs	Obs	Int_1	2^{m^2}	$\Theta(n^2)$	§4.2
Modular XOR	$\mathsf{Int}_{\mathrm{all}}$	CF_1	2^m	$\Theta(n)$	§5.1
Masked parity	$\mathsf{Int}_{\mathrm{all}}$	CF_1	$\left(\sum_{k=0}^{d} \binom{m}{k}\right)^m$	$\Theta(nd \log \frac{en}{d})$	§5.2
Confounded modules	$\mathsf{Obs}, \mathcal{G}$	Int_1	2^m	$\Theta(n)$	§7

Table 1: Lower rung: answers given; higher rung: answers to specify. Ambiguity counts compatible higher-rung answers; gaps are in bits. Obs: observations; Int_1: single-node interventions; $\mathsf{Int}_{\mathrm{all}}$: all hard interventions; CF_1: single-node counterfactuals; $\mathcal{G}$: observed mixed graph. The bipartite family has $n = 2m + 1$ variables; the XOR, parity, and confounded families have $n = 2m$. For masked parity, $1 \leq d \leq m$.

2 Related Work

Pearl's causal hierarchy. Pearl's three-level hierarchy [11, 12] is generically strict: the Causal Hierarchy Theorem [1] establishes non-identifiability, showing that some higher-rung queries are not functions of lower-rung information. Shpitser and Pearl [17] give complete identification algorithms; Ibeling and Icard [6] compare frameworks. We quantify this ambiguity: even for simple binary acyclic SCMs, specifying a higher-rung answer given the lower-rung answer can require $\Theta(n^2)$ additional bits.

Causal discovery. Constraint- and score-based methods identify causal graphs up to Markov equivalence from observations [5, 14, 18]. Eberhardt et al. [3] and Shanmugam et al. [15] bound the number/sample complexity of interventions for identification. We give description-length lower bounds: even with infinite observational data, the obs–int information gap is $\Theta(n^2)$ in our family.

Description length and MDL. Kolmogorov complexity [9, 10] and minimum description length (MDL) [4] measure the information needed to describe objects or models. Our use is different: the compressed object is neither the dataset nor the SCM parameterization, but the complete collection of answers to a query class, or *answer oracle*. This distinction is what enables our separations: the observational answer oracle can have constant description length even when the corresponding interventional oracle requires $\Theta(n^2)$ bits.

Causal reasoning in AI. Recent benchmarks find uneven causal-reasoning performance in LLMs [7, 8]; our results give one information-theoretic reason why observational predictive training alone need not induce interventional competence in worst-case structured families.

Complexity across Pearl's hierarchy. Dörfler et al. [2] study satisfiability across probabilistic, interventional, and counterfactual languages in Pearl's hierarchy, proving strictly increasing computational complexity for certain languages (e.g., $\mathrm{NP}^{\mathrm{PP}}$, PSPACE, and NEXP completeness under addition and marginalization). They ask how hard it is to decide whether there exists an SCM satisfying a given collection of causal statements; we bound the *information content* of higher-rung answer oracles given lower-rung ones.

3 Framework: Query-Class Description Length

We first define the models and their causal answers, then measure the additional bits needed to specify higher-rung answers given lower-rung answers. The lemma proofs are in Appendix A.

3.1 Models and Causal Answers

Definition 3.1 (Structural Causal Model)**.** A *structural causal model* (SCM) on binary variables $X_1, \ldots, X_n \in \{0, 1\}$ is a tuple $M = (G, \{f_i\}, P_U)$ where:

- G is a directed acyclic graph (DAG) on $[n] := \{1, \ldots, n\}$.
- $U_1, \ldots, U_n$ are independent finite-valued exogenous (noise) variables; their product law P_U is specified by finite rational probability tables.
- Structural equations specify each variable as a function of its parents $\mathrm{pa}_G(i)$ and noise:

$$X_i = f_i\big((X_j)_{j \in \mathrm{pa}_G(i)}, U_i\big).$$

We require each f_i to be computable. Every distribution defined below is then a finite table of rational probabilities computable from the model.

Sampling $(U_1, \ldots, U_n) \sim P_U$ and evaluating the equations in topological order gives the **observational distribution** $P_M(X_1, \ldots, X_n)$.

Definition 3.2 (Hard intervention)**.** A *hard intervention* $\mathrm{do}(X_i = x)$ modifies M by replacing the equation for X_i with the constant $X_i := x$, leaving all other equations unchanged. The resulting distribution is $P_M^{\mathrm{do}(X_i = x)}$.

Definition 3.3 (Counterfactual)**.** For a fixed exogenous realization $u = (u_1, \ldots, u_n)$, the *counterfactual* $X^{(i \leftarrow b)}(u)$ is the value of the entire system $(X_1, \ldots, X_n)$ when we intervene with $\mathrm{do}(X_i = b)$ but use the *same* noise realization u.

Counterfactual queries concern joint distributions such as $P(X, X^{(i \leftarrow 0)}, X^{(i \leftarrow 1)})$ of three n-bit vectors: the actual outcome and those with X_i forced to 0 and to 1, computed from the same noise.

We collect these distributions into answer objects, also called answer oracles.

Definition 3.4 (Query Families)**.** For an SCM M on n variables:

- **Observational:** $\mathsf{Obs}(M) := P_M$, the joint distribution over $(X_1, \ldots, X_n)$.
- **Single-node interventional:** $\mathsf{Int}_1(M) := \big(P_M, (P_M^{\mathrm{do}(X_i=0)}, P_M^{\mathrm{do}(X_i=1)})_{i=1}^n\big)$, the observational distribution plus the $2n$ single-variable interventional distributions.
- **All hard interventions:** $\mathsf{Int}_{\mathrm{all}}(M) := \big(P_M^{\mathrm{do}(X_S = x_S)}\big)_{S \subseteq [n],\, x_S \in \{0,1\}^S}$, the laws under all hard interventions on subsets of variables, single-node and multi-node. We call it the *full hard-do interventional oracle* (it excludes soft, stochastic, edge, and exogenous interventions); $\mathsf{Int}_1(M)$ is computable from it.
- **Single-node counterfactual:** $\mathsf{CF}_1(M) := \big(P(X, X^{(i \leftarrow 0)}, X^{(i \leftarrow 1)})\big)_{i=1}^n$. This collects the counterfactual joint distributions for all variables.

3.2 Description Lengths and Causal Gaps

Definition 3.5 (Kolmogorov Complexity)**.** Fix a universal prefix-free Turing machine. The *Kolmogorov complexity* $K(z)$ is the length of the shortest program that outputs string z. The conditional complexity $K(z \mid w)$ is the shortest program length when w is given as auxiliary input. By the invariance theorem [10, Section 2.1] and the same argument for encodings, neither the choice of universal machine nor the choice of computably invertible encoding changes K by more than an additive $O(1)$; we encode a distribution as the string of its lowest-terms rational probabilities.

Definition 3.6 (Query-Class Description Length: General Form)**.** Let Q be a query class with a computable finite encoding $\mathrm{Ans}_Q(M)$ of all its answers under M. The *query-class description length* of M with respect to Q is

$$\mathrm{DL}(Q;\, M) \;:=\; K\big(\mathrm{Ans}_Q(M) \mid n\big).$$

The *conditional gap* between query classes Q_2 and Q_1 is

$$\Delta(Q_2 \mid Q_1;\, M) := K\big(\mathrm{Ans}_{Q_2}(M) \mid \mathrm{Ans}_{Q_1}(M),\, n\big).$$

The lower-rung answers and the number of variables are supplied as inputs; the gap counts the additional program bits needed to produce the higher-rung answers.

For the observational and single-node query families, write $\mathrm{DL}_1(M) := K(\mathsf{Obs}(M) \mid n)$, $\mathrm{DL}_2(M) := K(\mathsf{Int}_1(M) \mid n)$, and $\mathrm{DL}_3(M) := K(\mathsf{CF}_1(M) \mid n)$. The corresponding *causal description gaps* are $\Delta_{2|1}(M) := K(\mathsf{Int}_1(M) \mid \mathsf{Obs}(M), n)$ and $\Delta_{3|2}(M) := K(\mathsf{CF}_1(M) \mid \mathsf{Int}_1(M), n)$. We also consider $K(\mathsf{CF}_1(M) \mid \mathsf{Int}_{\mathrm{all}}(M), n)$, the counterfactual gap given all hard-intervention distributions.

The following lemmas supply the counting and reduction tools for our separations.

Lemma 3.7 (Counting bound for Kolmogorov complexity)**.** *Let $z_1, \ldots, z_N$ be N distinct strings and w an arbitrary auxiliary string. Then:*

1. *At least one z_i satisfies $K(z_i \mid w) \geq \log_2 N - O(1)$.*
2. *For every $c \geq 0$, all but at most a 2^{-c} fraction of the z_i satisfy $K(z_i \mid w) \geq \log_2 N - c - O(1)$.*

We use this bound when distinct hidden structures give distinct answers.

Lemma 3.8 (Hierarchy Monotonicity)**.** $\mathrm{DL}_1(M) \leq \mathrm{DL}_2(M) + O(1) \leq \mathrm{DL}_3(M) + O(1)$.

Selecting components and taking marginals recovers lower-rung answers from higher-rung answers.

Lemma 3.9 (Conditioning on a constant-complexity object is free)**.** *If a family satisfies* $K(\mathsf{Obs}(M) \mid n) = O(1)$ *with a bound independent of* M *and* n*, then* $K(\mathsf{Int}_1(M) \mid \mathsf{Obs}(M), n) = K(\mathsf{Int}_1(M) \mid n) \pm O(1)$*, and consequently* $\Delta_{2|1}(M) = \mathrm{DL}_2(M) \pm O(1)$*.*

In the tree and bipartite families, observations have constant complexity, so by Lemma 3.9 interventional description length gives the gap. For arbitrary lower-rung answers, the conditional definition applies directly.

4 Observational–Interventional Separations

We apply the counting bound to families whose members share one observational law: trees give the bounded-indegree separation, and bipartite graphs give the quadratic separation. Full proofs are in Appendix B, and Figure 1(a) and (b) there sketch the two constructions.

4.1 Rooted Trees

Let T be a rooted labeled tree on $[n]$, with root r and parent $\mathrm{pa}_T(v)$ for each $v \neq r$. Set $X_r := U_r \sim \mathrm{Bernoulli}(1/2)$ and $X_v := X_{\mathrm{pa}_T(v)}$ for $v \neq r$. All trees share the observational law $P^\star$, assigning mass $1/2$ to each of 0^n and 1^n.

Under $\mathrm{do}(X_i = 0)$, every descendant of i, including i itself, becomes 0; non-descendants remain fair. The distributions for these n intervention settings therefore identify the descendant sets, which determine the root and parent links.

Theorem 4.1 (Rooted-tree separation)**.** *Let* $\mathcal{M}^n_{\mathrm{tree}}$ *be the family of rooted-tree SCMs* M_T *above. All members share* $\mathsf{Obs}(M_T) = P^\star$ *with* $\mathrm{DL}_1(M_T) = O(1)$*. Moreover:*

1. ***Upper bound.*** *For every* T*,* $\mathrm{DL}_2(M_T) \leq (n-1)\log_2 n + O(\log n)$*.*
2. ***Lower bound.*** *For uniformly random* T *on* $[n]$ *and every* $c \geq 0$*,*

$$\Pr\big[\Delta_{2|1}(M_T) \geq (n-1)\log_2 n - c - O(1)\big] \;\geq\; 1 - 2^{-c}.$$

Thus a uniformly chosen tree has a $\Theta(n \log n)$ gap with high probability, although every node has at most one parent.

Proof sketch. Parent labels describe T in $(n-1)\log_2 n + O(\log n)$ bits. Cayley's formula gives n^{n-1} rooted trees; $\mathrm{do}(X_i = 0)$ zeroes exactly i and its descendants, so the intervention tables determine T. Lemma 3.7 gives the lower bound; Lemma 3.9 moves it to $\Delta_{2|1}$ since $P^\star$ has constant complexity.

4.2 Bipartite Graphs

To obtain a quadratic gap, we use two layers of m variables and a common root, so $n = 2m + 1$. Each of the m^2 possible edges between the layers can be present or absent.

Definition 4.2 (Bipartite-graph SCM)**.** Let $m \geq 1$ and $n = 2m + 1$. Use a fixed computable partition into a root r, layer $A = \{a_1, \ldots, a_m\}$, and layer $B = \{b_1, \ldots, b_m\}$. For a bipartite graph $G \subseteq A \times B$, the SCM M_G has $U_r \sim \mathrm{Bernoulli}(1/2)$ as the only random exogenous variable, and structural equations

$$X_r := U_r, \qquad X_{a_i} := X_r \;\text{ for all } a_i \in A, \qquad X_{b_j} := X_r \wedge \bigwedge_{(a_i, b_j) \in G} X_{a_i} \;\text{ for all } b_j \in B,$$

where the empty AND is defined as 1.

Every variable equals the root observationally. Setting a layer-A variable to zero forces its neighbors in B to zero, while non-neighbors remain fair. The following lemma makes this distinction precise.

Lemma 4.3 (Bipartite decoding)**.** *For every bipartite graph* $G \subseteq A \times B$*:*

(i) $\mathsf{Obs}(M_G) = P^\star$*, assigning mass* $1/2$ *to each of* 0^n *and* 1^n*.*

(ii) For every $a_i \in A$*,* $P^{\mathrm{do}(X_{a_i}=0)}_{M_G}(X_{b_j} = 0)$ *equals* 1 *if* $(a_i, b_j) \in G$ *and* $1/2$ *otherwise, so the neighborhood* $N_G(a_i) := \{b_j : (a_i, b_j) \in G\}$ *is determined by* $P^{\mathrm{do}(X_{a_i}=0)}_{M_G}$*.*

(iii) Consequently the map $G \mapsto \mathsf{Int}_1(M_G)$ is injective.

