LOGARITHMIC INTERVENTIONS SUFFICE FOR NONLINEAR CAUSAL REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal representation learning aims to recover latent variables and their causal graph from nonlinear observations. However, existing identifiability results leave open full recovery under hidden multi-node intervention targets and replacement laws that vary across target-environment pairs. In this work, we show that same-unit factual-interventional pairs separate target identification from causal ordering. A minimum-code criterion over candidate residual charts and binary target codes recovers residual coordinates and intervention signatures (binary environment-target patterns) by selecting a minimum Hamming weight code consistent with paired invariances. After this alignment, coordinatewise normalization of environment density ratios isolates causal-order information from replacement strength. We then introduce Conditional-foliation peeling (CFP), which compares the oriented-foliation structure of the resulting conditional CDFs to identify and recursively remove sinks, variables with no remaining children. For smooth scalar invertible-noise SCMs with an invertible decoder and fixed conditional supports, the paired laws identify endogenous variables up to permutation and componentwise strictly monotone transformations and recover the entire DAG, assuming positive paired-support densities and Terminal-response faithfulness (TRF). Any distinct nonzero signatures suffice, including codes of arbitrary rank and nesting. Matching upper and lower bounds show that the minimax number of labeled intervention environments, excluding the factual environment, is exactly $\lceil \log_2(d+1) \rceil$ for $d \geq 2$.

1 INTRODUCTION

Causal representation learning seeks latent variables whose relationships support causal reasoning. Same-unit factual–interventional pairs provide an especially informative source: every untargeted mechanism reuses its factual exogenous draw, so each pair records exact coordinatewise invariances even though an unknown nonlinear decoder hides the coordinates (Brehmer et al., 2022). Existing results establish identification from paired atomic changes (Brehmer et al., 2022), recover abstractions from arbitrary-subset changes (Li et al., 2025), or use collections of unpaired environments (von Kügelgen et al., 2023; Varıcı et al., 2025; Ng et al., 2025; Jin & Syrgkanis, 2024). This leaves a central combined setting: recover both the nonlinear causal variables and their full Directed Acyclic Graph (DAG) when every labeled environment changes a hidden subset of mechanisms and the replacement law may vary across every target–environment pair.

The difficulty separates into two coupled obstacles. First, each node carries a binary target signature across environments, and these signatures may form a low-rank code with nested columns. Pairwise support equalities must therefore align latent coordinates for arbitrary code rank and nesting. Second, each environment ratio multiplies several node-specific changes of unrestricted strength. Raw density variation consequently entangles replacement strength with causal direction. Full recovery requires both a coordinate assignment for general binary codes and an order comparison that is invariant to replacement strength.

The paired laws provide one invariant for each obstacle. Support invariances determine which residual coordinates can change, while coordinatewise-normalized environment ratios determine how those coordinates propagate through the causal system. Keeping these two signals separate turns hidden intervention design into a code-alignment problem and causal ordering into a geometric comparison.

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

The support invariant identifies the residual chart and the hidden target signatures. A candidate chart maps the factual law to the uniform cube, and its binary code records which coordinates may change in each pair. The minimum-code criterion selects, among all support-consistent charts, a target code of least Hamming weight. Positive paired-support densities and distinct nonzero signatures ensure that every resulting chart recovers the true residual coordinates up to permutation and coordinate reflection. The same permutation aligns the target code, including codes of arbitrary rank and nesting.

The ratio invariant identifies causal order after this alignment. Each environment ratio factorizes over its targets, and each factor integrates to one along its own residual; coordinatewise normalization therefore removes replacement strength. The resulting conditional CDFs give an observable sink test: an untargeted sink has the identity CDF, whereas its targeted CDFs trace a shared oriented foliation. Terminal-response faithfulness (TRF) makes every non-sink detectable through a current sink. Conditional-foliation peeling (CFP) repeatedly applies this test, recovers an endogenous coordinate at each step, and reads every direct edge from the inverse recovered map.

These two invariants identify scalar nonlinear structural causal models with fixed conditional supports. They recover residual variables up to permutation and reflection, endogenous variables up to the corresponding componentwise monotone transformations, and the entire DAG under the matched labels. They also make intervention complexity a sharp coding question: m environments provide $2^m - 1$ nonzero signatures, and any d distinct signatures suffice. Matching collisions of the observed laws for zero or repeated columns prove that, for $d \geq 2$, exactly $m^\star(d) = \lceil \log_2(d+1) \rceil$ labeled intervention environments are necessary and sufficient.

The same analysis exposes a design–mechanism tradeoff. Compressed multi-target codes use terminal responses to reveal causal order; TRF follows from sign-coherent nonzero edge responses and holds generically in an explicit real-analytic nonlinear family. With one atomic environment per node, each endogenous coordinate is recovered directly, so the intervention design itself supplies the ordering information. The population comparisons also admit a selective finite-sample realization in an affine–probit observation model with quadratic mechanisms. Atomic pairs align the residual chart, simultaneous intervals determine the sink order and direct parents, and the estimator returns a graph once every alternative has been separated. In exact arithmetic it returns the true graph or abstains with probability at least $1 - \delta$. Across 800 numerical evaluations, every one of the 288 returned graphs is correct, and correct returns increase from 0/160 to 160/160 over the tested sample-size range.

Our contributions are summarized as follows:

1. *Paired-support identification:* We show that same unit factual-interventional pairs identify residual coordinates and hidden multi-target signatures, even for arbitrary-rank and nested intervention codes.
2. *Conditional-foliation peeling:* We introduce CFP, an observable sink-identification procedure that recursively recovers endogenous coordinates and causal order from normalized environment responses.
3. *Full and optimal causal recovery:* We recover the entire DAG up to the natural representation ambiguities and prove the sharp intervention complexity $m^\star(d) = \lceil \log_2(d+1) \rceil$.

2 Problem formulation

2.1 Latent structural model

Let $[d] = \{1, \ldots, d\}$ index the nodes of a DAG $\mathcal{G} = ([d], E)$, with $d \geq 1$. Independent exogenous variables $\mathbf{U} \sim \mathrm{Unif}((0,1)^d)$ generate the endogenous variables through

$$Z_i = f_i(Z_{\mathrm{pa}(i)}, U_i), \qquad i \in [d]. \tag{1}$$

For every parent value, $u_i \mapsto f_i(z_{\mathrm{pa}(i)}, u_i)$ is a strictly increasing C^2 diffeomorphism from $(0,1)$ onto the same open interval I_i. Thus each mechanism has fixed conditional support. Every displayed parent is *structurally active*, meaning that f_i varies with that parent. Acyclicity then defines the triangular C^2 solution map

$$F = (F_1, \ldots, F_d) : (0,1)^d \to \mathcal{Z} := \prod_{i=1}^{d} I_i. \tag{2}$$

An unknown C^2 diffeomorphism $g : \mathcal{Z} \to \mathcal{X}$ produces the observation $\mathbf{X} = g(\mathbf{Z})$. The inverse generative map $q := F^{-1} \circ g^{-1} : \mathcal{X} \to (0,1)^d$ is the *residual chart*.

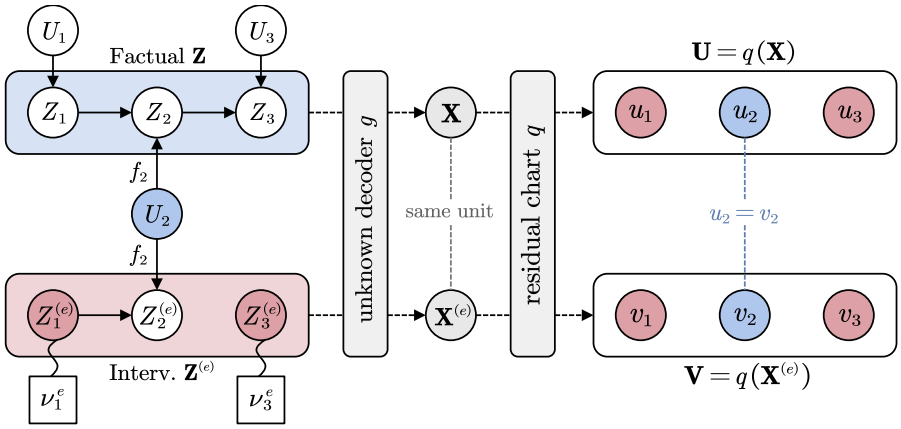

Figure 1: **Hidden multi-target intervention and same-unit residual invariance.** In environment e, $S_e = \{1, 3\}$ and $S_e^c = \{2\}$: nodes 1 and 3 are replaced, while node 2 reuses the factual mechanism and residual U_2. The common residual chart q maps same-unit pair $(\mathbf{X}, \mathbf{X}^{(e)})$ to $(\mathbf{U}, \mathbf{V})$, yielding $u_2 = v_2$ even though $Z_2^{(e)}$ may change through the intervened ancestor $Z_1^{(e)}$. Across environments, these paired-support equalities provide the signal for recovering hidden target signatures.

2.2 Labeled paired interventions

Each of the m labeled intervention environments has a hidden target set $S_e \subseteq [d]$. Given a factual draw $\mathbf{U}$, environment e keeps the factual mechanism and residual at $i \in S_e^c$ and replaces each target $i \in S_e$ by an independent draw $Z_i^{(e)}$ with density ν_i^e, where $\nu_i^e : I_i \to (0, \infty)$. Every ν_i^e is a positive C^1 probability density. Replacement draws are mutually independent and independent of $\mathbf{U}$, and their laws may vary freely across target–environment pairs.

The observed law family consists of the factual law and the labeled paired laws

$$\mathcal{T} = \left(P_0, \{P_e^{\text{pair}}\}_{e=1}^m\right), \qquad P_0 := \mathcal{L}(\mathbf{X}), \qquad P_e^{\text{pair}} = \mathcal{L}(\mathbf{X}, \mathbf{X}^{(e)}). \tag{3}$$

Write $\mathbf{U} = q(\mathbf{X})$ and $\mathbf{V} = q(\mathbf{X}^{(e)})$. Their paired support is

$$\mathcal{M}_e = \{(u, v) \in (0,1)^{2d} : u_{S_e^c} = v_{S_e^c}\}. \tag{4}$$

Let $\iota_e : (0,1)^d \times (0,1)^{S_e} \to \mathcal{M}_e$ map (u, w) to (u, v), where $v_{S_e} = w$ and $v_{S_e^c} = u_{S_e^c}$, and let $\mu_e := (\iota_e)_{\#}\text{Leb}^{d+|S_e|}$. Assume that $(q, q)_{\#}P_e^{\text{pair}}$ has a positive density relative to μ_e almost everywhere on $\mathcal{M}_e$. Consequently, $\mathcal{M}_e$ is the relative support of this pushforward law, and continuity extends pair equalities from almost everywhere to the full support.

The *target signatures* of nodes $i \in [d]$ are the binary columns $\boldsymbol{\sigma}_i = (\mathbf{1}\{i \in S_e\})_{e=1}^m \in \{0,1\}^m$, which are distinct and nonzero. Their containments are unrestricted, so the code matrix $\boldsymbol{\Sigma} = [\boldsymbol{\sigma}_1 \cdots \boldsymbol{\sigma}_d]$ may have rank below d.

2.3 Terminal-response faithfulness

Peeling distinguishes a non-sink through its effect on at least one remaining sink. Terminal-response faithfulness formalizes this requirement. A set $R \subseteq [d]$ is *ancestor closed* if it contains every strict ancestor of its nodes. Write $\mathcal{G}[R]$ for the induced subgraph, and let $\text{an}_{\mathcal{G}[R]}(j)$ and $\text{de}_{\mathcal{G}[R]}(j)$ denote strict ancestors and descendants in that graph. When the graph is omitted, $\text{an}(j)$ and $\text{de}(j)$ refer to strict ancestors and descendants in $\mathcal{G}$. Recursive sink deletion produces exactly such sets. Ancestor closure makes F_i a function of $\mathbf{U}_R$ for every $i \in R$; denote this restriction by $F_i^R : (0,1)^R \to I_i$. Later formulas suppress the superscript when R is fixed.

Assumption 2.1 (Terminal-response faithfulness (TRF))**.** For every ancestor-closed R and every non-sink j of $\mathcal{G}[R]$, there is a sink $\ell \in \text{sink}(\mathcal{G}[R]) \cap \text{de}_{\mathcal{G}[R]}(j)$ such that $\partial_{u_j} F_\ell^R \not\equiv 0$ on $(0,1)^R$.

TRF asks each non-sink for one responsive terminal descendant. Structural activity guarantees response along direct edges, while TRF preserves a nonzero aggregate response along at least one path family to a current sink. Proposition 4.6 gives two sufficient mechanism regimes, and Proposition 4.5 replaces TRF with one atomic environment per node.

2.4 IDENTIFICATION TARGET

Three transformations preserve the observed law family: relabeling residual coordinates, reflecting a uniform residual $u_i \mapsto 1 - u_i$, and applying a strictly monotone scalar transformation to each endogenous variable. We call two models *causally equivalent* if one can be obtained from the other by these transformations and the same node relabeling maps one DAG to the other. Our target is to identify this equivalence class from $\mathcal{T}$ and to identify the entire DAG within the matched labels.

3 CONDITIONAL-FOLIATION PEELING

CFP extracts two complementary invariants from the paired laws. Support invariance and minimum code align residual coordinates with hidden target signatures. Coordinatewise normalization then removes replacement strength from environment ratios, exposing a sink test that depends on causal order. Recursive sink removal constructs endogenous coordinates, whose inverse map reveals the direct edges. Figure 2 summarizes this pipeline, from normalized environment responses and sink identification to recursive coordinate recovery and final DAG reconstruction.

3.1 MINIMUM-CODE CHART

Minimum code resolves the coordinate assignment hidden by the nonlinear decoder. Let μ_d be the uniform probability measure on $(0,1)^d$. We optimize over C^1 diffeomorphisms $h : \mathcal{X} \to (0,1)^d$ and binary code matrices $\mathbf{\Lambda} = [\boldsymbol{\lambda}_1 \ \cdots \ \boldsymbol{\lambda}_d] \in \{0,1\}^{m \times d}$ with distinct nonzero columns. Write λ_{ej} for a candidate code entry and σ_{ei} for its true counterpart. The population problem is

$$\underset{h,\mathbf{\Lambda}}{\text{minimize}} \quad \|\mathbf{\Lambda}\|_0 \tag{5}$$

$$\text{subject to} \quad h_{\#}P_0 = \mu_d, \tag{6}$$

$$\mathbb{E}_{P_e^{\text{pair}}}\left[(h_j(\mathbf{X}) - h_j(\mathbf{X}^{(e)}))^2\right] = 0, \quad (e,j) \text{ such that } \lambda_{ej} = 0. \tag{7}$$

Constraint equation 6 maps the factual law to the uniform cube, and equation 7 makes coordinate j invariant in every environment with $\lambda_{ej} = 0$. The true pair $(q, \mathbf{\Sigma})$ is feasible. If output coordinate j depends nontrivially on residual i, every environment targeting i must mark j as changeable. The candidate signature of j therefore contains the union of the signatures of all residuals on which it depends. Summing these containments gives the code-weight lower bound, and minimum weight selects the residual coordinates even when the signature matrix has low rank or nested columns. Theorem 4.1 formalizes this argument.

3.2 ENVIRONMENT-RATIO FACTORIZATION

Coordinatewise normalization isolates causal-order information from the node-specific strength of each replacement law. The formulas below use the true residual order; permutation and reflection transport them to every recovered chart. Let p_0 be the density of $q(\mathbf{X}) = \mathbf{U}$ and let p_e be the density of $\mathbf{V} = q(\mathbf{X}^{(e)})$. Since $p_0 \equiv 1$, the environment ratio is $r_e(u) := p_e(u)/p_0(u) = p_e(u)$.

The intervention construction and the triangular Jacobian of F factorize this ratio as

$$r_e(u) = \prod_{i \in S_e} \rho_i^e(u), \qquad \rho_i^e(u) := \nu_i^e(F_i(u))\, \partial_{u_i} F_i(u). \tag{8}$$

Each factor depends on u_i and the residuals of its ancestors. Fixed conditional support gives it unit mass along coordinate i:

$$\int_0^1 \rho_i^e(u_{-i}, t)\, \mathrm{d}t = 1. \tag{9}$$

This identity puts factors from every environment on a common unit-mass scale. The positive C^1 product in equation 8 is the unique continuous representative of the density ratio.

Sink marginalization preserves the normalized product structure. Let R be the ancestor-closed set remaining after recovered sinks have been deleted, and let r_e^R be the ratio after marginalizing those sinks. Then equation 9 preserves the product form: $r_e^R(u_R) = \prod_{i \in S_e \cap R} \rho_i^e(u_R)$. Thus every peeling step retains the same factorization. Section D establishes the factorization and its preservation under recursive marginalization.

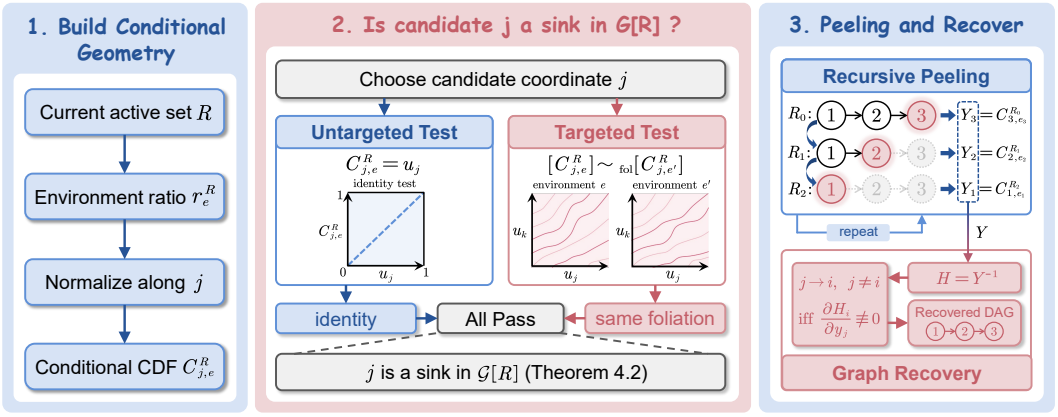

Figure 2: **Conditional-foliation peeling (CFP) overview**. Starting from the residual coordinates and target signatures, CFP normalizes environment ratios to construct conditional CDFs, identifies current sinks through untargeted identity and targeted oriented-foliation tests, and recursively recovers and peels one endogenous coordinate at a time. After all coordinates are recovered, direct causal edges are read from the nonzero derivative pattern (i.e. inverse-Jacobian) of the inverse map $H = Y^{-1}$.

3.3 Conditional CDF foliations

Conditional CDFs retain the geometry of a mechanism while discarding the environment-specific parameterization of its replacement law. The sink test therefore compares one coordinate across environments after conditioning on the other active coordinates. For $j \in R$ and environment e, define the fiber normalizer $\mathcal{N}_{j,e}^R$, conditional density $K_{j,e}^R$, and conditional CDF $C_{j,e}^R$:

$$\mathcal{N}_{j,e}^R(u_{R\setminus\{j\}}) := \int_0^1 r_e^R(u_{R\setminus\{j\}}, t)\,\mathrm{d}t, \qquad K_{j,e}^R(u_R) := \frac{r_e^R(u_R)}{\mathcal{N}_{j,e}^R(u_{R\setminus\{j\}})}$$

$$C_{j,e}^R(u_R) := \int_0^{u_j} K_{j,e}^R(u_{R\setminus\{j\}}, t)\ \mathrm{d}t = \int_0^{u_j} \frac{r_e^R(u_{R\setminus\{j\}}, t)}{\int_0^1 r_e^R(u_{R\setminus\{j\}}, s)\,\mathrm{d}s}\ \mathrm{d}t.$$

Positivity makes the normalizer finite on almost every fiber and yields a conditional CDF that is absolutely continuous, strictly increasing in u_j, and ranges from zero to one. The comparisons below use the resulting almost-everywhere CDF class, which is determined by the observed law family.

Definition 3.1 (Common oriented foliation)**.** Two conditional-CDF classes have the same oriented foliation, written $[C_{j,e}^R] \sim_{\text{fol}} [C_{j,e'}^R]$, if there is an increasing homeomorphism $\Gamma_{e,e'} : (0,1) \to (0,1)$ such that

$$C_{j,e}^R = \Gamma_{e,e'} \circ C_{j,e'}^R \qquad \text{Lebesgue-a.e. on } (0,1)^R. \tag{10}$$

The equivalence class records the shared leaves and their orientation while allowing each environment to parameterize those leaves differently.

The foliation comparison also has a geometric differential form. For C^1 versions with positive u_j derivatives, finite normalizers, and connected level sets of each reference conditional CDF, the equivalent criterion is

$$\mathrm{d}C_{j,e}^R \wedge \mathrm{d}C_{j,e'}^R = 0 \quad \text{on } (0,1)^R.$$

Thus the smooth form checks whether the two CDF gradients are collinear. The class-level definition remains the basis of the identification theorem. Section E makes the conditional-law classes version invariant and relates the differential criterion to a global oriented foliation.

The sink test combines two observable consequences of terminality. Fix any minimum-code chart $(h, \mathbf{\Lambda})$. Coordinate j passes on active set R when

$$C_{j,e}^R(u_R) = u_j \qquad \text{Lebesgue-a.e. and for every } e \text{ with } \lambda_{ej} = 0, \tag{11}$$

$$[C_{j,e}^R] \sim_{\text{fol}} [C_{j,e'}^R] \qquad \text{for every } e, e' \text{ with } \lambda_{ej} = \lambda_{e'j} = 1. \tag{12}$$

For a non-sink j, TRF selects a responsive terminal descendant ℓ. Because $\boldsymbol{\sigma}_j \neq \boldsymbol{\sigma}_\ell$, either some environment targets ℓ but not j, which violates the identity comparison, or $\mathrm{supp}(\boldsymbol{\sigma}_\ell) \subsetneq \mathrm{supp}(\boldsymbol{\sigma}_j)$, in which case two environments targeting j induce incompatible foliations. Thus every non-sink violates at least one comparison.

3.4 A THREE-NODE WORKED EXAMPLE

The chain $1 \to 2 \to 3$ makes the sink test concrete. Two environments target nodes $\{1, 3\}$ with different node-3 replacement laws, while a third targets $\{1, 2\}$. Separate normalization makes node 3 pass both comparisons; after it is marginalized, node 2 passes, giving the order $3, 2, 1$. Foliation geometry remains stable as replacement magnitudes vary. The calculation is in Section F.

3.5 PEELING AND GRAPH RECOVERY

CFP applies the sink test recursively and turns each accepted sink into an endogenous coordinate. Algorithm 1 gives the population procedure. Every conditional CDF is evaluated in the chosen minimum-code chart; when a pointwise value is needed, we select a continuous monotone representative of its almost-everywhere class.

By Theorem 4.2, $\mathcal{S}_R$ is exactly the set of current sinks and is nonempty at every iteration. At termination, $Y = M \circ F$ for a componentwise strictly monotone diffeomorphism M. Letting $\Psi := F^{-1}$ and $H := Y^{-1} = \Psi \circ M^{-1}$ gives

$$j \to i, \;\; j \neq i \iff \partial_{y_j} H_i \not\equiv 0. \quad (13)$$

Componentwise M preserves the derivative zero pattern of Ψ, so equation 13 identifies direct edges (Section H). We use $\mathbf{A}_{ij} = \mathbf{1}\{i \to j\}$, with parents in rows and children in columns.

Algorithm 1 Conditional-foliation peeling (CFP)

Require: observed law family $\mathcal{T}$; a minimizer $(h, \mathbf{\Lambda})$
Ensure: sink order, endogenous coords. Y, DAG $\widehat{E}$

$R \leftarrow [d]$, $\pi \leftarrow ()$ ▷ active nodes and sink order
while $R \neq \varnothing$ **do**
  1. Identify a current sink
  Form $\{[C^R_{j,e}] : j \in R,\ e \in [m]\}$
  $\mathcal{S}_R \leftarrow \{j \in R : j$ passes equation 11 and 12$\}$
  2. Recover its coordinate and peel
  Choose $j \in \mathcal{S}_R$ and e_j such that $\lambda_{e_j,j} = 1$
  $Y_j \leftarrow C^R_{j,e_j}$; $\pi \leftarrow (\pi, j)$
  $R \leftarrow R \setminus \{j\}$ ▷ peel the recovered sink
end while
  3. Recover direct edges
$Y \leftarrow (Y_1, \ldots, Y_d)$; $H \leftarrow Y^{-1}$
$\widehat{E} \leftarrow \{\, j \to i : j \neq i,\ \partial_{y_j} H_i \not\equiv 0 \,\}$
return π, Y, $\widehat{E}$ ▷ sink order, coordinates, DAG

Section 4.6 replaces the population equalities with simultaneous intervals and returns a graph once the residual chart, sink order, and direct parents are uniquely determined.

4 IDENTIFICATION RESULTS

4.1 RESIDUAL COORDINATES AND TARGET SIGNATURES

Paired supports determine both the residual chart and the hidden target signatures through the minimum-code chart.

Theorem 4.1 (Minimum-code residual identification)**.** *Suppose the model satisfies the paired-support density condition in Section 2, and its target signatures are distinct and nonzero. Every global minimizer $(h, \mathbf{\Lambda})$ of equation 5–equation 7 satisfies*

$$h = \mathsf{R} \circ \Pi \circ q, \quad (14)$$

where Π is a coordinate permutation and R independently applies either $u_i \mapsto u_i$ or $u_i \mapsto 1 - u_i$. The same permutation maps the columns of $\mathbf{\Lambda}$ to the target signatures in $\mathbf{\Sigma}$.

The proof uses a nonzero Jacobian monomial to match candidate coordinates with true residuals. Code containment and minimum weight then align the target signatures, while uniform-distribution rigidity reduces each remaining scalar map to a reflection. Crucially, the argument uses distinct signatures rather than linear independence, so it covers codes of arbitrary rank and nesting (Section C).

4.2 SINK CHARACTERIZATION

Normalized environment ratios turn heterogeneous interventions into the observable sink test. Conditional CDFs are interpreted as the almost-everywhere classes defined in Section 3.3.

Theorem 4.2 (Conditional-foliation sink characterization)**.** *Assume the model in Section 2, including TRF. Fix any minimum-code chart and use its recovered coordinate labels. Let R be an active set obtained by deleting sinks. For every $j \in R$,*

$$j \in \mathrm{sink}(\mathcal{G}[R]) \iff \textit{equation 11 and equation 12 hold.} \quad (15)$$

The two comparisons are understood in the sense of Section 3.3.

For a sink, normalization yields the identity in untargeted environments and a shared oriented foliation in targeted environments. TRF guarantees that every non-sink violates at least one comparison, making the test both observable and exact; Section G gives the argument.

4.3 Full causal identification

The two invariants now compose: minimum code supplies the residual labels, while conditional foliations supply a sink order and one endogenous coordinate per deletion.

Theorem 4.3 (Full identification from paired laws)**.** *Under the assumptions of Theorem 4.1 and Theorem 4.2, the observed law family $\mathcal{T}$ identifies:*

(i) the residual variables up to a common permutation and independent coordinate reflections;

(ii) the endogenous variables up to the same permutation and independent componentwise strictly monotone transformations;

(iii) and the entire DAG under the matched node labels.

In particular, any two admissible models with the same observed law family are causally equivalent in the sense of Section 2. The identified causal class and labeled DAG are invariant to the chosen minimum-code chart, valid sink order, and targeting environment e_j with $\lambda_{e_j,j} = 1$.

Each recovered sink supplies one monotone endogenous coordinate, and marginalization preserves the product form on the next active set. Induction recovers all coordinates; the inverse-map derivative pattern then distinguishes direct edges from longer ancestral paths. The proof appears in Section H.

4.4 Optimal intervention complexity

Full identification requires only logarithmically many labeled environments. Let $m^\star(d)$ be the smallest number in a target design that identifies every d-node model of Theorem 4.3 while its target sets remain hidden. The design is chosen before observation, and its target signatures are the minimax design variables.

Corollary 4.4 (Optimal number of environments)**.** *For $d \geq 2$ under the structural, support, and terminal-response conditions of Theorem 4.3, the minimax number of labeled interventional environments, excluding the factual environment, is*

$$m^\star(d) = \lceil \log_2(d+1) \rceil . \tag{16}$$

For $d = 1$, the allowed scalar monotone ambiguity makes the observational law sufficient, and hence $m^\star(1) = 0$.

With m environments, the binary codebook offers $2^m - 1$ nonzero target signatures, so any d of them can be assigned to the nodes and accepted by Theorem 4.2. This counting limit is attainable, while every smaller design forces a zero or repeated column and hence a collision of observed laws (Section I). The logarithmic count is therefore both sufficient and necessary.

4.5 Mechanism and design regimes

TRF supplies a terminal response at every step for compressed multi-target codes. Atomic environments recover each coordinate directly.

Proposition 4.5 (Identification from atomic environments)**.** *Under the structural, decoder, replacement, and paired-support conditions in Section 2, suppose the design contains a labeled singleton environment e_i for every node i. Then the observed law family identifies the residual variables up to permutation and reflection, the endogenous variables up to the same permutation and componentwise strictly monotone transformations, and the entire DAG under the matched labels.*

Atomic environments give distinct nonzero target signatures. Normalizing each atomic environment ratio directly recovers its endogenous coordinate, after which equation 13 identifies the DAG; see Section H.1.

Two mechanism regimes guarantee TRF.

Proposition 4.6 (Mechanism regimes implying terminal response)**.** *Either of the following conditions implies Assumption 2.1.*

(a) Every displayed edge derivative $\partial_{z_k} f_i$ has a constant nonzero sign. All directed paths with the same ordered endpoints have the same product of edge-derivative signs. Then every ancestor has an everywhere nonzero response on every descendant.

(b) Fix a DAG, real-analytic increasing diffeomorphisms $\tau_i : \mathbb{R} \to I_i$ with $\tau_i' > 0$, and real-analytic edge maps $\phi_{ik} : I_k \to \mathbb{R}$ with $\phi_{ik}' > 0$. Consider

$$Z_i = \tau_i\left(\operatorname{logit}(U_i) + \sum_{k \in \mathrm{pa}(i)} \theta_{ik}\phi_{ik}(Z_k)\right). \tag{17}$$

Fix any $u^\circ \in (0,1)^d$. If the edge-weight vector $\boldsymbol{\theta}$ has an absolutely continuous law, then with probability one every displayed edge is active and $\partial_{u_j} F_\ell(u^\circ) \neq 0$ simultaneously for every strict ancestor–descendant pair $j \in \mathrm{an}(\ell)$.

Together, the two regimes cover sign-coherent mechanisms and almost every model in an explicit nonlinear analytic family. The path expansion and analytic zero-set arguments appear in Section J, with the linear specialization in Proposition J.1.

4.6 Finite-sample realization

At finite samples, the population argument becomes two statistical questions: 1) do paired differences determine the residual directions, 2) and do the estimated responses determine a unique graph? Simultaneous confidence intervals answer both questions. The rule returns the graph selected by every interval and abstains while any sink or edge decision remains ambiguous.

For linear mixtures of probit-transformed residuals, pair differences estimate the residual directions and target signatures. Confidence intervals for the normalized environment ratios then certify sinks, and fresh samples certify direct edges. Section L develops this estimator and its finite-sample bounds.

The numerical study uses an atomic quadratic model whose bounded-moment score has the same zero set as the population conditional-CDF sink score. For the environment targeting node j, write $\mathbf{V} = q(\mathbf{X}^{(e_j)})$ and let γ_{kj} be the marginal-response moment defined in equation 262. On an active set R, the score

$$s_{k,R} = \sum_{j \in R \setminus \{k\}} \gamma_{kj}^2, \qquad \gamma_{kj} := 3\mathbb{E}_{e_j}[(2V_k - 1)(2V_j - 1)] \tag{18}$$

vanishes exactly when k is a sink. An independent covariance system then separates direct parents from more distant ancestors. The proof and execution order appear in Section O and Algorithm 3.

Assumption 4.7 (Finite-sample model)**.** The theorem instantiates the population model with the orthogonal affine–probit chart and quadratic SCM in Assumption O.1. It uses one labeled atomic environment per node, maximum indegree two, and response and edge margins that separate the population decisions.

Independent samples estimate the chart, calibrate the response intervals, and recover the graph after the sink order is fixed. All thresholds are chosen before sampling. Each interval combines sampling, representation, and roundoff error; its sampling contribution shrinks with N. Section O.5 derives the numerical enclosure.

Corollary 4.8 (Selective finite-sample graph recovery)**.** *Under the finite-sample model, let $\widehat{\mathbf{A}}_N$ be the interval rule's output from N samples per split and environment, and let $\mathbf{A}_\star$ be the true adjacency matrix. Choose the simultaneous intervals to have total coverage error at most δ. Then*

$$\mathbb{P}\Big\{\widehat{\mathbf{A}}_N \in \{\mathbf{A}_\star, \perp\}\Big\} \geq 1 - \delta. \tag{19}$$

On the simultaneous coverage event, every resolved sink and parent decision equals its population counterpart. Resolving all decisions therefore gives $\widehat{\mathbf{A}}_N = \mathbf{A}_\star$; see Theorem O.10.

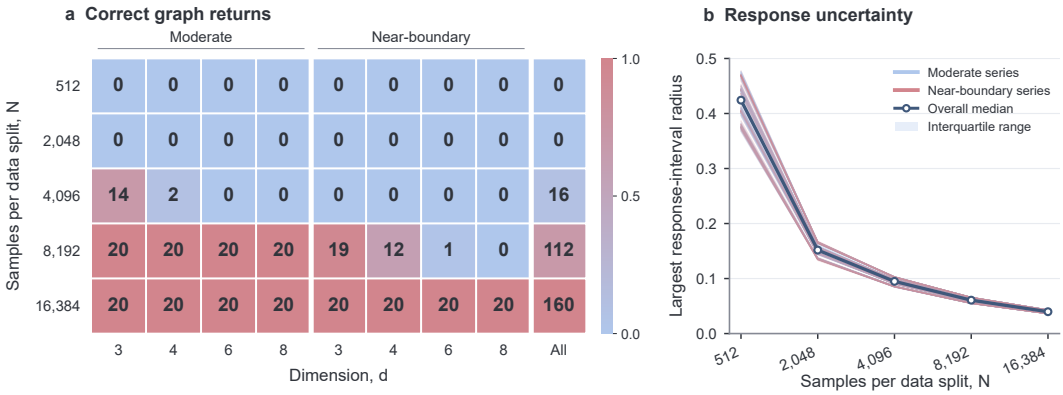

Figure 3: **More data turn abstention into correct graph recovery. a,** Correct returns by dimension and response regime (20 per cell; 160 overall). At 16,384, all 160 evaluations return the correct graph; every earlier return is also correct. **b,** Largest marginal-response interval radius for each replicate series; the dark blue line and band summarize the median and interquartile range across 160 series.

The numerical study measures how often additional data turn abstention into a correct graph; Corollary 4.8 supplies the soundness guarantee. Across four dimensions, four graph families, two response regimes, and five sample sizes, the empirical correct-return rate rises from 0% $(0/160)$ to 100% $(160/160)$. The rule returns 288 graphs over 800 evaluations, and all 288 have structural Hamming distance zero. A matched 2,700 evaluation study isolates the contribution of same-unit pairing to chart and signature recovery. Valid pairing reaches $50/50$ exact signature recovery from $N = 16$ for $d = 3$ and from $N = 1024$ for $d = 8$; cyclic rematching produces $3/900$ recoveries. Sections O.7 to O.9 report the study design and per-setting results. Section K converts observed-law accuracy into target accuracy on compact model classes, while Section L gives the constructive probit–linear estimator. Section M identifies the response scale required for reliable graph recovery, and Section N extends selective guarantees to contaminated channel laws.

5 Discussion and Conclusion

We establish that logarithmically many labeled intervention environments suffice for full causal identification in the nonlinear model class studied here. Same-unit pairs make this possible through two complementary signals. Paired invariances provide the alignment signal: minimum code recovers residual coordinates and hidden target signatures. Normalized environment responses provide the ordering signal even when replacement laws vary across targets and environments. Conditional-foliation peeling (CFP) recovers causal order and endogenous coordinates, and the inverse coordinate map reveals direct edges. Together, these signals identify the causal representation up to permutation and componentwise strictly monotone transformations, along with entire DAG.

This separation makes intervention design an optimal coding problem. Distinct nonzero signatures suffice even for low-rank or nested codes. At the population level, matching upper and lower bounds show that $\lceil \log_2(d+1) \rceil$ labeled intervention environments are necessary and sufficient for uniform identification when $d \geq 2$. This exposes a design-mechanism tradeoff: compressed multi-target designs use the terminal responses guaranteed by TRF to reveal causal order, while atomic environments recover each endogenous coordinate directly.

The finite-sample analysis shows how these identification steps lead to reliable graph recovery. In the atomic affine-probit model with quadratic mechanisms, simultaneous confidence intervals ensure that the procedure returns the true graph or abstains with probability at least $1 - \delta$. Across 800 evaluations, all 288 returned graphs are correct; at the largest tested sample size, all 160 evaluations recover the correct graph. By combining what interventions preserve with how their effects propagate, same-unit pairing recovers nonlinear causal variables together with the graph that connects them. Full nonlinear causal identification is therefore achievable at the counting limit of intervention codes.

AI USE STATEMENT

Generative AI tools were used to polish parts of the writing, assist with portions of the experiment code, and help formalize parts of the exposition. All AI-assisted outputs were reviewed and verified by the authors, who take responsibility for final content. Further details are provided in Appendix A.

REFERENCES

Kartik Ahuja, Jason S. Hartford, and Yoshua Bengio. Weakly supervised representation learning with sparse perturbations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15516–15528, 2022. doi: 10.52202/068431-1129. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/63d3bae2c1f525745003f679e45bcf7b-Abstract-Conference.html.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 372–407. PMLR, 2023. URL https://proceedings.mlr.press/v202/ahuja23a.html.

Simon Bing, Urmi Ninad, Jonas Wahl, and Jakob Runge. Identifying linearly-mixed causal representations from multi-node interventions. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 843–867. PMLR, 2024. URL https://proceedings.mlr.press/v236/bing24a.html.

Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco S. Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38319–38331, 2022. doi: 10.52202/068431-2776. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/fa567e2b2c870f8f09a87b6e73370869-Abstract-Conference.html.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep K. Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. In *Advances in Neural Information Processing Systems*, volume 36, pp. 45419–45462, 2023. doi: 10.52202/075280-1968. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8e5de4cb639ef718f44060dc257cb04f-Abstract-Conference.html.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://jmlr.org/papers/v13/gretton12a.html.

Jikai Jin and Vasilis Syrgkanis. Learning linear causal representations from general environments: Identifiability and intrinsic ambiguity. In *Advances in Neural Information Processing Systems*, volume 37, pp. 63466–63509, 2024. doi: 10.52202/079017-2027. URL https://doi.org/10.52202/079017-2027.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. *Journal of Machine Learning Research*, 27(71):1–90, 2026. URL https://jmlr.org/papers/v27/24-0771.html.

Inbeom Lee, Tongtong Jin, and Bryon Aragam. Beyond identifiability: Learning causal representations with few environments and finite samples. *arXiv preprint arXiv:2603.25796*, 2026. doi: 10.48550/arXiv.2603.25796. URL https://arxiv.org/abs/2603.25796.

Xiusi Li, Sékou-Oumar Kaba, and Siamak Ravanbakhsh. On the identifiability of causal abstractions. In *Proceedings of the Twenty-Eighth International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3241–3249. PMLR, 2025. URL https://proceedings.mlr.press/v258/li25g.html.

Wendong Liang, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal component analysis. In *Advances in Neural Information Processing Systems*, volume 36, pp. 32481–32520, 2023. doi: 10.52202/075280-1411. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/67089958e98b243d5cc1881ad60418b8-Abstract-Conference.html.

Francesco Montagna. On the identifiability of causal graphs with the invariance principle. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ta8BKRa1bl.

Ignavier Ng, Shaoan Xie, Xinshuai Dong, Peter Spirtes, and Kun Zhang. Causal representation learning from general environments under nonparametric mixing. In *Proceedings of the Twenty-Eighth International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3700–3708. PMLR, 2025. URL https://proceedings.mlr.press/v258/ng25a.html.

Chandler Squires, Anna Seigal, Salil S. Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32540–32560. PMLR, 2023. URL https://proceedings.mlr.press/v202/squires23a.html.

Burak Varıcı, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. Linear causal representation learning from unknown multi-node interventions. In *Advances in Neural Information Processing Systems*, volume 37, pp. 111614–111648, 2024. doi: 10.52202/079017-3545. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ca70528fb11dc8086c6a623da9f3fee6-Abstract-Conference.html.

Burak Varıcı, Emre Acartürk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning: Linear and general transformations. *Journal of Machine Learning Research*, 26(112):1–90, 2025. URL https://jmlr.org/papers/v26/24-0194.html.

Julius von Kügelgen, Michel Besserve, Wendong Liang, Luigi Gresele, Armin Kekić, Elias Bareinboim, David M. Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. In *Advances in Neural Information Processing Systems*, volume 36, pp. 48603–48638, 2023. doi: 10.52202/075280-2110. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/97fe251c25b6f99a2a23b330a75b11d4-Abstract-Conference.html.

Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/956e5427549a82a7472e02adc88360e9-Abstract-Conference.html.

Appendix

APPENDIX CONTENTS

A AI use statement . . . 13
B Related work . . . 13
C Residual and signature identification . . . 14
D Environment ratios and marginal peeling . . . 18
E Conditional-law and rigidity lemmas . . . 21
F Three-node peeling example: calculations . . . 26
G Proof of the sink characterization . . . 27
H Full causal identification from the observed laws . . . 29
H.1 Identification under atomic designs . . . 31
I Sharp intervention lower bound . . . 32
J Mechanism regimes for terminal response . . . 35
J.1 Genericity in linear SCMs . . . 36
K Finite-sample bounds from observed-law distances . . . 37
L Finite-sample recovery with a probit–linear chart . . . 39
L.1 Residual-chart recovery . . . 40
L.2 Independent-sample estimator . . . 44
L.3 Legendre ratio estimation . . . 47
L.4 Chart-error propagation and conditional-CDF approximation . . . 51
L.5 From fitted ratios to certified sink scores . . . 60
L.6 Adaptive peeling . . . 67
L.7 DAG recovery from fresh samples . . . 72
M Weak-response lower bound . . . 74
N Robust selective recovery . . . 77
N.1 Robustness across sample groups . . . 78
O Atomic finite-sample recovery . . . 81
O.1 Atomic finite-sample model and selective rule . . . 81
O.2 Residual-chart recovery from atomic pairs . . . 83
O.3 Simultaneous response intervals and sink peeling . . . 84
O.4 Direct-edge recovery . . . 86
O.5 Finite-precision soundness . . . 87
O.6 Selective finite-sample theorem . . . 87
O.7 Finite-sample results . . . 88
O.8 Same-unit pairing supplies the chart signal . . . 90
O.9 Implementation details . . . 91
P Response faithfulness and conditional-law regularity . . . 92
P.1 Terminal-response faithfulness . . . 92
P.2 Conditional-law regularity . . . 92

A AI USE STATEMENT

We appreciate the rapid progress of LLMs and the valuable assistance they provide to researchers in the machine learning community. In this work, LLMs were used in the following ways:

1. LLMs were occasionally used to polish and improve the writing.
2. LLMs assisted in writing parts of the code for the finite-sample experiments.
3. In Section L, due to the complexity of the presentation and the multiple levels of notation involved, LLMs assisted in formalizing parts of the exposition more rigorously to avoid notational conflicts and ambiguities.

All AI-assisted outputs and revisions were carefully reviewed and verified by the authors. The theoretical contributions and experimental analyses were developed and carried out by the authors.

B RELATED WORK

Paired changes. Paired methods use sample correspondence to turn unchanged mechanisms into observable invariances. Sparse before–after perturbations identify latent blocks from target-overlap patterns (Ahuja et al., 2022), while random atomic interventions identify nonlinear latent variables and their causal model (Brehmer et al., 2022). Arbitrary target sets support recovery of causal abstractions, with finer resolution supplied by singleton intersections (Li et al., 2025); support geometry gives related atomic and block-affine guarantees (Ahuja et al., 2023). We consider same-unit pairs with hidden multi-node targets and replacement laws that vary across node–environment pairs. The signature matrix may have arbitrary rank and nested columns. The recovery target is the nonlinear endogenous representation together with the full DAG.

Unpaired nonlinear identification. Unpaired methods exchange sample correspondence for additional environment structure or distributional variation. Two atomic domains per node, together with supplied groups that share a target, identify variables up to componentwise diffeomorphism and the DAG up to isomorphism (von Kügelgen et al., 2023). Causal Component Analysis takes the DAG and target sets as observed and identifies components under atomic interventions or target blocks under repeated multi-target regimes (Liang et al., 2023). Score-based causal representation learning recovers variables and graph from two stochastic hard interventions per node with latent target groupings (Varıcı et al., 2025). General-environment results recover the DAG under nonparametric mixing while identifying some variables only up to transformations that involve surrounding parents (Ng et al., 2025); mechanism-sparsity objectives yield partial or, under a graph condition, complete disentanglement for temporally mediated effects (Lachapelle et al., 2026). Hidden single-node interventions also identify Gaussian linear SCMs behind nonlinear decoders (Buchholz et al., 2023). Our result instead uses same-unit pairing to recover hidden multi-target signatures and resolves the nonlinear variables and matched DAG from the same paired laws.

Multi-target and auxiliary environments. Multi-target identifiability depends jointly on the observation interface and the requested recovery target. Under linear mixing, variance diversity identifies variables from multi-node interventions (Bing et al., 2024), and hard versus soft interventions support different targets under their respective diversity conditions (Varıcı et al., 2024). General unpaired environments identify a linear DAG while some variables remain mixed with surrounding parents; the nonlinear extension uses supplied groups of environments with a shared target and retains the same ambiguity (Jin & Syrgkanis, 2024). Logarithmic target-set designs give finite-sample recovery of variables, targets, and the DAG in a linear factor–SEM (Lee et al., 2026). Linear disentanglement from nodewise coverage (Squires et al., 2023), multi-view conditional-CDF uniformization (Yao et al., 2024), and observed-SCM graph recovery from auxiliary Gaussian-noise rescalings (Montagna, 2026) provide distinct data and recovery interfaces. Our logarithmic design result applies to hidden multi-target replacements in a paired nonlinear SCM and gives the exact minimax number of labeled environments for full-DAG identification.

Table 1 positions the closest results by the information exposed to the learner and the causal object each method identifies. We recover nonlinear variables and their matched DAG from same-unit pairs with hidden multi-node targets and heterogeneous replacement laws.

Table 1: Closest identifiability results by data interface, intervention information, model class, and recovery target. Blue ✓ marks recovered/identified quantities; red ✓ marks information assumed known or given; ✗ means not recovered; elsewhere, – means not applicable or not separately specified.

Work	Data / interventions					Model		Identifiability			Condition
	P/U	Int.	Type	Tgt.	Aux.	Map	Causal	Vars.	DAG	Extra / qual.	
Brehmer et al. (2022)	P	A	perf.	–	–	dec.	NL-SCM	✓	✓	SCM	Atomic coverage
von Kügelgen et al. (2023)	U	A	perf.	?	–	mix.	NL-SCM	✓	✓	–	2 discr. perf./node + meta.
Liang et al. (2023)	U	A/MT	perf.	✓	–	mix.	NL-CBN	A: ✓comp.	✓	MT: blk.	Discr. + MT blk. coverage
Bing et al. (2024)	U	MT	–	–	–	L-mix.	NL-ANM	✓	✗	–	Variance diversity
Varıcı et al. (2024)	U	MT	H/S	?	–	L-mix.	gen.-SCM	✓H	✓H	anc.S	Diverse environments
Jin & Syrgkanis (2024)	U	Gen.	–	–	–	L-mix.	L-SCM	✓SNA	✓	–	$K \geq d$; nodewise diversity
Li et al. (2025)	P	Sub.	–	–	–	–	NL-CRL	✗	✗	abstr.	Singleton ∩ refine variables
Varıcı et al. (2025)	U	A	hard	?	–	gen.-map	NL-SCM	✓	✓	–	two hard interventions/node
Ng et al. (2025)	U	Gen.	–	–	–	mix.	NL-SCM	✓SPM	✓	–	Third-derivative changes
Lachapelle et al. (2026)	U	MT	–	?	temp.	NL-mix.	temp. eff.	partial → complete†			Mechanism sparsity
Lee et al. (2026)	U	MT	–	?	–	L-fact.	SEM	✓	✓	targets	Log. designs; 2 scales/design
Ours	**P**	**MT**	**–**	**hidden**	**–**	**dec.**	**NL-SCM**	✓endog.	✓full	–	**min-code + ratios + opt. log**

Key. **P/U:** paired/unpaired samples; **Int.:** intervention class; **Tgt.:** target information; **Aux.:** auxiliary information; **A/MT:** atomic/multi-target; **Gen.:** general; **Sub.:** subset; **H/S:** hard/soft; **temp.:** temporal auxiliary observations; **NL/L:** nonlinear/linear; **ANM:** additive-noise model; **dec./mix./fact.:** decoder/mixing/factor; **SCM/CBN/SEM/CRL:** structural causal model / causal Bayesian network / structural equation model / causal representation learning; **comp.:** componentwise ambiguity; **blk.:** block ambiguity; **anc.:** ancestors; **SNA/SPM:** surrounded-node ambiguity / surrounding-parent mixing; **abstr.:** abstraction; **endog.:** endogenous. † Complete recovery holds under the stated graph condition.

C Residual and signature identification

The proof isolates the three ingredients behind residual identification: 1) full-support lifting, 2) minimum-code signature matching, 3) and rigidity of product-uniform triangular maps. Throughout this appendix, $\mathrm{supp}(a)$ and $\|a\|_0$ denote the support and Hamming weight of a binary vector a; for a binary matrix, $\|\mathbf{\Lambda}\|_0 := \sum_{j=1}^{d} \|\boldsymbol{\lambda}_j\|_0$ is the elementwise Hamming weight. We use the intrinsic measure μ_e on $\mathcal{M}_e$ defined in Section 2. Pair equalities first lift to all of $\mathcal{M}_e$ and imply Jacobian zeros. Minimum code weight then matches the unknown signatures, while product-uniformity eliminates the remaining triangular shears.

Lemma C.1 (Full-support lifting)**.** *Fix an environment e and an output coordinate j. Let $T : (0,1)^d \to \mathbb{R}^d$ be continuous. If $T_j(\mathbf{U}) = T_j(\mathbf{V})$ almost surely under environment e, then*

$$T_j(u) = T_j(v) \qquad \text{for every } (u,v) \in \mathcal{M}_e. \tag{20}$$

If T is C^1, then

$$i \in S_e \quad \Longrightarrow \quad \partial_{u_i} T_j \equiv 0. \tag{21}$$

Proof. Let p_e^{pair} be the density of $(\mathbf{U}, \mathbf{V})$ with respect to μ_e. By assumption, $p_e^{\mathrm{pair}}(u,v) > 0$ for μ_e-almost every $(u,v) \in \mathcal{M}_e$. Fix an output coordinate j and define the nonnegative continuous function

$$\Delta_j(u,v) := \big(T_j(u) - T_j(v)\big)^2, \qquad (u,v) \in \mathcal{M}_e.$$

The almost-sure equality in the lemma is equivalent to

$$0 = \mathbb{E}\big[\Delta_j(\mathbf{U},\mathbf{V})\big] = \int_{\mathcal{M}_e} \Delta_j(u,v) p_e^{\mathrm{pair}}(u,v)\, \mathrm{d}\mu_e(u,v). \tag{22}$$

Suppose, for contradiction, that $\Delta_j(u^\circ, v^\circ) > 0$ at some point of $\mathcal{M}_e$. Continuity of Δ_j in the relative topology gives a relatively open neighborhood $O \subset \mathcal{M}_e$ of (u°, v°) and a number $\varepsilon > 0$ such that

$$\Delta_j(u,v) \geq \varepsilon \qquad \text{for every } (u,v) \in O. \tag{23}$$

The parametrization ι_e in Section 2 is a homeomorphism onto $\mathcal{M}_e$. Hence $\iota_e^{-1}(O)$ is a nonempty open subset of $(0,1)^d \times (0,1)^{S_e}$ and has positive Lebesgue measure. By the definition $\mu_e = (\iota_e)_{\#}\mathrm{Leb}^{d+|S_e|}$,

$$\mu_e(O) = \mathrm{Leb}^{d+|S_e|}(\iota_e^{-1}(O)) > 0. \tag{24}$$

For $n \geq 1$, set $O_n := O \cap \{(u,v) : p_e^{\mathrm{pair}}(u,v) \geq 1/n\}$.

Because $p_e^{\text{pair}} > 0$ almost everywhere on O, $O \setminus \bigcup_{n\geq 1} O_n$ is μ_e-null. Equation 24 and countable subadditivity imply that $\mu_e(O_n) > 0$ for at least one n. For that n,

$$\int_O p_e^{\text{pair}}(u,v)\ \mathrm{d}\mu_e(u,v) \geq \frac{1}{n}\mu_e(O_n) > 0. \tag{25}$$

Combining equation 23 and equation 25 yields

$$\int_{\mathcal{M}_e} \Delta_j p_e^{\text{pair}}\ \mathrm{d}\mu_e \geq \varepsilon \int_O p_e^{\text{pair}}\ \mathrm{d}\mu_e > 0,$$

contradicting equation 22. Therefore $\Delta_j \equiv 0$ on $\mathcal{M}_e$, which proves equation 20.

For the differential conclusion, fix an arbitrary $u \in (0,1)^d$ and an arbitrary target coordinate $i \in S_e$. Choose $\delta > 0$ such that $u_i + t \in (0,1)$ whenever $|t| < \delta$, and set $v(t) := u + t\mathbf{e}_i,\ |t| < \delta$. Because $i \in S_e$, the curve satisfies $u_{S_e^c} = v(t)_{S_e^c}$, hence $(u, v(t)) \in \mathcal{M}_e$. The pointwise identity already proved gives

$$T_j(v(t)) = T_j(u) \qquad \text{for every } |t| < \delta.$$

The left side is differentiable at $t = 0$, and the chain rule gives

$$0 = \frac{\mathrm{d}}{\mathrm{d}t} T_j(v(t))\bigg|_{t=0} = \mathrm{D}T_j(u)\mathbf{e}_i = \partial_{u_i} T_j(u).$$

Since u was arbitrary, $\partial_{u_i} T_j \equiv 0$, proving equation 21. □

Lemma C.2 (Jacobian–code containment)**.** *Let $(h, \mathbf{\Lambda})$ be feasible for equation 5–equation 7, put $T = h \circ q^{-1}$, and let $\boldsymbol{\lambda}_j$ be column j of $\mathbf{\Lambda}$. Then*

$$\partial_{u_i} T_j \not\equiv 0 \quad \Longrightarrow \quad \operatorname{supp}(\boldsymbol{\sigma}_i) \subseteq \operatorname{supp}(\boldsymbol{\lambda}_j). \tag{26}$$

Moreover, every feasible pair obeys

$$\|\mathbf{\Lambda}\|_0 \geq \|\mathbf{\Sigma}\|_0. \tag{27}$$

Proof. Fix input coordinate i, output coordinate j, and environment $e \in \operatorname{supp}(\boldsymbol{\sigma}_i)$. By definition of the true signature, $e \in \operatorname{supp}(\boldsymbol{\sigma}_i)$ is equivalent to $i \in S_e$. If $\lambda_{ej} = 0$, the feasibility constraint equation 7 states that $\mathbb{E}_{P_e^{\text{pair}}}\left[(h_j(\mathbf{X}) - h_j(\mathbf{X}^{(e)}))^2\right] = 0$. In true residual coordinates this is $T_j(\mathbf{U}) = T_j(\mathbf{V})$ almost surely. Applying Lemma C.1 therefore gives

$$\lambda_{ej} = 0 \ \text{ and } \ e \in \operatorname{supp}(\boldsymbol{\sigma}_i) \quad \Longrightarrow \quad \partial_{u_i} T_j \equiv 0. \tag{28}$$

Taking the contrapositive of equation 28 for each $e \in \operatorname{supp}(\boldsymbol{\sigma}_i)$ gives

$$\begin{aligned} \partial_{u_i} T_j \not\equiv 0 \quad &\Longrightarrow \quad \lambda_{ej} = 1 \quad \text{for every } e \in \operatorname{supp}(\boldsymbol{\sigma}_i) \\ &\Longrightarrow \quad \operatorname{supp}(\boldsymbol{\sigma}_i) \subseteq \operatorname{supp}(\boldsymbol{\lambda}_j), \end{aligned}$$

which is equation 26.

For the weight lower bound, fix an arbitrary $u^\circ \in (0,1)^d$. Since T is a C^1 diffeomorphism, the inverse function theorem gives

$$\det \mathrm{D}T(u^\circ) \neq 0. \tag{29}$$

The Leibniz expansion is

$$\det \mathrm{D}T(u^\circ) = \sum_{\pi \in \mathfrak{S}_d} \operatorname{sgn}(\pi) \prod_{i=1}^{d} \partial_{u_i} T_{\pi(i)}(u^\circ). \tag{30}$$

If every product in equation 30 were zero, then the determinant would be zero, contradicting equation 29. Hence there exists a permutation $\pi \in \mathfrak{S}_d$ such that $\prod_{i=1}^{d} \partial_{u_i} T_{\pi(i)}(u^\circ) \neq 0$. The selected product is nonzero, so every factor is nonzero. Hence, for each i,

$$\partial_{u_i} T_{\pi(i)}(u^\circ) \neq 0 \quad \Longrightarrow \quad \partial_{u_i} T_{\pi(i)} \not\equiv 0.$$

The nonzero value at u° shows that each matched derivative is a nonzero function on the cube. Thus the global containment implication equation 26 applies to every matched pair. Applying equation 26 coordinate by coordinate gives

$$\operatorname{supp}(\boldsymbol{\sigma}_i) \subseteq \operatorname{supp}(\boldsymbol{\lambda}_{\pi(i)}), \qquad i = 1, \ldots, d. \tag{31}$$

Taking cardinalities in equation 31, summing, and using bijectivity of π yields

$$\|\mathbf{\Sigma}\|_0 = \sum_{i=1}^{d} \|\boldsymbol{\sigma}_i\|_0 \le \sum_{i=1}^{d} \|\boldsymbol{\lambda}_{\pi(i)}\|_0 = \sum_{j=1}^{d} \|\boldsymbol{\lambda}_j\|_0 = \|\mathbf{\Lambda}\|_0.$$

This is equation 27. □

Lemma C.3 (Exact code matching). *If $(h, \mathbf{\Lambda})$ is a global minimizer of the constrained problem equation 5–equation 7, then for some permutation π,*

$$\boldsymbol{\lambda}_{\pi(i)} = \boldsymbol{\sigma}_i \qquad \text{for all } i \in [d]. \tag{32}$$

After applying this output permutation, any linear extension of strict signature inclusion makes $\mathrm{D}T$ *lower triangular.*

Proof. The true pair $(q, \mathbf{\Sigma})$ is feasible. Because $q(\mathbf{X}) = \mathbf{U}$ under the factual law and $\mathbf{U} \sim \mu_d$, $q_{\#}P_0 = \mu_d$, so equation 6 holds. If $\sigma_{ej} = 0$, then $j \notin S_e$. The paired construction reuses U_j, and the factual structural inverse applied to the counterfactual returns the same residual:

$$q_j(\mathbf{X}) = U_j = V_j = q_j(\mathbf{X}^{(e)}).$$

Thus every constraint in equation 7 also holds. The assumed nonzero, pairwise distinct true signatures ensure that $\mathbf{\Sigma}$ belongs to the admissible code class. Consequently,

$$\min_{(h,\mathbf{\Lambda}) \text{ feasible}} \|\mathbf{\Lambda}\|_0 \le \|\mathbf{\Sigma}\|_0. \tag{33}$$

On the other hand, Lemma C.2 applies to every feasible pair and gives

$$\|\mathbf{\Lambda}\|_0 \ge \|\mathbf{\Sigma}\|_0. \tag{34}$$

For a global minimizer, equation 33 and equation 34 therefore force

$$\|\mathbf{\Lambda}\|_0 = \|\mathbf{\Sigma}\|_0. \tag{35}$$

Let π be the permutation selected from the nonzero determinant monomial in the proof of Lemma C.2. Define the coordinate gaps $\delta_i := \|\boldsymbol{\lambda}_{\pi(i)}\|_0 - \|\boldsymbol{\sigma}_i\|_0,\ i \in [d]$. The support inclusions equation 31 imply $\delta_i \ge 0$. Moreover,

$$\sum_{i=1}^{d} \delta_i = \sum_{i=1}^{d} \|\boldsymbol{\lambda}_{\pi(i)}\|_0 - \sum_{i=1}^{d} \|\boldsymbol{\sigma}_i\|_0 = \|\mathbf{\Lambda}\|_0 - \|\mathbf{\Sigma}\|_0 = 0,$$

where the final equality is equation 35. Since every $\delta_i \ge 0$, the zero sum gives $\delta_i = 0$ for every i, hence $\|\boldsymbol{\lambda}_{\pi(i)}\|_0 = \|\boldsymbol{\sigma}_i\|_0$ for every i. Each $\mathrm{supp}(\boldsymbol{\sigma}_i)$ is a finite subset of $\mathrm{supp}(\boldsymbol{\lambda}_{\pi(i)})$ with the same cardinality. The two supports are therefore equal, and binary-valuedness gives $\boldsymbol{\lambda}_{\pi(i)} = \boldsymbol{\sigma}_i$ for every i, which proves equation 32.

For the triangular form, relabel the outputs by π so that $\boldsymbol{\lambda}_i = \boldsymbol{\sigma}_i$. Then equation 26 becomes

$$\partial_{u_k} T_i \not\equiv 0 \quad \Longrightarrow \quad \mathrm{supp}(\boldsymbol{\sigma}_k) \subseteq \mathrm{supp}(\boldsymbol{\sigma}_i). \tag{36}$$

Define a strict partial order by $k \prec i \iff \mathrm{supp}(\boldsymbol{\sigma}_k) \subsetneq \mathrm{supp}(\boldsymbol{\sigma}_i)$.

The signatures are pairwise distinct, so whenever $k \ne i$, the inclusion in equation 36 is strict. Choose a linear extension of $\prec$. Apply the same reindexing to the matched input and output coordinates; this simultaneous conjugation by a permutation preserves both product-uniformity and the matched code identities. In this common order, if $k > i$ and $\partial_{u_k} T_i \not\equiv 0$, then equation 36 would imply $k \prec i$, which requires $k < i$, a contradiction. Therefore

$$\partial_{u_k} T_i \equiv 0 \qquad \text{whenever } k > i. \tag{37}$$

Equation 37 says precisely that $\mathrm{D}T(u)$ is lower triangular for every $u \in (0,1)^d$. □

Lemma C.4 (Dependence implied by a triangular Jacobian). *Let $G : (0,1)^d \to \mathbb{R}^d$ be C^1 and suppose $\partial_{u_k} G_i \equiv 0$ whenever $k > i$. Then there is a C^1 function $g_i : (0,1)^i \to \mathbb{R}$ such that*

$$G_i(u) = g_i(u_1, \ldots, u_i).$$

Proof. Fix i and two points $u, v \in (0,1)^d$ with $u_{1:i} = v_{1:i}$. Replace the tail coordinates of u by those of v one at a time. At the replacement of coordinate $k > i$, the fundamental theorem of calculus gives

$$G_i(\dots, v_k, \dots) - G_i(\dots, u_k, \dots) = \int_{u_k}^{v_k} \partial_{u_k} G_i(\dots, t, \dots)\, \mathrm{d}t = 0.$$

After all tail coordinates have been replaced, we obtain $G_i(u) = G_i(v)$. Thus G_i is constant on every fiber of the projection $u \mapsto u_{1:i}$. Fix $c \in (0,1)^{d-i}$ and define $g_i(a) := G_i(a, c), a \in (0,1)^i$.

For $i = d$, c is the unique empty tuple, so the same definition applies. Fiber constancy gives $G_i(u) = g_i(u_{1:i})$, and restriction of the C^1 map G_i to this fixed tail section gives $g_i \in C^1$. □

Lemma C.5 (Scalar uniform-preserving rigidity)**.** *Let $\tau : (0,1) \to (0,1)$ be C^1 with $\tau'(t) \neq 0$ for every t. If $W \sim \mathrm{Unif}(0,1)$ and $\tau(W) \sim \mathrm{Unif}(0,1)$, then*

$$\tau(t) = t \quad \text{for all } t, \qquad \text{or} \qquad \tau(t) = 1 - t \quad \text{for all } t. \tag{38}$$

Proof. Continuity and nonvanishing of τ' on the connected interval imply that τ' has a constant sign. First suppose $\tau' > 0$. Strict monotonicity then gives, for every $t \in (0,1)$,

$$\begin{aligned} \tau(t) &= \mathbb{P}\big(\tau(W) \le \tau(t)\big) && \text{because } \tau(W) \text{ is uniform} \\ &= \mathbb{P}(W \le t) && \text{because } \tau \text{ is increasing} \\ &= t. \end{aligned}$$

If $\tau' < 0$, then τ is strictly decreasing and $\tau(t) = \mathbb{P}\big(\tau(W) \le \tau(t)\big) = \mathbb{P}(W \ge t) = 1 - t$.