The distributions for the m settings $\mathrm{do}(X_{a_i} = 0)$ therefore determine all m^2 adjacency bits.

Theorem 4.4 (Quadratic separation: $\Theta(n^2)$)**.** *Let $n = 2m + 1$ and $\mathcal{M}^n_{\mathrm{bip}} := \{M_G : G \subseteq A \times B\}$ as in Definition 4.2. Then:*

1. ***Upper bound.*** *For every G, $\mathrm{DL}_2(M_G) \le m^2 + O(1)$.*
2. ***Lower bound.*** *For $G \sim \mathrm{Unif}(2^{A\times B})$ and every $c \ge 0$,*
$$\Pr\big[\Delta_{2|1}(M_G) \ge m^2 - c - O(1)\big] \ge 1 - 2^{-c}.$$
In particular, taking $c = m^2/2$ shows that a uniformly random graph has $\Delta_{2|1}(M_G) = \Omega(n^2)$ except with probability $\exp(-\Omega(n^2))$.

Proof sketch. Encoding the adjacency matrix uses m^2 bits and lets a fixed program simulate M_G. Conversely, Lemma 4.3(iii) gives 2^{m^2} distinct interventional answers with the same observation $P^\star$. Applying Lemma 3.7 conditional on $(P^\star, n)$ gives the lower bound.

4.3 Degree-Sensitive Optimality

The tree and bipartite constructions differ in their maximum indegree. To bound the gap in terms of indegree, we also fix the available mechanisms and noise laws; describing the model then reduces to encoding parent sets and choices from these libraries.

Definition 4.5 (Finite-gate-schema bounded-indegree SCM class)**.** A *gate schema* is a computable family $\gamma = (\gamma_k)_{k\ge 0}$ with $\gamma_k : \{0,1\}^k \times \mathcal{U}_\gamma \to \{0,1\}$, specified by a constant-length program independent of n and k (uniformly across input sizes). Here $\mathcal{U}_\gamma$ is the noise-input space. Examples include the constant, copy, negation, parity, and unbounded-AND/OR schemas.

Fix a *gate library*, a finite set Γ of gate schemas, and a *noise library*, a finite set Π of finite-support rational noise distributions, both independent of n. The class $\mathcal{M}_{n,d}(\Gamma, \Pi)$ consists of binary acyclic SCMs on n endogenous variables, each with at most d endogenous parents. Each equation applies a component γ_k, $\gamma \in \Gamma$, to its $k \le d$ parent values in increasing variable-label order and its local noise; each local noise law belongs to Π.

The bipartite-graph SCMs of Definition 4.2 belong to $\mathcal{M}_{n,m+1}(\Gamma_{\mathrm{bip}}, \Pi_{\mathrm{bip}})$ for $\Gamma_{\mathrm{bip}} = \{\mathrm{copy}, \mathrm{AND}\}$ and $\Pi_{\mathrm{bip}} = \{\delta_0, \mathrm{Bernoulli}(1/2)\}$, where δ_0 is the point mass at 0; the AND schema applies uniformly at every input size. The rooted-tree SCMs belong to $\mathcal{M}_{n,1}(\{\mathrm{copy}\}, \{\delta_0, \mathrm{Bernoulli}(1/2)\})$.

Theorem 4.6 (Degree-sensitive upper bound)**.** *For every fixed finite gate-schema library Γ and finite noise library Π, and every $1 \le d \le n-1$, there exists a constant $C_{\Gamma,\Pi}$ such that every $M \in \mathcal{M}_{n,d}(\Gamma, \Pi)$ satisfies*

$$\mathrm{DL}_2(M) \;\le\; C_{\Gamma,\Pi} \cdot n + n \log_2\Big(\sum_{k=0}^{d} \binom{n-1}{k}\Big) + O(n \log n).$$

In particular $\mathrm{DL}_2(M) = O(nd \log \frac{en}{d} + n \log n)$, and consequently $\Delta_{2|1}(M)$ obeys the same bound.

Proof sketch. An SCM description encodes a topological order, parent sets, and gate/noise choices; a fixed program then computes its intervention answers. The inequality $\sum_{k=0}^{d} \binom{n-1}{k} \le (d+1)(e(n-1)/d)^d$ for $1 \le d \le n-1$ gives the asymptotic bound.

Corollary 4.7 (Optimality of the tree and bipartite constructions)**.** *Every class $\mathcal{M}_{n,d}(\Gamma, \Pi)$ has gap at most $O(n \log n)$ when $d = O(1)$ and $O(n^2)$ when $d = \Theta(n)$. The rooted-tree and bipartite constructions, which lie in such classes, attain these rates by Theorems 4.1 and 4.4, so they are order-optimal within finite-gate-schema SCMs of maximum indegree d in the bounded-degree and dense regimes.*

Corollary 4.8 (The same rate for every oracle computable from the SCM)**.** *For every $M \in \mathcal{M}_{n,d}(\Gamma, \Pi)$ and every fixed query class Q with a finite answer computable from a description of M, including $\mathsf{Int}_{\mathrm{all}}$ and CF_1, $\mathrm{DL}(Q; M) = O(nd \log \frac{en}{d} + n \log n)$. For every SCM, $K(\mathsf{Int}_{\mathrm{all}} \mid \mathsf{Obs}, n) \ge K(\mathsf{Int}_1 \mid \mathsf{Obs}, n) - O(1)$, so all lower bounds for Int_1 transfer: $\Delta(\mathsf{Int}_{\mathrm{all}} \mid \mathsf{Obs})$ is $\Theta(n^2)$ for all but an $\exp(-\Omega(n^2))$ fraction of bipartite graphs and $\Theta(n \log n)$ for all but an exponentially small fraction of trees.*

Without the library restriction, the gap can be unbounded even for fixed n and d (Remark B.6).

4.4 Finite-Precision Robustness

We now allow total-variation error in the specified intervention distributions.

Definition 4.9 (Approximate query-class description length)**.** Let d_Q be a metric on answer objects for query class Q. For $\varepsilon \geq 0$, define

$$K_\varepsilon\big(\mathrm{Ans}_Q(M) \mid w\big) := \min\big\{|p| \ : \ d_Q\big(U(p,w), \mathrm{Ans}_Q(M)\big) \leq \varepsilon\big\},$$

where U is the fixed universal machine and w is auxiliary input. The minimum is over halting programs outputting valid encoded answer objects. The ε-approximate description gap is $\Delta_\varepsilon(Q_2 \mid Q_1; M) := K_\varepsilon(\mathrm{Ans}_{Q_2}(M) \mid \mathrm{Ans}_{Q_1}(M), n)$. For interventional answer objects we use the metric

$$d_{\mathsf{Int}}(I, I') := \max\left\{ \mathrm{TV}(P_I, P_{I'}),\ \max_{i\in[n],\, b\in\{0,1\}} \mathrm{TV}\big(P_I^{\mathrm{do}(X_i=b)},\ P_{I'}^{\mathrm{do}(X_i=b)}\big) \right\},$$

where TV denotes total variation distance and $P_I, P_{I'}$ are the observational components of I, I'.

Theorem 4.10 (Approximate quadratic separation)**.** *Consider the bipartite family $\mathcal{M}^n_{\mathrm{bip}}$ with $n = 2m+1$. For every fixed $0 \leq \varepsilon < 1/4$:*

1. ***Upper bound.*** *For every G, $\Delta_\varepsilon(\mathsf{Int}_1 \mid \mathsf{Obs};\ M_G) \leq m^2 + O(1)$.*
2. ***Lower bound (existence).*** *There exists $G \subseteq A \times B$ such that*
$$\Delta_\varepsilon(\mathsf{Int}_1 \mid \mathsf{Obs};\ M_G) \ \geq \ m^2 - O(1).$$
3. ***Lower bound (high probability).*** *For uniformly random $G \sim \mathrm{Unif}(2^{A\times B})$ and every $c \geq 0$,*
$$\Pr\big[\, \Delta_\varepsilon(\mathsf{Int}_1 \mid \mathsf{Obs};\ M_G) \geq m^2 - c - O(1) \,\big] \ \geq \ 1 - 2^{-c}.$$

The quadratic information requirement therefore survives constant error in every intervention distribution.

Proof sketch. Exact encoding gives the upper bound. If $G \neq G'$, some edge differs, so the corresponding intervention gives marginal probabilities differing by $1/2$. Their answer objects are therefore at least $1/2$ apart in d_{Int}, making the ε-balls disjoint for $\varepsilon < 1/4$. Counting conditional on $(P^\star, n)$ gives the lower bounds.

5 Counterfactual Separations: Modular XOR and Masked Parity

Section 4 quantified what observations leave unspecified about interventions. We now construct models that agree on every hard-intervention distribution but differ in counterfactual answers. Modular XOR gives a linear gap; varying parent sets gives larger gaps. Full proofs for this section are in Appendix C.

5.1 The Modular-XOR Family

Consider a pair with a fair-bit input and an output that either equals independent fair noise or XORs that noise with the input. Both mechanisms give identical hard-intervention distributions. With the noise held fixed, changing the input leaves the output unchanged in the first mechanism and flips it in the second.

Definition 5.1 (Modular-XOR SCM)**.** Let $m \geq 1$ and $n = 2m$. For a hidden string $s \in \{0,1\}^m$, the SCM M_s has m independent modules (X_t, Y_t) with i.i.d. exogenous $U_{X_t}, U_{Y_t} \sim \mathrm{Bernoulli}(1/2)$ and

$$X_t := U_{X_t}, \qquad Y_t := \begin{cases} U_{Y_t} & \text{if } s_t = 0, \\ X_t \oplus U_{Y_t} & \text{if } s_t = 1. \end{cases}$$

Here $\oplus$ denotes XOR, addition modulo two.

Lemma 5.2 (Modular-XOR invariances)**.** *For all $s, s' \in \{0,1\}^m$:*

(i) $\mathsf{Obs}(M_s)$ *is uniform on* $\{0,1\}^{2m}$.

(ii) $\mathsf{Int}_1(M_s) = \mathsf{Int}_1(M_{s'})$.

(iii) $\mathsf{Int}_{\mathrm{all}}(M_s) = \mathsf{Int}_{\mathrm{all}}(M_{s'})$*, so the counterfactual gap below survives complete knowledge of all hard do-distributions.*

(iv) The map $s \mapsto \mathsf{CF}_1(M_s)$ *is injective.*

Theorem 5.3 (Counterfactual gap after the full hard-do interventional oracle)**.** *Let* $n = 2m$ *and* $\mathcal{M}^n_{\mathrm{xor}} := \{M_s : s \in \{0,1\}^m\}$ *as in Definition 5.1. For every* s, $\mathrm{DL}(\mathsf{Int}_{\mathrm{all}}; M_s) = O(1)$ *and* $\mathrm{DL}_3(M_s) \le m + O(1)$*. For uniform* s *and every* $c \ge 0$*,*

$$\Pr[K(\mathsf{CF}_1(M_s) \mid \mathsf{Int}_{\mathrm{all}}(M_s), n) \ge m - c - O(1)] \ge 1 - 2^{-c}.$$

For uniform s, the gap is $\Theta(n)$ with high probability, even given $\mathsf{Int}_{\mathrm{all}}$; this also holds for $\Delta_{3|2}$.

Proof sketch. Given n, a fixed program generates the common oracle: intervention targets are fixed and all remaining variables are independent fair bits. Lemma 5.2(iv) gives 2^m distinct counterfactual answers. Conditional counting gives the lower bound; encoding s in m bits gives the upper bound.

5.2 Masked Parity: A Degree-Sensitive Counterfactual Gap

Each modular-XOR output has two parent-set choices: empty or its designated input. Allowing any input gives $m + 1$ choices at $d = 1$; allowing sets of up to d inputs increases the ambiguity further. Each output is the parity of its chosen inputs, masked by an independent fair bit.

Let $n = 2m$, $1 \le d \le m$, and $S = (S_1, \ldots, S_m)$ with $S_j \subseteq [m]$ and $|S_j| \le d$. Define M_S by $X_i := U_i$ and $Y_j := \left(\bigoplus_{i \in S_j} X_i\right) \oplus V_j$, with independent fair bits U_i, V_j and the empty XOR equal to 0.

The family $\mathcal{M}^{n,d}_{\mathrm{par}}$ has $N_{m,d} := \left(\sum_{k=0}^{d} \binom{m}{k}\right)^m$ members. It uses only the parity schema and fair-bit noise, and contains modular XOR when $S_j \subseteq \{j\}$.

Theorem 5.4 (Degree-sensitive counterfactual gap)**.** *Let* $L_{m,d} := \log_2 N_{m,d} = \Theta(nd \log \frac{en}{d})$*. Every* M_S *has uniform* $\mathsf{Obs}(M_S)$ *on* $\{0,1\}^{2m}$ *and the same* $\mathsf{Int}_{\mathrm{all}}(M_S)$*, with* $\mathrm{DL}(\mathsf{Int}_{\mathrm{all}}; M_S) = O(1)$*.*

For every S, $\mathrm{DL}_3(M_S) \le L_{m,d} + O(\log n)$*. For uniform* S *and every* $c \ge 0$*,*

$$\Pr[K(\mathsf{CF}_1(M_S) \mid \mathsf{Int}_{\mathrm{all}}(M_S), n) \ge L_{m,d} - c - O(1)] \ge 1 - 2^{-c}.$$

With high probability under uniform S*, the gap is* $\Theta(n \log n)$ *at* $d = 1$ *and* $\Theta(n^2)$ *at* $d = \Theta(n)$*, order-optimal within finite-gate-schema SCMs of maximum indegree* d *by Corollary 4.8.*

Quadratic counterfactual gaps can therefore persist even given every hard-intervention law, and with uniform, full-support observations.