These are the two alternatives in equation 38. □

Lemma C.6 (Triangular product-uniform rigidity)**.** *Let $T : (0,1)^d \to (0,1)^d$ be a C^1 diffeomorphism with lower-triangular Jacobian. If $T_{\#}\mu_d = \mu_d$, then there are fixed signs $s_i \in \{-1, 1\}$ such that*

$$T_i(u) = \begin{cases} u_i, & s_i = 1, \\ 1 - u_i, & s_i = -1. \end{cases} \tag{39}$$

Proof. By Lemma C.4, for each i there is a C^1 map τ_i such that

$$T_i(u) = \tau_i(u_{1:i-1}, u_i). \tag{40}$$

Because $\mathrm{D}T(u)$ is lower triangular,

$$\det \mathrm{D}T(u) = \prod_{k=1}^{d} \partial_{u_k} T_k(u). \tag{41}$$

The diffeomorphism assumption gives $\det \mathrm{D}T(u) \neq 0$ for every u. Equation 41 therefore implies

$$\partial_{u_i} T_i(u) \neq 0 \qquad \text{for every } i \text{ and every } u. \tag{42}$$

For fixed i, the sign of $\partial_{u_i} T_i$ is a continuous map from the connected cube to $\{-1, 1\}$, and is therefore constant. Denote this sign by s_i.

Let $\mathbf{U} \sim \mu_d$. We prove equation 39 by induction on i. For $i = 1$, equation 40 gives $T_1(\mathbf{U}) = \tau_1(U_1)$. The first marginal of $T(\mathbf{U})$ is uniform because $T_{\#}\mu_d = \mu_d$. Applying Lemma C.5 gives $T_1(\mathbf{U}) = U_1$ if $s_1 = 1$, $T_1(\mathbf{U}) = 1 - U_1$ if $s_1 = -1$.

Fix $i \ge 2$ and assume that the claimed formula holds for $T_1, \dots, T_{i-1}$. Put $\mathbf{A} := \mathbf{U}_{1:i-1}$. The induction hypothesis makes the map $\mathbf{A} \longmapsto T_{1:i-1}(\mathbf{U})$ a coordinatewise identity/reflection and hence a bimeasurable bijection. Consequently, $\sigma(\mathbf{A}) = \sigma(T_{1:i-1}(\mathbf{U}))$, so conditioning on either random vector gives the same conditional laws.

The output vector $T(\mathbf{U})$ has the product-uniform law. In particular, $T_i(\mathbf{U})$ is independent of $T_{1:i-1}(\mathbf{U})$ and is uniform. Therefore, for Lebesgue-almost every $a \in (0,1)^{i-1}$,

$$\mathcal{L}\big(T_i(\mathbf{U}) \mid \mathbf{A} = a\big) = \mathcal{L}\big(\tau_i(a, U_i) \mid \mathbf{A} = a\big) = \tau_i(a, \cdot)_{\#}\, \mathrm{Unif}(0,1) = \mathrm{Unif}(0,1).$$

Here the middle equality uses independence of U_i and $\mathbf{A} = \mathbf{U}_{1:i-1}$, which gives $\mathcal{L}(U_i \mid \mathbf{A} = a) = \mathrm{Unif}(0,1)$ for almost every a. The derivative of $t \mapsto \tau_i(a,t)$ has the fixed sign s_i by equation 42. Hence there is a full-measure set $E_i \subset (0,1)^{i-1}$ such that, for every $a \in E_i$, Lemma C.5 gives simultaneously for all $t \in (0,1)$,

$$\tau_i(a,t) = \begin{cases} t, & s_i = 1, \\ 1-t, & s_i = -1. \end{cases} \tag{43}$$

The set E_i is dense: every nonempty open subset of $(0,1)^{i-1}$ has positive Lebesgue measure. For an arbitrary a, choose a sequence $a_n \in E_i$ with $a_n \to a$. Continuity of τ_i in its first argument and equation 43 imply the same formula at a. This completes the induction for every i, proving equation 39 pointwise on the cube. □

Proof of Theorem 4.1. Let $(h, \mathbf{\Lambda})$ be an arbitrary global minimizer and put $T = h \circ q^{-1}$. By Lemma C.3, choose π such that $\boldsymbol{\lambda}_{\pi(i)} = \boldsymbol{\sigma}_i$ for every i. Let Π_π be the permutation map defined by $(\Pi_\pi y)_i = y_{\pi(i)}$; then $\Pi_\pi T$ has matched input and output labels. Let Π_Q simultaneously place those matched labels in a linear extension of strict signature inclusion. The conjugated map

$$\overline{T} := \Pi_Q \Pi_\pi T \Pi_Q^{-1} \tag{44}$$

has a lower-triangular Jacobian everywhere by equation 37.

The factual pushforward constraints give

$$\begin{aligned} T_\# \mu_d &= (h \circ q^{-1})_\# (q_\# P_0) \\ &= h_\# P_0 \\ &= \mu_d. \end{aligned}$$

Permutation maps preserve μ_d, so $\overline{T}_\# \mu_d = \mu_d$. By Lemma C.6, there is a coordinatewise reflection R_0 such that $\overline{T} = \mathsf{R}_0$. Solving equation 44 for the successive relabelings gives

$$\begin{aligned} \Pi_\pi T = \Pi_Q^{-1} \mathsf{R}_0 \Pi_Q &=: \mathsf{R}_1, \\ T = \Pi_\pi^{-1} \mathsf{R}_1 & \\ = (\Pi_\pi^{-1} \mathsf{R}_1 \Pi_\pi) \Pi_\pi^{-1} &=: \mathsf{R}\,\Pi, \end{aligned}$$

where R_1 and R remain coordinatewise reflections because a permutation only reorders their signs, and $\Pi := \Pi_\pi^{-1}$ is a coordinate permutation. Since $T = h \circ q^{-1}$,

$$h = \mathsf{R} \circ \Pi \circ q.$$

The definition of Π_π also shows that output coordinate $\pi(i)$ matches true coordinate i, so the same Π matches the recovered columns to the true target signatures. □

D Environment ratios and marginal peeling

Ratio factorization isolates heterogeneous replacement laws into nodewise terms, and reverse-topological integration makes sink peeling exact.

Lemma D.1 (Reflection conjugacy). *Let $\mathsf{R} : (0,1)^d \to (0,1)^d$ be a coordinatewise identity/reflection map and let $\widetilde{u} = \mathsf{R}(u)$. Conjugating the residual chart and every density by R preserves (i) zero/nonzero derivative statements, (ii) the non-target conditional-CDF identity, and (iii) almost-everywhere common oriented foliations. If the jth coordinate is reflected, the conditional-CDF classes transform as*

$$\widetilde{C}_{j,e}(\widetilde{u}) = 1 - C_{j,e}\big(\mathsf{R}(\widetilde{u})\big) \qquad \textit{almost everywhere}; \tag{45}$$

if it is unchanged, it transforms as $\widetilde{C}_{j,e}(\widetilde{u}) = C_{j,e}(\mathsf{R}(\widetilde{u}))$.

Proof. Every coordinate reflection is an involution, so $\mathsf{R}^{-1} = \mathsf{R}$. The Jacobian determinant of R has absolute value one, so the conjugated density is

$$\widetilde{r}_e(\widetilde{u}) = r_e(\mathsf{R}(\widetilde{u})). \tag{46}$$

For every k, the chain rule multiplies a u_k derivative by the nonzero constant $\partial_{\widetilde{u}_k}\mathsf{R}_k \in \{-1, 1\}$. Thus a derivative is identically zero in one orientation exactly when it is identically zero in the other. For either orientation of coordinate j, substituting equation 46 into the conditional normalizer and using $s = t$ (unchanged) or $s = 1 - t$ (reflected) gives

$$\widetilde{\mathcal{N}}_{j,e}(\widetilde{u}_{-j}) = \int_0^1 r_e\big(\mathsf{R}_{-j}(\widetilde{u}_{-j}), \mathsf{R}_j(t)\big)\,\mathrm{d}t = \int_0^1 r_e\big(\mathsf{R}_{-j}(\widetilde{u}_{-j}), s\big)\,\mathrm{d}s = \mathcal{N}_{j,e}\big(\mathsf{R}_{-j}(\widetilde{u}_{-j})\big). \quad (47)$$

Here R_{-j} is the coordinatewise reflection with coordinate j omitted. It preserves Lebesgue-null sets, so the exceptional conditioning set in the new orientation is exactly the pullback of the old one. Dividing equation 46 by equation 47 gives $\widetilde{K}_{j,e}(\widetilde{u}) = K_{j,e}(\mathsf{R}(\widetilde{u}))$ on every good fiber. If coordinate j is unchanged, integrating from zero to $\widetilde{u}_j$ gives $\widetilde{C}_{j,e}(\widetilde{u}) = C_{j,e}(\mathsf{R}(\widetilde{u}))$.

Suppose instead that $\mathsf{R}_j(\widetilde{u}_j) = 1 - \widetilde{u}_j$. On the corresponding full-measure good set, the conditional normalizer is unchanged because the substitution $s = 1 - t$ reverses the limits, and therefore

$$\widetilde{C}_{j,e}(\widetilde{u}) = \int_0^{\widetilde{u}_j} K_{j,e}((\mathsf{R}(\widetilde{u}))_{-j}, 1 - t)\,\mathrm{d}t = \int_{1-\widetilde{u}_j}^1 K_{j,e}((\mathsf{R}(\widetilde{u}))_{-j}, s)\,\mathrm{d}s = 1 - C_{j,e}(\mathsf{R}(\widetilde{u})).$$

This proves equation 45 as an equality of conditional-law classes. The non-target identity is preserved in both cases: if $C_{j,e}(u) = u_j$, then the unchanged formula gives $\widetilde{C}_{j,e}(\widetilde{u}) = \widetilde{u}_j$, while the reflected formula gives $\widetilde{C}_{j,e}(\widetilde{u}) = 1 - (1 - \widetilde{u}_j) = \widetilde{u}_j$. In particular, if $C_{j,e} = \Gamma \circ C_{j,e'}$ almost everywhere, then the reflected CDFs obey

$$\widetilde{C}_{j,e} = \widetilde{\Gamma} \circ \widetilde{C}_{j,e'}, \qquad \widetilde{\Gamma}(s) := 1 - \Gamma(1 - s),$$

almost everywhere. The map $\widetilde{\Gamma}$ is again an increasing homeomorphism. Hence the orientation and all tests are invariant. □

Reflection invariance permits the remaining arguments to use the true residual orientation and then transfer each conclusion to the recovered orientation.

Lemma D.2 (Heterogeneous ratio factorization)**.** *For every environment e, the marginal density of $\mathbf{V} = q(\mathbf{X}^{(e)})$ is*

$$p_e(v) = \prod_{i \in S_e} \nu_i^e(F_i(v))\,\partial_{v_i} F_i(v). \quad (48)$$

Each factor ρ_i^e in equation 8 is positive, depends only on v_i and residual ancestors of i, and satisfies equation 9.

Proof. Fix a topological ordering in which every parent precedes its child. Define the mixed-coordinate map A_e by

$$[A_e(v)]_i := \begin{cases} v_i, & i \notin S_e, \\ F_i(v), & i \in S_e. \end{cases}$$

For each coordinate, set

$$J_{e,i} := \begin{cases} (0,1), & i \notin S_e, \\ I_i, & i \in S_e, \end{cases} \qquad \mathcal{A}_e := \prod_{i=1}^d J_{e,i}.$$

Thus $A_e : (0,1)^d \to \mathcal{A}_e$ preserves the original coordinate order.

We first construct the inverse. Let a be a point in this codomain and recover v in topological order.

If $i \notin S_e$, set $v_i := a_i$.

If $i \in S_e$, all residual ancestors $v_{\mathrm{an}(i)}$ have already been recovered. The scalar map $t \longmapsto F_i(v_{\mathrm{an}(i)}, t) = f_i\big(F_{\mathrm{pa}(i)}(v), t\big)$ is a strictly increasing C^2 diffeomorphism from $(0,1)$ onto the fixed interval I_i. It therefore has a unique inverse, and we set $v_i := F_i(v_{\mathrm{an}(i)}, \cdot)^{-1}(a_i)$.

This recursion proves both existence and uniqueness of $A_e^{-1}(a)$. It remains to verify joint regularity in the already recovered ancestors and in a_i. For $i \in S_e$, define $G_i(x, t, a_i) := F_i(x, t) - a_i,\ x = v_{\mathrm{an}(i)}$. The parent solution coordinates are functions of $v_{\mathrm{an}(i)}$, so

$$\partial_t G_i(x, t, a_i) = \partial_{v_i} F_i(x, t) = \partial_{u_i} f_i(F_{\mathrm{pa}(i)}(x), t) > 0.$$

The parametric implicit-function theorem therefore gives a local C^1 solution $t = \psi_i(x, a_i)$. Scalar injectivity makes this solution unique, so local branches agree on every overlap and define a global C^1 inverse in (x, a_i). Induction in topological order, composing ψ_i with the previously recovered ancestor coordinates, makes every coordinate of A_e^{-1} jointly C^1. Hence A_e is a C^1 diffeomorphism.

The Jacobian is triangular in the same topological order. Indeed, for $i \notin S_e$, the ith row has diagonal entry $\partial_{v_i}[A_e(v)]_i = 1$. For $i \in S_e$, acyclicity makes F_i depend on v_i and residual ancestors only, and hence

$$\partial_{v_i}[A_e(v)]_i = \partial_{v_i} F_i(v) > 0.$$

Thus

$$\det \mathrm{D}A_e(v) = \prod_{i \in S_e} \partial_{v_i} F_i(v). \tag{49}$$

Let $\mathbf{V} = q(\mathbf{X}^{(e)})$. For $i \notin S_e$, the residual inverse recovers the reused exogenous coordinate. Explicitly, $F(\mathbf{V}) = \mathbf{Z}^{(e)}$, and a non-targeted equation satisfies

$$\begin{aligned} f_i(F_{\mathrm{pa}(i)}(\mathbf{V}), V_i) &= F_i(\mathbf{V}) \\ &= Z_i^{(e)} \\ &= f_i(Z^{(e)}_{\mathrm{pa}(i)}, U_i) \\ &= f_i(F_{\mathrm{pa}(i)}(\mathbf{V}), U_i). \end{aligned}$$

Injectivity of f_i in its last argument gives $V_i = U_i$. For $i \in S_e$, the definition of $q = F^{-1} \circ g^{-1}$ gives $F_i(\mathbf{V}) = Z_i^{(e)}$, the independent replacement draw. Hence the mixed vector $A_e(\mathbf{V})$ consists of independent unit-density uniforms in non-target coordinates and independent replacement variables in target coordinates. Its density is therefore

$$p_{A_e(\mathbf{V})}(a) = \prod_{i \in S_e} \nu_i^e(a_i).$$

Applying the change-of-variables theorem to $a = A_e(v)$ and using equation 49 gives

$$\begin{aligned} p_e(v) &= p_{A_e(\mathbf{V})}(A_e(v)) |\det \mathrm{D}A_e(v)| \\ &= \prod_{i \in S_e} \nu_i^e(F_i(v)) \prod_{i \in S_e} \partial_{v_i} F_i(v) = \prod_{i \in S_e} \nu_i^e(F_i(v))\, \partial_{v_i} F_i(v). \end{aligned}$$

This is equation 48.

For the dependence claim, $F_i(v)$ is obtained by solving the equations for the ancestors of i followed by the equation at i. Consequently, $\rho_i^e(v) := \nu_i^e(F_i(v))\partial_{v_i} F_i(v)$ depends on v_i and $v_{\mathrm{an}(i)}$ and is independent of descendant residuals. It is positive and C^1 because ν_i^e is positive C^1, F_i is C^2, and $\partial_{v_i} F_i > 0$.

Finally fix v_{-i} and write $\varphi_i(t) := F_i(v_{-i}, t)$. The fixed-support assumption says that φ_i is an increasing diffeomorphism from $(0, 1)$ onto I_i. For $0 < a < b < 1$, the ordinary change of variables $z = \varphi_i(t)$ and $\mathrm{d}z = \varphi_i'(t)\,\mathrm{d}t$ gives

$$\int_a^b \rho_i^e(v_{-i}, t)\ \mathrm{d}t = \int_a^b \nu_i^e(\varphi_i(t))\varphi_i'(t)\ \mathrm{d}t = \int_{\varphi_i(a)}^{\varphi_i(b)} \nu_i^e(z)\ \mathrm{d}z. \tag{50}$$

Both integrands are nonnegative. Letting $a \downarrow 0$ and $b \uparrow 1$ and applying monotone convergence to equation 50 yields

$$\int_0^1 \rho_i^e(v_{-i}, t)\ \mathrm{d}t = \int_{I_i} \nu_i^e(z)\ \mathrm{d}z = 1. \tag{51}$$

Here onto monotonicity gives $\varphi_i((0, 1)) = I_i$, including finite or infinite open endpoints. The fixed conditional support I_i makes the final integral independent of v_{-i}. □

Lemma D.3 (Sink marginalization). *Let R be ancestor closed and let $\overline{R} = [d] \setminus R$. Marginalizing p_e over $u_{\overline{R}}$ gives*

$$p_e^R(u_R) = \prod_{i \in S_e \cap R} \rho_i^e(u_R). \tag{52}$$

The factual marginal on R remains uniform, so $r_e^R = p_e^R$.

Proof. Let $\overline{R} = [d] \setminus R$. Since R is ancestor closed, $\overline{R}$ is descendant closed: if $i \in \overline{R}$ and $i \to \cdots \to k$, then $k \in \overline{R}$. Set $n_R := |\overline{R}|$ and choose an ordering $i_1, \ldots, i_{n_R}$ of $\overline{R}$ in reverse topological order, so every proper descendant of i_t occurs among $i_1, \ldots, i_{t-1}$. Define the partial marginal integrals

$$G_0(u) := \prod_{k \in S_e} \rho_k^e(u),$$

and, recursively,

$$G_t(u_{[d] \setminus \{i_1, \ldots, i_t\}}) := \int_0^1 G_{t-1}(u_{[d] \setminus \{i_1, \ldots, i_{t-1}\}}) \, \mathrm{d}u_{i_t}.$$

The function $G_0 = p_e$ is nonnegative and integrates to one. Tonelli's theorem therefore permits these repeated integrations in the chosen order, including temporarily infinite fiber integrals on a null set. We prove by induction that

$$G_t = \prod_{k \in S_e \setminus \{i_1, \ldots, i_t\}} \rho_k^e, \tag{53}$$

with all remaining variables held fixed.

The case $t = 0$ is the definition of G_0. Assume the invariant for $t-1$, set $L_{t-1} := S_e \setminus \{i_1, \ldots, i_{t-1}\}$, and consider $i = i_t$. If a remaining factor ρ_k^e with $k \neq i$ depended on u_i, then i would be an ancestor of k. If $k \in R$, ancestor closure would imply $i \in R$, contradicting $i \in \overline{R}$. If $k \in \overline{R}$, then k is a proper descendant of i and would already have been removed from the product because of the reverse topological ordering. Thus among the factors remaining in equation 53, dependence on u_i is confined to the factor indexed by i.

The induction step therefore separates the u_i factor explicitly:

$$G_t = \underbrace{\prod_{k \in L_{t-1} \setminus \{i\}} \rho_k^e}_{\text{independent of } u_i} \begin{cases} \displaystyle\int_0^1 \rho_i^e(u_{-i}, s) \, \mathrm{d}s, & i \in S_e, \\ \displaystyle\int_0^1 1 \, \mathrm{d}s, & i \notin S_e \end{cases} = \prod_{k \in S_e \setminus \{i_1, \ldots, i_t\}} \rho_k^e.$$

The first integral equals one by equation 51; the second equals one because $(0, 1)$ has unit length. This proves the invariant at step t. At $t = n_R$,

$$p_e^R(u_R) = G_{n_R}(u_R) = \prod_{k \in S_e \cap R} \rho_k^e(u_R),$$

which is equation 52. Finally, the factual residual law is product uniform, so its marginal on R has density one. Therefore $r_e^R = p_e^R$. For every remaining $k \in R$, ancestor closure gives $\mathrm{an}(k) \subseteq R$, so each displayed ρ_k^e genuinely depends only on u_R. We interpret an empty product as one; this covers both $R = [d]$ and $R = \varnothing$. □

E Conditional-law and rigidity lemmas

Version-invariant conditional laws and rigidity convert ratio factors into exact identity and foliation criteria.

Lemma E.1 (Version-invariant conditional CDF). *Let $r : (0, 1)^R \to (0, \infty)$ be a positive probability density, so $\int_{(0,1)^R} r(u) \, \mathrm{d}u = 1$, and fix $j \in R$. Define*

$$\mathcal{N}_j^r(u_{R \setminus \{j\}}) := \int_0^1 r(u_{R \setminus \{j\}}, t) \, \mathrm{d}t. \tag{54}$$

Then the set $\Omega_j^r := \{u_{R\setminus\{j\}} : 0 < \mathcal{N}_j^r(u_{R\setminus\{j\}}) < \infty\}$ *has full Lebesgue measure. On* $\Omega_j^r \times (0,1)$,

$$K_j^r(u) := \frac{r(u)}{\mathcal{N}_j^r(u_{R\setminus\{j\}})}, \qquad C_j^r(u) := \int_0^{u_j} K_j^r(u_{R\setminus\{j\}}, t)\ \mathrm{d}t \tag{55}$$

are respectively a regular conditional density and CDF. For almost every conditioning value, C_j^r *is absolutely continuous and strictly increasing in* u_j*, has boundary limits zero and one, and satisfies*

$$\partial_{u_j} C_j^r = K_j^r > 0 \quad \textit{for almost every } u_j. \tag{56}$$

Arbitrary measurable conditional-CDF extensions outside Ω_j^r *produce the same Lebesgue-a.e. equivalence class, and every regular conditional version belongs to that class.*

Proof. Write $w = u_{R\setminus\{j\}}$. Because $r \geq 0$, Tonelli's theorem applies directly and gives

$$\int_{(0,1)^{R\setminus\{j\}}} \mathcal{N}_j^r(w)\ \mathrm{d}w = \int_{(0,1)^R} r(u)\ \mathrm{d}u < \infty.$$

Thus $\mathcal{N}_j^r < \infty$ almost everywhere. Positivity of r gives $\mathcal{N}_j^r(w) > 0$ whenever it is finite because the integrand is positive throughout $(0,1)$. Hence Ω_j^r has full Lebesgue measure.

Let $A \subseteq (0,1)^{R\setminus\{j\}}$ and $B \subseteq (0,1)$ be Borel sets. On Ω_j^r, the definition of K_j^r gives

$$\begin{aligned}\int_{A\cap\Omega_j^r} \left[\int_B K_j^r(w,t)\ \mathrm{d}t\right] \mathcal{N}_j^r(w)\ \mathrm{d}w &= \int_{A\cap\Omega_j^r}\int_B \frac{r(w,t)}{\mathcal{N}_j^r(w)} \mathcal{N}_j^r(w)\ \mathrm{d}t\ \mathrm{d}w \\ &= \int_A \int_B r(w,t)\ \mathrm{d}t\ \mathrm{d}w. \end{aligned} \tag{57}$$

The exceptional conditioning set is Lebesgue-null and therefore contributes zero to both sides. Equation 57 is the defining disintegration identity for the conditional law of u_j given w; its conditioning marginal has density $\mathcal{N}_j^r$. Thus K_j^r is a regular conditional density on the good conditioning set. On its null complement, set $K_j^r(w,t) = 1$ and $C_j^r(w,t) = t$; this measurable probability-kernel extension completes a regular conditional version and preserves its almost-everywhere class. Integrating K_j^r over $(0, u_j]$ gives the CDF C_j^r in equation 55.

On each good fiber, K_j^r is positive and integrates to one. The fundamental theorem for Lebesgue integrals therefore makes $C_j^r(w,\cdot)$ absolutely continuous, gives its endpoint limits zero and one, and yields equation 56. Since $\mathcal{N}_j^r$ is positive and finite Lebesgue-almost everywhere, the conditioning marginal and Lebesgue measure have the same null sets. Almost-sure uniqueness of regular conditional kernels then makes every regular conditional version agree with equation 55 for Lebesgue-almost every conditioning value. In particular, two explicit extensions can differ only on $((0,1)^{R\setminus\{j\}} \setminus \Omega_j^r) \times (0,1)$, which is a null set. □

Lemma E.2 (Conditional identity forces score independence)**.** *Let* r *be as in Lemma E.1 and additionally* C^1*. If its conditional-CDF class satisfies*

$$C_j^r(u) = u_j \quad \textit{for Lebesgue-almost every } u, \tag{58}$$

then

$$\partial_{u_j} \log r \equiv 0 \quad \textit{on } (0,1)^R. \tag{59}$$

Proof. By Fubini, for almost every $w \in \Omega_j^r$ the equality in equation 58 holds for almost every u_j. Both sides are continuous in u_j on that fiber, so they agree everywhere there. Their almost-everywhere derivatives therefore agree, and equation 56 gives $K_j^r(w,u_j) = 1$ for almost every u_j. Hence $r(w,u_j) = \mathcal{N}_j^r(w)$ for almost every u_j. Continuity of r in u_j upgrades this equality to every u_j on every such good fiber. Therefore $\partial_{u_j} \log r = 0$ on a full-measure subset of $(0,1)^R$. This score is continuous because r is positive and C^1; a continuous function that vanishes almost everywhere on an open set vanishes everywhere, proving equation 59. □

Lemma E.3 (Foliation transfers absent-coordinate scores)**.** *Let* r *and* s *be positive* C^1 *probability densities on* $(0,1)^R$*, and let* $j, \ell \in R$ *be distinct. Suppose that* s *is independent of* u_ℓ *and that*

$$[C_j^r] \sim_{\mathrm{fol}} [C_j^s]. \tag{60}$$

Then $\partial_{u_j} \log r$ *is independent of* u_ℓ *on the entire open cube.*

Proof. Write $w = u_{R\setminus\{j,\ell\}}$. Since s is independent of u_ℓ, both its fiber normalizer in the u_j direction and its conditional density are independent of u_ℓ wherever they are finite. We may therefore choose one measurable conditional-CDF representative with values in $(0,1)$, $\widetilde{C}_j^s(w,u_j)$ for all u_ℓ, setting it equal to u_j on exceptional w-fibers. By equation 60, there is an increasing homeomorphism Γ such that $G(w,u_j) := \Gamma(\widetilde{C}_j^s(w,u_j))$ is a representative of C_j^r that is independent of u_ℓ. Indeed, changing $\widetilde{C}_j^s$ on a null set changes its composition with the fixed continuous map Γ only on that same null set.

Fubini's theorem supplies a full-measure set of w such that the set

$$A_w := \left\{ a \in (0,1) : \begin{array}{l} \text{the } r\text{-conditional fiber at } (w,a) \text{ is good, and} \\ C_j^r(w,a,u_j) = G(w,u_j) \text{ for almost every } u_j \end{array} \right\}$$

has full measure in $(0,1)$. Equivalently, for every $a \in A_w$,

$$C_j^r(w,a,u_j) = G(w,u_j) \quad \text{for almost every } u_j. \tag{61}$$

Fix such a w and any $a, b \in A_w$. The corresponding two fibers of C_j^r are absolutely continuous in u_j by Lemma E.1, and both agree almost everywhere with $G(w,\cdot)$ by equation 61. They therefore agree everywhere, and their derivatives K_j^r agree almost everywhere. Because r is positive and C^1 and the fiber normalizers are positive constants in u_j, the two conditional densities are positive C^1 functions of u_j. Their almost-everywhere equality therefore holds everywhere, as does the equality of their log derivatives. On a good fiber,

$$K_j^r = \frac{r}{\mathcal{N}_j^r}, \qquad \partial_{u_j} \log K_j^r = \partial_{u_j} \log r,$$

because $\mathcal{N}_j^r$ is finite, positive, and independent of u_j. The u_j-log derivatives are thus equal for almost every pair of good u_ℓ-fibers. Put $\psi = \partial_{u_j} \log r$ and define the pairwise difference

$$\Delta_\psi(w,a,b,u_j) := \psi(w,a,u_j) - \psi(w,b,u_j).$$

The preceding argument and Fubini's theorem give $\Delta_\psi = 0$ almost everywhere on its open product domain. Since r is positive and C^1, ψ and hence Δ_ψ are continuous. A continuous function that is zero almost everywhere on an open set is zero everywhere. Thus ψ is independent of u_ℓ at every point. □

Lemma E.4 (Normalized-score rigidity)**.** *Let $\rho(s,t,w) > 0$ be C^1 on a product of open intervals, with*

$$\int_0^1 \rho(s,t,w)\ \mathrm{d}t = 1 \quad \textit{for every } (s,w).$$

If $\partial_s \log \rho(s,t,w)$ is independent of t, then it is identically zero.

Proof. Fix an arbitrary value of w and an arbitrary base point s_0 in the s-interval. By hypothesis there is a function b such that

$$\partial_s \log \rho(s,t,w) = b(s,w) \qquad \text{for every } t. \tag{62}$$

Define

$$\Xi(s,w) := \int_{s_0}^s b(\xi,w)\ \mathrm{d}\xi. \tag{63}$$

Since $\rho > 0$ and ρ is C^1, $\log \rho$ is C^1. The fundamental theorem of calculus applied to equation 62 gives, for every t,

$$\log \rho(s,t,w) - \log \rho(s_0,t,w) = \int_{s_0}^s \partial_\xi \log \rho(\xi,t,w)\ \mathrm{d}\xi = \int_{s_0}^s b(\xi,w)\ \mathrm{d}\xi = \Xi(s,w).$$

Exponentiating yields

$$\rho(s,t,w) = e^{\Xi(s,w)} \rho(s_0,t,w).$$

Integrate both sides over $t \in (0,1)$. The normalization at s and at s_0 gives

$$1 = \int_0^1 \rho(s,t,w)\ \mathrm{d}t = e^{\Xi(s,w)} \int_0^1 \rho(s_0,t,w)\ \mathrm{d}t = e^{\Xi(s,w)}.$$

Because Ξ is real valued, $e^{\Xi(s,w)} = 1$ implies $\Xi(s,w) = 0$ for every s. Differentiating equation 63 now gives $b(s,w) = \partial_s \Xi(s,w) = 0$. Hence equation 62 vanishes identically. □

Write $\alpha^e_{ji} = \partial_{u_j} \log \rho^e_i$.

Lemma E.5 (Terminal response transfer)**.** *Suppose j is a strict ancestor of ℓ and $\partial_{u_j} F_\ell \not\equiv 0$. For every environment e targeting ℓ,*

$$\alpha^e_{j\ell} \not\equiv 0, \qquad \alpha^e_{j\ell} \text{ varies with } u_\ell.$$

Proof. Write $I_\ell = (\alpha_\ell, \beta_\ell)$, allowing infinite endpoints, and define $N^e_\ell(z) := \int_{\alpha_\ell}^{z} \nu^e_\ell(y)\, \mathrm{d}y$.

Positivity and normalization of ν^e_ℓ make $N^e_\ell : I_\ell \to (0,1)$ a strictly increasing C^1 diffeomorphism.

Fix $u_{-\ell}$ and put $\varphi(t) := F_\ell(u_{-\ell}, t)$. Scalar noise solvability and the fixed conditional range imply

$$\varphi'(t) = \partial_{u_\ell} F_\ell(u_{-\ell}, t) > 0, \qquad \varphi(0+) = \alpha_\ell, \qquad \varphi(1-) = \beta_\ell.$$

Using the definition $\rho^e_\ell = \nu^e_\ell(F_\ell)\partial_{u_\ell} F_\ell$ and substituting $y = \varphi(t)$ gives, for every $u_\ell \in (0,1)$,

$$\int_0^{u_\ell} \rho^e_\ell(u_{-\ell}, t)\, \mathrm{d}t = \int_0^{u_\ell} \nu^e_\ell(\varphi(t))\varphi'(t)\, \mathrm{d}t = \int_{\alpha_\ell}^{\varphi(u_\ell)} \nu^e_\ell(y)\, \mathrm{d}y = N^e_\ell(F_\ell(u)). \tag{64}$$

We prove the first assertion by contradiction. Suppose $\alpha^e_{j\ell} \equiv 0$. Since $\rho^e_\ell > 0$,

$$0 = \alpha^e_{j\ell} = \partial_{u_j} \log \rho^e_\ell = \frac{\partial_{u_j} \rho^e_\ell}{\rho^e_\ell} \quad \Longrightarrow \quad \partial_{u_j} \rho^e_\ell \equiv 0. \tag{65}$$

For every fixed value of the other coordinates, the u_j-interval is connected. The fundamental theorem of calculus applied to equation 65 therefore shows that ρ^e_ℓ is independent of u_j. The integral on the left side of equation 64 is then independent of u_j, and so is $N^e_\ell(F_\ell(u))$. Applying the inverse of the injective map N^e_ℓ gives that $F_\ell(u)$ is independent of u_j. Hence $\partial_{u_j} F_\ell \equiv 0$, contradicting the response hypothesis. We conclude that $\alpha^e_{j\ell} \not\equiv 0$.