Proof sketch. Private fair noise makes all hard-intervention laws identical across S. Holding the noise fixed gives $Y_j^{(X_i \leftarrow 0)} \oplus Y_j^{(X_i \leftarrow 1)} = \mathbf{1}[i \in S_j]$, so CF_1 determines S. Counting the $N_{m,d}$ distinct answers gives the lower bound; encoding the index of S and d gives the upper bound.

6 Ambiguity-to-Bits: A General Principle

The constructions in Sections 4–5 follow a common principle: count the distinct higher-rung answers compatible with a lower-rung answer. The logarithm of this count, up to an additive constant, lower-bounds the worst-case additional description length. We formalize this principle and its Shannon counterpart, then derive consequences for observational learning and intervention design. Full proofs for this section are in Appendix D.

Definition 6.1 (General ambiguity class)**.** For query classes Q_1, Q_2, a finite SCM family $\mathcal{M}$ on a common set of n variables, and an attained lower-rung answer a, define

$$\mathcal{F}_{Q_2|Q_1}(a; \mathcal{M}) := \{\mathrm{Ans}_{Q_2}(M) : M \in \mathcal{M},\ \mathrm{Ans}_{Q_1}(M) = a\}.$$

Theorem 6.2 (Ambiguity-to-bits, Kolmogorov form)**.** *For any finite family $\mathcal{M}$ on n variables, query classes Q_1, Q_2, and attained lower-rung answer a,*

$$\max_{M\in\mathcal{M}:\,\mathrm{Ans}_{Q_1}(M)=a} K\big(\mathrm{Ans}_{Q_2}(M)\mid a, n\big) \;\geq\; \log_2 |\mathcal{F}_{Q_2|Q_1}(a;\mathcal{M})| - O(1).$$

Moreover, for every $c \geq 0$, all but a 2^{-c} fraction of the distinct strings in $\mathcal{F}_{Q_2|Q_1}(a;\mathcal{M})$ have conditional complexity at least $\log_2 |\mathcal{F}_{Q_2|Q_1}(a;\mathcal{M})| - c - O(1)$.

The probability statement is uniform over distinct higher-rung answers; injectivity transfers it to uniform hidden parameters in our constructions.

Theorem 6.3 (Ambiguity-to-entropy, Shannon form)**.** *Let Θ be a finite-valued random hidden parameter and M_Θ its SCM. Write $L(\Theta) := \mathrm{Ans}_{Q_1}(M_\Theta)$ and $U(\Theta) := \mathrm{Ans}_{Q_2}(M_\Theta)$ for the lower- and higher-rung answers, and let H denote discrete Shannon entropy. If U distinguishes all parameter values consistent with each lower-rung answer, then $\mathsf{H}(U(\Theta) \mid L(\Theta) = \ell) = \mathsf{H}(\Theta \mid L(\Theta) = \ell)$ for every ℓ with $\Pr[L(\Theta) = \ell] > 0$. In particular, if Θ, conditional on $L(\Theta) = \ell$, is uniform over N compatible parameter values, this conditional entropy equals $\log_2 N$.*

Under this condition, identifying the higher-rung answer resolves all remaining uncertainty about the hidden parameter.

In Sections 4–5, the explicit encodings match log ambiguity up to lower-order terms, as summarized in Table 1. Under uniform hidden parameters, the same separation rates therefore hold for Shannon conditional entropy.

Remark 6.4 (Computability)**.** Kolmogorov complexity is uncomputable, and exact ambiguity counting can be hard for implicit families. Our constructions use explicit counts and encodings; efficient enumeration in general remains open.

6.1 Learning-Theoretic Consequence

In the bipartite family, passive data reveal nothing about the hidden graph.

Corollary 6.5 (No-free-lunch for observational learners)**.** *Let G be uniform in the bipartite family and D an i.i.d. observational sample of any size. Their mutual information is $I(G; D) = 0$. Every learner using only D and private randomness independent of G has probability at most 2^{-m^2} of exactly recovering $\mathsf{Int}_1(M_G)$. Given also an independent uniform queried pair $(i, j) \in [m]^2$, every such learner predicting $P_{M_G}^{\mathrm{do}(X_{a_i}=0)}(X_{b_j} = 0)$ has expected absolute error at least $1/4$.*

More observational data cannot reduce this family's causal ambiguity. Benchmarks supplying causal graphs or narratives, such as CLadder [7], give the learner additional information, so Corollary 6.5 does not directly apply to that setting.

6.2 From Bits to Experiments

To translate ambiguity into experiment counts, fix a finite SCM family and a lower-rung answer with N compatible target oracles. Each *experiment* is a hard-intervention setting that returns its exact post-interventional law; let R be the maximum number of distinct returned distributions across compatible models for any one setting, so one setting reveals at most $\log_2 R$ bits. We count settings; estimating each law may require many samples.

Proposition 6.6 (From bits to experiments)**.** *Any adaptive strategy guaranteed to identify the target on every compatible model in at most t settings satisfies $R^t \geq N$. Hence $t \geq \lceil \log_2 N / \log_2 R \rceil$ when $R > 1$, and identification is impossible when $R = 1 < N$.*

In the bipartite family $N = 2^{m^2}$ and $R \leq 2^m$, so m settings are necessary, and the m single-node interventions $\mathrm{do}(X_{a_i} = 0)$ suffice. Adaptivity and multi-node interventions cannot reduce this worst-case setting count. In the modular-XOR and masked-parity families $R = 1$, so no number of hard-intervention experiments identifies the counterfactual oracle.

Proof sketch. A depth-t strategy has at most R^t outcome sequences. A hard intervention on a bipartite SCM sees G only through one bit per non-intervened layer-B node. In the XOR and parity families, all hard-intervention laws coincide, so these experiments cannot resolve the counterfactual ambiguity.

7 How Assumptions Change the Interventional Gap

Experiments can reduce ambiguity by supplying new distributions; assumptions can reduce it by restricting the compatible models. We now examine graph knowledge, positivity, faithfulness, and shared latent causes. Full statements and proofs are in Appendix E.

Proposition 7.1 (What assumptions buy)**.** *Parts 1–2 concern binary SCMs with independent local noises (Definition 3.1) and do not require fixed finite gate libraries. Part 3 permits shared latent causes.*

1. ***Graph knowledge.*** *In the tree and bipartite families, $\Delta_{2|1}(M_G) = K(G \mid n) \pm O(1)$. More generally, a known graph G and positive probability for every state give $K(\mathsf{Int}_{\mathrm{all}} \mid \mathsf{Obs}, G, n) = O(1)$. Without positivity, this conclusion can fail.*
2. ***Positivity and faithfulness.*** *If $P = \mathsf{Obs}(M)$ is strictly positive and faithful to the graph, let $\mathcal{E}(P)$ be the set of DAGs implying exactly the conditional independences of P. Then $K(\mathsf{Int}_{\mathrm{all}} \mid \mathsf{Obs}, n) \le \log_2 |\mathcal{E}(P)| + O(\log n) \le \log_2(n!) + O(\log n)$. For generic positive rational P (outside a proper algebraic exceptional set), the $n!$ orderings yield faithful SCMs on complete DAGs sharing P but with distinct Int_1 oracles. Under uniform orderings, their conditional Shannon entropy is $\log_2(n!)$ and their conditional complexity is $\Theta(n \log n)$ with high probability.*
3. ***Confounding.*** *We construct 2^m SCMs on $n = 2m$ variables, indexed by m-bit strings, using independent finite latent roots shared by pairs of variables. All have the same uniform observations and observed mixed graph $\mathcal{G}$, which records directed edges and pairs sharing a latent cause. For every hidden string, $K(\mathsf{Int}_1 \mid \mathsf{Obs}, \mathcal{G}, n) \le m + O(1)$. Under a uniform hidden string, $\mathsf{H}(\mathsf{Int}_1 \mid \mathsf{Obs}, \mathcal{G}) = m$ and, for every $c \ge 0$, the Kolmogorov gap is at least $m - c - O(1)$ with probability at least $1 - 2^{-c}$. The lower bound also holds for ε-accurate answers in the total-variation metric of Definition 4.9 when $0 \le \varepsilon < 1/8$.*

A known graph identifies interventions under positivity and independent noise, while unknown orientations or shared latent causes can leave growing gaps.

Linear models. The $\Theta(n \log n)$ worst-case gap also occurs in faithful zero-mean linear Gaussian models with rational covariance matrices. Computable identification from a finite observational encoding reduces the gap to $O(1)$. Linear Gaussian models with independent errors of equal positive variance satisfy these conditions under rational covariance encoding [13]. Linear non-Gaussian acyclic models (LiNGAM) with independent errors are identifiable [16]; the same description-gap conclusion additionally requires a finite encoding and computable recovery.

8 Conclusion

We quantify the additional information needed to move between levels of Pearl's causal hierarchy. Our binary SCM families exhibit $\Theta(n^2)$ gaps from observations to interventions and from all hard-intervention laws to counterfactuals, in the worst case and with high probability under uniform hidden parameters. The former starts from an $O(1)$-bit observational description, survives total-variation error $\varepsilon < 1/4$, and is order-optimal within finite-gate-schema SCMs of maximum indegree $d = \Theta(n)$. General ambiguity bounds and matching encodings explain these gaps in Kolmogorov and Shannon terms. Experiment counts also depend on how many distinct distributions each setting can return. A known graph with positivity closes the interventional gap; positivity and faithfulness alone leave $\Theta(n \log n)$ bits; shared latent causes keep $\Theta(n)$ bits even given the observed graph. Further work could extend these bounds to nonlinear additive-noise and smooth models and to soft and stochastic interventions. Finite-precision identification stability and efficient ambiguity counting also warrant study.

Acknowledgments and Disclosure of Funding

The author thanks the anonymous reviewers and the area chair for constructive comments that improved the results and presentation. This work received no third-party funding. The author declares no competing interests.

References

[1] Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On Pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 507–556. ACM Books, 2022.

[2] Julian Dörfler, Benito van der Zander, Markus Bläser, and Maciej Liskiewicz. From probability to counterfactuals: The increasing complexity of satisfiability in Pearl's causal hierarchy. In *International Conference on Learning Representations (ICLR)*, 2025.

[3] Frederick Eberhardt, Clark Glymour, and Richard Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among N variables. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 178–184. AUAI Press, 2005.

[4] Peter D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

[5] Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464, 2012.

[6] Duligur Ibeling and Thomas Icard. Comparing causal frameworks: Potential outcomes, structural models, graphs, and abstractions. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[7] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. CLadder: Assessing causal reasoning in language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[8] Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2024.

[9] Andrey N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1):1–7, 1965.

[10] Ming Li and Paul Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, 3rd edition, 2008.

[11] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

[12] Judea Pearl and Dana Mackenzie. *The Book of Why*. Basic Books, 2018.

[13] Jonas Peters and Peter Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

[14] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

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

[16] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

[17] Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

[18] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.

A Proofs for Section 3

We prove Lemmas 3.7, 3.8, and 3.9 in their main-text order, using the answer objects and description lengths of Definitions 3.4 and 3.6.

Proof of Lemma 3.7. For any fixed w, there are at most $2^k - 1$ programs of length less than k (summing over lengths $0, 1, \ldots, k-1$). Setting $k = \lfloor \log_2 N \rfloor$ shows that fewer than N strings can have conditional complexity $K(\cdot \mid w)$ below $\log_2 N - O(1)$, so at least one must satisfy $K(z_i \mid w) \geq \log_2 N - O(1)$. For part (2), at most $2^{\log_2 N - c} = N \cdot 2^{-c}$ strings can have conditional complexity below $\log_2 N - c - O(1)$, so the remaining fraction $\geq 1 - 2^{-c}$ satisfies the bound. □

Proof of Lemma 3.8. Direct from Definition 3.4: $\mathsf{Obs}(M)$ is a component of $\mathsf{Int}_1(M)$, and $\mathsf{Int}_1(M)$ is a marginal of $\mathsf{CF}_1(M)$. Both reductions are computable. □

Proof of Lemma 3.9. The upper bound $K(\mathsf{Int}_1 \mid \mathsf{Obs}, n) \leq K(\mathsf{Int}_1 \mid n) + O(1)$ holds by ignoring the auxiliary input. For the lower bound, concatenate a program of uniformly bounded length that outputs $\mathsf{Obs}(M)$ from n with a program that outputs $\mathsf{Int}_1(M)$ from $(\mathsf{Obs}(M), n)$: this yields $K(\mathsf{Int}_1 \mid n) \leq K(\mathsf{Int}_1 \mid \mathsf{Obs}, n) + O(1)$. □

B Proofs for Section 4

This appendix contains all full proofs for Section 4, in the order of the main-text results.

B.1 Rooted Trees

We give the formal construction and full proofs underlying the $\Theta(n \log n)$ rooted-tree separation. Theorem B.5 below restates and proves the main-text Theorem 4.1; together with Corollary 4.7 it shows the bound is order-optimal within finite-gate-schema SCMs of maximum indegree $d = O(1)$.

Intuition. Consider n variables that are always perfectly correlated: either all 0 or all 1, with equal probability. A passive observer sees only this correlation. But *how* does the correlation arise?

Imagine the variables arranged in a tree, with one root variable X_r that is uniformly random, and all other variables copying their parent. The tree structure determines the *direction of information flow*, but this structure is invisible to passive observation: all trees produce the same "all equal" distribution.

Interventions reveal the hidden structure. If we force $X_i = 0$, then:

- All *descendants* of i in the tree become 0 (they copy from ancestors, and i is now 0).
- All *non-descendants* remain random (they don't depend on i).

The interventional distributions reveal the descendant structure, which determines the tree. Cayley's formula gives n^{n-1} rooted labeled trees, all observationally identical but interventionally distinct. The counting argument below therefore gives a $\Theta(n \log n)$ gap in the worst case and with high probability under a uniform tree.

Formal construction. For a rooted tree T on $[n]$ with root r, we write $\mathrm{pa}_T(v)$ for the parent of v in T (defined for all $v \neq r$), and $\mathrm{Desc}_T(i)$ for the set of *descendants* of node i in T, including i itself. That is, $j \in \mathrm{Desc}_T(i)$ if and only if i lies on the unique path from j to r.

Definition B.1 (Rooted-Tree SCM)**.** Let T be a rooted labeled tree on vertex set $[n]$ with root r. Define the SCM M_T:

- **Exogenous variables:** $U_r \sim \mathrm{Bernoulli}(1/2)$. All other U_i are constant (say, 0) and play no role.
- **Structural equations:**

$$X_r := U_r,$$
$$X_v := X_{\mathrm{pa}_T(v)} \quad \text{for each } v \neq r,$$

  where $\mathrm{pa}_T(v)$ denotes the parent of v in the rooted tree T.

Only the root label and parent links vary across this family.

Lemma B.2 (All Trees Have the Same Observations)**.** *For every rooted tree T, the observational distribution of M_T is:*

$$P^\star(x_1, \ldots, x_n) = \begin{cases} 1/2 & \text{if } x_1 = x_2 = \cdots = x_n = 0, \\ 1/2 & \text{if } x_1 = x_2 = \cdots = x_n = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, $\mathrm{DL}_1(M_T) = K(P^\star \mid n) = O(1)$.

Passive observations are identical for every rooted tree.

Proof. Since T is a connected tree and every non-root node copies its parent, any value assigned to the root propagates to all nodes. Specifically:

- If $U_r = 0$: Then $X_r = 0$. For any node v at distance 1 from r, we have $X_v = X_{\mathrm{pa}_T(v)} = X_r = 0$. By induction on distance from r, all nodes have value 0.
- If $U_r = 1$: By the same argument, all nodes have value 1.

Since $U_r \sim \mathrm{Bernoulli}(1/2)$, the outcome is 0^n with probability $1/2$ and 1^n with probability $1/2$.

The distribution $P^\star$ can be specified by a constant-length program: "output $1/2$ for 0^n and 1^n, output 0 otherwise." Thus $K(P^\star \mid n) = O(1)$. □

Lemma B.3 (Interventions Reveal Descendants)**.** *For any node $i \in [n]$ and the intervention $\mathrm{do}(X_i = 0)$:*

$$P_{M_T}^{\mathrm{do}(X_i=0)}(X_j = 0) = \begin{cases} 1 & \text{if } j \in \mathrm{Desc}_T(i), \\ 1/2 & \text{if } j \notin \mathrm{Desc}_T(i), \end{cases}$$

where $\mathrm{Desc}_T(i)$ is the set of descendants of i in T (including i itself).

An intervention identifies which variables descend from its target.

Proof. Under $\mathrm{do}(X_i = 0)$, the equation for X_i becomes $X_i := 0$. All other equations remain unchanged.

Case 1: $j \in \mathrm{Desc}_T(i)$**.** Every descendant of i lies on a path from i going away from the root. Along this path, each node copies its parent. Since $X_i = 0$ and descendants copy from ancestors (with i being an ancestor), we have $X_j = 0$ deterministically.

Case 2: $j \notin \mathrm{Desc}_T(i)$**.** If j is not a descendant of i, then the path from the root r to j does not pass through i. The value of X_j depends only on nodes along this path, which are unaffected by the intervention on X_i. Thus X_j still equals $X_r = U_r \sim \mathrm{Bernoulli}(1/2)$, so $P(X_j = 0) = 1/2$. □

Lemma B.4 (Descendant Sets Determine the Tree)**.** *The collection of descendant sets $\{\mathrm{Desc}_T(i) : i \in [n]\}$ uniquely determines the rooted tree T.*

The descendant relation determines both the root and every parent edge.

Proof. We show how to reconstruct T from the descendant sets.

Finding the root: The root r is the unique node with $\mathrm{Desc}_T(r) = [n]$ (every node is a descendant of the root).

Finding parent-child relationships: For any non-root node v, its parent $\mathrm{pa}_T(v)$ is characterized as follows. Consider all nodes $u \neq v$ such that $v \in \mathrm{Desc}_T(u)$ (i.e., u is an ancestor of v). Among these, $\mathrm{pa}_T(v)$ is the one with the *smallest* descendant set.

To see why: along the unique path from v to the root r, descendant sets strictly increase:

$$\mathrm{Desc}_T(v) \subsetneq \mathrm{Desc}_T(\mathrm{pa}_T(v)) \subsetneq \mathrm{Desc}_T(\mathrm{pa}_T(\mathrm{pa}_T(v))) \subsetneq \cdots \subsetneq \mathrm{Desc}_T(r) = [n].$$

The strict inclusions hold because each step toward the root adds at least the parent to the descendant set. Thus the parent has the smallest descendant set among all ancestors.

This procedure reconstructs all parent-child edges, hence the entire tree. □

Theorem B.5 (Rooted-Tree Separation: $\Theta(n \log n)$)**.** *Let $\mathcal{M}^n_{\text{tree}}$ be the family of rooted-tree SCMs M_T of Definition B.1, over all rooted labeled trees T on $[n]$. Then:*

1. *The map $T \mapsto \mathsf{Int}_1(M_T)$ is injective on $\mathcal{M}^n_{\text{tree}}$.*
2. ***Lower bound:** There exists T with $\Delta_{2|1}(M_T) \geq (n-1)\log_2 n - O(1)$.*
3. ***Upper bound:** For all T, $\mathrm{DL}_2(M_T) \leq (n-1)\log_2 n + O(\log n)$.*
4. ***High probability:** For a uniformly random rooted labeled tree T and every $c \geq 0$,*

$$\mathbb{P}\big[\Delta_{2|1}(M_T) \geq (n-1)\log_2 n - c - O(1)\big] \geq 1 - 2^{-c}.$$

Consequently, $\Delta_{2|1}(M_T) = \Theta(n \log n)$ for all but an exponentially small fraction of trees.

Uniformly selected trees typically require a description of order $n \log n$.

Proof. **Part 1 (Injectivity):** By Lemma B.3, from $\mathsf{Int}_1(M_T)$ we can read off:

$$\mathrm{Desc}_T(i) = \{j \in [n] : P^{\mathrm{do}(X_i=0)}_{M_T}(X_j = 0) = 1\}$$

for each i. By Lemma B.4, the descendant sets determine T. Thus different trees yield different interventional families.

Part 2 (Lower bound): By Cayley's formula, there are n^{n-1} rooted labeled trees on $[n]$. (Cayley's formula gives n^{n-2} labeled unrooted trees on $[n]$; choosing a root from n vertices gives n^{n-1} rooted labeled trees.) By injectivity, there are n^{n-1} distinct elements in $\{\mathsf{Int}_1(M_T) : T \in \mathcal{M}^n_{\text{tree}}\}$.

By the counting bound for Kolmogorov complexity: among N distinct strings, at least one has complexity $\geq \log_2 N - O(1)$. Applying this with $N = n^{n-1}$:

$$\max_T K(\mathsf{Int}_1(M_T) \mid n) \geq \log_2(n^{n-1}) - O(1) = (n-1)\log_2 n - O(1).$$

By Lemma B.2, $\mathrm{DL}_1(M_T) = O(1)$ for all T. By Lemma 3.9, $\Delta_{2|1}(M_T) = \mathrm{DL}_2(M_T) \pm O(1)$.

Part 3 (Upper bound): For $n \geq 2$, encode the base-n rank of the Prüfer sequence in $\lceil (n-2)\log_2 n \rceil$ bits, plus $O(\log n)$ bits for the root. The case $n = 1$ has one constant-complexity oracle. Total: $(n-1)\log_2 n + O(\log n)$ bits.

A constant-length program can then simulate M_T and output $\mathsf{Int}_1(M_T)$. Thus $K(\mathsf{Int}_1(M_T) \mid n) \leq (n-1)\log_2 n + O(\log n)$.

Part 4 (High probability): By part (2) of the counting bound (Lemma 3.7), among the n^{n-1} distinct strings $\{\mathsf{Int}_1(M_T)\}$, all but a 2^{-c} fraction satisfy $K(\mathsf{Int}_1(M_T) \mid n) \geq (n-1)\log_2 n - c - O(1)$. Transferring to $\Delta_{2|1}$ via Lemma 3.9 incurs an additional $O(1)$ term. □

B.2 Bipartite Graphs

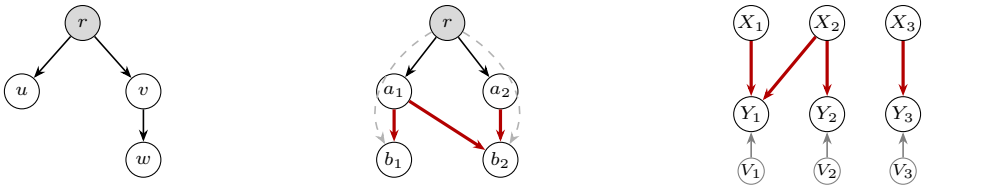
Figure 1 places the rooted-tree, bipartite, and masked-parity constructions side by side; Figure 2 illustrates Definition 4.2; Figure 3 compares the bipartite and rooted-tree rates. We then prove the decoding lemma and quadratic separation.

Proof of Lemma 4.3(i). We analyze the two cases for the root's value.

Case $U_r = 0$: Then $X_r = 0$. Since each $X_{a_i} = X_r$, we have $X_{a_i} = 0$ for all i. For each b_j, the equation is:

$$X_{b_j} = X_r \wedge \bigwedge_{a_i:(a_i,b_j)\in G} X_{a_i} = 0 \wedge (\cdots) = 0.$$

Thus all variables are 0.

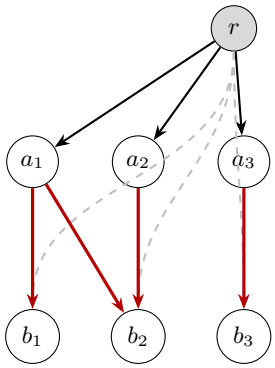


(a) rooted tree
$\mathrm{do}(X_v{=}0)$: $w \to 0$, u stays random

(b) bipartite
$\mathrm{do}(X_{a_1}{=}0)$: $b_1, b_2 \to 0$, reveals row 1

(c) masked parity: unclamped bits stay fair
$Y_j^{(X_i \leftarrow 0)} \oplus Y_j^{(X_i \leftarrow 1)} = \mathbf{1}[i \in S_j]$

Figure 1: **Three hidden-parameter constructions.** In (b) and (c), red edges mark the hidden structure. (a) Rooted tree: the root is a fair bit and every other node copies its parent, so observations see only "all equal" while single-node interventions reveal descendant sets. (b) Bipartite: each layer-B node is the AND of the root and its hidden neighbors in layer A, so $\mathrm{do}(X_{a_i}{=}0)$ reveals row i of the adjacency matrix. (c) Masked parity: each Y_j is the parity of a hidden set S_j of inputs masked by a private fair bit V_j, so unclamped coordinates remain independent fair bits under hard interventions, while counterfactuals with the noise held fixed reveal the sets.



Observations:
All variables equal.
$P(0^n) = P(1^n) = 1/2$

$\mathrm{DL}_1(M_G) = O(1)$

Intervention $\mathrm{do}(X_{a_1} = 0)$**:**
$b_1, b_2 \to 0$ (edges from a_1)
$b_3 \to$ random (no edge)

Reveals $N_G(a_1) = \{b_1, b_2\}$

Red edges = hidden graph G (revealed by interventions)

Figure 2: **The Bipartite-Graph Construction.** Each layer-A node copies the root r; each layer-B node takes the AND of the root and its hidden layer-A neighbors (red edges). Observationally, all 2^{m^2} choices of G produce the same distribution. Interventions on layer A reveal which layer-B nodes are neighbors.

Case $U_r = 1$**:** Then $X_r = 1$ and $X_{a_i} = 1$ for all i. For each b_j:

$$X_{b_j} = X_r \wedge \bigwedge_{a_i:(a_i,b_j)\in G} X_{a_i} = 1 \wedge 1 \wedge \cdots \wedge 1 = 1.$$

(If b_j has no neighbors, the empty AND is 1, so $X_{b_j} = 1 \wedge 1 = 1$.)

Thus all variables are 1.

Since $U_r \sim \mathrm{Bernoulli}(1/2)$, the outcome is 0^n or 1^n each with probability $1/2$. This is $P^\star$, independent of G. □

Proof of Lemma 4.3(ii). Under $\mathrm{do}(X_{a_i} = 0)$, the equation for X_{a_i} becomes $X_{a_i} := 0$. The root $X_r = U_r$ remains random, and all other layer-A nodes satisfy $X_{a_k} = X_r$ for $k \neq i$.