For the second assertion, suppose that $\alpha^e_{j\ell}$ is independent of u_ℓ. In Lemma E.4, set

$$s = u_j, \qquad t = u_\ell, \qquad w = u_{[d]\setminus\{j,\ell\}}, \qquad \rho = \rho^e_\ell.$$

The variables j and ℓ are distinct because j is a strict ancestor of the terminal descendant ℓ. The required normalization is exactly equation 9. The rigidity lemma would therefore imply $\alpha^e_{j\ell} \equiv 0$, contradicting the first part. Thus $\alpha^e_{j\ell}$ varies with u_ℓ. □

Lemma E.6 (Within-environment response independence)**.** *Fix (e, j) and an ancestor-closed active set R. The nonzero functions $\{\alpha^e_{ji} : i \in S_e \cap R,\ \alpha^e_{ji} \not\equiv 0\}$ are linearly independent over constant coefficients.*

Proof. Let $\mathcal{I}^R_{e,j} := \{i \in S_e \cap R : \alpha^e_{ji} \not\equiv 0\}$, $J := \{i \in \mathcal{I}^R_{e,j} : c_i \neq 0\}$. Because ρ^e_i depends only on u_i and residual ancestors of i,

$$i \in \mathcal{I}^R_{e,j} \quad \Longrightarrow \quad i = j \ \text{ or } \ j \in \mathrm{an}(i). \tag{66}$$

Suppose, for contradiction, that the family is linearly dependent over constant coefficients. Then there are constants c_i, not all zero, such that

$$\sum_{i \in \mathcal{I}^R_{e,j}} c_i \alpha^e_{ji}(u_R) = 0 \quad \text{for every } u_R \in (0,1)^R. \tag{67}$$

The set J defined above is finite and nonempty. Choose $i^\star \in J$ that is maximal under the DAG ancestor order restricted to J.

First consider $i^\star = j$. By equation 66, every other element of J would be a proper descendant of j, contradicting maximality of j. Therefore $J = \{j\}$, and equation 67 reduces to $c_j \alpha^e_{jj} \equiv 0$. Since $c_j \neq 0$, this contradicts $\alpha^e_{jj} \not\equiv 0$, the definition of $j \in \mathcal{I}^R_{e,j}$.

It remains to consider $i^\star \neq j$. Fix any $k \in J \setminus \{i^\star\}$. If α^e_{jk} depended on $u_{i^\star}$, then ρ^e_k would depend on that coordinate. The ancestral dependence of ρ^e_k would imply $i^\star = k$ or $i^\star \in \mathrm{an}(k)$.

The first alternative is excluded by the choice of k, and the second says that k is a proper descendant of $i^\star$, contradicting maximality. Thus every α^e_{jk} with $k \in J \setminus \{i^\star\}$ is independent of $u_{i^\star}$. Solving equation 67 for the maximal term gives

$$\alpha^e_{ji^\star} = -\frac{1}{c_{i^\star}} \sum_{k \in J \setminus \{i^\star\}} c_k \alpha^e_{jk}, \tag{68}$$

so $\alpha^e_{ji^\star}$ is also independent of $u_{i^\star}$.

Apply Lemma E.4 to $\rho^e_{i^\star}$ with $s = u_j,\ \ t = u_{i^\star},\ \ w = u_{R\setminus\{j,i^\star\}}$. Separate normalization supplies the lemma's integral hypothesis, and equation 68 supplies score independence from t. The lemma gives $\alpha^e_{ji^\star} \equiv 0$, contradicting $i^\star \in \mathcal{I}^R_{e,j}$. Thus only the trivial constant-coefficient relation exists, and the family is linearly independent. □

Lemma E.7 (Smooth foliation criterion)**.** *Let $C_1, C_2 : (0,1)^R \to (0,1)$ be C^1 conditional CDFs with positive u_j derivatives and boundary limits zero and one as u_j tends to zero and one. Then*

$$\mathrm{d}C_1 \wedge \mathrm{d}C_2 = 0 \quad \Longleftrightarrow \quad [C_1] \sim_{\mathrm{fol}} [C_2].$$

Whenever either condition holds, the relation has a pointwise representative $C_1 = \Gamma \circ C_2$ with an increasing C^1 diffeomorphism $\Gamma : (0,1) \to (0,1)$.

Proof. Suppose first that $[C_1] \sim_{\mathrm{fol}} [C_2]$. By definition there is an increasing homeomorphism Γ such that $C_1 = \Gamma \circ C_2$ almost everywhere. Both sides are continuous, so the equality holds everywhere. Fix a conditioning value $w_0 \in (0,1)^{R\setminus\{j\}}$, interpreted as the unique point when $R = \{j\}$, and write $C_{2,w_0}(t) = C_2(w_0, t)$. Positivity of $\partial_{u_j} C_2$ and the boundary limits make C_{2,w_0} a C^1 diffeomorphism from $(0,1)$ onto $(0,1)$. The pointwise equality identifies the original homeomorphism as

$$\Gamma(c) = C_1\big(w_0, C^{-1}_{2,w_0}(c)\big).$$

The one-dimensional inverse function theorem now gives $\Gamma \in C^1$ and

$$\Gamma'(c) = \frac{\partial_{u_j} C_1\big(w_0, C^{-1}_{2,w_0}(c)\big)}{\partial_{u_j} C_2\big(w_0, C^{-1}_{2,w_0}(c)\big)} > 0. \tag{69}$$

The map Γ was already a homeomorphism of $(0,1)$ onto itself, and equation 69 shows that its derivative is everywhere nonzero. Applying the inverse function theorem at every point gives a C^1 inverse, so Γ is a C^1 diffeomorphism. The chain rule gives $\mathrm{d}C_1 = \Gamma'(C_2)\ \mathrm{d}C_2$. The wedge product of a one-form with a scalar multiple of itself is zero, so $\mathrm{d}C_1 \wedge \mathrm{d}C_2 = \Gamma'(C_2)\ \mathrm{d}C_2 \wedge \mathrm{d}C_2 = 0$.

Conversely, assume $\mathrm{d}C_1 \wedge \mathrm{d}C_2 = 0$. At an arbitrary point u and for every coordinate $k \neq j$, the coefficient of $\mathrm{d}u_k \wedge \mathrm{d}u_j$ is zero:

$$\partial_{u_k} C_1(u)\, \partial_{u_j} C_2(u) - \partial_{u_j} C_1(u)\, \partial_{u_k} C_2(u) = 0. \tag{70}$$

Because $\partial_{u_j} C_2 > 0$, define $\alpha(u) := \frac{\partial_{u_j} C_1(u)}{\partial_{u_j} C_2(u)} > 0$. Dividing equation 70 by $\partial_{u_j} C_2(u)$ gives $\partial_{u_k} C_1(u) = \alpha(u) \partial_{u_k} C_2(u)$ $(k \neq j)$. The same identity holds for $k = j$ by the definition of α, hence

$$\mathrm{d}C_1(u) = \alpha(u)\ \mathrm{d}C_2(u) \qquad \text{for every } u. \tag{71}$$

Let $\mathcal{W} := (0,1)^{R\setminus\{j\}}$; when $|R| = 1$, interpret $\mathcal{W}$ as a singleton. Fix $c \in (0,1)$ and $w \in \mathcal{W}$. The map $t \longmapsto C_2(w,t)$ is strictly increasing, with limits zero and one at the two endpoints. Therefore there is a unique number $s_c(w) \in (0,1)$ satisfying

$$C_2(w, s_c(w)) = c. \tag{72}$$

Since $\partial_{u_j} C_2 > 0$, the implicit function theorem makes s_c locally C^1. Uniqueness in equation 72 makes the local solutions agree on overlaps, so s_c is a globally defined C^1 function. Consequently,

$$C_2^{-1}(c) = \{(w, s_c(w)) : w \in \mathcal{W}\}.$$

For any $w_0, w_1 \in \mathcal{W}$, convexity of $\mathcal{W}$ and the C^1 regularity of s_c make

$$\gamma(t) := \big((1-t)w_0 + tw_1,\, s_c((1-t)w_0 + tw_1)\big)$$

a C^1 path in $C_2^{-1}(c)$. Hence every fiber is C^1 path connected. Along such a path, equation 71 gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}C_1(\gamma(t)) &= (\mathrm{d}C_1)_{\gamma(t)}[\gamma'(t)] \\ &= \alpha(\gamma(t))(\mathrm{d}C_2)_{\gamma(t)}[\gamma'(t)] \\ &= \alpha(\gamma(t))\frac{\mathrm{d}}{\mathrm{d}t}C_2(\gamma(t)) = 0.\end{aligned}$$

Thus C_1 is constant on every fiber of C_2.

Choose any fixed base point $w_0 \in \mathcal{W}$ and define

$$\Gamma(c) := C_1(w_0, s_c(w_0)). \tag{73}$$

Constancy on fibers makes this definition independent of the selected point in $C_2^{-1}(c)$, and therefore

$$C_1(u) = \Gamma(C_2(u)) \qquad \text{for every } u. \tag{74}$$

The implicit function theorem applied to $C_2(w_0, s_c(w_0)) = c$ gives

$$\frac{\mathrm{d}}{\mathrm{d}c}s_c(w_0) = \frac{1}{\partial_{u_j}C_2(w_0, s_c(w_0))}.$$

Differentiating equation 73 yields

$$\Gamma'(c) = \partial_{u_j}C_1(w_0, s_c(w_0))\frac{\mathrm{d}}{\mathrm{d}c}s_c(w_0) = \frac{\partial_{u_j}C_1(w_0, s_c(w_0))}{\partial_{u_j}C_2(w_0, s_c(w_0))} > 0.$$

Thus Γ is C^1 and strictly increasing. To prove surjectivity, fix $y \in (0,1)$. The conditional-CDF boundary values and positivity of $\partial_{u_j}C_1$ give a unique t_y with $C_1(w_0, t_y) = y$. Put $c_y = C_2(w_0, t_y)$. Then equation 74 gives $\Gamma(c_y) = y$. Hence Γ maps $(0,1)$ onto $(0,1)$. Its positive derivative and the inverse function theorem show that its inverse is C^1, so Γ is the required increasing diffeomorphism. □

F THREE-NODE PEELING EXAMPLE: CALCULATIONS

The chain $1 \to 2 \to 3$ verifies each population invariant used by the peeling argument. Recall its parameters $0 < a, b < 1/2$ and

$$\begin{aligned}F_1(u) &= u_1, \\ F_2(u) &= u_2 + a(2u_1 - 1)u_2(1 - u_2), \\ F_3(u) &= u_3 + b(2F_2(u) - 1)u_3(1 - u_3).\end{aligned}$$

The three target sets are $S_{\mathrm{A}} = \{1,3\},\ S_{\mathrm{B}} = \{1,3\},\ S_{\mathrm{C}} = \{1,2\}$. Node 1 is replaced uniformly whenever targeted; node 3 is uniform in A and has density $\nu_3^{\mathrm{B}}(z) = 1 + \delta(2z-1)$ in B; node 2 has density $\nu_2^{\mathrm{C}}(z) = 1 + \gamma(2z-1)$ in C, where $0 < \delta, \gamma < 1$. The relevant derivatives and ratios are:

$$\begin{aligned}&\partial_{u_2}F_2 = 1 + a(2u_1 - 1)(1 - 2u_2) > 1 - a > 0, \\ &\partial_{u_1}F_2 = 2au_2(1 - u_2) > 0, \\ &\partial_{u_3}F_3 = 1 + b(2F_2 - 1)(1 - 2u_3) > 1 - b > 0, \\ &\partial_{u_2}F_3 = 2bu_3(1 - u_3)\partial_{u_2}F_2 > 0, \\ &\partial_{u_1}F_3 = 4abu_2(1 - u_2)u_3(1 - u_3) > 0.\end{aligned}$$

Thus terminal node 3 responds to both ancestors on $\{1,2,3\}$, and after it is peeled, terminal node 2 responds to node 1 on $\{1,2\}$. The node signatures, in the order $(\mathrm{A}, \mathrm{B}, \mathrm{C})$, are $\boldsymbol{\sigma}_1 = (1,1,1)^\top$, $\boldsymbol{\sigma}_2 = (0,0,1)^\top$, and $\boldsymbol{\sigma}_3 = (1,1,0)^\top$.

Uniform replacements of node 1 contribute the unit ratio factor $\rho_1^e(u) = 1 \cdot \partial_{u_1}F_1(u) = 1$ but remain visible in the paired support. Ratio factorization therefore gives

$$\begin{aligned}r_{\mathrm{A}}(u) &= \partial_{u_3}F_3(u), \\ r_{\mathrm{B}}(u) &= \big[1 + \delta(2F_3(u) - 1)\big]\partial_{u_3}F_3(u), \\ r_{\mathrm{C}}(u) &= \big[1 + \gamma(2F_2(u) - 1)\big]\partial_{u_2}F_2(u).\end{aligned}$$

For node 3, the environment-C ratio is independent of u_3. The other two ratios are normalized derivatives along u_3. With $N_{\mathrm{B}}(z) = z + \delta(z^2 - z)$,

$$C_{3,\mathrm{C}} = u_3, \qquad C_{3,\mathrm{A}} = F_3, \qquad C_{3,\mathrm{B}} = N_{\mathrm{B}}(F_3).$$

Since $N'_{\mathrm{B}}(z) > 1 - \delta$ and its endpoint limits are zero and one, N_{B} is an increasing homeomorphism. The target CDFs therefore induce the same oriented foliation.

Node 2 remains a non-sink while node 3 is active. On the slice $u_1 = 1/2$, we have $F_2(1/2, t) = t$. Hence the environment-A ratio, viewed as a function of the integration variable t, is

$$r_{\mathrm{A}}(1/2, t, u_3) = \partial_{u_3} F_3(1/2, t, u_3) = 1 + b(2t - 1)(1 - 2u_3).$$

Its u_2-normalizer is

$$\mathcal{N}_{2,\mathrm{A}}(1/2, u_3) = \int_0^1 [1 + b(2t - 1)(1 - 2u_3)] \, \mathrm{d}t = 1 + b(1 - 2u_3)[t^2 - t]_{t=0}^{t=1} = 1.$$

The normalized conditional CDF is therefore

$$C_{2,\mathrm{A}}(1/2, u_2, u_3) = \int_0^{u_2} [1 + b(2t - 1)(1 - 2u_3)] \, \mathrm{d}t = u_2 + b(1 - 2u_3)(u_2^2 - u_2).$$

At $(u_2, u_3) = (1/2, 1/4)$ the value is $1/2 - b/8$. Continuity extends the inequality from the identity to a positive-measure neighborhood.

Marginalizing the recovered sink removes its separately normalized factor:

$$r_{\mathrm{A}}^{\{1,2\}} = r_{\mathrm{B}}^{\{1,2\}} = 1, \qquad r_{\mathrm{C}}^{\{1,2\}} = \big[1 + \gamma(2F_2 - 1)\big]\partial_{u_2} F_2.$$

Node 2 is now a sink: $C_{2,\mathrm{A}}^{\{1,2\}} = C_{2,\mathrm{B}}^{\{1,2\}} = u_2$ and $C_{2,\mathrm{C}}^{\{1,2\}} = N_{\mathrm{C}}(F_2)$, where $N_{\mathrm{C}}(z) = z + \gamma(z^2 - z)$. Integrating the last factor over u_2 leaves all one-dimensional ratios equal to one, completing order 3,2,1.

All conditional normalizers in this smooth example are finite on every fiber. For a general positive integrable ratio, Tonelli's theorem guarantees finiteness for almost every conditioning value. The conditional-law definition in Section 3.3 absorbs arbitrary choices on exceptional fibers, while a true sink retains the unique continuous representative u_j or $N_j^e(F_j)$.

G Proof of the sink characterization

The two conditional-law tests are jointly sufficient and necessary for sinkhood: factorization proves the forward direction, while terminal response excludes every active non-sink.

Proof of Theorem 4.2. Fix the minimum-code minimizer in the theorem. By Theorem 4.1, relabel the recovered coordinates so that $\lambda_{ej} = \sigma_{ej}$ for all (e, j). The reflection conjugacy in Lemma D.1 preserves both tests, so it suffices to work in the true residual orientation $h = q$. We first prove the forward implication. Assume that j is a sink of $\mathcal{G}[R]$. By the ancestral dependence established in Lemma D.2, a factor ρ_i^e can depend on u_j only if $i = j$ or if j is an ancestor of i. Sinkhood places every proper descendant of j outside R, so each factor with $i \neq j$ is independent of u_j. Define

$$\beta_{j,e}^R(u_{R\setminus\{j\}}) := \prod_{i \in S_e \cap R,\ i \neq j} \rho_i^e(u_R).$$

The preceding argument shows that $\beta_{j,e}^R$ does not depend on u_j. By Lemma D.3,

$$r_e^R(u_R) = \beta_{j,e}^R(u_{R\setminus\{j\}}) \begin{cases} 1, & j \notin S_e, \\ \rho_j^e(u_R), & j \in S_e. \end{cases}$$

Let

$$\mathcal{N}_{j,e}^R(u_{R\setminus\{j\}}) := \int_0^1 r_e^R(u_{R\setminus\{j\}}, t) \, \mathrm{d}t$$

be the conditional normalizer. If $j \notin S_e$, then

$$\mathcal{N}^R_{j,e} = \beta^R_{j,e} \int_0^1 1 \, \mathrm{d}t = \beta^R_{j,e},$$

and hence $K^R_{j,e} = \beta^R_{j,e}/\beta^R_{j,e} = 1$. If $j \in S_e$, separate normalization instead gives

$$\mathcal{N}^R_{j,e} = \beta^R_{j,e} \int_0^1 \rho^e_j(u_{R\setminus\{j\}}, t) \, \mathrm{d}t = \beta^R_{j,e},$$

and therefore $K^R_{j,e} = \rho^e_j$. Combining the cases gives

$$K^R_{j,e} = \begin{cases} 1, & j \notin S_e, \\ \rho^e_j, & j \in S_e. \end{cases}$$

Here $\beta^R_{j,e}$ is finite and positive at every interior point. Thus a true sink has a pointwise conditional formula on every fiber.

Integrating these conditional densities, the non-target case gives

$$C^R_{j,e}(u_R) = \int_0^{u_j} 1 \, \mathrm{d}t = u_j.$$

In the target case, the substitution already derived in equation 64, now with $\ell = j$, gives

$$C^R_{j,e}(u_R) = \int_0^{u_j} \nu^e_j(F_j(u_{R\setminus\{j\}}, t)) \partial_t F_j(u_{R\setminus\{j\}}, t) \, \mathrm{d}t = N^e_j(F_j(u_R)).$$

Thus

$$C^R_{j,e} = \begin{cases} u_j, & j \notin S_e, \\ N^e_j(F_j(u_R)), & j \in S_e. \end{cases}$$

These pointwise functions are representatives of the regular conditional-CDF classes, so the almost-everywhere non-target identity holds.

For two environments e, e' targeting j, define $\Gamma_{e,e'} := N^e_j \circ (N^{e'}_j)^{-1}$. Both replacement CDFs are increasing diffeomorphisms, so $\Gamma_{e,e'}$ is an increasing diffeomorphism of $(0, 1)$. Moreover,

$$\begin{aligned} C^R_{j,e} &= N^e_j(F_j) \\ &= \Gamma_{e,e'}(N^{e'}_j(F_j)) \\ &= \Gamma_{e,e'} \circ C^R_{j,e'}. \end{aligned}$$

Thus the target-foliation test holds. A recovered residual reflection is handled by Lemma D.1. If coordinate j is reflected, then, in the reflected input $\widetilde{u}$,

$$\widetilde{C}^R_{j,e}(\widetilde{u}_R) = 1 - N^e_j\big(F^R_j(\mathsf{R}_R(\widetilde{u}_R))\big).$$

This conjugates $\Gamma_{e,e'}$ by $s \mapsto 1 - s$ and preserves its orientation.

Conversely, suppose j is not a sink. By Assumption 2.1, choose a sink descendant ℓ in $\mathcal{G}[R]$ such that $\partial_{u_j} F^R_\ell \not\equiv 0$. Because R is ancestor closed and $\ell \in R$, every residual ancestor of ℓ also lies in R. Thus the full solution coordinate factors through the projection $\mathrm{pr}_R : (0,1)^d \to (0,1)^R$: $F_\ell = F^R_\ell \circ \mathrm{pr}_R$, $\partial_{u_j} F_\ell = \big(\partial_{u_j} F^R_\ell\big) \circ \mathrm{pr}_R$. Hence the TRF response is nonzero both on the restricted active cube and in the full-coordinate notation used by Lemma E.5. There are two exhaustive cases.

If $\mathrm{supp}(\boldsymbol{\sigma}_\ell) \not\subseteq \mathrm{supp}(\boldsymbol{\sigma}_j)$, choose $e \in \mathrm{supp}(\boldsymbol{\sigma}_\ell) \setminus \mathrm{supp}(\boldsymbol{\sigma}_j)$. Then $\ell \in S_e$ and $j \notin S_e$. By Lemma E.5, $\alpha^e_{j\ell} \not\equiv 0$. Positivity of all ratio factors allows us to differentiate the logarithm of the finite product:

$$\partial_{u_j} \log r^R_e = \partial_{u_j} \sum_{i \in S_e \cap R} \log \rho^e_i = \sum_{i \in S_e \cap R} \alpha^e_{ji}.$$

After zero functions are omitted, this is a constant-coefficient relation with every coefficient equal to one. The response family is linearly independent by Lemma E.6, and it contains the nonzero ℓ term. Therefore

$$\partial_{u_j} \log r_e^R \not\equiv 0. \tag{75}$$

Suppose the almost-everywhere non-target identity held. Lemma E.2, applied to the positive C^1 density r_e^R, would give

$$\partial_{u_j} \log r_e^R \equiv 0,$$

contradicting equation 75. Hence the non-target identity fails under the regular-conditional formulation.

It remains to consider $\mathrm{supp}(\boldsymbol{\sigma}_\ell) \subseteq \mathrm{supp}(\boldsymbol{\sigma}_j)$. Distinct signatures make this inclusion strict. Choose environments $e \in \mathrm{supp}(\boldsymbol{\sigma}_\ell), e' \in \mathrm{supp}(\boldsymbol{\sigma}_j) \setminus \mathrm{supp}(\boldsymbol{\sigma}_\ell)$. Environment e targets both j and ℓ, whereas e' targets j but not ℓ. Since ℓ is an active sink, dependence on u_ℓ could arise from the ℓ-indexed factor or a descendant-indexed factor. The former is absent because $\ell \notin S_{e'}$, and sinkhood excludes the latter. Hence $\partial_{u_\ell} r_{e'}^R \equiv 0$. Since the u_ℓ-fiber is an interval and $r_{e'}^R$ is C^1, the fundamental theorem of calculus makes $r_{e'}^R$ independent of u_ℓ. Suppose, for contradiction, that the almost-everywhere target-foliation test held for e and e':

$$C_{j,e}^R = \Gamma \circ C_{j,e'}^R \qquad \text{almost everywhere}$$

for an increasing homeomorphism Γ. Applying Lemma E.3 with $r = r_e^R$ and $s = r_{e'}^R$ shows that $\partial_{u_j} \log r_e^R$ is independent of u_ℓ on the entire cube. On the other hand, the product factorization gives the pointwise identity

$$\partial_{u_j} \log r_e^R = \sum_{i \in S_e \cap R} \alpha_{ji}^e.$$

Because ℓ is a sink, the term $\alpha_{j\ell}^e$ is the only term on the right that can depend on u_ℓ: every other factor is indexed by a node outside $\{\ell\} \cup \mathrm{de}(\ell)$. Thus

$$\alpha_{j\ell}^e = \partial_{u_j} \log r_e^R - \sum_{i \in (S_e \cap R) \setminus \{\ell\}} \alpha_{ji}^e$$

is a difference of functions that are all independent of u_ℓ. Consequently $\alpha_{j\ell}^e$ is independent of u_ℓ on the entire cube. This contradicts Lemma E.5. Hence the target-foliation test fails. The two signature cases are exhaustive, so every non-sink fails at least one test, proving the reverse implication. □

H Full causal identification from the observed laws

Iterating the sink characterization preserves ancestor closure, recovers one endogenous coordinate per step, and exposes direct edges through the inverse Jacobian.

Proof of Theorem 4.3. Fix an arbitrary global minimizer of the minimum-code problem and an arbitrary sequence of admissible peeling choices. Here *admissible* means choosing a coordinate that passes both observable sink tests and then an environment e_j with $\lambda_{e_j,j} = 1$. By Theorem 4.1, relabel coordinates by the minimizer's matched permutation. By Lemma D.1, it is enough to prove the statement in the true residual orientation; the reflected conclusion follows by conjugation.

Let $R_0 = [d]$. After t nodes have been removed, let $\overline{R}_t = [d] \setminus R_t$. We maintain the following four invariants:

(I1) R_t is ancestor closed;

(I2) for every environment e,

$$r_e^{R_t}(u_{R_t}) = \prod_{i \in S_e \cap R_t} \rho_i^e(u_{R_t}); \tag{76}$$

(I3) for every removed node $k \in \overline{R}_t$, the constructed coordinate has the form $Y_k = M_k \circ F_k$ for a scalar strictly monotone diffeomorphism M_k;

(I4) every choice through step t belongs to a nonempty set determined by the observed law family, and the recovered coordinate has the form in (I3) independently of that choice.

At $t = 0$, $R_0 = [d]$ is ancestor closed, equation 76 is Lemma D.2, the third invariant is vacuous, and the fourth follows from the arbitrary global minimizer fixed above. Thus the base case holds.

Assume the invariants at some $t < d$. By Theorem 4.2, the observable tests on R_t identify exactly the sinks of $\mathcal{G}[R_t]$. Choose one, denoted j_t, according to the arbitrary choice sequence fixed at the outset. Its recovered target signature is nonzero, so the set of targeting environments is nonempty. Choose any e_{j_t} with $\lambda_{e_{j_t},j_t} = 1$. The target-CDF class contains the continuous function $N_{j_t}^{e_{j_t}} \circ F_{j_t}$. Its continuous representative is unique: two continuous functions that agree almost everywhere on the open cube agree everywhere. Select this representative and define

$$Y_{j_t}(u_{R_t}) := C_{j_t,e_{j_t}}^{R_t}(u_{R_t}) = N_{j_t}^{e_{j_t}}(F_{j_t}(u_{R_t})) = M_{j_t}(F_{j_t}(u_{R_t})), \tag{77}$$

where $M_{j_t} := N_{j_t}^{e_{j_t}}$ is a scalar increasing diffeomorphism from I_{j_t} to $(0, 1)$. In a reflected residual orientation, Lemma D.1 replaces this with a decreasing scalar diffeomorphism, which remains inside the allowed ambiguity. We regard Y_{j_t} as a function on the full cube by composition with the projection onto u_{R_t}. Ancestor closure ensures that F_{j_t} depends on these retained coordinates, so this extension preserves equation 77.

Sinkhood places every descendant of j_t outside $R_t \setminus \{j_t\}$. Removing j_t therefore preserves all ancestors of the remaining nodes, and $R_{t+1} := R_t \setminus \{j_t\}$ is ancestor closed. Lemma D.3 applied to this new ancestor-closed set gives $r_e^{R_{t+1}} = \prod_{i \in S_e \cap R_{t+1}} \rho_i^e$ for every e.

The construction equation 77 establishes the third invariant for the newly removed node, and the fixed choices establish the fourth because every admissible sink and targeting environment gives a scalar monotone transform of the same endogenous coordinate. Assumption 2.1 is quantified over every ancestor-closed active set, so the sink theorem remains applicable at the next step. This completes the induction step.

Exactly one node is removed per iteration, so $R_d = \varnothing$. Collecting the d scalar identities yields, after the common permutation and possible residual reflections,

$$Y(u) = M(F(u)), \qquad M(z) := (M_1(z_1), \ldots, M_d(z_d)). \tag{78}$$

Each M_i is a scalar diffeomorphism and F is a triangular diffeomorphism. Their composition Y is therefore a diffeomorphism, proving claims (i) and (ii).

For direct-edge identification, let $\Psi := F^{-1}$ and fix node i. The inverse structural equation is

$$f_i(z_{\mathrm{pa}(i)}, \Psi_i(z)) = z_i. \tag{79}$$

For fixed $(z_{\mathrm{pa}(i)}, z_i)$, the scalar equation in equation 79 has a unique solution because $u_i \mapsto f_i(z_{\mathrm{pa}(i)}, u_i)$ is a diffeomorphism. Denote this inverse by

$$\eta_i(z_{\mathrm{pa}(i)}, z_i) := f_i(z_{\mathrm{pa}(i)}, \cdot)^{-1}(z_i), \qquad \Psi_i(z) = \eta_i(z_{\mathrm{pa}(i)}, z_i). \tag{80}$$

The parametric implicit-function theorem gives $\eta_i \in C^1$ because $\partial_{u_i} f_i > 0$. For a displayed parent $j \in \mathrm{pa}(i)$, differentiate equation 79 with respect to z_j, holding z_i fixed:

$$\begin{aligned} 0 &= \partial_{z_j} f_i(z_{\mathrm{pa}(i)}, \Psi_i(z)) + \partial_{u_i} f_i(z_{\mathrm{pa}(i)}, \Psi_i(z)) \partial_{z_j} \Psi_i(z), \\ \partial_{z_j} \Psi_i(z) &= -\frac{\partial_{z_j} f_i}{\partial_{u_i} f_i}\bigg|_{u_i = \Psi_i(z)}. \end{aligned} \tag{81}$$

Structural activity means that f_i varies with z_j. Since the parent domain is connected and f_i is C^1, this implies $\partial_{z_j} f_i \not\equiv 0$. Every mechanism point is attained by the inverse chart: for any $(z_{\mathrm{pa}(i)}, u_i)$, choose

$$z_i = f_i(z_{\mathrm{pa}(i)}, u_i).$$

Since F maps onto $\mathcal{Z} = \prod_k I_k$, the remaining endogenous coordinates can be chosen arbitrarily in their intervals, and equation 80 then gives $\Psi_i(z) = u_i$. Consequently a point where $\partial_{z_j} f_i$ is nonzero maps to a point where the numerator in equation 81 is nonzero. The denominator in equation 81 is everywhere positive, so

$$j \in \mathrm{pa}(i) \quad \Longrightarrow \quad \partial_{z_j} \Psi_i \not\equiv 0. \tag{82}$$

If $j \notin \mathrm{pa}(i)$ and $j \neq i$, then the inverse equation defining Ψ_i contains no z_j, and hence

$$\partial_{z_j} \Psi_i \equiv 0. \tag{83}$$

Equation 78 gives $H := Y^{-1} = \Psi \circ M^{-1}$. For $j \neq i$, the chain rule yields

$$\partial_{y_j} H_i(y) = \partial_{z_j} \Psi_i(M^{-1}(y))\,(M_j^{-1})'(y_j). \tag{84}$$

The derivative of the scalar diffeomorphism M_j^{-1} is nonzero everywhere. Thus equation 84 is identically zero if and only if $\partial_{z_j} \Psi_i$ is identically zero. Combining equation 82 and 83 proves

$$j \to i \quad \Longleftrightarrow \quad j \neq i \ \text{ and } \ \partial_{y_j} H_i \not\equiv 0. \tag{85}$$

Finally, let two admissible models induce the same observed law family. Their minimum-code feasible set and objective coincide. Fix any common global minimizer and any common admissible sequence of sink and environment choices; the argument above applies because all relevant candidate sets are determined by those laws. Minimum-code recovery relates the common chart to each model's residuals by a permutation and reflections, and the unique continuous representative at every chosen sink constructs one common Y. If their endogenous vectors are $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$, then, under matched labels,

$$Y_i = M_i^{(1)}(Z_i^{(1)}) = M_i^{(2)}(Z_i^{(2)}).$$

Solving for the second model gives

$$Z_i^{(2)} = (M_i^{(2)})^{-1} \circ M_i^{(1)}(Z_i^{(1)}),$$

a strictly monotone scalar transformation. The common inverse chart H and equation 85 give the same labeled DAG for both. Therefore the models are causally equivalent. □

H.1 Identification under atomic designs

Atomic environments expose every endogenous coordinate directly, eliminating the terminal-response condition required by compressed designs.

Proof of Proposition 4.5. For each node i, the row indexed by e_i contains a one in column i and zeros in every other column. Hence the target signatures are nonzero and pairwise distinct. Theorem 4.1 therefore applies to every global minimizer. Relabel by the recovered column permutation; by Lemma D.1, it suffices to work in the true residual orientation.