Case $(a_i, b_j) \in G$**:** The equation for X_{b_j} includes X_{a_i} as an AND input:

$$X_{b_j} = X_r \wedge X_{a_i} \wedge \bigwedge_{a_k \neq a_i:(a_k,b_j)\in G} X_{a_k}.$$

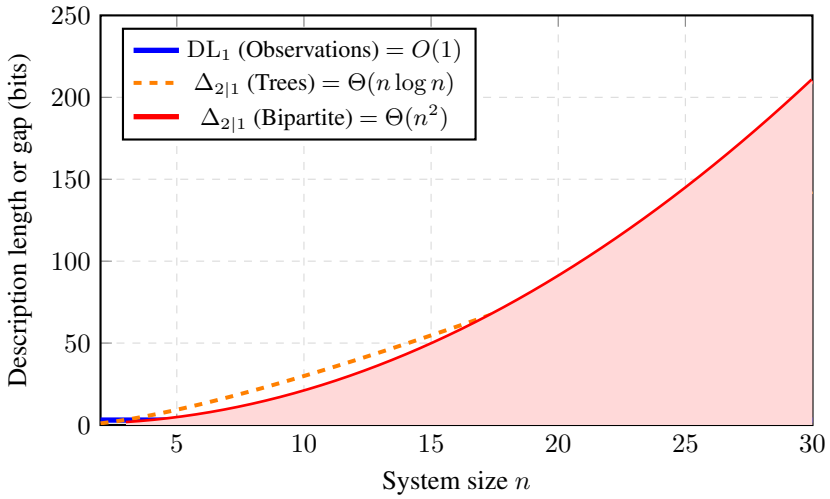


Figure 3: **Illustrative leading-order gap rates.** The constant blue line and the tree and bipartite curves depict asymptotic rates, omitting encoding constants and lower-order terms. They are not computed Kolmogorov complexities of individual models; the gap rates hold in the worst case and with high probability under uniform hidden graphs.

Since $X_{a_i} = 0$, the AND evaluates to 0 regardless of other inputs. Thus $X_{b_j} = 0$ deterministically, so $P(X_{b_j} = 0) = 1$.

Case $(a_i, b_j) \notin G$**:** The equation for X_{b_j} does not include X_{a_i}:

$$X_{b_j} = X_r \wedge \bigwedge_{a_k : (a_k, b_j) \in G} X_{a_k}.$$

All terms in this AND equal X_r (since $X_{a_k} = X_r$ for $k \neq i$), so

$$X_{b_j} = X_r \wedge X_r \wedge \cdots \wedge X_r = X_r.$$

This also holds for an empty neighbor set, since the empty AND is 1.

Thus $X_{b_j} = X_r \sim \mathrm{Bernoulli}(1/2)$, so $P(X_{b_j} = 0) = 1/2$. □

Proof of Lemma 4.3(iii). By Lemma 4.3(ii), from $\mathsf{Int}_1(M_G)$ we can extract:

$$N_G(a_i) = \{b_j \in B : P_{M_G}^{\mathrm{do}(X_{a_i}=0)}(X_{b_j} = 0) = 1\}$$

for each $a_i \in A$. The neighborhoods $\{N_G(a_i)\}_{i=1}^m$ determine all edges of G:

$$(a_i, b_j) \in G \iff b_j \in N_G(a_i).$$

Thus $\mathsf{Int}_1(M_G)$ determines G uniquely. □

Proof of Theorem 4.4. **Lower bound:** There are 2^{m^2} bipartite graphs $G \subseteq A \times B$ (each of the m^2 potential edges is present or absent). By Lemma 4.3(iii), there are 2^{m^2} distinct interventional families. By Lemma 4.3(i), all share $\mathsf{Obs}(M_G) = P^\star$ with $\mathrm{DL}_1 = O(1)$.

By the counting bound: among 2^{m^2} distinct strings, at least one has Kolmogorov complexity $\geq m^2 - O(1)$. By Lemma 3.9, $\Delta_{2|1}(M_G) = \mathrm{DL}_2(M_G) \pm O(1)$.

The high-probability statement follows from the stronger counting bound: all but a 2^{-c} fraction have complexity $\geq m^2 - c - O(1)$.

Upper bound: The graph G can be encoded as an $m \times m$ binary adjacency matrix, requiring m^2 bits. Given G and n, a constant-length program simulates M_G and outputs $\mathsf{Int}_1(M_G)$. Thus $K(\mathsf{Int}_1(M_G) \mid n) \leq m^2 + O(1)$. □

B.3 Degree-Sensitive Optimality

Proof of Theorem 4.6. Parent inputs follow the canonical order of Definition 4.5, so no input permutation needs encoding. For the stated library memberships, take $\mathrm{copy}_0(u) = u$, $\mathrm{copy}_k(x, u) = x_1$ for $k \geq 1$, and $\mathrm{parity}_k(x, u) = (\bigoplus_{i=1}^{k} x_i) \oplus u$, including $k = 0$. An SCM in $\mathcal{M}_{n,d}(\Gamma, \Pi)$ can be encoded by listing:

1. a topological ordering of the n variables, using $O(n \log n)$ bits;
2. for each variable, its parent set, chosen from at most $\sum_{k=0}^{d} \binom{n-1}{k}$ possibilities;
3. for each variable, a gate from the fixed library Γ and a noise distribution from the fixed library Π, using $O(1)$ bits per variable.

The entire SCM thus has a description of length

$$O(n \log n) + n \log_2\Big(\sum_{k=0}^{d} \binom{n-1}{k}\Big) + O_{\Gamma,\Pi}(n).$$

Given this description and n, a fixed program can compute all single-node interventional distributions and output $\mathsf{Int}_1(M)$. Hence the same expression upper-bounds $K(\mathsf{Int}_1(M) \mid n) = \mathrm{DL}_2(M)$.

Finally,

$$\sum_{k=0}^{d} \binom{n-1}{k} \;\leq\; (d+1) \left(\frac{e(n-1)}{d}\right)^{d}$$

for $1 \leq d \leq n-1$, yielding the stated asymptotic form. Since $\Delta_{2|1}(M) \leq \mathrm{DL}_2(M) + O(1)$, the same upper bound applies to the gap. □

Proof of Corollary 4.7. Substitute $d = O(1)$ and $d = \Theta(n)$ into Theorem 4.6; the rooted-tree and bipartite constructions lie in $\mathcal{M}_{n,1}$ and $\mathcal{M}_{n,m+1}$ under the gate-schema libraries identified after Definition 4.5, realizing the rates via Theorems 4.1 and 4.4. □

Proof of Corollary 4.8. The encoding in the proof of Theorem 4.6 describes the SCM M itself: its topological order, parent sets, gate choices, and noise distributions. From this description and n, a fixed program simulates M under any hard intervention $\mathrm{do}(X_S = x_S)$ and jointly across all required worlds using the same noise realization, so it outputs $\mathsf{Int}_{\mathrm{all}}(M)$ and $\mathsf{CF}_1(M)$, and more generally $\mathrm{Ans}_Q(M)$ for any query class Q whose answer is computable from M. Hence $\mathrm{DL}(Q; M)$ obeys the same bound. For the transfer inequality, a fixed program maps $\mathsf{Int}_{\mathrm{all}}(M)$ to $\mathsf{Int}_1(M)$ by selecting the single-node components, so $K(\mathsf{Int}_1 \mid \mathsf{Obs}, n) \leq K(\mathsf{Int}_{\mathrm{all}} \mid \mathsf{Obs}, n) + O(1)$. The bipartite statement combines the transferred lower bound of Theorem 4.4 (with $c = m^2/2$) and the upper bound at $d = m + 1 = \Theta(n)$, which is $O(n^2)$; the rooted-tree statement combines the transferred lower bound of Theorem 4.1 and the upper bound at $d = 1$, which is $O(n \log n)$. □

Remark B.6 (Why a bounded library is necessary)**.** The restriction to finite libraries Γ and Π cannot be dropped, even with n and d fixed. Take a bipartite SCM M_G with an edge $(a_i, b_j) \in G$ and let the response of X_{b_j} on the configuration $(X_r, X_{a_i}) = (1, 0)$, which never occurs observationally, be $\mathrm{Bernoulli}(q)$ for an arbitrary rational q instead of 0; call the result $M_{G,q}$. Its observational oracle is still $P^\star$, but under $\mathrm{do}(X_{a_i} = 0)$ the probability of $X_{b_j} = 1$ equals $q/2$, so the single-node interventional oracle reveals q and the gap is at least $K(q) - O(1)$, which is unbounded over rational q. Hence no gap bound depending only on n and d exists over unrestricted SCM classes. For any finite SCM family, the log-ambiguity lower bound of Theorem 6.2 still applies; equality up to lower-order terms requires a matching encoding of the relevant ambiguity class, as provided by the explicit codes for our constructions.

B.4 Finite-Precision Robustness

Proof of Theorem 4.10. **Upper bound.** An exact m^2-bit encoding of G as an $m \times m$ binary adjacency matrix lets a fixed program output $\mathsf{Int}_1(M_G)$ exactly, hence also ε-approximately for any $\varepsilon \geq 0$. So $K_\varepsilon(\mathsf{Int}_1(M_G) \mid \mathsf{Obs}(M_G), n) \leq m^2 + O(1)$.

Lower bound. Let $G \neq G'$ be two distinct bipartite graphs. They differ on some edge (a_i, b_j); WLOG $(a_i, b_j) \in G$ and $(a_i, b_j) \notin G'$. By Lemma 4.3(ii),

$$P_{M_G}^{\mathrm{do}(X_{a_i}=0)}(X_{b_j} = 0) = 1, \qquad P_{M_{G'}}^{\mathrm{do}(X_{a_i}=0)}(X_{b_j} = 0) = 1/2.$$

Total variation distance does not increase under marginalization, so the full interventional distributions satisfy

$$\mathrm{TV}\big(P_{M_G}^{\mathrm{do}(X_{a_i}=0)},\ P_{M_{G'}}^{\mathrm{do}(X_{a_i}=0)}\big) \ \geq \ \tfrac{1}{2}.$$

Hence $d_{\mathsf{Int}}(\mathsf{Int}_1(M_G), \mathsf{Int}_1(M_{G'})) \geq 1/2$. For $\varepsilon < 1/4$, the closed ε-balls around the 2^{m^2} answer objects are pairwise disjoint, so any ε-accurate description must still distinguish among 2^{m^2} possibilities. Applying the conditional counting bound (Lemma 3.7, part 2) with the common observation $P^\star$ as side information yields the existence and high-probability lower bounds. □

C Proofs for Section 5

We give full proofs for Section 5, following the main-text order.

C.1 The Modular-XOR Family

We prove parts (i)–(iv) of Lemma 5.2 in order, followed by Theorem 5.3.

Proof of Lemma 5.2(i). We show each module (X_t, Y_t) is uniformly distributed on $\{0,1\}^2$.

Case $s_t = 0$: $X_t = U_{X_t}$ and $Y_t = U_{Y_t}$ are independent $\mathrm{Bernoulli}(1/2)$ variables.

Case $s_t = 1$: $X_t = U_{X_t} \sim \mathrm{Bernoulli}(1/2)$. For $Y_t = X_t \oplus U_{Y_t}$: for any fixed $x \in \{0,1\}$,

$$\mathbb{P}(Y_t = y \mid X_t = x) = \mathbb{P}(U_{Y_t} = x \oplus y) = 1/2.$$

Thus Y_t is uniform and independent of X_t. (Formally: $\mathbb{P}(X_t = x, Y_t = y) = \mathbb{P}(X_t = x)\mathbb{P}(Y_t = y) = 1/4$.)

Since modules are independent and each is uniform on $\{0,1\}^2$, the full distribution is uniform on $\{0,1\}^{2m}$. □

Proof of Lemma 5.2(ii). Consider any single-variable intervention $\mathrm{do}(X_t = x)$ or $\mathrm{do}(Y_t = y)$.

Intervention $\mathrm{do}(X_t = x)$: The equation $X_t := x$ replaces whatever X_t was. In module t:

- If $s_t = 0$: $Y_t = U_{Y_t}$, uniform.
- If $s_t = 1$: $Y_t = x \oplus U_{Y_t}$, also uniform (since U_{Y_t} is uniform).

Either way, Y_t is uniform and independent of the intervention value. Other modules are unaffected.

Intervention $\mathrm{do}(Y_t = y)$: The equation $Y_t := y$ replaces whatever Y_t was. In both cases ($s_t = 0$ or 1), $X_t = U_{X_t}$ remains uniform. Other modules unaffected.

Since all single-variable interventional distributions are the same across all s, the entire family $\mathsf{Int}_1(M_s)$ is independent of s. □

Proof of Lemma 5.2(iii). It suffices to consider a single module (X_t, Y_t), since modules are independent and interventions factor across modules. Fix an arbitrary intervention on any subset of $\{X_t, Y_t\}$.

If Y_t is intervened on, then Y_t is fixed by the intervention. The remaining variable X_t, if not intervened on, equals U_{X_t} and is uniform in both the no-effect and XOR mechanisms.

If Y_t is not intervened on but X_t is set to x, then

$$Y_t = \begin{cases} U_{Y_t}, & s_t = 0, \\ x \oplus U_{Y_t}, & s_t = 1. \end{cases}$$

In either case Y_t is uniform Bernoulli and independent of all other modules.

Finally, if neither variable is intervened on, then for $s_t = 0$ we have $(X_t, Y_t) = (U_{X_t}, U_{Y_t})$, uniform on $\{0,1\}^2$; for $s_t = 1$ we have $(X_t, Y_t) = (U_{X_t}, U_{X_t} \oplus U_{Y_t})$, also uniform on $\{0,1\}^2$.

Thus every possible intervention produces the same distribution regardless of s_t. Taking the product over independent modules proves the claim. □

Proof of Lemma 5.2(iv). For each module t, consider the counterfactual query: "What is $\mathbb{P}(Y_t^{(X_t \leftarrow 0)} = Y_t^{(X_t \leftarrow 1)})$?"

Let $Y_t^{(X_t \leftarrow b)}$ denote the value of Y_t when we intervene with $\mathrm{do}(X_t = b)$, evaluated on the same noise U_{Y_t}.