In the atomic environment e_i, ratio factorization contains exactly one factor:

$$r_{e_i}(u) = \rho_i^{e_i}(u) = \nu_i^{e_i}(F_i(u))\,\partial_{u_i} F_i(u).$$

Separate normalization gives the conditional normalizer along u_i:

$$\mathcal{N}_{i,e_i}^{[d]}(u_{-i}) = \int_0^1 r_{e_i}(u_{-i}, t)\ \mathrm{d}t = \int_0^1 \rho_i^{e_i}(u_{-i}, t)\ \mathrm{d}t = 1.$$

Write $I_i = (\alpha_i, \beta_i)$, allowing infinite endpoints, and let $N_i^{e_i}(z) := \int_{\alpha_i}^{z} \nu_i^{e_i}(y)\ \mathrm{d}y$ be the replacement CDF. Since $F_i(u_{-i}, \cdot)$ is an increasing diffeomorphism from $(0,1)$ onto I_i, the substitution $z = F_i(u_{-i}, t)$ gives

$$\begin{aligned} C_{i,e_i}^{[d]}(u) &= \frac{1}{\mathcal{N}_{i,e_i}^{[d]}(u_{-i})} \int_0^{u_i} \rho_i^{e_i}(u_{-i}, t)\ \mathrm{d}t \\ &= \int_0^{u_i} \nu_i^{e_i}(F_i(u_{-i}, t))\partial_t F_i(u_{-i}, t)\ \mathrm{d}t = \int_{\alpha_i}^{F_i(u)} \nu_i^{e_i}(z)\ \mathrm{d}z = N_i^{e_i}(F_i(u)). \end{aligned}$$

This identity holds for every node i. Thus the observed laws simultaneously supply

$$Y_i = M_i \circ F_i, \qquad M_i := N_i^{e_i},$$

for every node. Each M_i is an increasing scalar diffeomorphism, so $Y = M \circ F$ is a diffeomorphism. If residual coordinate i is reflected, Lemma D.1 replaces this coordinate by $1 - N_i^{e_i}(F_i)$, which is still a strictly monotone scalar transform.

Writing $H = Y^{-1}$, the inverse-mechanism argument in equations 79–85 identifies every direct edge. After column permutation, the rows e_i are recognizable as singleton rows of the recovered code matrix, so the observed laws also identify their target labels. Hence every global minimizer identifies the endogenous coordinates and direct-edge DAG under atomic environments. □

I SHARP INTERVENTION LOWER BOUND

The collision constructions below establish the exact necessity of nonzero, distinct target signatures for the worst-case design complexity $m^\star(d)$ in Corollary 4.4. The learner observes environment labels while the target sets remain latent.

Proposition I.1 (Signature collisions)**.** *For every $d \geq 2$ and every labeled target design, uniform causal identification requires:*

(a) every node has a nonzero signature;

(b) every pair of nodes has distinct signatures.

Violating either requirement yields two models with the same paired law family that satisfy scalar noise solvability, global diffeomorphic decoding, positive intrinsic density on each paired support, and terminal-response faithfulness, and they belong to different causal equivalence classes.

Proof. Zero signature. Let node j have the zero signature and choose any $p \neq j$. Define

$$\ell_{\mathrm{sig}}(t) = \frac{1}{1+e^{-t}}, \qquad b_a(z,u) = \ell_{\mathrm{sig}}\big(\log(u/(1-u)) + az\big), \quad a \neq 0.$$

Write $\mathrm{logit}(u) = \log(u/(1-u))$. Since $\ell'_{\mathrm{sig}}(t) = \ell_{\mathrm{sig}}(t)(1-\ell_{\mathrm{sig}}(t))$, direct differentiation gives

$$\partial_u b_a(z,u) = \frac{b_a(z,u)(1-b_a(z,u))}{u(1-u)} > 0, \tag{86}$$

$$\partial_z b_a(z,u) = a\, b_a(z,u)(1-b_a(z,u)) \neq 0. \tag{87}$$

Moreover,

$$\lim_{u\downarrow 0} b_a(z,u) = 0, \qquad \lim_{u\uparrow 1} b_a(z,u) = 1, \tag{88}$$

for every $z \in (0,1)$. Solving $x = b_a(z,u)$ for u gives

$$u = \ell_{\mathrm{sig}}\big(\mathrm{logit}(x) - az\big). \tag{89}$$

Equations 86–89 show that, for each fixed z, the map $u \mapsto b_a(z,u)$ is a smooth increasing diffeomorphism from $(0,1)$ onto the fixed interval $(0,1)$. Equation 87 makes the displayed parent structurally active.

In model A, let all variables except j be independent uniform roots, set

$$Z_j = b_a(Z_p, U_j),$$

and use the identity decoder $g_{\mathrm{A}}(z) = z$. Its only edge is $p \to j$. In model B, let all $W_i = U_i$ be independent uniform roots, use the empty DAG, and use the triangular decoder

$$(g_{\mathrm{B}}(w))_p = w_p, \qquad (g_{\mathrm{B}}(w))_j = b_a(w_p, w_j), \qquad (g_{\mathrm{B}}(w))_k = w_k \;\; (k \notin \{p,j\}).$$

For an observed point $x = g_{\mathrm{B}}(w)$, the inverse coordinates are

$$w_p = x_p, \qquad w_j = \ell_{\mathrm{sig}}\big(\log(x_j/(1-x_j)) - ax_p\big) \tag{90}$$

and $w_k = x_k$ for $k \notin \{p,j\}$. The Jacobian of g_{B} is triangular and its jth diagonal entry is the positive derivative in equation 86. Together with the explicit inverse equation 90, this proves that g_{B} is a smooth global diffeomorphism of the cube.

Choose, for every targeted root and environment, the same positive C^1 replacement density in both models. Couple the factual models with the same $\mathbf{U}$ and couple corresponding replacement draws. For environment e, write

$$\widetilde{U}_k^{(e)} := \begin{cases} \xi_k^{(e)}, & k \in S_e, \\ U_k, & k \notin S_e, \end{cases} \qquad k \neq j,$$

where the $\xi_k^{(e)}$ are the common replacement draws. Since $\boldsymbol{\sigma}_j = \mathbf{0}$, $j \notin S_e$ for every e, so the noise U_j is always reused.

Under model A, the counterfactual observation is

$$X_{\mathrm{A},k}^{(e)} = \widetilde{U}_k^{(e)} \quad (k \neq j), \qquad X_{\mathrm{A},j}^{(e)} = b_a(\widetilde{U}_p^{(e)}, U_j).$$

Under model B, its empty-DAG counterfactual latent is $W_k^{(e)} = \widetilde{U}_k^{(e)}$ for $k \neq j$ and $W_j^{(e)} = U_j$. Applying the decoder gives

$$X_{\mathrm{B},k}^{(e)} = \widetilde{U}_k^{(e)} \quad (k \neq j), \qquad X_{\mathrm{B},j}^{(e)} = b_a(\widetilde{U}_p^{(e)}, U_j).$$

Thus

$$\mathbf{X}_{\mathrm{A}}^{(e)} = \mathbf{X}_{\mathrm{B}}^{(e)} \qquad \text{almost surely for every labeled environment } e.$$

The same calculation in the factual setting gives $\mathbf{X}_{\mathrm{A}} = \mathbf{X}_{\mathrm{B}}$ almost surely. Hence, under this coupling,

$$(\mathbf{X}_{\mathrm{A}}, \mathbf{X}_{\mathrm{A}}^{(e)}) = (\mathbf{X}_{\mathrm{B}}, \mathbf{X}_{\mathrm{B}}^{(e)}) \qquad \text{almost surely for every } e.$$

This pointwise coupling proves equality of the paired law families, including the within-pair dependence.

Both models satisfy the remaining assumptions. Every target replacement has positive density on the full fixed conditional support $(0, 1)$. To verify the intrinsic paired density explicitly, fix an environment e and collect its replacement draws as $\xi_{S_e}^{(e)}$. In the residual chart of either model, the calculation above gives

$$(q(\mathbf{X}), q(\mathbf{X}^{(e)})) = \iota_e(\mathbf{U}, \xi_{S_e}^{(e)}), \tag{91}$$

because every non-target residual, including U_j, is reused and every targeted node is a root with identity residual coordinate. The source vector $(\mathbf{U}, \xi_{S_e}^{(e)})$ has density

$$(u, w) \longmapsto \prod_{k \in S_e} \nu_k^e(w_k) > 0 \tag{92}$$

with respect to $\mathrm{Leb}^{d+|S_e|}$. Since $\mu_e = (\iota_e)_{\#}\mathrm{Leb}^{d+|S_e|}$, equations 91–92 prove positivity of the paired density with respect to μ_e on the whole support relation.

Model A also satisfies TRF. Its ancestor-closed active sets are obtained by choosing any subset of the isolated roots, together with one of the three possibilities $\varnothing$, $\{p\}$, or $\{p, j\}$ for the nontrivial two-node component. Every active set with a non-sink contains $\{p, j\}$; its only non-sink is p, whose terminal descendant is j, and

$$\partial_{u_p} F_j(u) = a\, b_a(u_p, u_j)(1 - b_a(u_p, u_j)) \neq 0.$$

After j is deleted, p is a sink and TRF becomes vacuous. Model B also has vacuous TRF because its DAG is empty. The two DAGs differ by the edge $p \to j$, placing the models in different causal equivalence classes.

Duplicate signature. Let nodes p and j have the same signature. Take an empty DAG with independent standard Gaussian roots

$$Z_i = \Phi^{-1}(U_i),$$

where Φ is the standard Gaussian CDF. In every environment targeting the duplicate block, replace both coordinates independently from $\mathsf{N}(0, \tau_e^2)$ for some $\tau_e > 0$; an environment targets either both coordinates or neither because their signatures agree. Other root coordinates and their positive replacement densities can be chosen identically in the two models.

Let $\mathbf{O}$ be a two-dimensional orthogonal matrix with every entry nonzero. Model A uses the identity decoder. Model B has independent standard Gaussian roots $\mathbf{W}$ and uses the decoder

$$(g_{\mathrm{B}}(w))_{\{p,j\}} = \mathbf{O}^\top w_{\{p,j\}}, \qquad (g_{\mathrm{B}}(w))_k = w_k \quad (k \notin \{p, j\}).$$

The inverse applies $\mathbf{O}$ to the displayed block, so this decoder is a global smooth diffeomorphism.

Let $\mathbf{G} \sim \mathsf{N}(0, \mathbf{I}_2)$ be the factual block in model A and $\mathbf{W} \sim \mathsf{N}(0, \mathbf{I}_2)$ the factual block in model B. The factual laws already agree, including when $m = 0$. On the duplicate block,

$$\mathbf{X}_{\mathrm{A},\{p,j\}} = \mathbf{G} \sim \mathsf{N}(0, \mathbf{I}_2), \qquad \mathbf{X}_{\mathrm{B},\{p,j\}} = \mathbf{O}^\top \mathbf{W} \sim \mathsf{N}(0, \mathbf{I}_2),$$

and all coordinates outside the block are chosen identically in the two models. Hence $P_0^{\mathrm{A}} = P_0^{\mathrm{B}}$.

The paired block laws coincide environment by environment. When environment e leaves the duplicate block untargeted, both models reuse the block:

$$(\mathbf{X}_{\mathrm{A},\{p,j\}}, \mathbf{X}^{(e)}_{\mathrm{A},\{p,j\}}) = (\mathbf{G}, \mathbf{G}), \qquad (\mathbf{X}_{\mathrm{B},\{p,j\}}, \mathbf{X}^{(e)}_{\mathrm{B},\{p,j\}}) = (\mathbf{O}^\top \mathbf{W}, \mathbf{O}^\top \mathbf{W}). \tag{93}$$

Orthogonal invariance gives $\mathbf{O}^\top \mathbf{W} \sim \mathsf{N}(0, \mathbf{I}_2)$, so the two diagonal pair laws in equation 93 are equal.

If environment e targets the block, let $\mathbf{E}^{\mathrm{A}}_e \sim \mathsf{N}(0, \tau_e^2 \mathbf{I}_2)$ and $\mathbf{E}^{\mathrm{B}}_e \sim \mathsf{N}(0, \tau_e^2 \mathbf{I}_2)$ be independent of the respective factual blocks. Then

$$(\mathbf{X}_{\mathrm{A},\{p,j\}}, \mathbf{X}^{(e)}_{\mathrm{A},\{p,j\}}) = (\mathbf{G}, \mathbf{E}^{\mathrm{A}}_e), \qquad (\mathbf{X}_{\mathrm{B},\{p,j\}}, \mathbf{X}^{(e)}_{\mathrm{B},\{p,j\}}) = (\mathbf{O}^\top \mathbf{W}, \mathbf{O}^\top \mathbf{E}^{\mathrm{B}}_e). \tag{94}$$

The first pair in equation 94 is jointly Gaussian with mean zero and block covariance

$$\begin{pmatrix} \mathbf{I}_2 & 0 \\ 0 & \tau_e^2 \mathbf{I}_2 \end{pmatrix}.$$

The second pair has the same covariance because $\mathbf{O}^\top \mathbf{I}_2 \mathbf{O} = \mathbf{I}_2$ and $\mathbf{O}^\top (\tau_e^2 \mathbf{I}_2) \mathbf{O} = \tau_e^2 \mathbf{I}_2$; independence makes its cross-covariance zero. Equality of means and covariances therefore gives equality of the two Gaussian pair laws. All coordinates outside the block are chosen identically and independently in the two models, including their reused/replaced coupling. Thus every complete labeled pair law is the same.

Each scalar root map $u \mapsto \Phi^{-1}(u)$ is a smooth increasing diffeomorphism from $(0, 1)$ onto the fixed interval $\mathbb{R}$. Gaussian replacement densities are positive on that full interval. In non-target coordinates the residual pair lies on the equality constraint. In a target coordinate, a replacement $E \sim \mathsf{N}(0, \tau_e^2)$ becomes the residual value $V = \Phi(E)$, whose density on $(0, 1)$ is

$$\zeta_e(v) = \frac{\phi_{\tau_e}(\Phi^{-1}(v))}{\phi(\Phi^{-1}(v))} > 0,$$

where ϕ is the standard Gaussian density and ϕ_{τ_e} is the $\mathsf{N}(0, \tau_e^2)$ density. Thus, under the support parametrization ι_e, the full source density is the product of unit factual densities and positive factors of the form ζ_e (and the corresponding positive factors for nodes outside the block). Its pushforward has positive density with respect to μ_e on the paired support relation. Both DAGs are empty, so structural activity and terminal-response faithfulness are vacuous.

The representations belong to different causal equivalence classes. For the same observation x, the latent blocks satisfy

$$\mathbf{W}_{\{p,j\}} = \mathbf{O}\mathbf{Z}_{\{p,j\}}.$$

Suppose a permutation π and scalar monotone maps m_p, m_j represented this relation componentwise. For either output coordinate r, the value $m_r(Z_{\pi(r)})$ remains fixed when the other input coordinate varies. Every row of $\mathbf{O}$ has two nonzero entries, so $(\mathbf{O}\mathbf{Z})_r$ changes under that same variation. This contradiction rules out every componentwise monotone representation and places the models in different causal equivalence classes. □

Proof of Corollary 4.4. The argument combines a design upper bound, a minimax lower bound, and the scalar boundary case.

Sufficiency for $d \geq 2$. With m labeled environments, a node signature is a vector in $\{0, 1\}^m$. Exactly one of the 2^m vectors is zero, so there are $2^m - 1$ distinct nonzero signatures. If

$$d \leq 2^m - 1, \tag{95}$$

choose any injection from the d labeled nodes into these nonzero vectors and define the target sets row by row from the resulting signature matrix. For every SCM and replacement family satisfying the other model assumptions, this design has pairwise distinct nonzero target signatures. The paired laws recover these signatures through Theorem 4.1, and Theorem 4.3 then recovers the causal equivalence class. Thus any m satisfying equation 95 is sufficient.

Necessity for $d \geq 2$. Fix an arbitrary labeled design with m environments. Suppose

$$d > 2^m - 1. \tag{96}$$

If all d columns were nonzero and pairwise distinct, they would be d distinct elements of a set of cardinality $2^m - 1$, contradicting equation 96. Therefore every such design has at least one of the following: **1) a zero column**, **2) or two equal columns**.

For either alternative, Proposition I.1 constructs two admissible models with the same labeled paired laws but outside the causal equivalence class. Hence every design satisfying equation 96 fails uniform identification over the admissible models. This proves the minimax lower bound.

Combining the two directions, the smallest integer m satisfying equation 95 obeys

$$\begin{aligned} 2^m &\geq d+1, \\ m &\geq \log_2(d+1), \\ m^\star(d) &= \lceil \log_2(d+1) \rceil . \end{aligned}$$

The case $d = 1$. The one-node DAG is uniquely empty. Consider any two scalar admissible model explanations of the same observational variable X:

$$X = g_1(Z^{(1)}) = g_2(Z^{(2)}).$$

Since both decoders are scalar diffeomorphisms,

$$Z^{(2)} = g_2^{-1} \circ g_1(Z^{(1)}). \tag{97}$$

A scalar diffeomorphism on a connected interval is strictly increasing or strictly decreasing, so equation 97 is exactly the allowed componentwise monotone ambiguity. For the residual charts, let $T := q_2 \circ q_1^{-1}$. Both charts map the common observational law to $\mathrm{Unif}(0,1)$, so $T_{\#}\,\mathrm{Unif}(0,1) = \mathrm{Unif}(0,1)$. Lemma C.5 gives $T(u) = u$ in the increasing case and $T(u) = 1 - u$ in the decreasing case. This is exactly the allowed residual identity/reflection ambiguity. The observational law alone therefore identifies the equivalence class, and $m^\star(1) = 0$. □

J Mechanism regimes for terminal response

Proof of Proposition 4.6. The two parts follow from the sign-coherent recursion and the analytic zero-set argument.

Sign-coherent mechanisms. Because the domain of each structural mechanism is connected, a continuous nowhere-zero edge derivative $\partial_{z_k} f_i$ has constant sign. The path expansion follows by repeated substitution. Fix a source node j and write $D_i(u) := \partial_{u_j} F_i(u)$. Differentiating $F_i(u) = f_i(F_{\mathrm{pa}(i)}(u), u_i)$ gives the recursion

$$D_i(u) = \mathbf{1}\{i = j\}\, \partial_{u_i} f_i(F_{\mathrm{pa}(i)}(u), u_i) + \sum_{k \in \mathrm{pa}(i)} \partial_{z_k} f_i(F_{\mathrm{pa}(i)}(u), u_i) D_k(u). \tag{98}$$

If $i = j$, no parent of j depends on the later residual u_j, so $D_j(u) = \partial_{u_j} f_j(F_{\mathrm{pa}(j)}(u), u_j) > 0$.

If i precedes j or is not a descendant of j, the recursion and topological induction give $D_i = 0$. For a descendant, substitute equation 98 recursively into each nonzero parent term. Every substitution appends one directed edge, and acyclicity terminates the expansion after finitely many steps at D_j. Conversely, every directed path from j to the descendant occurs exactly once. Therefore, for every strict ancestor–descendant pair $j \in \mathrm{an}(\ell)$,

$$\partial_{u_j} F_\ell(u) = \partial_{u_j} F_j(u) \sum_{p : j \rightsquigarrow \ell} \prod_{(k \to i) \in p} \partial_{z_k} f_i(F_{\mathrm{pa}(i)}(u), u_i). \tag{99}$$

The leading derivative is positive by scalar noise solvability. Every path product is nonzero, and the sign-coherence condition makes all summands have the same sign. Hence equation 99 is nowhere zero. Now fix an ancestor-closed active set R and a non-sink j of $\mathcal{G}[R]$. Following directed edges inside the finite active DAG reaches a sink descendant $\ell \in R$. Every ancestor of ℓ lies in R, so restriction of the full response factors through $\mathrm{pr}_R : (0,1)^d \to (0,1)^R$ and gives

$$\partial_{u_j} F_\ell = \left(\partial_{u_j} F_\ell^R\right) \circ \mathrm{pr}_R . \tag{100}$$

The left side is nowhere zero by equation 99, and pr_R is onto. Therefore $\partial_{u_j} F_\ell^R$ is nowhere zero on $(0,1)^R$. This is TRF.

Analytic fixed-support family. For fixed parent values, $\mathrm{logit}(u_i)$ ranges over $\mathbb{R}$. Thus equation 17 is an increasing analytic diffeomorphism from $(0,1)$ onto the fixed interval I_i, with

$$\partial_{u_i} f_i = \frac{\tau_i'(\cdot)}{u_i(1-u_i)} > 0.$$

Moreover, for an edge $k \to i$,

$$\partial_{z_k} f_i = \tau_i'(\cdot)\, \theta_{ik} \phi_{ik}'(z_k),$$

so the displayed edge is active whenever $\theta_{ik} \neq 0$.

Fix $u^\circ \in (0,1)^d$. The maps logit, τ_i, and ϕ_{ik} are real analytic on their open domains. For a root, the formula in equation 17 is therefore jointly analytic in (u, θ). Inductively, if the solution coordinates of every parent of i are jointly analytic, then the finite weighted sum inside τ_i and its composition with τ_i are jointly analytic as well. A topological induction proves that every $F_i(u;\theta)$ is jointly real analytic. Partial differentiation preserves real analyticity, so for a fixed ancestor–descendant pair the response

$$G_{j\ell}(\theta) := \partial_{u_j} F_\ell(u^\circ; \theta)$$

is real analytic on $\mathbb{R}^{|E|}$. Choose one directed path $p : j = v_0 \to \cdots \to v_s = \ell$, set its edge weights to one, and set all off-path weights to zero. At this parameter vector θ^p, the chain rule reduces the response to

$$G_{j\ell}(\theta^p) = \partial_{u_j} F_j(u^\circ; \theta^p) \prod_{r=1}^{s} \partial_{z_{v_{r-1}}} f_{v_r}(F_{\mathrm{pa}(v_r)}(u^\circ; \theta^p), u^\circ_{v_r}; \theta^p) > 0.$$

Hence $G_{j\ell}$ is a nonzero real-analytic function. The zero set of a nonzero real-analytic function has Lebesgue measure zero. A finite union over all ancestor–descendant pairs, together with the coordinate hyperplanes $\{\theta_{ik} = 0\}$, is still null. Absolute continuity of the edge-weight law therefore proves the simultaneous response and edge-activity statements.

Finally, let R be ancestor closed and let j be a non-sink of $\mathcal{G}[R]$. Following a directed path inside R reaches a sink descendant ℓ. All ancestors of ℓ lie in R. Define $u^\circ_R := (u^\circ)_R$. Restricting the solution map therefore gives the explicit equality

$$\partial_{u_j} F^R_\ell(u^\circ_R) = \partial_{u_j} F_\ell(u^\circ) \neq 0.$$

In particular, the restricted response is a nonzero function, which is TRF. □

J.1 Genericity in linear SCMs

Proposition J.1 (Genericity of terminal response in linear SCMs)**.** *Fix a DAG and draw its nonzero edge weights from any joint distribution with a density. Then, with probability one, every ancestor has a nonzero total effect on every descendant. Hence Assumption 2.1 holds almost surely.*

Proof of Proposition J.1. Let $\mathbf{B}$ be the weighted adjacency matrix in a topological ordering. Then $\mathbf{B}$ is strictly triangular and therefore nilpotent: $\mathbf{B}^d = 0$. Write the scalar noise maps as $\varepsilon_i : (0,1) \to \mathbb{R}$ and put $\boldsymbol{\varepsilon}(\mathbf{U}) = (\varepsilon_1(U_1), \ldots, \varepsilon_d(U_d))$. The linear SCM and its solution are

$$\mathbf{Z} = \mathbf{B}\mathbf{Z} + \boldsymbol{\varepsilon}(\mathbf{U}), \qquad \mathbf{Z} = (\mathbf{I}_d - \mathbf{B})^{-1} \boldsymbol{\varepsilon}(\mathbf{U}). \tag{101}$$

Nilpotence gives the finite Neumann expansion

$$(\mathbf{I}_d - \mathbf{B})^{-1} = \mathbf{I}_d + \mathbf{B} + \mathbf{B}^2 + \cdots + \mathbf{B}^{d-1}. \tag{102}$$

Fix a strict ancestor–descendant pair (j, i). For every $k \geq 1$, matrix multiplication gives

$$(\mathbf{B}^k)_{ij} = \sum_{j = v_0 \to v_1 \to \cdots \to v_k = i} \prod_{r=1}^{k} \mathbf{B}_{v_r v_{r-1}}, \tag{103}$$

where the sum is over length-k directed paths from j to i. Combining equation 102 and equation 103 yields the total-effect polynomial

$$[(\mathbf{I}_d - \mathbf{B})^{-1}]_{ij} = \sum_{p \in \mathrm{Path}(j \rightsquigarrow i)} m_p(\mathbf{B}), \qquad m_p(\mathbf{B}) := \prod_{(a \to b) \in p} \mathbf{B}_{ba}. \tag{104}$$

At least one path exists because j is an ancestor of i. Distinct paths in a DAG use distinct edge sets and hence contribute distinct monomials. Therefore the right side of equation 104 is a nonzero polynomial.

The zero set of a nonzero real polynomial has Lebesgue measure zero. Since the joint edge-weight law has a density, $\mathbb{P}\left([(\mathbf{I}_d - \mathbf{B})^{-1}]_{ij} = 0\right) = 0$ for this fixed pair. There are finitely many ancestor–descendant pairs, so union of their cancellation events also has probability 0. Hence, almost surely,

$$[(\mathbf{I}_d - \mathbf{B})^{-1}]_{ij} \neq 0 \quad \text{for every strict ancestor–descendant pair } (j, i). \tag{105}$$

The total effect determines the residual derivative required by TRF. Differentiating equation 101 with respect to u_j gives

$$\partial_{u_j} F_i(u) = [(\mathbf{I}_d - \mathbf{B})^{-1}]_{ij} \varepsilon_j'(u_j). \tag{106}$$

Scalar noise solvability gives $\varepsilon_j'(u_j) > 0$, while equation 105 makes the matrix factor nonzero. Therefore $\partial_{u_j} F_i \not\equiv 0$ for every strict ancestor–descendant pair.

Finally fix an ancestor-closed active set R and a non-sink j of $\mathcal{G}[R]$. Following any directed edge from j inside the finite active DAG must eventually reach a sink descendant $\ell \in R$. Every vertex on a directed path from j to ℓ is an ancestor of ℓ, and ancestor closure places every such vertex in R. Thus the total-effect polynomial from j to ℓ in the restricted SCM is the same as in the full SCM. Equation 106 gives

$$\partial_{u_j} F_\ell \not\equiv 0 \quad \text{on } (0,1)^R,$$

which is exactly Assumption 2.1. □

K Finite-sample bounds from observed-law distances

The population identification result admits two complementary finite-sample realizations. Observed-law minimum distance converts distributional accuracy into target accuracy on compact model classes, while a constructive probit–linear analysis certifies the residual chart, adaptive peeling, and direct-edge decisions. A weak-response lower bound identifies the necessary signal scale, and selective robustness extends the guarantee to perturbed data laws. Appendix O specializes the atomic estimator used in the numerical study.

Let $\mathfrak{M}$ denote the admissible paired-intervention models in Section 2. For $\mathfrak{m} \in \mathfrak{M}$, write

$$P_c^{\mathfrak{m}} := \begin{cases} \mathcal{L}_{\mathfrak{m}}(X), & c = 0, \\ \mathcal{L}_{\mathfrak{m}}(X, X^{(c)}), & 1 \le c \le m, \end{cases} \qquad \mathcal{T}(\mathfrak{m}) := (P_c^{\mathfrak{m}})_{c=0}^{m}.$$

Let $\vartheta(\mathfrak{m})$ denote the causal equivalence class and DAG under the matched node labels identified in Theorem 4.3, and set $\mathcal{Y} := \vartheta(\mathfrak{M})$.

Assumption K.1 (Population identifiability)**.** For every $\mathfrak{m}, \mathfrak{m}' \in \mathfrak{M}$,

$$\mathcal{T}(\mathfrak{m}) = \mathcal{T}(\mathfrak{m}') \quad \Longrightarrow \quad \vartheta(\mathfrak{m}) = \vartheta(\mathfrak{m}').$$

Observed-data formulation. Two sampling regimes support the results below. The constructive estimator uses mutually independent samples for chart fitting, ratio estimation, score calibration, and graph recovery. The robust distributional estimator requires i.i.d. rows within each sample group and permits dependence between groups. Both estimators observe factual rows, same-unit pairs, and environment labels; they recover the residual chart, intervention targets, and graph.

K.1 Compactification and inverse modulus

For each channel c, let k_c be a measurable characteristic kernel on the channel sample space, with reproducing-kernel Hilbert space (RKHS) $\mathcal{H}_c$, feature map φ_c, and

$$\sup_x k_c(x, x) \le K_c < \infty. \tag{107}$$

This is the standard maximum mean discrepancy (MMD) construction Gretton et al. (2012), applied channelwise. For a law P, its RKHS mean embedding is $\mu_c^{\varphi}(P) = \mathbb{E}_P \varphi_c(X)$. Define the observed-law metric

$$d_{\mathcal{T}}(\mathfrak{m}, \mathfrak{m}') = \max_{0 \le c \le m} \left\| \mu_c^{\varphi}(P_c^{\mathfrak{m}}) - \mu_c^{\varphi}(P_c^{\mathfrak{m}'}) \right\|_{\mathcal{H}_c}.$$

Assumption K.2 (Compact finite-sample class)**.** The restricted class $\mathfrak{M}_B \subset \mathfrak{M}$ is a compact metric space. For every c, $\mathfrak{m} \mapsto \mu_c^{\varphi}(P_c^{\mathfrak{m}})$ is continuous. The target image $\vartheta(\mathfrak{M}_B)$ is metrized by d_ϑ, and $\mathfrak{m} \mapsto \vartheta(\mathfrak{m})$ is continuous. Assumption K.1 holds on $\mathfrak{M}_B$.

Sufficient conditions for Assumption K.2 fix the graph and signature alphabets, extend primitive maps to compact closures, bound them uniformly in C^{s+1}, impose uniform lower singular-value and density bounds, and take a closed class. A C^{s+1} bound gives precompactness in C^s by Arzela–Ascoli; lower singular-value and density bounds keep the closure inside the admissible class. If multiple graph strata are included, a positive edge margin is needed to make those strata closed.

Definition K.3 (Inverse modulus)**.** For $t \geq 0$, define

$$\omega_B(t) = \sup\left\{d_\vartheta(\vartheta(\mathfrak{m}), \vartheta(\mathfrak{m}')) : \mathfrak{m}, \mathfrak{m}' \in \mathfrak{M}_B,\ d_{\mathcal{T}}(\mathfrak{m}, \mathfrak{m}') \leq t\right\}.$$

Lemma K.4 (Compact identification gives a uniform inverse modulus)**.** *Under Assumptions K.1 and K.2, $\omega_B(t) \to 0$ as $t \downarrow 0$.*

Proof. Assume for contradiction that the conclusion fails. Then there are $\epsilon > 0$, sequences $t_n \downarrow 0$, and $\mathfrak{m}_n, \mathfrak{m}_n' \in \mathfrak{M}_B$ such that $d_{\mathcal{T}}(\mathfrak{m}_n, \mathfrak{m}_n') \leq t_n$ but $d_\vartheta(\vartheta(\mathfrak{m}_n), \vartheta(\mathfrak{m}_n')) \geq \epsilon$. After passing to a subsequence, compactness gives $\mathfrak{m}_n \to \mathfrak{m}$ and $\mathfrak{m}_n' \to \mathfrak{m}'$. Continuity of the channel embeddings gives $d_{\mathcal{T}}(\mathfrak{m}, \mathfrak{m}') = 0$. Since every kernel is characteristic, $P_c^{\mathfrak{m}} = P_c^{\mathfrak{m}'}$ for all c, hence $\mathcal{T}(\mathfrak{m}) = \mathcal{T}(\mathfrak{m}')$. Assumption K.1 gives $\vartheta(\mathfrak{m}) = \vartheta(\mathfrak{m}')$. Continuity of ϑ then makes $d_\vartheta(\vartheta(\mathfrak{m}_n), \vartheta(\mathfrak{m}_n')) \to 0$, a contradiction. □

Lemma K.4 converts finite empirical observed-law error into finite target error. A polynomial rate follows under a quantitative condition such as

$$\omega_B(t) \leq L_B t^{\alpha_B}, \qquad 0 < \alpha_B \leq 1. \tag{108}$$

The Hölder condition is an additional regularity assumption.

K.2 Empirical embedding concentration

Lemma K.5 (Simultaneous empirical mean-embedding bound)**.** *Let $X_{c,1}, \ldots, X_{c,n_c}$ be i.i.d. from P_c, independently across channels, and let $\widehat{\mu^{\varphi}}_c = n_c^{-1} \sum_i \varphi_c(X_{c,i})$. For $\delta \in (0,1)$, set*

$$a_c(\delta) = \sqrt{\frac{K_c}{n_c}} + \sqrt{\frac{2K_c \log((m+1)/\delta)}{n_c}}, \qquad a(\delta) = \max_c a_c(\delta). \tag{109}$$

Then, with probability at least $1-\delta$,

$$\left\|\widehat{\mu^{\varphi}}_c - \mu_c^{\varphi}(P_c)\right\|_{\mathcal{H}_c} \leq a_c(\delta) \quad \textit{simultaneously for every } c. \tag{110}$$

Proof. Write $\boldsymbol{\xi}_i = \varphi_c(X_{c,i}) - \mu_c^{\varphi}(P_c)$. Independence and centering give the exact second-moment calculation

$$\mathbb{E}\left\|\widehat{\mu^{\varphi}}_c - \mu_c^{\varphi}(P_c)\right\|_{\mathcal{H}_c}^2 = \frac{1}{n_c^2}\mathbb{E}\left\|\sum_{i=1}^{n_c} \boldsymbol{\xi}_i\right\|_{\mathcal{H}_c}^2 = \frac{1}{n_c^2}\sum_{i=1}^{n_c}\mathbb{E}\|\boldsymbol{\xi}_i\|_{\mathcal{H}_c}^2 \leq \frac{K_c}{n_c}.$$

The last inequality uses $\mathbb{E}\|\varphi_c(X)\|^2 = \mathbb{E}k_c(X,X) \leq K_c$ and discards the nonpositive centering correction. Jensen therefore gives $\mathbb{E}\|\widehat{\mu^{\varphi}}_c - \mu_c^{\varphi}(P_c)\|_{\mathcal{H}_c} \leq \sqrt{K_c/n_c}$.

If one observation is replaced by $X_{c,i}'$, the norm of the change in the mean embedding is at most

$$\frac{1}{n_c}\|\varphi_c(X_{c,i}) - \varphi_c(X_{c,i}')\|_{\mathcal{H}_c} \leq \frac{2\sqrt{K_c}}{n_c}.$$

McDiarmid's inequality applied to the scalar function $f = \|\widehat{\mu^{\varphi}}_c - \mu_c^{\varphi}(P_c)\|$ thus yields, for every $t > 0$,

$$\mathbb{P}\left(f - \mathbb{E}f \geq t\right) \leq \exp\left(-\frac{n_c t^2}{2K_c}\right).$$

Set $t = \sqrt{2K_c \log((m+1)/\delta)/n_c}$. The failure probability for one channel is $\delta/(m+1)$, and the union bound over $c = 0, \ldots, m$ proves equation 110. □

K.3 Minimum-distance estimator and target error

Define

$$\mathcal{J}_n(\mathfrak{m}) = \max_{0\le c\le m} \left\|\widehat{\mu^{\varphi}}_c - \mu_c^{\varphi}(P_c^{\mathfrak{m}})\right\|_{\mathcal{H}_c}.$$

Compactness and continuity give an exact minimizer. An η_{opt}-accurate numerical solution is represented by any measurable η_{opt}-minimizer:

$$\mathcal{J}_n(\widehat{\mathfrak{m}}) \le \inf_{\mathfrak{m}\in\mathfrak{M}_B} \mathcal{J}_n(\mathfrak{m}) + \eta_{\mathrm{opt}}.$$

Theorem K.6 (Compact-class inverse-modulus bound)**.** *Assume the data are generated by $\mathfrak{m}_\star \in \mathfrak{M}_B$ and Assumptions K.1–K.2 hold. Then, with probability at least $1-\delta$,*

$$d_\vartheta(\vartheta(\widehat{\mathfrak{m}}), \vartheta(\mathfrak{m}_\star)) \le \omega_B(2a(\delta) + \eta_{\mathrm{opt}}). \tag{111}$$

If the quantitative inverse bound equation 108 holds, the right side is at most $L_B(2a(\delta)+\eta_{\mathrm{opt}})^{\alpha_B}$.

Proof. Work on the simultaneous event in Lemma K.5. The true model has $\mathcal{J}_n(\mathfrak{m}_\star) \le a(\delta)$, so $\mathcal{J}_n(\widehat{\mathfrak{m}}) \le a(\delta) + \eta_{\mathrm{opt}}$. For every channel,

$$\begin{aligned}\left\|\mu_c^{\varphi}(P_c^{\widehat{\mathfrak{m}}}) - \mu_c^{\varphi}(P_c^{\mathfrak{m}_\star})\right\|_{\mathcal{H}_c} &\le \left\|\mu_c^{\varphi}(P_c^{\widehat{\mathfrak{m}}}) - \widehat{\mu^{\varphi}}_c\right\|_{\mathcal{H}_c} + \left\|\widehat{\mu^{\varphi}}_c - \mu_c^{\varphi}(P_c^{\mathfrak{m}_\star})\right\|_{\mathcal{H}_c} \\ &\le 2a(\delta) + \eta_{\mathrm{opt}}.\end{aligned}$$

Thus $d_{\mathcal{T}}(\widehat{\mathfrak{m}}, \mathfrak{m}_\star) \le 2a(\delta) + \eta_{\mathrm{opt}}$. The definition of ω_B proves equation 111. □

Corollary K.7 (Exact recovery of a separated discrete target)**.** *Let $\vartheta_{\mathrm{disc}}(\mathfrak{m})$ be a finite-valued target such as the matched labeled DAG, and define*

$$\Delta_{\mathrm{disc}} = \inf\{d_{\mathcal{T}}(\mathfrak{m}, \mathfrak{m}') : \mathfrak{m}, \mathfrak{m}' \in \mathfrak{M}_B,\ \vartheta_{\mathrm{disc}}(\mathfrak{m}) \ne \vartheta_{\mathrm{disc}}(\mathfrak{m}')\}.$$

If $\Delta_{\mathrm{disc}} > 0$ and $2a(\delta) + \eta_{\mathrm{opt}} < \Delta_{\mathrm{disc}}$, then $\vartheta_{\mathrm{disc}}(\widehat{\mathfrak{m}}) = \vartheta_{\mathrm{disc}}(\mathfrak{m}_\star)$ with probability at least $1-\delta$.

Proof. The proof of Theorem K.6 gives $d_{\mathcal{T}}(\widehat{\mathfrak{m}}, \mathfrak{m}_\star) \le 2a(\delta) + \eta_{\mathrm{opt}}$. A different discrete target would contradict the definition of Δ_{disc}. □

Remark K.8 (Two complementary upper bounds). Theorem K.6 gives a compact-class guarantee for observed-law minimum distance. The general estimator below provides an explicit probit–linear construction. Candidate embeddings for minimum distance may be computed analytically or by an independent simulator; a uniform simulator error adds to η_{opt} and the observed-law radius.

L Finite-sample recovery with a probit–linear chart

The population argument identifies causal variables and their graph from exact intervention laws. Here we ask when finitely many paired observations support the same decisions. Theorem L.34 gives the result: under its stated conditions, the estimator returns the correct DAG or abstains with probability at least $1-\delta$. Recovery requires the uncertainty in each sink and edge decision to be smaller than the corresponding signal margin.

Two coordinate systems organize the construction. The residual chart $q : X \mapsto U$ recovers the independent factual noise inputs. Causal dependence lies in the structural map F, which turns these inputs into endogenous variables. Peeling recovers those variables up to componentwise monotone transformations, forming a coordinate map Y. We first estimate q and the intervention targets from paired differences. We then estimate environment density ratios and normalize them into conditional CDFs, whose comparisons identify and remove sinks. Finally, the inverse of Y reveals the direct edges. Sections L.6 and L.7 assemble the intermediate bounds into the peeling and graph guarantees.

Notation. The active set $R \subseteq [d]$ contains the nodes still to be recovered; $j \in R$ denotes a coordinate, and e, f denote intervention environments. The signature entry σ_{ej} equals one when environment e targets node j, and zero otherwise. The polynomial degree p and trimming level η are fixed throughout the construction. We display them when defining an estimator or comparing approximation errors, and otherwise suppress them in the CDFs and scores. A hat denotes a fitted or empirical quantity, while a check denotes its numerical evaluation. A superscript $\sharp$ distinguishes an exactly represented input from its ideal real-valued counterpart when that distinction enters a bound. Within individual proofs, fixed indices are suppressed after their scope is stated.

L.1 RESIDUAL-CHART RECOVERY

Probit–linear observation model. The probit–linear model makes the residual chart estimable from second moments while leaving the endogenous structural map nonlinear. The probit transform Φ^{-1} turns uniform noise coordinates into Gaussian ones. Whitening the observations removes their covariance, leaving an unknown orthogonal transformation for the paired interventions to identify.

Let Φ be the standard normal CDF, applied coordinatewise, and set $\boldsymbol{\zeta}^0 = \Phi^{-1}(\mathbf{U}) \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_d)$. The probit-linear chart model assumes

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{A}_{\mathrm{mix}}\boldsymbol{\zeta}^0, \qquad \mathbf{A}_{\mathrm{mix}} \in \mathbb{R}^{d\times d} \text{ invertible},$$

so that

$$q(\mathbf{x}) = \Phi\big(\mathbf{A}_{\mathrm{mix}}^{-1}(\mathbf{x} - \boldsymbol{\mu})\big)\,. \tag{112}$$

The endogenous structural map F may remain nonlinear. Equivalently, choose the decoder $g(\mathbf{z}) = \boldsymbol{\mu} + \mathbf{A}_{\mathrm{mix}}\Phi^{-1}(F^{-1}(\mathbf{z}))$, which remains nonlinear whenever F is.

A target is a node whose mechanism is replaced in a particular environment. A sink has no children among the remaining nodes. The target signatures tell us which environments to compare when searching for sinks; the sink order is recovered later.

For a paired observation in environment e, write $\boldsymbol{\zeta}^{0,e} = \Phi^{-1}(q(\mathbf{X}^{(e)}))$ and $\boldsymbol{\Delta}_e^{\zeta} = \boldsymbol{\zeta}^{0,e} - \boldsymbol{\zeta}^0$. Exact pairing gives $(\boldsymbol{\Delta}_e^{\zeta})_j = 0$ for $j \notin S_e$. Define

$$\boldsymbol{\Omega}_e = \mathbb{E}[\boldsymbol{\Delta}_e^{\zeta}(\boldsymbol{\Delta}_e^{\zeta})^\top].$$

Assumption L.1 (Target-subspace excitation)**.** For every e, the principal block $(\boldsymbol{\Omega}_e)_{S_eS_e}$ is positive definite. Equivalently,

$$\mathrm{range}(\boldsymbol{\Omega}_e) = \mathrm{span}\{\mathbf{e}_j : j \in S_e\}.$$

For finite-sample bounds, assume its smallest nonzero eigenvalue is at least $\gamma > 0$ and $\big\|\boldsymbol{\Delta}_e^{\zeta}\big\|_{\psi_2} \le K_\Delta$.

Assumption L.1 ensures that paired differences explore every targeted direction. Their covariance can therefore reveal the whole target subspace.

Population whitening and binary projector algebra. After whitening, paired differences span the directions changed by each intervention. The projector onto this span keeps those directions and removes the others. All environments share the same unknown orthogonal transformation, so their projectors share an eigenbasis. A weighted sum encodes each direction's target signature as its eigenvalue, recovering coordinates and signatures together.

Let $\boldsymbol{\Sigma}_X = \mathrm{Cov}(\mathbf{X}) = \mathbf{A}_{\mathrm{mix}}\mathbf{A}_{\mathrm{mix}}^\top$ and

$$\mathbf{O} = \boldsymbol{\Sigma}_X^{-1/2}\mathbf{A}_{\mathrm{mix}}. \tag{113}$$

Then $\mathbf{O}$ is orthogonal and the whitened observation is $\widetilde{\mathbf{X}} = \boldsymbol{\Sigma}_X^{-1/2}(\mathbf{X} - \boldsymbol{\mu}) = \mathbf{O}\boldsymbol{\zeta}^0$. Let $\mathbf{D}_e = \mathrm{diag}(\sigma_{e1}, \ldots, \sigma_{ed})$.

Proposition L.2 (Population residual chart and target signatures)**.** *Under Assumption L.1, the range projector of*

$$\mathbf{B}_e = \mathbb{E}[(\widetilde{\mathbf{X}}^{(e)} - \widetilde{\mathbf{X}})(\widetilde{\mathbf{X}}^{(e)} - \widetilde{\mathbf{X}})^\top]$$

is

$$\mathbf{P}_e = \mathbf{O}\mathbf{D}_e\mathbf{O}^\top. \tag{114}$$

For weights $w_e = 2^{e-1}$, define

$$\mathbf{C}_{\mathrm{sig}} = \sum_{e=1}^{m} w_e\mathbf{P}_e. \tag{115}$$

Then

$$\mathbf{C}_{\mathrm{sig}} = \mathbf{O}\,\mathrm{diag}(c_1, \ldots, c_d)\mathbf{O}^\top, \qquad c_i = \sum_{e=1}^{m} 2^{e-1}\sigma_{ei}.$$

If signatures are nonzero and distinct, the eigenvalues c_i are distinct positive integers. Hence $\mathbf{O}$ is identified up to a signed permutation, the binary signatures are recovered by reading the bits of c_i, and equation 112 is recovered up to coordinate permutation and reflections.

Proof. Since $\widetilde{\mathbf{X}}^{(e)} - \widetilde{\mathbf{X}} = \mathbf{O}\boldsymbol{\Delta}_e^{\zeta}$, $\mathbf{B}_e = \mathbf{O}\boldsymbol{\Omega}_e\mathbf{O}^\top$. Assumption L.1 says that the range projector of $\boldsymbol{\Omega}_e$ is $\mathbf{D}_e$, proving equation 114. Summing the projectors yields

$$\mathbf{C}_{\text{sig}} = \mathbf{O}\left(\sum_e 2^{e-1}\mathbf{D}_e\right)\mathbf{O}^\top = \mathbf{O}\,\mathrm{diag}(c_1,\dots,c_d)\mathbf{O}^\top.$$

Distinct signatures have distinct binary encodings. Each eigenspace is therefore one dimensional, so its normalized eigenvector is the corresponding column of $\mathbf{O}$ up to sign and ordering. Finally, $\mathbf{A}_{\text{mix}} = \boldsymbol{\Sigma}_X^{1/2}\mathbf{O}$ and $\mathbf{A}_{\text{mix}}^{-1} = \mathbf{O}^\top\boldsymbol{\Sigma}_X^{-1/2}$, which gives the chart. A sign flip before Φ maps u_i to $1-u_i$. □

L.1.1 Chart perturbation bounds

The sample estimator follows the same steps: estimate the whitened difference moments, extract their range projectors, and diagonalize their weighted sum. The next proposition bounds the error at each step.

Let $\widehat{\boldsymbol{\mu}}$, $\widehat{\boldsymbol{\Sigma}}_X$, and $\widehat{\widetilde{\mathbf{X}}} = (\widehat{\boldsymbol{\Sigma}}_X)^{-1/2}(\mathbf{X} - \widehat{\boldsymbol{\mu}})$ be empirical estimates. Let $\widehat{\mathbf{B}}_e$ be the empirical second moment of whitened paired differences. Threshold the spectrum of $\widehat{\mathbf{B}}_e$ to obtain a range projector $\widehat{\mathbf{P}}_e$, form $\widehat{\mathbf{C}}_{\text{sig}} = \sum_e 2^{e-1}\widehat{\mathbf{P}}_e$, and diagonalize $\widehat{\mathbf{C}}_{\text{sig}}$.

Proposition L.3 (Deterministic chart perturbation)**.** *Assume*

$$\left\|\widehat{\mathbf{B}}_e - \mathbf{B}_e\right\|_{\text{op}} \le \epsilon_B < \gamma/4 \quad \textit{for all } e,$$

and use any projector threshold in $[2\epsilon_B, \gamma - 2\epsilon_B]$*. Then*

$$\left\|\widehat{\mathbf{P}}_e - \mathbf{P}_e\right\|_{\text{op}} \le \epsilon_{\text{proj}} := \frac{2\epsilon_B}{\gamma - \epsilon_B}.$$

Consequently,

$$\left\|\widehat{\mathbf{C}}_{\text{sig}} - \mathbf{C}_{\text{sig}}\right\|_{\text{op}} \le \epsilon_{\text{sig}} := (2^m - 1)\epsilon_{\text{proj}}.$$

If $\epsilon_{\text{sig}} < 1/2$*, nearest-integer rounding recovers all code integers. If also* $\epsilon_{\text{sig}} < g_c/4$*, where* $g_c = \min_{i\neq j}|c_i - c_j| \ge 1$*, then there are a permutation matrix* $\boldsymbol{\Pi}$ *and diagonal reflection matrix* $\mathbf{J}$ *such that the recovered eigenvector matrix* $\widehat{\mathbf{O}}$ *satisfies*

$$\left\|\widehat{\mathbf{O}} - \mathbf{O}\boldsymbol{\Pi}\mathbf{J}\right\|_{\text{op}} \le C_d\frac{\epsilon_{\text{sig}}}{g_c}.$$

One may take $C_d = 2\sqrt{2d}$*; the proof gives the stronger Frobenius-norm bound with this constant.*

Proof. Fix e and order the eigenvalues of $\mathbf{B}_e$ increasingly. The range projector $\mathbf{P}_e$ projects onto the eigenspace with eigenvalues at least γ, while the nullspace has eigenvalue zero. Weyl's inequality gives, for every eigenvalue index, $\left|\widehat{\lambda}_k(\mathbf{B}_e) - \lambda_k(\mathbf{B}_e)\right| \le \left\|\widehat{\mathbf{B}}_e - \mathbf{B}_e\right\|_{\text{op}} \le \epsilon_B$.

Consequently the empirical null cluster lies in $[0, \epsilon_B]$ and the empirical positive cluster lies in $[\gamma - \epsilon_B, \infty)$. A threshold $\xi_e \in [2\epsilon_B, \gamma - 2\epsilon_B]$ therefore selects exactly the same rank as $\mathbf{P}_e$; the interval is nonempty because $\epsilon_B < \gamma/4$.

The two range subspaces have spectral separation at least $\gamma - \epsilon_B$. The Davis–Kahan sin-$\boldsymbol{\Theta}$ bound applied to them gives

$$\|\sin\boldsymbol{\Theta}_e(\mathrm{range}(\widehat{\mathbf{P}}_e), \mathrm{range}(\mathbf{P}_e))\|_{\text{op}} \le \frac{\epsilon_B}{\gamma - \epsilon_B}.$$

For orthogonal projectors of equal rank, $\|\widehat{\mathbf{P}}_e - \mathbf{P}_e\|_{\text{op}} = \|\sin\boldsymbol{\Theta}_e\|_{\text{op}}$. In particular, the slightly looser bound used below is $\|\widehat{\mathbf{P}}_e - \mathbf{P}_e\|_{\text{op}} \le 2\epsilon_B/(\gamma - \epsilon_B) = \epsilon_{\text{proj}}$.

Summing weighted errors and using the triangle inequality gives

$$\|\widehat{\mathbf{C}}_{\mathrm{sig}} - \mathbf{C}_{\mathrm{sig}}\|_{\mathrm{op}} = \left\| \sum_{e=1}^{m} 2^{e-1}(\widehat{\mathbf{P}}_e - \mathbf{P}_e) \right\|_{\mathrm{op}} \leq \sum_{e=1}^{m} 2^{e-1} \|\widehat{\mathbf{P}}_e - \mathbf{P}_e\|_{\mathrm{op}} \leq (2^m - 1)\epsilon_{\mathrm{proj}} = \epsilon_{\mathrm{sig}}.$$

The eigenvalues of $\mathbf{C}_{\mathrm{sig}}$ are the integers c_i. Weyl's inequality applied to $\widehat{\mathbf{C}}_{\mathrm{sig}}$ gives $|\widehat{c}_i - c_i| \leq \epsilon_{\mathrm{sig}}$. If $\epsilon_{\mathrm{sig}} < 1/2$, the nearest integer to each $\widehat{c}_i$ is uniquely c_i. For the eigenvectors, let $\mathbf{v}_i$ be the eigenvector for c_i and let $\widehat{\mathbf{v}}_i$ be the eigenvector matched to the nearest integer. The separation from the other population eigenvalues is at least $g_c - \epsilon_{\mathrm{sig}}$. A one-dimensional Davis–Kahan bound therefore gives

$$\sin \angle(\widehat{\mathbf{v}}_i, \mathbf{v}_i) \leq \frac{\epsilon_{\mathrm{sig}}}{g_c - \epsilon_{\mathrm{sig}}} \leq \frac{4}{3}\frac{\epsilon_{\mathrm{sig}}}{g_c} \quad \text{when } \epsilon_{\mathrm{sig}} < g_c/4.$$

Choosing signs and assembling the columns yields the signed permutation $\mathbf{\Pi J}$. For example, converting the columnwise angle bounds to Frobenius norm gives

$$\|\widehat{\mathbf{O}} - \mathbf{O\Pi J}\|_{\mathrm{F}} \leq 2\sqrt{2d}\frac{\epsilon_{\mathrm{sig}}}{g_c},$$

which is the stated finite-dimensional constant. □

Corollary L.4 (Chart error on the perturbation event)**.** *Suppose* $\lambda_{\min}(\mathbf{\Sigma}_X) \geq \lambda_- > 0$ *and*

$$\left\|\widehat{\mathbf{\Sigma}}_X - \mathbf{\Sigma}_X\right\|_{\mathrm{op}} \leq \epsilon_\Sigma < \lambda_-/2, \qquad \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \leq \epsilon_\mu.$$

Let $\mathbf{L} = \mathbf{A}_{\mathrm{mix}}^{-1} = \mathbf{O}^\top \mathbf{\Sigma}_X^{-1/2}$ *and* $\widehat{\mathbf{L}} = \widehat{\mathbf{O}}^\top (\widehat{\mathbf{\Sigma}}_X)^{-1/2}$*. Under the alignment in Proposition L.3,*

$$\left\|\widehat{\mathbf{L}} - \mathbf{J\Pi}^\top \mathbf{L}\right\|_{\mathrm{op}} \leq C\left(\lambda_-^{-1/2}\frac{C_d \epsilon_{\mathrm{sig}}}{g_c} + \lambda_-^{-3/2}\epsilon_\Sigma\right) =: \epsilon_L. \tag{116}$$

For a fresh factual observation,

$$\left(\mathbb{E}\left\|\widehat{\mathbf{L}}(\mathbf{X} - \widehat{\boldsymbol{\mu}}) - \mathbf{J\Pi}^\top \mathbf{L}(\mathbf{X} - \boldsymbol{\mu})\right\|_2^2\right)^{1/2} \leq \epsilon_L \sqrt{\operatorname{tr} \mathbf{\Sigma}_X} + \left\|\widehat{\mathbf{L}}\right\|_{\mathrm{op}} \epsilon_\mu.$$

The same bound multiplied by $1/\sqrt{2\pi}$ *controls residual-coordinate root-mean-square error because* Φ *is* $1/\sqrt{2\pi}$*-Lipschitz.*

Proof. First, $\widehat{\mathbf{\Sigma}}_X \succeq (\lambda_-/2)\mathbf{I}_d$ by Weyl. The resolvent representation of the inverse square root gives

$$\begin{aligned}(\widehat{\mathbf{\Sigma}}_X)^{-1/2} - \mathbf{\Sigma}_X^{-1/2} &= \frac{1}{\pi}\int_0^\infty t^{-1/2}\left[(\widehat{\mathbf{\Sigma}}_X + t\mathbf{I}_d)^{-1} - (\mathbf{\Sigma}_X + t\mathbf{I}_d)^{-1}\right] \mathrm{d}t \\ &= \frac{1}{\pi}\int_0^\infty t^{-1/2}(\widehat{\mathbf{\Sigma}}_X + t\mathbf{I}_d)^{-1}(\mathbf{\Sigma}_X - \widehat{\mathbf{\Sigma}}_X)(\mathbf{\Sigma}_X + t\mathbf{I}_d)^{-1}\,\mathrm{d}t. \end{aligned} \tag{117}$$

The minimum-eigenvalue bounds make the integral finite and imply, with a universal numerical C_{isr},

$$\|(\widehat{\mathbf{\Sigma}}_X)^{-1/2} - \mathbf{\Sigma}_X^{-1/2}\|_{\mathrm{op}} \leq C_{\mathrm{isr}}\lambda_-^{-3/2}\epsilon_\Sigma.$$

Let $\widetilde{\mathbf{L}} = \mathbf{J\Pi}^\top \mathbf{O}^\top \mathbf{\Sigma}_X^{-1/2}$ be the aligned population chart. Add and subtract $\widehat{\mathbf{O}}^\top \mathbf{\Sigma}_X^{-1/2}$:

$$\begin{aligned}\|\widehat{\mathbf{L}} - \widetilde{\mathbf{L}}\|_{\mathrm{op}} &\leq \|\widehat{\mathbf{O}}^\top - \mathbf{J\Pi}^\top \mathbf{O}^\top\|_{\mathrm{op}} \|\mathbf{\Sigma}_X^{-1/2}\|_{\mathrm{op}} \;+\; \|\widehat{\mathbf{O}}\|_{\mathrm{op}} \|(\widehat{\mathbf{\Sigma}}_X)^{-1/2} - \mathbf{\Sigma}_X^{-1/2}\|_{\mathrm{op}} \\ &\leq \lambda_-^{-1/2} C_d \frac{\epsilon_{\mathrm{sig}}}{g_c} \;+\; C_{\mathrm{isr}}\lambda_-^{-3/2}\epsilon_\Sigma. \end{aligned}$$

This is equation 116 after absorbing fixed numerical factors.

For a fresh $\mathbf{X}$, expand the centered chart discrepancy as

$$\widehat{\mathbf{L}}(\mathbf{X}-\widehat{\boldsymbol{\mu}})-\mathbf{J}\boldsymbol{\Pi}^{\top}\mathbf{L}(\mathbf{X}-\boldsymbol{\mu})=\left(\widehat{\mathbf{L}}-\mathbf{J}\boldsymbol{\Pi}^{\top}\mathbf{L}\right)(\mathbf{X}-\boldsymbol{\mu})+\widehat{\mathbf{L}}(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}).$$

Cauchy–Schwarz and $\mathbb{E}\|\mathbf{X}-\boldsymbol{\mu}\|_2^2=\operatorname{tr}\boldsymbol{\Sigma}_X$ give

$$\left(\mathbb{E}\|\cdot\|_2^2\right)^{1/2}\le\epsilon_L\sqrt{\operatorname{tr}\boldsymbol{\Sigma}_X}+\|\widehat{\mathbf{L}}\|_{\mathrm{op}}\epsilon_\mu.$$

Finally, $|\Phi(a)-\Phi(b)|\le|a-b|/\sqrt{2\pi}$ follows from the maximum of the standard normal density, giving the residual-coordinate statement. □

Assumption L.5 (Explicit covariance moment-generating bounds)**.** Let $\mathbf{X}^\circ=\mathbf{X}-\boldsymbol{\mu}$. For every unit vector $\mathbf{v}$ and every real λ,

$$\mathbb{E}\exp(\lambda\mathbf{v}^{\top}\mathbf{X}^\circ)\le\exp(\lambda^2\kappa_X^2/2). \tag{118}$$

For $|\lambda|\le\kappa_X^{-2}$, also assume

$$\mathbb{E}\exp\left(\lambda\{(\mathbf{v}^{\top}\mathbf{X}^\circ)^2-\mathbb{E}(\mathbf{v}^{\top}\mathbf{X}^\circ)^2\}\right)\le\exp(\lambda^2\kappa_X^4/2). \tag{119}$$

For the raw paired difference $\boldsymbol{\Delta}_e^X=\mathbf{X}^{(e)}-\mathbf{X}$, assume the analogue of equation 119 with parameter κ_e. The factual sample used for $(\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Sigma}}_X)$ is independent of all paired samples. Finally,

$$\lambda_{\min}(\boldsymbol{\Sigma}_X)\ge\lambda_->0,\qquad\|\mathbf{C}_e\|_{\mathrm{op}}\le L_{C,e},\qquad\mathbf{C}_e=\mathbb{E}[\boldsymbol{\Delta}_e^X(\boldsymbol{\Delta}_e^X)^{\top}].$$

For $n\ge1$, $\kappa>0$, and $q>0$, define

$$t(n,\kappa,q)=\kappa^2\max\left\{\sqrt{\frac{2q}{n}},\frac{2q}{n}\right\}.$$

Theorem L.6 (Simultaneous residual-chart and signature recovery)**.** *Under Assumptions L.1 and L.5, set*

$$\chi_0=\log\frac{8\cdot9^d}{\delta},\qquad\chi_{\mathrm{pair}}=\log\frac{4m\cdot9^d}{\delta}, \tag{120}$$

$$\epsilon_\mu=2\kappa_X\sqrt{\frac{2\chi_0}{n_0}},\qquad\epsilon_\Sigma=2t(n_0,\kappa_X,\chi_0)+\frac{8\kappa_X^2\chi_0}{n_0}, \tag{121}$$

$$\epsilon_{C,e}=2t(n_e,\kappa_e,\chi_{\mathrm{pair}}). \tag{122}$$

If $\epsilon_\Sigma<\lambda_-/2$, define

$$\epsilon_{\mathrm{wh}}=\lambda_-^{-3/2}\epsilon_\Sigma,\quad\epsilon_{B,e}=\frac{2\epsilon_{C,e}}{\lambda_-}+\epsilon_{\mathrm{wh}}L_{C,e}\left(\sqrt{\frac{2}{\lambda_-}}+\frac{1}{\sqrt{\lambda_-}}\right),\quad\epsilon_B=\max_e\epsilon_{B,e}. \tag{123}$$

Then, with probability at least $1-\delta$, simultaneously,

$$\|\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|_2\le\epsilon_\mu,\quad\|\widehat{\boldsymbol{\Sigma}}_X-\boldsymbol{\Sigma}_X\|_{\mathrm{op}}\le\epsilon_\Sigma,\quad\|\widehat{\mathbf{B}}_e-\mathbf{B}_e\|_{\mathrm{op}}\le\epsilon_{B,e}\quad(e\in[m]).$$

Consequently, if $\epsilon_B<\gamma/4$ and the deterministic separation conditions of Proposition L.3 hold, code and chart recovery obey that proposition and Corollary L.4 with confidence $1-\delta$.

Proof. The mgf bound equation 119 and Chernoff's method give, for a fixed unit $\mathbf{v}$ and $W_s=(\mathbf{v}^{\top}\mathbf{X}_s^\circ)^2-\mathbb{E}(\mathbf{v}^{\top}\mathbf{X}^\circ)^2$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{s=1}^n W_s\right|>x\right)\le2\exp\left[-\frac{n}{2}\min\left\{\frac{x^2}{\kappa_X^4},\frac{x}{\kappa_X^2}\right\}\right]. \tag{124}$$

Indeed, for $x\le\kappa_X^2$ choose $\lambda=x/\kappa_X^4$; for $x>\kappa_X^2$ choose the endpoint $\lambda=\kappa_X^{-2}$ and loosen $-nx/\kappa_X^2+n/2\le-nx/(2\kappa_X^2)$. The same argument applies to $\boldsymbol{\Delta}_e^X$ with κ_e.

A $1/4$-net $\mathcal{V}_{1/4}$ of the unit sphere has size at most 9^d, and every symmetric matrix $\mathbf{A}$ satisfies

$$\|\mathbf{A}\|_{\mathrm{op}}\le2\max_{\mathbf{v}\in\mathcal{V}_{1/4}}|\mathbf{v}^{\top}\mathbf{A}\mathbf{v}|. \tag{125}$$

Substitute $x = t(n_0, \kappa_X, \chi_0)$ in equation 124, take a union over the net, and use equation 125. The raw second-moment error is at most $2t(n_0, \kappa_X, \chi_0)$ except on an event of probability $\delta/4$. For the mean, equation 118 gives

$$\mathbb{P}(|\mathbf{v}^\top(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})| > s) \le 2\exp\left(-\frac{n_0 s^2}{2\kappa_X^2}\right).$$

The same net with $s = \kappa_X\sqrt{2\chi_0/n_0}$ gives $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \le 2s = \epsilon_\mu$ except with probability $\delta/4$. Since

$$\widehat{\boldsymbol{\Sigma}}_X = \frac{1}{n_0}\sum_s \mathbf{X}_s^\circ(\mathbf{X}_s^\circ)^\top - (\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})^\top,$$

the rank-one correction has norm at most $\epsilon_\mu^2 = 8\kappa_X^2\chi_0/n_0$. This proves the first two radii.

For every pair channel, use $x = t(n_e, \kappa_e, \chi_{\text{pair}})$, union over m nets, and use the raw pair second moment. The total pair failure probability is at most $\delta/2$, and $\|\widehat{\mathbf{C}}_e - \mathbf{C}_e\|_{\text{op}} \le \epsilon_{C,e}$ for every e.