Case $s_t = 0$: $Y_t^{(X_t \leftarrow 0)} = U_{Y_t}$ and $Y_t^{(X_t \leftarrow 1)} = U_{Y_t}$. These are identical, so $\mathbb{P}(Y_t^{(X_t \leftarrow 0)} = Y_t^{(X_t \leftarrow 1)}) = 1$.

Case $s_t = 1$: $Y_t^{(X_t \leftarrow 0)} = 0 \oplus U_{Y_t} = U_{Y_t}$ and $Y_t^{(X_t \leftarrow 1)} = 1 \oplus U_{Y_t} = 1 - U_{Y_t}$. These are always different, so $\mathbb{P}(Y_t^{(X_t \leftarrow 0)} = Y_t^{(X_t \leftarrow 1)}) = 0$.

Thus the counterfactual query for module t outputs 1 if $s_t = 0$ and 0 if $s_t = 1$. The m queries together determine s.

Since $\mathsf{CF}_1(M_s)$ contains $P(X, X^{(t \leftarrow 0)}, X^{(t \leftarrow 1)})$ for all t, which includes the joint distribution of $Y_t^{(X_t \leftarrow 0)}$ and $Y_t^{(X_t \leftarrow 1)}$, it determines s. □

Proof of Theorem 5.3. **Common interventional oracle.** By Lemma 5.2(iii), $\mathsf{Int}_{\mathrm{all}}(M_s)$ is identical for all s. Its description is constant: every intervention fixes the intervened variables and leaves each non-intervened module uniformly distributed in the manner described in the proof of Lemma 5.2(iii).

Upper bound. Given s (m bits) and n, a constant-length program simulates M_s and outputs $\mathsf{CF}_1(M_s)$. Thus $K(\mathsf{CF}_1(M_s) \mid n) \le m + O(1)$.

Lower bound. By Lemma 5.2(iv), the map $s \mapsto \mathsf{CF}_1(M_s)$ is injective. Hence there are 2^m distinct counterfactual answer objects consistent with the same full hard-do interventional oracle. The conditional counting bound (Lemma 3.7, part 2), applied with $w = (\mathsf{Int}_{\mathrm{all}}(M_s), n)$, yields the high-probability claim, since this auxiliary input is common to all s. □

C.2 Masked Parity

We now prove Theorem 5.4. Fix $n = 2m$, $1 \le d \le m$, and a hidden tuple S as in §5.2.

Lemma C.1 (Hard interventions see fair coins)**.** *Fix any hard intervention* $\mathrm{do}(X_T = x_T, Y_W = y_W)$ *with $T, W \subseteq [m]$, including the empty intervention. Under M_S the post-interventional law is the product of the point masses δ_{x_T} and δ_{y_W} with independent* $\mathrm{Bernoulli}(1/2)$ *laws for every non-intervened coordinate, and in particular does not depend on S.*

The private masks keep every nonclamped output uniformly random.

Proof. Non-intervened inputs satisfy $X_i = U_i$, independent fair bits. For a non-intervened output Y_j, write $Y_j = Z_j \oplus V_j$ where $Z_j := \bigoplus_{i \in S_j} X_i$ is a function of (U, x_T). Since V_j is a fair bit independent of (U, V_{-j}), conditionally on any value of (U, V_{-j}) the bit Y_j is uniform; hence Y_j is uniform and independent of all other coordinates, intervened or not. Taking the product over the non-intervened outputs gives the stated law. The empty intervention yields the uniform law on $\{0,1\}^{2m}$. □

Lemma C.2 (Counterfactuals reveal the sets)**.** *For all $i, j \in [m]$, $Y_j^{(X_i \leftarrow 0)} \oplus Y_j^{(X_i \leftarrow 1)} = \mathbf{1}[i \in S_j]$ with probability one under M_S. Consequently the map $S \mapsto \mathsf{CF}_1(M_S)$ is injective.*

Comparing two worlds with the same noise reveals each hidden parent membership.

Proof. Both counterfactual worlds use the same exogenous realization (u, v). In the world $\mathrm{do}(X_i = b)$, $Y_j = \big(\bigoplus_{i' \in S_j \setminus \{i\}} u_{i'}\big) \oplus b\,\mathbf{1}[i \in S_j] \oplus v_j$, so the two worlds differ exactly in the term $b\,\mathbf{1}[i \in S_j]$. The component of $\mathsf{CF}_1(M_S)$ for the variable X_i is the joint law of $(X, X^{(X_i \leftarrow 0)}, X^{(X_i \leftarrow 1)})$ over the whole system, whose marginal on $(Y_j^{(X_i \leftarrow 0)}, Y_j^{(X_i \leftarrow 1)})$ is supported on the diagonal if $i \notin S_j$ and on the anti-diagonal if $i \in S_j$. Reading this off for all pairs (i, j) recovers S. □

Lemma C.3 (Rate of $L_{m,d}$)**.** *For $m \geq 2$ and $1 \leq d \leq m$, $L_{m,d} = m \log_2 \sum_{k=0}^{d} \binom{m}{k} = \Theta\big(nd \log \frac{en}{d}\big)$ with $n = 2m$.*

The number of possible hidden parent sets determines the degree-sensitive rate.

Proof. For $1 \leq d \leq m$ the standard bounds $(m/d)^d \leq \binom{m}{d} \leq \sum_{k=0}^{d} \binom{m}{k} \leq (em/d)^d$ give $d \log_2(m/d) \leq L_{m,d}/m \leq d \log_2(em/d)$. The upper bound is $O(md \log(em/d)) = O(nd \log(en/d))$. For $d \leq m/2$ we have $\log_2(m/d) \geq 1$, so $\log_2(em/d) \leq (1 + \log_2 e) \log_2(m/d)$ and $L_{m,d} = \Omega(md \log(em/d)) = \Omega(nd \log(en/d))$, using $\log(en/d) \leq 2 \log(em/d)$. For $m/2 < d \leq m$, $\sum_{k \leq d} \binom{m}{k} \geq 2^{m-1}$ gives $L_{m,d} \geq m(m-1) \geq m^2/2$, while $nd \log(en/d) \leq 2m \cdot m \log(2e) = O(m^2)$. □

Proof of Theorem 5.4. **Common observational and interventional laws.** Lemma C.1 with the empty intervention gives the uniform observational law, and for every hard intervention it gives a law that depends only on (T, x_T, W, y_W) and n. A constant-length program therefore outputs $\mathsf{Int}_{\mathrm{all}}(M_S)$ from n, so $\mathrm{DL}(\mathsf{Int}_{\mathrm{all}}; M_S) = O(1)$.

Upper bound. The tuple S is one of $N_{m,d}$ possibilities and can be specified by its index in a canonical enumeration, using $\lceil L_{m,d} \rceil$ bits, plus $O(\log n)$ bits for d and a self-delimiting header. Given S, d, and n, a fixed program simulates M_S in each pair of counterfactual worlds and outputs $\mathsf{CF}_1(M_S)$. Hence $\mathrm{DL}_3(M_S) \leq L_{m,d} + O(\log n)$.

Lower bound and optimality. By Lemma C.2 the $N_{m,d}$ strings $\{\mathsf{CF}_1(M_S)\}$ are distinct. As shown above, $\mathsf{Int}_{\mathrm{all}}(M_S)$ is a fixed string computable from n, so conditioning on it changes conditional complexity by at most $O(1)$, by the argument of Lemma 3.9. The counting bound (Lemma 3.7, part 2) with auxiliary input $(\mathsf{Int}_{\mathrm{all}}(M_S), n)$ then gives $K(\mathsf{CF}_1(M_S) \mid \mathsf{Int}_{\mathrm{all}}(M_S), n) \geq L_{m,d} - c - O(1)$ for all but a 2^{-c} fraction of S, which is the stated probability bound for uniform S. The rate statement is Lemma C.3, and order-optimality follows from Corollary 4.8 applied to CF_1 in the class $\mathcal{M}_{n,d}(\{\mathrm{parity}\}, \{\mathrm{Bernoulli}(1/2)\})$. Since $nd \log(en/d) \geq n \log n$ for $1 \leq d \leq m$, the additional $n \log n$ term in that upper bound is absorbed. □

D Proofs for Section 6

This appendix contains all full proofs for Section 6, in the order of the main-text results.

D.1 Ambiguity and Entropy

Proof of Theorem 6.2. The set $\mathcal{F}_{Q_2|Q_1}(a; \mathcal{M})$ is by construction a set of distinct finite strings. Apply the counting bound (Lemma 3.7) with auxiliary input $w = (a, n)$. □

Proof of Theorem 6.3. Conditional on $L(\Theta) = \ell$, the map $\Theta \mapsto U(\Theta)$ is a bijection on the support; bijections preserve Shannon entropy. If the conditional parameter law is uniform on N values, its entropy is $\log_2 N$. □

For an explicitly listed finite family with computable answer maps and a specified computable prior, the Shannon conditional entropy in Theorem 6.3 can be approximated to arbitrary accuracy by finite summation. This calculation does not compute K; its cost depends on the representation of the family and prior.

D.2 Learning-Theoretic Consequences

The following two bounds and mutual-information observation prove Corollary 6.5 for the bipartite family $\mathcal{M}^n_{\text{bip}}$ (Definition 4.2).

Corollary D.1 (No-free-lunch for observational learners)**.** *Let $G \sim \mathrm{Unif}(2^{A\times B})$ in the bipartite family $\mathcal{M}^n_{\text{bip}}$. Any learner $\mathcal{A}$, with private randomness independent of G, that receives an i.i.d. sample of any size from $\mathsf{Obs}(M_G)$ and outputs a prediction $\widehat{\mathsf{Int}}_1$ satisfies*

$$\mathbb{P}[\widehat{\mathsf{Int}}_1 = \mathsf{Int}_1(M_G)] \;\le\; 2^{-m^2}.$$

Independent observational data cannot improve exact recovery over guessing a uniform graph.

Proof. All M_G share the observational distribution $P^\star$, so any sample from $P^\star$ is statistically independent of G. The learner's output, a function of the sample and independent private randomness, is therefore independent of G. With 2^{m^2} equiprobable targets $\{\mathsf{Int}_1(M_G)\}$ (the map $G \mapsto \mathsf{Int}_1(M_G)$ is injective by Lemma 4.3(iii)) and output independent of the true target, the success probability is at most 2^{-m^2}. □

Corollary D.2 (Per-query prediction error)**.** *Let $G \sim \mathrm{Unif}(2^{A\times B})$ in the bipartite family $\mathcal{M}^n_{\text{bip}}$. Let $\hat{p}$ depend only on an i.i.d. observational sample, the queried indices, and private randomness independent of G. For an independent uniformly random pair $(i,j) \in [m]\times[m]$, let $p_{i,j}(G) := P^{\mathrm{do}(X_{a_i}=0)}_{M_G}(X_{b_j}=0)$. Then*

$$\mathbb{E}\big[|\hat{p}_{i,j} - p_{i,j}(G)|\big] \;\ge\; \tfrac{1}{4}.$$

This lower bound applies to the stated row-intervention marginal queries.

Proof. By Lemma 4.3(ii), $p_{i,j}(G) \in \{1/2, 1\}$ depending on whether $(a_i, b_j) \in G$. For $G \sim \mathrm{Unif}(2^{A\times B})$ the edge indicator is Bernoulli$(1/2)$, so $p_{i,j}(G)$ equals 1 and $1/2$ each with probability $1/2$. The observational sample is independent of G, hence so is the predictor output $\hat{p}_{i,j}$. Condition also on the queried indices (i,j) and on $\hat{p}_{i,j} = a$:

$$\mathbb{E}\big[|\hat{p}_{i,j} - p_{i,j}(G)| \,\big|\, \hat{p}_{i,j} = a\big] \;=\; \tfrac{1}{2}|a-1| + \tfrac{1}{2}|a-\tfrac{1}{2}| \;\ge\; \tfrac{1}{4},$$

where the last step uses the triangle inequality $|a-1| + |a-\frac{1}{2}| \ge |1-\frac{1}{2}| = \frac{1}{2}$. Averaging over a proves the claim. □

Remark D.3 (Shannon Mutual Information is Zero)**.** In the bipartite family $\mathcal{M}^n_{\text{bip}}$, a dataset $D \sim (P^\star)^N$ of N i.i.d. observational samples is independent of the hidden graph G, since all mechanisms share the same observational distribution. Therefore $I(G; D) = 0$ in the Shannon sense for every sample size N. The lower bounds in Corollary D.1 and Corollary D.2 follow from this information-theoretic independence and do not rely on uncomputability of Kolmogorov complexity.

D.3 From Bits to Experiments

We first make the objects of §6.2 precise.

Definition D.4 (Experiments and adaptive strategies)**.** Fix a finite family $\mathcal{M}$ on n variables, query classes Q_1, Q_2, and an attained lower-rung answer a; write $\mathcal{F} := \mathcal{F}_{Q_2|Q_1}(a; \mathcal{M})$ for the ambiguity class of Definition 6.1 and $N := |\mathcal{F}|$. An *experiment* is a permitted intervention setting q (for us, a hard intervention $\mathrm{do}(X_S = x_S)$) whose outcome is the exact population law $\mathrm{Ans}_q(M)$. A setting can require many fresh-unit samples to estimate; one setting is not one sampled unit. Let $R(q) := |\{\mathrm{Ans}_q(M) : M \in \mathcal{M},\, \mathrm{Ans}_{Q_1}(M) = a\}|$ be the number of distinct outcomes of q across models with lower-rung answer a and $R := \max_q R(q)$. An *adaptive strategy of depth t* chooses settings from previous outcomes and stops after at most t settings. The proposition requires correct target identification for every model with lower-rung answer a. Short paths can be padded by repeating a setting, so the depth bound applies to variable stopping times too.

Proof of Proposition 6.6. **General bound.** The outcome sequence of a depth-t strategy is its transcript. Each experiment is determined by earlier outcomes, so there are at most R^t transcripts. Since the output is a function of the transcript, at most R^t distinct targets can be produced. Correctness for every compatible model requires at least N distinct targets.

Bipartite family. $N = 2^{m^2}$ by Lemma 4.3(iii). Use the vertex layers A and B of Definition 4.2. Fix a hard intervention $\mathrm{do}(X_T = x_T)$ with $T \subseteq \{r\} \cup A \cup B$. We show that the post-interventional law of M_G depends on G only through the vector $z(G) \in \{0,1\}^{B \setminus T}$ with $z_j(G) := \mathbf{1}[\text{some } a_i \in N_G(b_j) \cap T \text{ has } x_{a_i} = 0]$, so that $R(q) \le 2^m$. Intervened coordinates are constants.

If $r \in T$ with $x_r = 0$, every non-intervened a_k equals 0 and every non-intervened b_j equals $0 \wedge (\cdots) = 0$, a law independent of G.

If $r \in T$ with $x_r = 1$, every non-intervened a_k equals 1, and a non-intervened b_j equals $1 \wedge \bigwedge_{a_i \in N_G(b_j)} X_{a_i}$, which is 0 if $z_j(G) = 1$ and 1 otherwise.

If $r \notin T$, then $X_r = U_r$, every non-intervened a_k equals X_r, and a non-intervened b_j equals $X_r \wedge \bigwedge_{a_i \in N_G(b_j) \cap T} x_{a_i} \wedge \bigwedge_{a_i \in N_G(b_j) \setminus T} X_r$, which is 0 if $z_j(G) = 1$ and X_r otherwise.

In all cases the law is a function of $(T, x_T, z(G))$, so at most $2^{|B \setminus T|} \le 2^m$ distinct outcomes occur. The general bound gives $t \ge \lceil m^2/m \rceil = m$. For sufficiency, Lemma 4.3(ii) shows that the outcomes of $\mathrm{do}(X_{a_i} = 0)$ for $i \in [m]$ determine the rows $N_G(a_i)$, hence G and $\mathsf{Int}_1(M_G)$, without adaptivity.

Prescribed finite precision. Fix an encoder ρ_q that depends only on the post-interventional law, and let the experimental outcome be $\rho_q(\mathrm{Ans}_q(M))$. Define R_ρ as above for these encoded outcomes. Deterministic encoding cannot increase the number of outcomes, so $R_\rho \le R$, and the same transcript bound holds whenever the encoded outcomes permit target identification. For the bipartite family, if each output encodes a distribution within total variation $\varepsilon < 1/4$ of the law, row interventions remain sufficient: their b_j-marginals lie strictly below or above $3/4$ according as the true probability of $X_{b_j} = 0$ is $1/2$ or 1. Thus the exact m-setting count also holds with such a fixed accurate encoder. Arbitrary model-dependent rounding or nonseparating encodings do not justify this conclusion.

Modular XOR and masked parity. By Lemma 5.2(iii) and Theorem 5.4, every hard intervention has the same outcome for every member of the family, so $R(q) = 1$ for all q, while $N = 2^m$ for modular XOR (Lemma 5.2(iv)) and $N = N_{m,d} \ge (m+1)^m \ge 2^m$ for masked parity (Lemma C.2). The general bound applies. The impossibility also holds for adaptively chosen samples from fresh units: conditional on every past transcript, the next setting has the same outcome distribution in every model. Induction gives identical finite transcript laws; these also determine the same law on infinite transcripts. Thus even unlimited fresh-unit samples cannot resolve the counterfactual target. □

E Proofs for Section 7

This appendix gives the full statements and proofs for Section 7, in main-text order. Propositions E.1 and E.2 establish part 1 of Proposition 7.1; Proposition E.3 establishes part 2; Definition E.4 and Theorem E.5 give the setting and construction for part 3. Proposition E.6 and Remark E.7 support the subsequent Gaussian discussion. Here, a Markovian SCM means an SCM with independent local noises, as in Definition 3.1.

E.1 Graph Knowledge

Proposition E.1 (Localization of the gap)**.** *In the rooted-tree and bipartite families,* $K(\mathsf{Int}_1(M_G) \mid \mathsf{Obs}(M_G), n) = K(G \mid n) \pm O(1)$*, where G is the hidden tree or bipartite graph.*

In these families the missing information is precisely the hidden graph description.

Proof of Proposition E.1. Let G range over rooted trees or bipartite graphs and M_G be the corresponding SCM. A fixed program maps G (with n) to $\mathsf{Int}_1(M_G)$ by simulation, so $K(\mathsf{Int}_1(M_G) \mid n) \le K(G \mid n) + O(1)$. Conversely, Lemmas B.3 and B.4 for trees, and Lemma 4.3(iii) for bipartite graphs, give fixed programs that recover G from $\mathsf{Int}_1(M_G)$, so $K(G \mid n) \le K(\mathsf{Int}_1(M_G) \mid n) + O(1)$. Since $\mathsf{Obs}(M_G) = P^\star$ has constant description length, Lemma 3.9 replaces $K(\mathsf{Int}_1(M_G) \mid n)$ by $K(\mathsf{Int}_1(M_G) \mid \mathsf{Obs}(M_G), n)$ at a cost of $O(1)$. □

Proposition E.2 (A known graph under positivity closes the gap)**.** *Let M be a Markovian SCM (Definition 3.1) whose observational law P_M is strictly positive on $\{0,1\}^n$, and let G be its graph. Then $K(\mathsf{Int}_{\mathrm{all}}(M) \mid \mathsf{Obs}(M), G, n) = O(1)$. Without positivity the conclusion fails: the SCMs $M_{G,q}$ of Remark B.6 share their graph and observational law, yet $K(\mathsf{Int}_1(M_{G,q}) \mid \mathsf{Obs}, G, n)$ is unbounded over rational q.*

Positivity exposes every local response needed for truncated factorization.

Proof of Proposition E.2. For a Markovian SCM with graph G and strictly positive observational law P, the truncated factorization [11] gives, for every $S \subseteq [n]$ and x_S,

$$P^{\mathrm{do}(X_S = x_S)}(x) \;=\; \mathbf{1}[x'_S = x_S] \prod_{i \notin S} P\big(x_i \mid x_{\mathrm{pa}_G(i)}\big),$$

where x'_S denotes the S-coordinates of x. Positivity makes every conditional $P(x_i \mid x_{\mathrm{pa}_G(i)})$ well defined and computable from the rational table of P. A fixed program therefore outputs $\mathsf{Int}_{\mathrm{all}}(M)$ from (P, G, n), which is the claim. For the failure without positivity, fix a bipartite graph G with an edge (a_i, b_j) and consider the SCMs $M_{G,q}$ of Remark B.6 for rational $q \in [0,1]$: X_{b_j} keeps its parents and gains an independent $\mathrm{Bernoulli}(q)$ noise that is consulted only on the configuration $(X_r, X_{a_i}) = (1, 0)$. All $M_{G,q}$ are Markovian with the same graph and the same observational law $P^\star$, but $P^{\mathrm{do}(X_{a_i}=0)}_{M_{G,q}}(X_{b_j} = 1) = q/2$, so $\mathsf{Int}_1(M_{G,q})$ determines q and $K(\mathsf{Int}_1(M_{G,q}) \mid \mathsf{Obs}, G, n) \geq K(q) - O(1)$ is unbounded over q. □

E.2 Positivity and Faithfulness

Proposition E.3 (Positivity and faithfulness bound the worst-case gap at $\Theta(n \log n)$)**.** *Consider Markovian SCMs on n binary variables.*

1. ***Upper bound.*** *If M is Markovian with a strictly positive observational law P that is faithful to its graph, then $K(\mathsf{Int}_{\mathrm{all}}(M) \mid \mathsf{Obs}(M), n) \leq \log_2 |\mathcal{E}(P)| + O(\log n) \leq \log_2(n!) + O(\log n)$, where $\mathcal{E}(P)$ is the Markov equivalence class of DAGs faithful to P.*
2. ***Lower bound.*** *Fix a strictly positive rational distribution P on $\{0,1\}^n$ with no conditional independences. For each total ordering π of $[n]$ let M_π be the Markovian SCM on the complete DAG oriented by π whose mechanisms are the chain-rule conditionals of P. All $n!$ models generate P, and P is faithful to each of their graphs. Outside a proper algebraic subset of the simplex of positive laws, whose complement contains rational points, the oracles $\mathsf{Int}_1(M_\pi)$ are pairwise distinct; for such P and π uniform over orderings, $\mathsf{H}(\mathsf{Int}_1(M_\pi) \mid P) = \log_2(n!)$, and for every $c \geq 0$, $\Pr[K(\mathsf{Int}_1(M_\pi) \mid P, n) \geq \log_2(n!) - c - O(1)] \geq 1 - 2^{-c}$.*

Faithfulness restricts the residual graph choices but does not generally identify an ordering.

Proof of Proposition E.3. **Part 1.** Given the rational table of P and n, a fixed program computes every conditional independence relation of P exactly, and under faithfulness these relations determine the Markov equivalence class $\mathcal{E}(P)$ of the true graph, for instance by running the PC algorithm [18] with exact independence oracles and enumerating the DAGs consistent with the resulting completed partially directed acyclic graph (CPDAG) in a canonical order. The true graph G is one of them and can be specified by its index using $\lceil \log_2 |\mathcal{E}(P)| \rceil$ bits; with an $O(\log n)$ self-delimiting header this is $\log_2 |\mathcal{E}(P)| + O(\log n)$ bits. Given P and G, Proposition E.2 yields $\mathsf{Int}_{\mathrm{all}}(M)$ by truncated factorization, which uses positivity. Finally $|\mathcal{E}(P)| \leq n!$: all members of the class share a skeleton, and a topological ordering of a member orients every skeleton edge from the earlier to the later endpoint, so the sets of topological orderings of distinct members are disjoint. Assigning each member its lexicographically first ordering therefore injects the class into the $n!$ orderings.

Part 2. For a total ordering π, write $\prec_\pi i$ for the set of predecessors of i. The chain rule $P(x) = \prod_i P(x_i \mid x_{\prec_\pi i})$ defines a Markovian SCM M_π on the complete DAG oriented by π whose mechanisms are these conditionals; the conditionals are rational because P is, and M_π generates P. Explicitly, for each variable choose U_i uniform on $\{0, \ldots, L_i - 1\}$, with L_i a common multiple of its conditional-probability denominators, and set $X_i = \mathbf{1}[U_i < L_i P(X_i = 1 \mid x_{\prec_\pi i})]$.

These independent finite noises realize the required rational conditionals. Since P has no conditional independences and a complete DAG implies none, P is faithful to each of these graphs.

For distinctness, fix two orderings $\pi \neq \pi'$ and a pair (i, j) with $i \prec_\pi j$ and $j \prec_{\pi'} i$. In $M_{\pi'}$ the variable X_j is a non-descendant of X_i, so $P_{M_{\pi'}}^{\mathrm{do}(X_i=b)}(X_j = 1) = P(X_j = 1)$ for $b \in \{0, 1\}$. In M_π the same quantity is a rational function of the entries of P whose denominators are positive marginals. The condition that the two single-node oracles agree is therefore a finite system of polynomial equations in the entries of P after clearing denominators, and its solution set $Z_{\pi,\pi'}$ is a proper algebraic subset of the simplex provided one strictly positive P violates it.

Such a witness exists: let $X_i \sim \mathrm{Bernoulli}(1/2)$, $X_j = X_i \oplus \mathrm{Bernoulli}(\delta)$ for some $\delta \in (0, 1/2)$, and all other coordinates independent fair bits. This P is strictly positive, $P(X_j = 1 \mid x_{\prec_\pi j}) = P(X_j = 1 \mid x_i)$ because X_j is independent of the remaining predecessors given X_i, and so $P_{M_\pi}^{\mathrm{do}(X_i=1)}(X_j = 1) = 1 - \delta \neq 1/2 = P(X_j = 1)$.

The set Z_0 of laws with at least one conditional independence is likewise a proper algebraic subset of the simplex, since each independence is polynomial and proper: for $X_i \perp X_j \mid X_C$, make X_i, X_j a correlated noisy-copy pair and all remaining coordinates independent fair bits. This positive law violates that independence. Every nontrivial set-valued independence implies such a singleton-pair independence, so finitely many proper sets suffice.

A finite union of proper algebraic subsets of the simplex is a proper algebraic subset and contains no open set; hence the complement of $Z_0 \cup \bigcup_{\pi \neq \pi'} Z_{\pi,\pi'}$ is open and dense in the simplex and contains rational points. For any rational P in this complement, the $n!$ oracles $\mathsf{Int}_1(M_\pi)$ are pairwise distinct.

For such P, Theorem 6.3 with $\Theta = \pi$ uniform gives $\mathsf{H}(\mathsf{Int}_1(M_\pi) \mid P) = \log_2(n!)$, and Lemma 3.7 with auxiliary input (P, n) applied to the $n!$ distinct strings gives the Kolmogorov bound. The upper bound of part 1 applies to each M_π with $|\mathcal{E}(P)| = n!$, so the conditional-complexity gap is $\Theta(n \log n)$ in the worst case and with high probability under uniform orderings, while its Shannon conditional entropy is $\log_2(n!)$. □

E.3 Confounding

Confounding is represented by explicitly shared latent roots, extending Definition 3.1.