Estimated whitening adds the following perturbation. Set $\mathbf{W} = \boldsymbol{\Sigma}_X^{-1/2}$ and $\widehat{\mathbf{W}} = (\widehat{\boldsymbol{\Sigma}}_X)^{-1/2}$. On the covariance event, $\widehat{\boldsymbol{\Sigma}}_X \succeq (\lambda_-/2)\mathbf{I}_d$. The resolvent identity in equation 117 gives

$$\begin{aligned}\|\widehat{\mathbf{W}} - \mathbf{W}\|_{\text{op}} &\le \frac{\epsilon_\Sigma}{\pi}\int_0^\infty \frac{t^{-1/2}}{(t + \lambda_-/2)(t + \lambda_-)}\,\mathrm{d}t \\ &= \frac{2}{1+\sqrt{2}}\lambda_-^{-3/2}\epsilon_\Sigma \le \epsilon_{\text{wh}}.\end{aligned}$$

Also $\|\widehat{\mathbf{W}}\|_{\text{op}} \le \sqrt{2/\lambda_-}$ and $\|\mathbf{W}\|_{\text{op}} \le 1/\sqrt{\lambda_-}$. With $\mathbf{B}_e = \mathbf{W}\mathbf{C}_e\mathbf{W}$ and $\widehat{\mathbf{B}}_e = \widehat{\mathbf{W}}\widehat{\mathbf{C}}_e\widehat{\mathbf{W}}$, add and subtract $\widehat{\mathbf{W}}\mathbf{C}_e\widehat{\mathbf{W}}$:

$$\begin{aligned}\|\widehat{\mathbf{B}}_e - \mathbf{B}_e\|_{\text{op}} &\le \|\widehat{\mathbf{W}}\|_{\text{op}}^2\|\widehat{\mathbf{C}}_e - \mathbf{C}_e\|_{\text{op}} \;+\; \|\widehat{\mathbf{W}} - \mathbf{W}\|_{\text{op}}\|\mathbf{C}_e\|_{\text{op}}(\|\widehat{\mathbf{W}}\|_{\text{op}} + \|\mathbf{W}\|_{\text{op}}) \\ &\le \epsilon_{B,e}.\end{aligned}$$

The three allocated failure probabilities sum to δ, completing the simultaneous proof. □

Remark L.7 (Fixed threshold and relative separation). Theorem L.6 uses explicit constants. Once its fixed population quantities are specified, choose

$$\xi_e = \gamma/2 \qquad (e \in [m]). \tag{126}$$

Indeed, $\epsilon_B < \gamma/4$ gives $2\epsilon_B < \gamma/2 < \gamma - 2\epsilon_B$, so equation 126 lies strictly inside the admissible interval of Proposition L.3. The estimator evaluates equation 120–equation 123 and assumes the four strict inequalities

$$\epsilon_\Sigma < \lambda_-/2, \quad \epsilon_B < \gamma/4, \quad \epsilon_{\text{sig}} < 1/2, \quad \epsilon_{\text{sig}} < g_c/4, \tag{127}$$

and uses the fixed threshold $\gamma/2$ in the projector estimator.

L.2 Independent-sample estimator

With the chart and signatures estimated, the next task is to recover one causal coordinate at a time. A sink's conditional response passes two comparisons: identity in environments that do not target it, and a shared oriented foliation in environments that do. The scores below measure departures from these population conditions.

We use separate samples to fit the chart, fit the ratios, and calibrate the scores. This lets each fitted object be treated as fixed when analyzing the next sample. Graph recovery uses fresh randomization after peeling fixes the coordinate map. Write

$$\mathcal{D}_{\text{ch}}, \quad \mathcal{D}_{\text{rat}}, \quad \mathcal{D}_{\text{cal}}, \quad \mathcal{D}_{\text{gr}} \tag{128}$$

for these three independent sample groups and the graph-query random bits, respectively. The corresponding information sets are $\mathcal{F}_{\text{ch}} := \sigma(\mathcal{D}_{\text{ch}})$ and $\mathcal{F}_{\text{fit}} := \sigma(\mathcal{D}_{\text{ch}}, \mathcal{D}_{\text{rat}})$. Conditional on the fitted chart, the ratio-training observations remain i.i.d. This separation will let us analyze each fitted

component in turn. For each nonempty active set, order its coordinates as $R = \{r_1 < \cdots < r_{d_R}\}$ and write $u_R = (u_{r_1}, \ldots, u_{r_{d_R}})$. Let π_R be the uniform probability law on $(0,1)^R$.

A calibration query is a transformed factual observation at which we evaluate the fitted scores. We keep queries away from the cube boundary, where probit derivatives become large. The factual calibration sample is transformed once in the full residual chart. We retain a query only if all d estimated coordinates lie in $\mathcal{U}_{[d]} := [\eta, 1-\eta]^d$. The same accepted queries, projected onto R, are then reused throughout peeling; acceptance is never recomputed on a smaller active set. Write $\widetilde{\mathbf{U}}^\sharp_s$ for the exact dyadic value represented by query s. Section L.5.2 explains this representation and its outward interval enclosures. The acceptance indicator, accepted count, and projected empirical measure are

$$\widehat{\pi}^\sharp_R = \frac{1}{N^\sharp_{\mathrm{cal}}} \sum_{s: B^\sharp_s = 1} \delta_{\widetilde{\mathbf{U}}^\sharp_{s,R}}, \qquad B^\sharp_s := \mathbf{1}\{\widetilde{\mathbf{U}}^\sharp_s \in \mathcal{U}_{[d]}\}, \quad N^\sharp_{\mathrm{cal}} := \sum_s B^\sharp_s. \tag{129}$$

Proposition L.27 accounts for both the estimated-chart query law and the random number of accepted observations. Fit the residual chart and target signatures on $\mathcal{D}_{\mathrm{ch}}$ using whitening, paired-difference range projectors, and equation 115. Write the recovered signature matrix as $\widehat{\boldsymbol{\Sigma}} = [\widehat{\boldsymbol{\sigma}}_1 \ \cdots \ \widehat{\boldsymbol{\sigma}}_d]$. At active set R, the density ratio compares the environment's joint residual density with the factual density. We approximate this ratio by a polynomial; the Legendre basis and its two-sample fit are developed in Section L.3. The fitted coefficients solve

$$\widehat{\mathbf{G}}_R \widehat{\boldsymbol{\beta}}^R_e = \widehat{\mathbf{b}}^R_e, \qquad \widehat{r}^R_e = \widehat{\boldsymbol{\beta}}^{R\top}_e \boldsymbol{\psi}_{R,p}, \qquad \widehat{\ell}^R_e := \log \widehat{r}^R_e, \tag{130}$$

where the Gram matrix and intervention moment are defined in equation 144. The full-cube positivity test provides the lower bound required before taking the logarithm.

Holding all coordinates except u_j fixed traces a one-dimensional fibre. In the true chart, whose factual density is uniform, normalizing the ratio along this fibre gives the environment's conditional density. Integrating that density up to u_j gives its conditional CDF. We apply the same operation to the fitted ratio on $[\eta, 1-\eta]$. The operator $\mathcal{C}_{j,\eta}$ performs this normalization and integration, as defined in Section L.4.1, so the fitted CDF is

$$\widehat{C}^R_{j,e} := \mathcal{C}_{j,\eta}[\widehat{r}^R_e].$$

Section L.5.2 evaluates this CDF and its gradient on the calibration queries, yielding $\check{C}^R_{j,e}$ and $\check{K}^R_{j,e}$, the evaluated conditional density $K = \partial_{u_j} C$. The gradient ∇C records how the CDF changes with all active coordinates. The exact averages of these evaluated values give the empirical identity and foliation scores:

$$\widehat{I}^R_{j,e} = \widehat{\pi}^\sharp_R \left[(\check{C}^R_{j,e} - u_j)^2 + \left\| \nabla \check{C}^R_{j,e} - \mathbf{e}_j \right\|_2^2 \right], \tag{131}$$

$$\widehat{F}^R_{j,ef} = \widehat{\pi}^\sharp_R \left[\frac{\left\| \nabla \check{C}^R_{j,e} - (\check{K}^R_{j,e}/\check{K}^R_{j,f}) \nabla \check{C}^R_{j,f} \right\|_2^2}{1 + \left\| \nabla \check{C}^R_{j,e} \right\|_2^2} \right]. \tag{132}$$

Their numerical evaluations are denoted $\check{I}^R_{j,e}$ and $\check{F}^R_{j,ef}$; Section L.5.2 bounds the error in forming the averages. The identity score asks whether the conditional CDF equals u_j, the CDF of a uniform variable. This is a condition on the joint law: a residual unchanged within every pair can still become dependent on the other residuals. For a targeted sink, normalization makes each population CDF a monotone transformation of the same causal coordinate. Different environments can therefore change CDF values while preserving the level sets and their ordering, which is the oriented foliation. The gradients are normals to these level sets. Their positive u_j components, $K^R_{j,e}$ and $K^R_{j,f}$, determine the scale factor used to compare their directions. Taking the largest applicable returned score gives the computed sink score

$$\check{s}^R_j = \max \left\{ \max_{e: \widehat{\sigma}_{ej} = 0} \check{I}^R_{j,e}, \max_{e<f: \widehat{\sigma}_{ej} = \widehat{\sigma}_{fj} = 1} \check{F}^R_{j,ef} \right\}, \tag{133}$$

where an empty inner maximum equals zero. At the population level, true sinks have zero score, while the sink-margin condition separates every non-sink from zero. These zero-score statements use population CDFs normalized over the full fibre $[0,1]$. The computed CDFs use $[\eta, 1-\eta]$, so their

Algorithm 2 General Estimator

Require: independent samples $\mathcal{D}_{\mathrm{ch}}, \mathcal{D}_{\mathrm{rat}}, \mathcal{D}_{\mathrm{cal}}, \mathcal{D}_{\mathrm{gr}}$; fixed estimator parameters

Ensure: adjacency matrix $\widehat{\mathbf{A}}_{\mathrm{out}}$ or $\perp$

1. Fit the residual chart and queries

Fit the residual chart and target signatures from $\mathcal{D}_{\mathrm{ch}}$
Form the accepted calibration queries from $\mathcal{D}_{\mathrm{cal}}$ by equation 129
if $N^{\sharp}_{\mathrm{cal}} = 0$ **then**
    return $\perp$
end if

2. Certify, recover, and peel sinks

$R \leftarrow [d]$, $\pi \leftarrow ()$; initialize $\widehat{Y}$ ▷ active nodes and sink order
while $R \neq \varnothing$ **do**
    Fit $\{\widehat{r}^R_e\}$ from $\mathcal{D}_{\mathrm{rat}}$ and evaluate $\{\check{s}^R_j : j \in R\}$ by equation 130–equation 133
    if a required comparison is unresolved **then**
        return $\perp$
    end if
    $\widehat{\mathcal{S}}^{\mathrm{pass}}_R \leftarrow \{j \in R : \check{s}^R_j + e^s_R \le t_R\}$
    Select (j, e_j) from $\{(j,e) : j \in \widehat{\mathcal{S}}^{\mathrm{pass}}_R, \widehat{\sigma}_{e,j} = 1\}$ by the fixed tie rules
    if no such (j, e_j) exists **then**
        return $\perp$
    end if
    $\widehat{Y}_j \leftarrow \widehat{C}^R_{j,e_j}$; $\pi \leftarrow (\pi, j)$
    $R \leftarrow R \setminus \{j\}$ ▷ peel the certified sink
end while

3. Recover direct edges

Estimate $\{\widehat{s}_{ij} : i \neq j\}$ from fresh $\mathcal{D}_{\mathrm{gr}}$ queries of $[\mathrm{D}\widehat{Y}]^{-1}$
if a graph-query condition is unresolved **then**
    return $\perp$
end if
$\widehat{\mathbf{A}}_{\mathrm{out},ji} \leftarrow \mathbf{1}\{\widehat{s}_{ij} > 2\epsilon_{\mathrm{gr}}\}$ for $i \neq j$; set the diagonal to zero
if $\widehat{\mathbf{A}}_{\mathrm{out}}$ is cyclic **then**
    return $\perp$
end if
return $\widehat{\mathbf{A}}_{\mathrm{out}}$ ▷ recovered adjacency matrix

truncation bias enters the score-error radius even when the ratio is fitted exactly. The estimator allows for approximation and sampling error before declaring a sink. Algorithm 2 collects these decisions and returns $\perp$ when a required comparison is unresolved.

The analysis compares these empirical quantities with population scores under the uniform query law π_R:

$$I^R_{j,e} = \pi_R\big[(C^R_{j,e} - u_j)^2 + \|\nabla C^R_{j,e} - \mathbf{e}_j\|_2^2\big], \tag{134}$$

$$F^R_{j,ef} = \pi_R\left[\frac{\|\nabla C^R_{j,e} - (K^R_{j,e}/K^R_{j,f})\nabla C^R_{j,f}\|_2^2}{1 + \|\nabla C^R_{j,e}\|_2^2}\right], \tag{135}$$

$$s^R_j = \max\left\{\max_{e:\sigma_{ej}=0} I^R_{j,e}, \max_{e<f:\sigma_{ej}=\sigma_{fj}=1} F^R_{j,ef}\right\}. \tag{136}$$

Here $\pi_R[h] = \int h \, \mathrm{d}\pi_R$, and empty inner maxima equal zero. s^R_j is a scalar functional of the population conditional laws. Its empirical counterpart uses the recovered signatures, fitted conditional laws, and calibration measure.

L.3 LEGENDRE RATIO ESTIMATION

We now bound the error of the polynomial ratio fit used by the estimator. Legendre polynomials are orthogonal under the uniform factual law of the true residual chart, so each coefficient is an intervention moment. The estimated chart requires a correction through the factual Gram matrix. In either case, the error has two parts: the omitted polynomial tail and the error in estimating the retained coefficients. We control both values and first derivatives because the sink scores use the CDFs and their gradients.

Let P_k be the ordinary degree-k Legendre polynomial on $[-1,1]$. Fix $R = \{r_1 < \cdots < r_{d_R}\}$, where $d_R := |R|$, and identify $u_R = (u_{r_1}, \ldots, u_{r_{d_R}}) \in (0,1)^{d_R}$. For multi-indices $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}_0^{d_R}$, define the shifted orthonormal basis:

$$\psi_{\boldsymbol{\alpha}}^R(u_R) = \prod_{j=1}^{d_R} \sqrt{2\alpha_j + 1}\, P_{\alpha_j}(2u_{r_j} - 1), \qquad \int_{(0,1)^{d_R}} \psi_{\boldsymbol{\alpha}}^R(u_R)\psi_{\boldsymbol{\beta}}^R(u_R)\, \mathrm{d}u_R = \mathbf{1}\{\boldsymbol{\alpha} = \boldsymbol{\beta}\}.$$

Put $\mathcal{I}_p^{(R)} = \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\|_1 \le p\}$, $\quad L_{R,p} = |\mathcal{I}_p^{(R)}| = \binom{d_R+p}{p}$,

$$B_{\boldsymbol{\alpha}} = \prod_{j=1}^{d_R} \sqrt{2\alpha_j + 1}, \qquad D_{\boldsymbol{\alpha}} = \left(\sum_{j=1}^{d_R} [\alpha_j(\alpha_j + 1)]^2 \right)^{1/2}.$$

The classical bounds $|P_k(x)| \le 1$ and $\sup_{[-1,1]} |P_k'(x)| = k(k+1)/2$ imply

$$\|\psi_{\boldsymbol{\alpha}}^R\|_\infty \le B_{\boldsymbol{\alpha}}, \qquad \|\nabla \psi_{\boldsymbol{\alpha}}^R\|_{2,\infty} \le B_{\boldsymbol{\alpha}} D_{\boldsymbol{\alpha}}. \tag{137}$$

Set $\|h\|_{C^1} := \max\{\|h\|_\infty, \|\nabla h\|_{2,\infty}\}$.

Assumption L.8 (Fixed Legendre tail)**.** In the true residual chart, an active density ratio has the expansion $r = \sum_{\boldsymbol{\alpha}} a_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}^R$ and fixed tail bounds

$$\sum_{\boldsymbol{\alpha} \notin \mathcal{I}_p^{(R)}} |a_{\boldsymbol{\alpha}}| B_{\boldsymbol{\alpha}} \le A_0(p), \qquad \sum_{\boldsymbol{\alpha} \notin \mathcal{I}_p^{(R)}} |a_{\boldsymbol{\alpha}}| B_{\boldsymbol{\alpha}} \max\{1, D_{\boldsymbol{\alpha}}\} \le A_1(p). \tag{138}$$

Also

$$0 < c_r \le r \le C_r, \qquad \|\nabla r\|_{2,\infty} \le G_r.$$

The tail bounds and envelopes A_0, A_1, c_r, C_r, G_r specify the model class; they are fixed before sampling.

Given n intervention residuals $\mathbf{V}_1, \ldots, \mathbf{V}_n \sim r$, define

$$\widehat{a}_{\boldsymbol{\alpha}} = \frac{1}{n} \sum_{s=1}^{n} \psi_{\boldsymbol{\alpha}}^R(\mathbf{V}_{s,R}), \qquad \widehat{r}_p = \sum_{\boldsymbol{\alpha} \in \mathcal{I}_p^{(R)}} \widehat{a}_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}^R.$$

The constant coefficient equals one exactly because $\psi_{\mathbf{0}}^R \equiv 1$.

Theorem L.9 (Uniform-reference C^0/C^1 Legendre ratio bound)**.** *Under Assumption L.8, let*

$$b_{\boldsymbol{\alpha}}(\delta) = B_{\boldsymbol{\alpha}} \sqrt{\frac{2\log(2L_{R,p}/\delta)}{n}}. \tag{139}$$

With probability at least $1 - \delta$, simultaneously,

$$|\widehat{a}_{\boldsymbol{\alpha}} - a_{\boldsymbol{\alpha}}| \le b_{\boldsymbol{\alpha}}(\delta) \qquad (\boldsymbol{\alpha} \in \mathcal{I}_p^{(R)}), \tag{140}$$

$$\|\widehat{r}_p - r\|_\infty \le A_0(p) + \sum_{\boldsymbol{\alpha} \in \mathcal{I}_p^{(R)}} b_{\boldsymbol{\alpha}} B_{\boldsymbol{\alpha}} =: \epsilon_{r,0}, \tag{141}$$

$$\|\widehat{r}_p - r\|_{C^1} \le A_1(p) + \sum_{\boldsymbol{\alpha} \in \mathcal{I}_p^{(R)}} b_{\boldsymbol{\alpha}} B_{\boldsymbol{\alpha}} \max\{1, D_{\boldsymbol{\alpha}}\} =: \epsilon_{r,1}. \tag{142}$$

If $\epsilon_{r,0} < c_r$, then $\widehat{r}_p \ge c_r - \epsilon_{r,0} > 0$ on the entire cube.

Proof. Orthonormality and the uniform factual reference measure give, coefficient by coefficient,

$$a_{\boldsymbol{\alpha}} = \int r(u_R)\psi^R_{\boldsymbol{\alpha}}(u_R)\,\mathrm{d}u_R = \mathbb{E}_r[\psi^R_{\boldsymbol{\alpha}}(\mathbf{V}_R)].$$

Each summand lies in $[-B_{\boldsymbol{\alpha}}, B_{\boldsymbol{\alpha}}]$. Hoeffding's inequality therefore gives

$$\mathbb{P}\left(|\widehat{a}_{\boldsymbol{\alpha}} - a_{\boldsymbol{\alpha}}| > t\right) \le 2\exp\left(-\frac{nt^2}{2B^2_{\boldsymbol{\alpha}}}\right).$$

Substitution of equation 139 makes this probability $\delta/L_{R,p}$; a union bound over the exactly $L_{R,p}$ retained coefficients proves equation 140.

On that event, split the estimation error into retained coefficients and the tail:

$$\widehat{r}_p - r = \sum_{\boldsymbol{\alpha}\in\mathcal{I}_p^{(R)}} (\widehat{a}_{\boldsymbol{\alpha}} - a_{\boldsymbol{\alpha}})\psi^R_{\boldsymbol{\alpha}} - \sum_{\boldsymbol{\alpha}\notin\mathcal{I}_p^{(R)}} a_{\boldsymbol{\alpha}}\psi^R_{\boldsymbol{\alpha}}.$$

For the value norm, apply the triangle inequality and the first envelope in equation 137 to each retained term, then use equation 138; this gives equation 141. For the gradient norm, differentiate the finite retained sum term by term, apply the second envelope, and use equation 138. Taking the maximum of value and gradient bounds gives equation 142. Finally, pointwise, $\widehat{r}_p(u) \ge r(u) - |\widehat{r}_p(u) - r(u)| \ge c_r - \epsilon_{r,0}$, which proves positivity conditional on the fixed lower envelope. □

In the estimated chart, the factual density can be nonuniform, so the same Legendre functions need not remain orthogonal under the factual law. The density ratio still satisfies the moment identity $\mathbb{E}_0[\widetilde{r}\,\boldsymbol{\psi}] = \mathbb{E}_e[\boldsymbol{\psi}]$. Substituting a polynomial for $\widetilde{r}$ gives the factual Gram system below. This estimates the ratio in the learned coordinates. Section L.4 then compares that ratio with the true-coordinate ratio used by the sink characterization.

Order $\mathcal{I}_p^{(R)}$ lexicographically in $(\alpha_1, \dots, \alpha_{d_R})$. This fixes both the rows and columns of every Gram matrix. Collect the basis functions into the deterministic feature vector

$$\boldsymbol{\psi}_{R,p}(u_R) := (\psi^R_{\boldsymbol{\alpha}}(u_R))_{\boldsymbol{\alpha}\in\mathcal{I}_p^{(R)}} \in \mathbb{R}^{L_{R,p}},$$

and define its Euclidean value and C^1 envelopes by

$$B_{0,R,p} := \left(\sum_{\boldsymbol{\alpha}\in\mathcal{I}_p^{(R)}} B^2_{\boldsymbol{\alpha}}\right)^{1/2}, B_{1,R,p} := \left(\sum_{\boldsymbol{\alpha}\in\mathcal{I}_p^{(R)}} B^2_{\boldsymbol{\alpha}} \max\{1, D_{\boldsymbol{\alpha}}\}^2\right)^{1/2}.$$

Thus $\|\boldsymbol{\psi}_{R,p}(u_R)\|_2 \le B_{0,R,p}$, and for every coefficient vector $\boldsymbol{\beta}$,

$$\|\boldsymbol{\beta}^\top\boldsymbol{\psi}_{R,p}\|_\infty \le B_{0,R,p}\|\boldsymbol{\beta}\|_2, \qquad \|\boldsymbol{\beta}^\top\boldsymbol{\psi}_{R,p}\|_{C^1} \le B_{1,R,p}\|\boldsymbol{\beta}\|_2. \tag{143}$$

Two-sample fitting with a fixed active set. Fix (R, e, p) for the following calculation. We suppress these fixed indices on the feature vector, Gram matrix, moment vector, and coefficients. Thus $\mathbf{G} = \mathbf{G}_R$ and $\widehat{r}_p = \widehat{r}^R_e$, for example. Write $L = L_{R,p}$, $B_0 = B_{0,R,p}$, $B_1 = B_{1,R,p}$. The same convention applies to their estimation radii, such as $\epsilon_G = \epsilon_{G,R,p}$ and $\epsilon_\beta = \epsilon_{\beta,R,e,p}$. The resulting ratio-error bounds retain their indices for use later in the analysis.

Assumption L.10 (Legendre projection under the residual-chart factual law)**.** For the fixed fitted chart $\widehat{q}$, write $\widetilde{\mathbf{U}} = \widehat{q}(\mathbf{X})$ and $\widetilde{\mathbf{V}} = \widehat{q}(\mathbf{X}^{(e)})$. Let P^R_0 and P^R_e be their respective R-marginal laws, conditional on the fitted chart. Their density ratio is $\widetilde{r} = \mathrm{d}P^R_e/\,\mathrm{d}P^R_0$, also written $\widetilde{r}^R_e$ when the chart and indices need to be explicit. Define

$$\mathbf{G} := \mathbb{E}_0[\boldsymbol{\psi}(\widetilde{\mathbf{U}}_R)\boldsymbol{\psi}(\widetilde{\mathbf{U}}_R)^\top], \qquad \mathbf{b} := \mathbb{E}_e[\boldsymbol{\psi}(\widetilde{\mathbf{V}}_R)].$$

Assume

$$\lambda_{\min}(\mathbf{G}) \ge \lambda_G > 0, \qquad \boldsymbol{\beta}^\circ := \mathbf{G}^{-1}\mathbf{b}, \qquad r^\circ_p := (\boldsymbol{\beta}^\circ)^\top\boldsymbol{\psi}.$$

Assume the projection and density bounds

$$\|r_p^\circ - \widetilde{r}\|_\infty \le \tau_0,$$
$$\|r_p^\circ - \widetilde{r}\|_{C^1} \le \tau_1,$$
$$\inf_{u_R} \widetilde{r}(u_R) \ge c > 0.$$

Here $\tau_k = A_{G,R,e,k}(p)$ for $k = 0, 1$ and $c = c_{G,r}^{R,e}$. The lower Gram bound $\lambda_G = \lambda_{G,R,p}$, projection errors τ_0, τ_1, and density floor c are fixed model-class inputs, chosen independently of these samples.

Given factual samples $\widetilde{\mathbf{U}}_{1,R}, \dots, \widetilde{\mathbf{U}}_{n_0,R} \sim P_0^R$ and intervention samples $\widetilde{\mathbf{V}}_{1,R}, \dots, \widetilde{\mathbf{V}}_{n_e,R} \sim P_e^R$, define

$$\widehat{\mathbf{G}} := \frac{1}{n_0}\sum_{s=1}^{n_0} \boldsymbol{\psi}(\widetilde{\mathbf{U}}_{s,R})\boldsymbol{\psi}(\widetilde{\mathbf{U}}_{s,R})^\top, \qquad \widehat{\mathbf{b}} := \frac{1}{n_e}\sum_{s=1}^{n_e} \boldsymbol{\psi}(\widetilde{\mathbf{V}}_{s,R}), \tag{144}$$
$$\widehat{\boldsymbol{\beta}} := \widehat{\mathbf{G}}^{-1}\widehat{\mathbf{b}}, \qquad \widehat{r}_p := \widehat{\boldsymbol{\beta}}^\top \boldsymbol{\psi}.$$

Theorem L.11 (Two-sample Legendre environment-ratio bound)**.** *Under Assumption L.10, let*

$$\epsilon_G(\delta) := (B_0)^2\sqrt{\frac{2\log(4L^2/\delta)}{n_0}}, \qquad \epsilon_b(\delta) := B_0\sqrt{\frac{2\log(4L/\delta)}{n_e}}.$$

If $\epsilon_G < \lambda_G$, define

$$B_\beta := \frac{B_0}{\lambda_G}, \qquad \epsilon_\beta := \frac{\epsilon_b + \epsilon_G B_\beta}{\lambda_G - \epsilon_G}.$$

With probability at least $1-\delta$, simultaneously,

$$\|\widehat{\mathbf{G}} - \mathbf{G}\|_{\mathrm{op}} \le \epsilon_G, \qquad \|\widehat{\mathbf{b}} - \mathbf{b}\|_2 \le \epsilon_b,$$
$$\lambda_{\min}(\widehat{\mathbf{G}}) \ge \lambda_G - \epsilon_G > 0, \qquad \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\circ\|_2 \le \epsilon_\beta.$$

Consequently,

$$\|\widehat{r}_p - \widetilde{r}\|_\infty \le \tau_0 + B_0\epsilon_\beta =: \epsilon_{G,r,0}^{R,e,p}, \tag{145}$$
$$\|\widehat{r}_p - \widetilde{r}\|_{C^1} \le \tau_1 + B_1\epsilon_\beta =: \epsilon_{G,r,1}^{R,e,p}. \tag{146}$$

The target-gradient envelope required by the log bridge may be taken as

$$\|\nabla\widetilde{r}\|_{2,\infty} \le \tau_1 + B_1 B_\beta =: G_{G,r}^{R,e,p}. \tag{147}$$

If the value-error bound in equation 145 is smaller than c, then $\widehat{r}_p \ge c - \epsilon_{G,r,0}^{R,e,p} > 0$ throughout the cube. The logarithm in equation 130 is therefore well defined.

Proof. The Radon–Nikodym identity gives

$$\mathbf{b} = \mathbb{E}_e[\boldsymbol{\psi}(\widetilde{\mathbf{V}}_R)] = \mathbb{E}_0[\widetilde{r}(\widetilde{\mathbf{U}}_R)\boldsymbol{\psi}(\widetilde{\mathbf{U}}_R)].$$

For an arbitrary $\boldsymbol{\beta} \in \mathbb{R}^L$, the projection objective expands as

$$\mathbb{E}_0\Big[\big(\widetilde{r}(\widetilde{\mathbf{U}}_R) - \boldsymbol{\beta}^\top\boldsymbol{\psi}(\widetilde{\mathbf{U}}_R)\big)^2\Big] = \mathbb{E}_0[(\widetilde{r})^2] - 2\boldsymbol{\beta}^\top\mathbf{b} + \boldsymbol{\beta}^\top\mathbf{G}\boldsymbol{\beta}.$$

Because $\mathbf{G} \succ 0$, this quadratic has a unique minimizer, which satisfies $\mathbf{G}\boldsymbol{\beta}^\circ = \mathbf{b}$. Thus r_p° is the $L^2(P_0^R)$ projection of the environment ratio, with the factual Gram system supplying the relevant geometry.

For each pair $\boldsymbol{\alpha}, \boldsymbol{\gamma} \in \mathcal{I}_p^{(R)}$, the factual Gram summand obeys $|\psi_{\boldsymbol{\alpha}}^R(\widetilde{\mathbf{U}}_{s,R})\psi_{\boldsymbol{\gamma}}^R(\widetilde{\mathbf{U}}_{s,R})| \le B_{\boldsymbol{\alpha}}B_{\boldsymbol{\gamma}}$.

Hoeffding's inequality and a union bound over the L^2 ordered entries therefore give, with failure probability at most $\delta/2$,

$$|(\widehat{\mathbf{G}} - \mathbf{G})_{\boldsymbol{\alpha}\boldsymbol{\gamma}}| \le B_{\boldsymbol{\alpha}}B_{\boldsymbol{\gamma}}\sqrt{\frac{2\log(4L^2/\delta)}{n_0}}. \tag{148}$$

On this event,

$$\|\widehat{\mathbf{G}} - \mathbf{G}\|_{\mathrm{op}} \le \|\widehat{\mathbf{G}} - \mathbf{G}\|_{\mathrm{F}} \le \left(\sum_{\boldsymbol{\alpha},\boldsymbol{\gamma}} B_{\boldsymbol{\alpha}}^2 B_{\boldsymbol{\gamma}}^2\right)^{1/2} \sqrt{\frac{2\log(4L^2/\delta)}{n_0}} = \epsilon_G.$$

Similarly, a union bound over the L intervention feature coordinates gives, with failure probability at most $\delta/2$,

$$|(\widehat{\mathbf{b}} - \mathbf{b})_{\boldsymbol{\alpha}}| \le B_{\boldsymbol{\alpha}} \sqrt{\frac{2\log(4L/\delta)}{n_e}}.$$

Taking the Euclidean norm and using $\sum_{\boldsymbol{\alpha}} B_{\boldsymbol{\alpha}}^2 = (B_0)^2$ gives $\|\widehat{\mathbf{b}} - \mathbf{b}\|_2 \le \epsilon_b$. A union bound combines the factual and intervention events and permits shared samples across the two channels.

Weyl's inequality gives $\lambda_{\min}(\widehat{\mathbf{G}}) \ge \lambda_G - \epsilon_G > 0,\ \|\widehat{\mathbf{G}}^{-1}\|_{\mathrm{op}} \le (\lambda_G - \epsilon_G)^{-1}$. Moreover,

$$\|\mathbf{b}\|_2 \le \mathbb{E}_e \|\boldsymbol{\psi}(\widetilde{\mathbf{V}}_R)\|_2 \le B_0, \qquad \|\boldsymbol{\beta}^\circ\|_2 \le \|\mathbf{G}^{-1}\|_{\mathrm{op}} \|\mathbf{b}\|_2 \le B_\beta. \tag{149}$$

Subtracting the two normal equations yields the exact identity

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\circ = \widehat{\mathbf{G}}^{-1}\left[(\widehat{\mathbf{b}} - \mathbf{b}) + (\mathbf{G} - \widehat{\mathbf{G}})\boldsymbol{\beta}^\circ\right]. \tag{150}$$

Taking norms and using the preceding bounds gives

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\circ\|_2 &\le \|\widehat{\mathbf{G}}^{-1}\|_{\mathrm{op}} \left(\|\widehat{\mathbf{b}} - \mathbf{b}\|_2 + \|\widehat{\mathbf{G}} - \mathbf{G}\|_{\mathrm{op}} \|\boldsymbol{\beta}^\circ\|_2\right) \\ &\le \frac{\epsilon_b + \epsilon_G B_\beta}{\lambda_G - \epsilon_G} = \epsilon_\beta.\end{aligned}$$

Finally,

$$\widehat{r}_p - \widetilde{r} = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\circ)^\top \boldsymbol{\psi} + (r_p^\circ - \widetilde{r}).$$

Combining the feature bounds equation 143 with the projection tails gives equation 145 and equation 146. The same C^1 projection tail and equation 149 give equation 147. The positivity claim follows pointwise from the C^0 bound and the fixed floor c. □

Corollary L.12 (Ratio-to-log-ratio bridge)**.** *On the event of either Theorem L.9 or Theorem L.11, let r denote the corresponding population or estimated-chart target ratio and $\widehat{r}_p$ its fitted polynomial. Suppose $(\epsilon_{r,0}, \epsilon_{r,1}, c_r, G_r)$ denotes the value-error, C^1-error, density-floor, and target-gradient envelopes from the applicable theorem. In the two-sample case, these are $(\epsilon_{G,r,0}^{R,e,p}, \epsilon_{G,r,1}^{R,e,p}, c_{G,r}^{R,e}, G_{G,r}^{R,e,p})$. Suppose $\epsilon_{r,0} < c_r$ and set $c_* = c_r - \epsilon_{r,0}$. Then*

$$\|\log \widehat{r}_p - \log r\|_\infty \le \frac{\epsilon_{r,0}}{c_*},$$

$$\|\nabla \log \widehat{r}_p - \nabla \log r\|_{2,\infty} \le \frac{\epsilon_{r,1}}{c_*} + \frac{G_r \epsilon_{r,0}}{c_r c_*}.$$

Consequently $\|\log \widehat{r}_p - \log r\|_{C^1} \le \epsilon_\ell$, where $\epsilon_\ell := \max\left\{\frac{\epsilon_{r,0}}{c_*}, \frac{\epsilon_{r,1}}{c_*} + \frac{G_r \epsilon_{r,0}}{c_r c_*}\right\}$.