Definition E.4 (Semi-Markovian SCM and observed graph)**.** A *semi-Markovian SCM* has an acyclic endogenous graph and mutually independent finite-valued latent roots with finite rational probability tables. Each latent root is an input to at most two endogenous equations; local noise U_i collects the latent inputs to equation i. The *observed graph* $\mathcal{G}$ has the endogenous directed edges and $i \leftrightarrow j$ when equations i, j share a latent root. Query answers remain finite rational tables, so the description-length and ambiguity results apply. Theorem 4.6 concerns the independent-local-noise subclass.

Theorem E.5 (Confounding gap given the observed graph)**.** *Let $m \geq 1$ and $n = 2m$. For a hidden string $s \in \{0, 1\}^m$ let M_s consist of m independent modules (X_t, Y_t) in which $X_t \sim \mathrm{Bernoulli}(1/2)$, the exogenous response type $R_t = (Y_t(0), Y_t(1)) \in \{0, 1\}^2$ has the conditional law given X_t shown in Table 2, and $Y_t := Y_t(X_t)$. Then:*

1. *every M_s has the same observed graph $\mathcal{G}$ ($X_t \to Y_t$ and $X_t \leftrightarrow Y_t$ in each module) and the same observational law, uniform on $\{0, 1\}^{2m}$;*
2. *$P_{M_s}(Y_t = 1 \mid \mathrm{do}(X_t = 1))$ equals $1/2$ if $s_t = 0$ and $3/4$ if $s_t = 1$, so $s \mapsto \mathsf{Int}_1(M_s)$ is injective;*
3. *for s uniform, $\mathsf{H}(\mathsf{Int}_1(M_s) \mid \mathsf{Obs}, \mathcal{G}) = m$; for every s, $K(\mathsf{Int}_1(M_s) \mid \mathsf{Obs}, \mathcal{G}, n) \leq m + O(1)$; and for every $c \geq 0$, $\Pr[K(\mathsf{Int}_1(M_s) \mid \mathsf{Obs}, \mathcal{G}, n) \geq m - c - O(1)] \geq 1 - 2^{-c}$;*
4. *the high-probability complexity lower bound of part 3 holds for the ε-approximate gap of Definition 4.9 for every $0 \leq \varepsilon < 1/8$.*

A shared latent root can hide a mechanism bit even when its observed mixed graph is known.

Proof of Theorem E.5. It suffices to analyze one module, since modules are independent and every quantity factorizes. Write $R = (Y(0), Y(1))$ and read the rows of Table 2.

	$R_t \mid X_t = 0$	$R_t \mid X_t = 1$
$s_t = 0$	$\frac{1}{2}(\delta_{00} + \delta_{11})$	$\frac{1}{2}(\delta_{01} + \delta_{10})$
$s_t = 1$	$\frac{1}{2}(\delta_{11} + \delta_{01})$	$\frac{1}{2}(\delta_{01} + \delta_{00})$

Table 2: Conditional law of the response type $R_t = (Y_t(0), Y_t(1))$ given X_t in the confounded module of Theorem E.5; δ_{ab} is the point mass at $(Y_t(0), Y_t(1)) = (a, b)$.

Part 1. Use a latent root L uniform on four values. Map these values to (X, R) as $(0, 00), (0, 11), (1, 01), (1, 10)$ for $s = 0$, and $(0, 11), (0, 01), (1, 01), (1, 00)$ for $s = 1$. Set $X = X(L)$ and $Y = R(L)_X$. Both equations use L, and Y uses X, giving the same observed bow $X \to Y$, $X \leftrightarrow Y$ in either case; independent module roots meet Definition E.4. For the observational law, $P(Y = 1 \mid X = 0) = P(Y(0) = 1 \mid X = 0) = P(R \in \{10, 11\} \mid X = 0)$ equals $1/2$ in both rows, and $P(Y = 1 \mid X = 1) = P(R \in \{01, 11\} \mid X = 1)$ equals $1/2$ in both rows; with X a fair bit, (X, Y) is uniform on $\{0, 1\}^2$.

Part 2. Under $\mathrm{do}(X = 1)$, $Y = Y(1)$ and the exogenous law is unchanged, so $P(Y = 1 \mid \mathrm{do}(X = 1)) = P(R \in \{01, 11\})$, computed from the marginal of R, which is the average of the two columns. For $s = 0$ the marginal is uniform on the four types, giving $1/2$. For $s = 1$ the marginal puts mass $1/4$, $1/2$, $1/4$ on the types 11, 01, 00, giving $3/4$. Hence $\mathsf{Int}_1(M_s)$ determines every s_t, and $s \mapsto \mathsf{Int}_1(M_s)$ is injective.

Part 3. The observational law is uniform on $\{0, 1\}^{2m}$ and $\mathcal{G}$ is the same graph for all s; both are computable from n by constant-length programs, so conditioning on them changes conditional complexity by $O(1)$. Given s (m bits) and n, a fixed program outputs $\mathsf{Int}_1(M_s)$, which gives the upper bound. The 2^m strings $\mathsf{Int}_1(M_s)$ are distinct by part 2, so Lemma 3.7 gives the high-probability lower bound, and Theorem 6.3 with s uniform gives $\mathsf{H}(\mathsf{Int}_1(M_s) \mid \mathsf{Obs}, \mathcal{G}) = m$.

Part 4. If $s \neq s'$ differ at module t, the laws $P_{M_s}^{\mathrm{do}(X_t=1)}$ and $P_{M_{s'}}^{\mathrm{do}(X_t=1)}$ have Y_t-marginals that differ by $1/4$, and total variation does not increase under marginalization, so $d_{\mathsf{Int}}(\mathsf{Int}_1(M_s), \mathsf{Int}_1(M_{s'})) \geq 1/4$. For $\varepsilon < 1/8$ the ε-balls around the 2^m oracles are pairwise disjoint, and the argument of Theorem 4.10 applies verbatim. □

E.4 Gaussian Models

For the continuous extension, encode the zero-mean observational Gaussian law by its rational covariance Σ. Encode the full hard-do oracle by the finite tuple $(A_S, C_S)_{S \subseteq [n]}$ such that its law under $\mathrm{do}(X_S = x_S)$ is $\mathcal{N}(A_S x_S, C_S)$, including clamped coordinates and possibly singular C_S. These rational matrices specify all real-valued settings symbolically; they replace the binary probability-table encoding of §3. SEM abbreviates structural equation model.

Proposition E.6 (Faithful linear Gaussian orderings)**.** *Let Σ be a rational positive-definite $n \times n$ matrix such that $P = \mathcal{N}(0, \Sigma)$ has no conditional independences. For each total ordering π let M_π be the linear Gaussian SEM on the complete DAG oriented by π obtained by regressing each variable on its predecessors. Then every M_π has rational coefficients and error variances and generates P, and P is faithful to each of their graphs. Outside a proper algebraic subset of the cone of positive-definite matrices, whose complement contains rational points, the full hard-do oracles $\mathsf{Int}_{\mathrm{all}}(M_\pi)$ are pairwise distinct. For such Σ and π uniform over orderings, $\mathsf{H}(\mathsf{Int}_{\mathrm{all}}(M_\pi) \mid \Sigma) = \log_2(n!)$, for every $c \geq 0$, $\Pr[K(\mathsf{Int}_{\mathrm{all}}(M_\pi) \mid \Sigma, n) \geq \log_2(n!) - c - O(1)] \geq 1 - 2^{-c}$, and for every π, $K(\mathsf{Int}_{\mathrm{all}}(M_\pi) \mid \Sigma, n) \leq \log_2(n!) + O(\log n)$.*

Generic covariance matrices leave an ordering ambiguity of $n!$ distinct causal oracles.

Proof. Regressing X_j on its predecessors $X_{\prec_\pi j}$ gives the coefficient vector $\Sigma_{j,\prec j}\Sigma_{\prec j,\prec j}^{-1}$ and residual variance $\Sigma_{jj} - \Sigma_{j,\prec j}\Sigma_{\prec j,\prec j}^{-1}\Sigma_{\prec j,j}$, which are rational because Σ is; equivalently, M_π is the $LDL^\top$ factorization of Σ in the order π. The SEM $X_j = \beta_j^\top X_{\prec_\pi j} + \epsilon_j$ with independent Gaussian ϵ_j of these residual variances has covariance Σ, so it generates P, and faithfulness holds because P has no conditional independences while a complete DAG implies none. Under $\mathrm{do}(X_S = x_S)$ the law of M_π is Gaussian with mean affine in x_S and a covariance that depends only on S, both with rational

coefficients, so $\mathsf{Int}_{\mathrm{all}}(M_\pi)$ is a finite rational object; given Σ, n, and the index of π among the $n!$ orderings, a fixed program computes it, which gives the upper bound.

For distinctness, fix $\pi \neq \pi'$ and a pair (i, j) with $i \prec_\pi j$ and $j \prec_{\pi'} i$. In $M_{\pi'}$, X_j is a non-descendant of X_i, so $\mathbb{E}[X_j \mid \mathrm{do}(X_i = x)] = 0$ for all x. In M_π, $\mathbb{E}[X_j \mid \mathrm{do}(X_i = x)] = \tau_{ij}(\pi)\, x$, where the total effect $\tau_{ij}(\pi)$ is a rational function of the entries of Σ. Agreement of the two oracles forces $\tau_{ij}(\pi) = 0$, a polynomial condition after clearing denominators, and its solution set is proper if one positive-definite Σ violates it.

A witness: $X_i \sim \mathcal{N}(0, 1)$, $X_j = X_i + \eta$ with $\eta \sim \mathcal{N}(0, 1)$ independent, and all other coordinates independent standard Gaussians, so $\Sigma_{ii} = 1$, $\Sigma_{jj} = 2$, $\Sigma_{ij} = 1$, and all other off-diagonal entries vanish. In M_π every variable other than X_j has zero regression coefficients on its predecessors, and X_j has coefficient 1 on X_i and 0 elsewhere, so $\tau_{ij}(\pi) = 1 \neq 0$.

Each nontrivial Gaussian conditional independence is a proper polynomial condition: a covariance with a correlated pair i, j and all other coordinates independent violates $X_i \perp X_j \mid X_C$. The finite union of these conditions is therefore a proper algebraic subset.

A finite union of proper algebraic subsets of the open cone contains no open set, so its complement is open and dense and contains rational positive-definite matrices. For such Σ the $n!$ oracles are pairwise distinct, and Theorem 6.3 and Lemma 3.7 with auxiliary input (Σ, n) give the entropy and Kolmogorov statements. □

Remark E.7 (Identifiable subclasses collapse the exact gap)**.** Fix a class $\mathcal{C}$ with a finite observational encoding and a computable map from that encoding to its identified model and interventional oracle. Then a fixed program gives $K(\mathsf{Int}_{\mathrm{all}} \mid \mathsf{Obs}, n) = O(1)$. Equal-error-variance linear Gaussian SEMs with rational covariance satisfy these conditions [13]: search finitely many orderings for coincident sequential residual variances, using exact rational arithmetic; any successful ordering yields the unique weighted DAG by identifiability. LiNGAM gives another identifiability result [16], but its continuous non-Gaussian laws require a specified finite encoding and computable recovery before the same description-gap conclusion follows. Thus faithfulness alone permits a worst-case $\Theta(n \log n)$ Gaussian gap, whereas computable exact identification removes it. Approximate identification additionally needs separation or stability.

A finite-precision extension could quantize the Gaussian parameters or compare laws under a specified metric. Quantitative bounds would require explicit regularity conditions, such as bounded coefficients and well-conditioned covariances on nonclamped coordinates, together with a specified range of intervention values.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and §1 summarize the principal results: the $\Theta(n^2)$ obs–int gap with its degree-sensitive upper bound (which also covers the full hard-do and counterfactual oracles), finite-precision robustness, the $\Theta(n)$ and degree-sensitive counterfactual gaps conditional on $\mathsf{Int}_{\mathrm{all}}$, the unifying ambiguity-to-bits / -entropy theorems, and the experiment-count, assumption, and confounding results of §6.2 and §7; each is proved as a numbered result in the paper (full proofs in Appendices A–E).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 1 specifies the binary, acyclic setting. The tree and bipartite constructions in Section 4 have support on only two states; the full-support XOR and masked-parity constructions in Section 5 are unfaithful except for their empty-structure members. Section 7 distinguishes general bounds under graph knowledge, positivity, and faithfulness from the explicit confounded construction, and discusses rationally encoded Gaussian models. Remark 6.4 explains that the information bounds do not provide an efficient procedure for enumerating compatible answers. The closing sentences of §8 list directions for further work: nonlinear additive-noise and smooth models, soft and stochastic interventions, finite-precision identification stability, and efficient ambiguity counting.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All definitions, lemmas, and theorems are numbered and cross-referenced. Every theoretical claim in the main text either has its proof in-line or has a complete formal proof in Appendices A–E, with the proof headers explicitly cross-referencing the main-text statement (e.g., "Proof of Theorem 4.4").

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper is purely theoretical and contains no experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: The paper is purely theoretical; no code or data is associated with the results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper is purely theoretical; no experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [N/A]

Justification: The paper is purely theoretical; no experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper is purely theoretical; no experiments and no compute requirements.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper is theoretical, involves no human subjects, no datasets, and no model releases, and conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper is foundational theory: it bounds the information content of causal answer oracles. It does not propose a new model, dataset, or deployable system, so it has no direct positive or negative societal impact pathway. The framework can in principle inform evaluation of AI systems' causal reasoning, which is plausibly societally beneficial, but this is downstream of the present work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases no data, models, or other assets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: The paper does not use any external code, data, or model assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper introduces no new assets (no datasets, code, or models).

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not part of the core methodology of this paper; any use was limited to writing assistance and editing.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.