Proof. Every point of the line segment joining $r(u)$ and $\widehat{r}_p(u)$ is at least c_*. The mean-value theorem for $\log$ therefore gives $|\log \widehat{r}_p - \log r| \le |\widehat{r}_p - r|/c_*$. For the gradient, use the exact decomposition

$$\nabla \log \widehat{r}_p - \nabla \log r = \frac{\nabla \widehat{r}_p - \nabla r}{\widehat{r}_p} + \nabla r \left(\frac{1}{\widehat{r}_p} - \frac{1}{r}\right). \tag{151}$$

The first denominator is at least c_*. For the second term, $|1/\widehat{r}_p - 1/r| = |\widehat{r}_p - r|/(\widehat{r}_p r) \le \epsilon_{r,0}/(c_* c_r)$. Combining these two pointwise bounds and taking the supremum proves the result. □

Corollary L.13 (Analytic quadrature remainder for the fitted sieve)**.** *For any conditioned coordinate j, if $2n_{\mathrm{GL}} - 1 \ge p$, order-$n_{\mathrm{GL}}$ Gauss–Legendre quadrature integrates the fitted denominator, numerator, and their non-j derivatives exactly in exact arithmetic. Hence the analytic quadrature remainder for the fitted polynomial is zero. Interval bounds control floating-point roundoff separately.*

Proof. Holding the other coordinates fixed, every term in $\widehat{r}_p$ and each of its non-j derivatives has degree at most p in u_j. The numerator interval $[0, u_j]$ is an affine image of $[-1, 1]$, which preserves polynomial degree. Gauss–Legendre quadrature of order n_{GL} is exact for degree at most $2n_{\mathrm{GL}} - 1$, proving the claim. □

The Legendre bounds separate ratio approximation from sampling error. Corollary L.13 removes analytic quadrature error for the fitted polynomial. Proposition L.26 later makes the two-sample bounds simultaneous across active sets and environments; the calibration and peeling results then propagate them to graph recovery. For finite-precision inputs, outward interval arithmetic additionally bounds errors in CDFs, densities, scores, and the accepted calibration count.

L.4 Chart-error propagation and conditional-CDF approximation

The preceding fit estimates a ratio in learned coordinates; the population sink test uses the true residual coordinates. To connect them, view chart error as a small movement of points, called a transport. We bound its effect on active density ratios, then account for the tails omitted by the computed CDF. Density and Jacobian singular-value lower bounds keep normalization and inversion stable. Sink and edge margins ensure that the population decisions remain distinguishable after these estimation and approximation errors are added.

Assumption L.14 (Smooth ratio and normalizer class)**.** For every true active set R and environment e, r_e^R extends to $[0, 1]^R$ with

$$0 < c_r \le r_e^R \le C_r, \qquad \left\|\log r_e^R\right\|_{C^2} \le B_r.$$

Every conditional normalizer and conditional density used by CFP is bounded between fixed positive constants. More explicitly, there are fixed $n_-, n_+, k_-, k_+, G_0 > 0$ such that, for every admissible (j, e, R),

$$n_- \le \mathcal{N}_{j,e}^R \le n_+, \qquad k_- \le K_{j,e}^R \le k_+, \qquad \|\nabla C_{j,e}^R\|_{2,\infty} \le G_0.$$

These bounds ensure that the conditional CDFs and their gradient-based scores are well defined on every fibre.

Let $\Omega_Y := [\eta, 1 - \eta]^d$ and let π_Y denote its normalized Lebesgue probability measure. Let $\mathcal{Y}_\star$ be the family of population coordinate maps obtained from correct chart alignment, correct peeling, the fixed tie breaking, and the allowed choices of targeting environment.

Assumption L.15 (Observable sink and graph margins)**.** For the population scores equation 134–equation 136, every true sink has $s_j^R = 0$ and every non-sink obeys $s_j^R \ge \Delta_{\mathrm{sink}} > 0$.

For every $Y \in \mathcal{Y}_\star$, assume

$$\inf_u \sigma_{\min}(\mathrm{D}Y(u)) \ge \mu_Y > 0, \qquad \sup_u \|\mathrm{D}Y(u)\|_{\mathrm{op}} \le L_Y,$$

and define the uniform root-mean-square edge margin on the common inner domain

$$\Delta_{\mathrm{edge}}^{\pi} := \inf_{Y \in \mathcal{Y}_\star} \min_{j \to i} \left(\mathbb{E}_{U \sim \pi_Y}[\{\mathrm{D}H_Y(Y(\mathbf{U}))\}_{ij}^2]\right)^{1/2} > 0, \qquad H_Y = Y^{-1}.$$

For these nonnegative scores, an empty maximum equals zero and an empty edge-margin minimum equals $+\infty$. Each non-sink therefore requires at least one population sink test supplied by the available target signatures.

Assumption L.16 (Uniform residual-chart stability)**.** After signed-permutation alignment, chart error reaches an active marginal in one of two ways. Under a block-local transport, the retained coordinates depend only on their true values. Under a coupled transport, they can also depend on the removed coordinates, so the marginalization must be included. Proposition L.19 controls the first case, and Proposition L.20 controls the second through the fibre integral. Write $k(R) = 2$ in the block-local case and $k(R) = 3$ in the coupled case. The corresponding smoothness, small-error, and density-floor conditions hold uniformly, and

$$\left\|\log \widetilde{r}_e^R - \log r_e^R\right\|_{C^2} \le C_{\mathrm{leak}}(R, k(R))\epsilon_q(k(R)).$$

A common invertible change of coordinates contributes the same Jacobian factor to both densities, so it cancels from their ratio. This cancellation also holds for an active marginal when its coordinates depend only on the retained true coordinates. If they depend on removed coordinates as well, the Jacobian remains inside two differently weighted marginal integrals and need not cancel. We then bound each marginal density before taking their ratio. Controlling those Jacobian derivatives requires one additional derivative in the coupled case.

Here $\epsilon_q(k)$ collects chart errors through derivative order k, and $C_{\mathrm{leak}}(R,k)$ converts them into log-ratio error on R. The next results derive these quantities for the two cases.

Let ϕ_{N} denote the standard normal density. Its first four derivative envelopes will be denoted by $c_0,\dots,c_3$:

$$c_0 := \sup_{z\in\mathbb{R}} |\phi_{\mathrm{N}}(z)| = \frac{1}{\sqrt{2\pi}}, \qquad c_1 := \sup_{z\in\mathbb{R}} |\phi'_{\mathrm{N}}(z)| = \frac{1}{\sqrt{2\pi \mathrm{e}}},$$
$$c_2 := \sup_{z\in\mathbb{R}} |\phi''_{\mathrm{N}}(z)| = c_0, \qquad c_3 := 3c_1 + 3\sqrt{3}\,\exp(-3/2)c_0 \geq \sup_{z\in\mathbb{R}} |\phi'''_{\mathrm{N}}(z)|. \tag{152}$$

For a scalar-, vector-, or tensor-valued map h on a domain $\mathcal{D}$, write

$$\|h\|_{2,\infty;\mathcal{D}} := \sup_{u\in\mathcal{D}} \|h(u)\|_2, \qquad \|\mathrm{D}^r h\|_{\mathrm{op},\infty;\mathcal{D}} := \sup_{u\in\mathcal{D}} \|\mathrm{D}^r h(u)\|_{\mathrm{op}},$$

and

$$\|h\|_{C^k(\mathcal{D})} := \max_{0\leq r\leq k} \sup_{u\in\mathcal{D}} \|\mathrm{D}^r h(u)\|_{\mathrm{op}}, \qquad \mathrm{D}^0 h(u) := h(u),$$

where the norm for a vector is Euclidean and the norm for a multilinear map is its induced operator norm. When the domain is the active trimmed cube, the domain argument is suppressed. In particular, the C^2 bounds below are with respect to the observed coordinate x in Lemma L.17, and with respect to the residual coordinate u_R in Proposition L.19.

Lemma L.17 (Trimmed probit chart perturbation)**.** *Let $\eta \in (0,1/2)$ and put $B_\eta := |\Phi^{-1}(\eta)|$. After the signed permutation alignment of Proposition L.3, suppose*

$$\|\widehat{\mathbf{L}} - \mathbf{L}\|_{\mathrm{op}} \leq \epsilon_L, \qquad \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \leq \epsilon_\mu, \qquad \|\mathbf{L}\|_{\mathrm{op}} \leq L_{\max}, \qquad \|\mathbf{A}_{\mathrm{mix}}\|_{\mathrm{op}} \leq A_{\max}.$$

For $u \in [\eta,1-\eta]^d$, define $x(u) = \boldsymbol{\mu} + \mathbf{A}_{\mathrm{mix}}\Phi^{-1}(u)$, $q(x(u)) := \Phi(\mathbf{L}(x(u)-\boldsymbol{\mu})) = u$, and $\widehat{q}(x(u)) := \Phi(\widehat{\mathbf{L}}(x(u)-\widehat{\boldsymbol{\mu}}))$. Then, with

$$E_{\mathrm{lin}} := \epsilon_L A_{\max}\sqrt{d}\,B_\eta + (L_{\max}+\epsilon_L)\epsilon_\mu, \qquad L^+_{\max} := L_{\max} + \epsilon_L,$$

one has

$$\|\widehat{q}(x(u)) - q(x(u))\|_{2,\infty;\,u\in[\eta,1-\eta]^d} \leq c_0 E_{\mathrm{lin}} =: \epsilon_{q,0}, \tag{153}$$
$$\|\mathrm{D}_x\widehat{q} - \mathrm{D}_x q\|_{\mathrm{op},\,x([\eta,1-\eta]^d)} \leq c_1 E_{\mathrm{lin}} L^+_{\max} + c_0\epsilon_L =: \epsilon_{q,1}, \tag{154}$$
$$\|\mathrm{D}^2_x\widehat{q} - \mathrm{D}^2_x q\|_{\mathrm{op},\,x([\eta,1-\eta]^d)} \leq \sqrt{d}\,\left[c_2 E_{\mathrm{lin}}(L^+_{\max})^2 + c_1(2L_{\max}\epsilon_L + \epsilon_L^2)\right] =: \epsilon_{q,2}. \tag{155}$$
$$\|\mathrm{D}^3_x\widehat{q} - \mathrm{D}^3_x q\|_{\mathrm{op},\,x([\eta,1-\eta]^d)} \leq \sqrt{d}\,\left[c_3 E_{\mathrm{lin}}(L^+_{\max})^3 + c_2(3L^2_{\max}\epsilon_L + 3L_{\max}\epsilon_L^2 + \epsilon_L^3)\right] =: \epsilon_{q,3}. \tag{156}$$

Proof. For u in the trimmed cube, $\|\Phi^{-1}(u)\|_2 \leq \sqrt{d}\,B_\eta$, hence

$$\|\widehat{\mathbf{L}}(x(u)-\widehat{\boldsymbol{\mu}}) - \mathbf{L}(x(u)-\boldsymbol{\mu})\|_2 \leq \epsilon_L A_{\max}\sqrt{d}\,B_\eta + L^+_{\max}\epsilon_\mu = E_{\mathrm{lin}}.$$

The scalar mean-value theorem and the first envelope in equation 152 prove equation 153. The Jacobian identity is

$$\mathrm{D}_x q = \mathrm{diag}(\phi_{\mathrm{N}}(\mathbf{L}(x-\boldsymbol{\mu})))\mathbf{L}, \qquad \mathrm{D}_x\widehat{q} = \mathrm{diag}(\phi_{\mathrm{N}}(\widehat{\mathbf{L}}(x-\widehat{\boldsymbol{\mu}})))\widehat{\mathbf{L}}.$$

Add and subtract the mixed product with the true diagonal factor. Its first term is at most $c_1 E_{\mathrm{lin}} L^+_{\max}$ and its second at most $c_0\epsilon_L$, proving equation 154. Finally, each row Hessian is $\phi'_{\mathrm{N}}(z_i)\boldsymbol{\ell}_i\boldsymbol{\ell}_i^\top$, where $\boldsymbol{\ell}_i^\top$ and $\widehat{\boldsymbol{\ell}}_i^\top$ are the ith rows of $\mathbf{L}$ and $\widehat{\mathbf{L}}$, respectively. The product difference is bounded by the two

terms displayed in equation 155; summing the d row norms in ℓ_2 gives the factor $\sqrt{d}$. The third derivative of the ith coordinate is $\phi_{\mathrm{N}}''(z_i)\boldsymbol{\ell}_i^{\otimes 3}$. The exact telescoping bound

$$\|\widehat{\boldsymbol{\ell}}_i^{\otimes 3} - \boldsymbol{\ell}_i^{\otimes 3}\|_{\mathrm{op}} \le 3L_{\max}^2\epsilon_L + 3L_{\max}\epsilon_L^2 + \epsilon_L^3$$

together with the c_3 Lipschitz envelope for ϕ_{N}'' proves equation 156 after the same Euclidean row aggregation. □

Corollary L.18 (Observed-coordinate to residual-transport conversion)**.** *Fix $\eta \in (0, 1/2)$ and apply Lemma L.17 with trim level $\eta/2$. Denote the four resulting right sides by $\epsilon_{q,0}^+, \dots, \epsilon_{q,3}^+$; the superscript $+$ records the wider cube. For this fixed trim level, set*

$$\phi_{\min} := \phi_{\mathrm{N}}(B_{\eta/2}), \quad \kappa_1 := \frac{A_{\max}}{\phi_{\min}}, \quad \kappa_2 := \frac{A_{\max}B_{\eta/2}}{\phi_{\min}^2}, \quad \kappa_3 := \frac{A_{\max}(1+2B_{\eta/2}^2)}{\phi_{\min}^3}.$$

For $\mathsf{T}(u) := \widehat{q}(x(u))$ on $[\eta/2, 1-\eta/2]^d$, one has

$$\|\mathsf{T} - \mathrm{id}\,\|_{2,\infty} \le \epsilon_{q,0}^+, \tag{157}$$

$$\|\mathrm{D}\mathsf{T} - \mathbf{I}_d\|_{\mathrm{op},\infty} \le \epsilon_{q,1}^+\kappa_1, \tag{158}$$

$$\|\mathrm{D}^2\mathsf{T}\|_{\mathrm{op},\infty} \le \epsilon_{q,2}^+\kappa_1^2 + \epsilon_{q,1}^+\kappa_2. \tag{159}$$

$$\|\mathrm{D}^3\mathsf{T}\|_{\mathrm{op},\infty} \le \epsilon_{q,3}^+\kappa_1^3 + 3\epsilon_{q,2}^+\kappa_1\kappa_2 + \epsilon_{q,1}^+\kappa_3. \tag{160}$$

If the R-coordinates of T are block-local, the same bounds hold for their restriction T_R. Write a, b, h, k for bounds on its value error and first three derivative errors, respectively. The two derivative orders needed below can be summarized by

$$\epsilon_q(2) := \max_{0\le k\le 2}\epsilon_{q,k}^+, \qquad C_{\mathrm{tr}}(2) := \max\left\{1, \kappa_1, \kappa_1^2 + \kappa_2\right\},$$

$$\epsilon_q(3) := \max_{0\le k\le 3}\epsilon_{q,k}^+, \qquad C_{\mathrm{tr}}(3) := \max\left\{C_{\mathrm{tr}}(2), \kappa_1^3 + 3\kappa_1\kappa_2 + \kappa_3\right\}.$$

Here the trim level remains fixed, so it is omitted from C_{tr}. For any active coordinate restriction, these bounds give

$$\max\{a, b, h\} \le C_{\mathrm{tr}}(2)\epsilon_q(2), \qquad \max\{a, b, h, k\} \le C_{\mathrm{tr}}(3)\epsilon_q(3),$$

The block-local proposition uses the first inequality, whereas the coupled proposition uses the second wherever the bounds hold uniformly along its fibres. Thus each result uses only the derivative order it needs.

Proof. For $z_i(u) = \Phi^{-1}(u_i)$,

$$z_i'(u_i) = \frac{1}{\phi_{\mathrm{N}}(z_i(u_i))}, \qquad z_i''(u_i) = \frac{z_i(u_i)}{\phi_{\mathrm{N}}(z_i(u_i))^2}, \qquad z_i'''(u_i) = \frac{1 + 2z_i(u_i)^2}{\phi_{\mathrm{N}}(z_i(u_i))^3}.$$

On the wider trimmed cube, $|z_i(u_i)| \le B_{\eta/2}$ and $\phi_{\mathrm{N}}(z_i(u_i)) \ge \phi_{\min}$. Hence the decoder $x(u) = \boldsymbol{\mu} + \mathbf{A}_{\mathrm{mix}}z(u)$ obeys

$$\|\mathrm{D}_u x(u)\|_{\mathrm{op}} \le \kappa_1, \qquad \|\mathrm{D}_u^2 x(u)\|_{\mathrm{op}} \le \kappa_2, \qquad \|\mathrm{D}_u^3 x(u)\|_{\mathrm{op}} \le \kappa_3. \tag{161}$$

The value bound is Lemma L.17. Since $q \circ x = \mathrm{id}$, the first-order chain rule yields the exact identity $\mathrm{D}\mathsf{T} - \mathbf{I}_d = (\mathrm{D}_x\widehat{q} - \mathrm{D}_x q)\mathrm{D}_u x$, which proves equation 158. Subtracting the second-order chain rules for $\widehat{q} \circ x$ and $q \circ x$ gives

$$\mathrm{D}^2\mathsf{T} = (\mathrm{D}_x^2\widehat{q} - \mathrm{D}_x^2 q)[\mathrm{D}_u x, \mathrm{D}_u x] + (\mathrm{D}_x\widehat{q} - \mathrm{D}_x q)\mathrm{D}_u^2 x.$$

Taking the induced multilinear norm and substituting equation 161 proves equation 159. At third order, subtracting the two Faà di Bruno identities gives

$$\mathrm{D}^3\mathsf{T} = (\mathrm{D}_x^3\widehat{q} - \mathrm{D}_x^3 q)[\mathrm{D}_u x, \mathrm{D}_u x, \mathrm{D}_u x] + \sum_{\mathrm{cyc}}(\mathrm{D}_x^2\widehat{q} - \mathrm{D}_x^2 q)[\mathrm{D}_u^2 x, \mathrm{D}_u x] + (\mathrm{D}_x\widehat{q} - \mathrm{D}_x q)\mathrm{D}_u^3 x,$$

where the cyclic sum contains the three placements of the second derivative. The triangle inequality gives equation 160. A coordinate restriction weakly decreases each Euclidean induced norm, and the last claim follows by taking the largest coefficient multiplying $\epsilon_q(2)$ or $\epsilon_q(3)$ at the corresponding derivative order. □

In the two transport results below, we fix the active set R and an environment e, and write $r = r_e^R$ and $\widetilde{r} = \widetilde{r}_e^R$. Local transport bounds omit the fixed R index. Constants used across active sets, such as $C_{\mathrm{leak}}(R, k)$, retain their arguments. When a bound is required uniformly over environments, take the maximum of its constants over the admissible choices of e.

Proposition L.19 (Block-local residual-chart error)**.** *Fix an active set R. Put*

$$\eta \in (0, 1/2), \qquad \mathcal{U}_R = [\eta, 1-\eta]^{d_R}, \qquad \mathcal{U}_R^+ = [\eta/2, 1-\eta/2]^{d_R}.$$

Assume that the estimated active coordinates depend only on the true active coordinates. More precisely, let T be a C^2 map on an open neighborhood of $\mathcal{U}_R^+$. The factual and active interventional marginals are transformed by this same map, so on their support,

$$\widehat{q}_R(x(u_R, u_{R^c})) = \mathsf{T}(u_R) \quad \text{for all } u_R \in \mathcal{U}_R^+ \text{ and admissible } u_{R^c}.$$

Measure its displacement and derivative errors on $\mathcal{U}_R^+$ by

$$\begin{aligned} \|\mathsf{T} - \mathrm{id}\,\|_{2,\infty} \le a, \qquad \|\mathrm{D}\mathsf{T} - \mathbf{I}_{d_R}\|_{\mathrm{op},\infty} \le b < 1, \\ \|\mathrm{D}^2\mathsf{T}\|_{\mathrm{op},\infty} \le h, \qquad\qquad a \le \eta/2. \end{aligned} \tag{162}$$

For every active environment ratio, assume

$$\|\log r\|_{C^3(\mathcal{U}_R^+)} \le B_3.$$

The common Jacobian cancels from the ratio, giving

$$\widetilde{r}(u) = r(\mathsf{T}^{-1}(u)). \tag{163}$$

Write $s = (1-b)^{-1}$, $\ell = \log r$, $B_1 = \|\ell\|_{C^1(\mathcal{U}_R^+)}$, and $B_2 = \|\ell\|_{C^2(\mathcal{U}_R^+)}$. The inverse derivative factor and the resulting value, gradient, and Hessian error bounds are

$$\begin{aligned} L_0 &:= B_1 a s, \\ L_1 &:= B_1 b s + B_2 a s, \\ L_2 &:= B_2(s^2 - 1) + B_3 a s^3 + B_1 h s^3. \end{aligned} \tag{164}$$

Then

$$\|\log \widetilde{r} - \log r\|_{C^2} \le \max\{L_0, L_1, L_2\}.$$

In particular, if $a, b, h \le C_{\mathrm{tr}}(2)\epsilon_q(2)$ and $\epsilon_q(2) \le (2C_{\mathrm{tr}}(2))^{-1}$, the right side is at most $C_{\mathrm{leak}}(R, 2)\epsilon_q(2)$ with

$$C_{\mathrm{leak}}(R, 2) := (8B_1 + 6B_2 + 8B_3)C_{\mathrm{tr}}(2). \tag{165}$$

Proof. For the affine–probit chart in Lemma L.17, block-locality removes the cross-coordinate linear block in probit coordinates. The bound $b < 1$ makes the active block nonsingular, so the active transport is injective on $(0,1)^{d_R}$. We now show that its inverse maps $\mathcal{U}_R$ into $\mathcal{U}_R^+$.

Write $\delta = \mathsf{T} - \mathrm{id}$. For fixed $u \in \mathcal{U}_R$, consider $\Psi_u(v) := u - \delta(v),\ v \in \mathcal{U}_R^+$. Each coordinate of u is at least η from the boundary of the full unit cube, while $\|\delta(v)\|_2 \le a \le \eta/2$ implies the same coordinatewise bound. Thus $\Psi_u(\mathcal{U}_R^+) \subseteq \mathcal{U}_R^+$. Moreover, the mean-value theorem and equation 162 give $\|\Psi_u(v) - \Psi_u(w)\|_2 \le b\|v - w\|_2$.

Banach's fixed-point theorem gives a unique $v \in \mathcal{U}_R^+$ with $v = \Psi_u(v)$, equivalently $\mathsf{T}(v) = u$. The same derivative bound also gives $\|\mathsf{T}(v) - \mathsf{T}(w)\|_2 \ge (1-b)\|v - w\|_2$, hence injectivity on $\mathcal{U}_R^+$. Therefore $\mathcal{U}_R \subseteq \mathsf{T}(\mathcal{U}_R^+)$ and the inverse is single valued there.

Let $p_{0,R}$ and $p_{e,R}$ be the true factual and interventional marginal densities, so $r = p_{e,R}/p_{0,R}$. The change-of-variables formula gives, with $v = \mathsf{T}^{-1}(u)$,

$$\widetilde{p}_{c,R}(u) = p_{c,R}(v)\left|\det \mathrm{D}\mathsf{T}^{-1}(u)\right|, \qquad c \in \{0, e\}.$$

The common inverse Jacobian determinant cancels in $\widetilde{p}_{e,R}/\widetilde{p}_{0,R}$, proving equation 163. Finally, the inverse function theorem and the Neumann-series bound imply

$$\|\mathrm{D}\mathsf{T}^{-1}(u)\|_{\mathrm{op}} \le (1-b)^{-1}, \qquad \|\mathsf{T}^{-1}(u) - u\|_2 \le a \le as.$$

The value bound is the mean-value theorem. For the gradient, use $\nabla_u(\ell \circ \mathsf{T}^{-1})(u) = (\mathrm{D}\mathsf{T}^{-1}(u))^\top \nabla \ell(v)$ and $\|(\mathrm{D}\mathsf{T}^{-1}(u))^\top - \mathbf{I}_{d_R}\|_{\mathrm{op}} \le b/(1-b)$, then compare $\nabla\ell(v)$ and $\nabla\ell(u)$ using B_2. This gives L_1.

For the Hessian, put $\mathbf{J}_v = \mathrm{D}\mathsf{T}^{-1}(u)$ and $\mathbf{H}_w = \mathrm{D}^2\ell(w)$. The second-order chain rule gives

$$\mathrm{D}^2(\ell \circ \mathsf{T}^{-1})(u) = \mathbf{J}_v^\top \mathbf{H}_v \mathbf{J}_v + \nabla\ell(v) \lrcorner \mathrm{D}^2\mathsf{T}^{-1}(u), \tag{166}$$

where $\lrcorner$ denotes contraction of the output index of the inverse-map Hessian. The first difference splits exactly as

$$\mathbf{J}_v^\top \mathbf{H}_v \mathbf{J}_v - \mathbf{H}_u = \mathbf{J}_v^\top (\mathbf{H}_v - \mathbf{H}_u)\mathbf{J}_v + (\mathbf{J}_v^\top \mathbf{H}_u \mathbf{J}_v - \mathbf{H}_u),$$

and therefore has operator norm at most

$$s^2 B_3 \|v-u\|_2 + B_2(s^2-1) \le B_3 a s^3 + B_2(s^2-1).$$

Differentiating the identity $\mathsf{T}^{-1} \circ \mathsf{T} = \mathrm{id}$ twice gives $\|\mathrm{D}^2\mathsf{T}^{-1}(u)\|_{\mathrm{op}} \le h s^3$, so the contraction term in equation 166 is at most $B_1 h s^3$. Combining the two bounds proves equation 164. Finally, $b \le 1/2$ implies $s \le 2$, $s^2 - 1 \le 6b$, and $as \le 2a$; substituting $a, b, h \le C_{\mathrm{tr}}(2)\epsilon_q(2)$ yields equation 165. The Hessian coefficient $8B_1 + 6B_2 + 8B_3$ also bounds the value and gradient coefficients. □

Proposition L.20 (Coupled residual-chart error along fibres)**.** *Fix an active set R, put $S = R^{\mathrm{c}}$, and write $d_R = |R|$ and $d_S = |S|$. Let*

$$\mathcal{D}_R = [0,1]^{d_R}, \quad \mathcal{U}_R = [\eta, 1-\eta]^{d_R}, \quad \mathcal{U}_R^+ = [\eta/2, 1-\eta/2]^{d_R}, \quad \mathcal{U}_S = [0,1]^{d_S}, \quad \eta \in (0, 1/2).$$

For $c \in \{0, e\}$, assume the true residual law has a joint density $p_c(x,z)$ on the common full product support $\mathcal{D}_R \times \mathcal{U}_S$, where $x = u_R$ and $z = u_S$. The same estimated chart acts on the factual and interventional samples. For each fixed removed-coordinate value z, its active component defines a map of the retained coordinates $\mathsf{g}_z(x) := \mathrm{pr}_R\big(\mathsf{T}(x,z)\big)$.

Here pr_R denotes projection onto the coordinates in R. Suppose the following derivative bounds hold uniformly over the source neighborhood $(x,z) \in \mathcal{U}_R^+ \times \mathcal{U}_S$:

$$\|\mathsf{g}_z(x) - x\|_2 \le a \le \eta/2, \qquad \|\mathrm{D}_x \mathsf{g}_z(x) - \mathbf{I}_{d_R}\|_{\mathrm{op}} \le b < 1, \tag{167}$$

$$\|\mathrm{D}_x^2 \mathsf{g}_z(x)\|_{\mathrm{op}} \le h, \qquad \|\mathrm{D}_x^3 \mathsf{g}_z(x)\|_{\mathrm{op}} \le k.$$

Assume the no-boundary-influx condition

$$\mathsf{g}_z^{-1}(\mathcal{U}_R) \cap \mathcal{D}_R \subseteq \mathcal{U}_R^+, \qquad z \in \mathcal{U}_S. \tag{168}$$

A stronger sufficient condition extends the Jacobian bound in equation 167 to the full convex support $\mathcal{D}_R$; the lower-Lipschitz argument then makes each fibre map globally injective. For $j = 0, 1, 2, 3$, define the active-direction joint-density envelopes

$$M_{c,j} := \int_{\mathcal{U}_S} \sup_{x \in \mathcal{U}_R^+} \|\mathrm{D}_x^j p_c(x,z)\|_{\mathrm{op}} \, \mathrm{d}z < \infty, \qquad c \in \{0, e\}.$$

These envelopes integrate over the removed coordinates and control derivatives only in the active coordinates; $\mathrm{D}_x^0 p_c = p_c$. Their integrability allows differentiation under the integral. Assume the true active marginals

$$p_{c,R}(x) := \int_{\mathcal{U}_S} p_c(x,z) \, \mathrm{d}z \quad \textit{satisfy} \quad \inf_{x \in \mathcal{U}_R} p_{c,R}(x) \ge m_c > 0.$$

The inverse map stretches directions by at most $\kappa = (1-b)^{-1}$. The following constants bound its Jacobian determinant, its deviation from one, and its first two derivatives:

$$J_1 := d_R h \kappa^{d_R+2}, \qquad J_0 := \kappa^{d_R}, \qquad J_2 := d_R k \kappa^{d_R+3} + (d_R^2 + 2d_R) h^2 \kappa^{d_R+4},$$
$$\kappa := (1-b)^{-1}, \qquad \delta_J := J_0 - 1.$$

For each $c \in \{0, e\}$, the resulting density errors at orders zero, one, and two are bounded by

$$\Delta_{c,0} := a\kappa^{d_R} M_{c,1} + \delta_J M_{c,0}, \tag{169}$$

$$\Delta_{c,1} := a\kappa^{d_R+1} M_{c,2} + \left(\delta_J + b\kappa^{d_R+1}\right) M_{c,1} + J_1 M_{c,0}, \tag{170}$$

$$\Delta_{c,2} := a\kappa^{d_R+2} M_{c,3} + \left[\delta_J + b\kappa^{d_R+1}(\kappa+1)\right] M_{c,2} + \left[h\kappa^{d_R+3} + 2\kappa J_1\right] M_{c,1} + J_2 M_{c,0}. \tag{171}$$

Then the estimated-chart active marginal $\widetilde{p}_{c,R}$ obeys

$$\|\mathrm{D}^j \widetilde{p}_{c,R} - \mathrm{D}^j p_{c,R}\|_{\mathrm{op},\infty;\mathcal{U}_R} \le \Delta_{c,j}, \qquad j = 0, 1, 2. \tag{172}$$

If $\Delta_{c,0} < m_c$ for both $c = 0, e$, the estimated marginal remains positive, with floor $m_c^* = m_c - \Delta_{c,0}$. The corresponding log-density error bounds are

$$\Lambda_{c,0} := \frac{\Delta_{c,0}}{m_c^*}, \tag{173}$$

$$\Lambda_{c,1} := \frac{\Delta_{c,1}}{m_c^*} + \frac{M_{c,1}\Delta_{c,0}}{m_c m_c^*}, \tag{174}$$

$$\Lambda_{c,2} := \frac{\Delta_{c,2}}{m_c^*} + \frac{M_{c,2}\Delta_{c,0}}{m_c m_c^*} + \frac{\Delta_{c,1}(2M_{c,1} + \Delta_{c,1})}{(m_c^*)^2} + \frac{M_{c,1}^2 \Delta_{c,0}}{m_c m_c^*}\left(\frac{1}{m_c^*} + \frac{1}{m_c}\right). \tag{175}$$

The log-ratio error is the sum of the two log-density errors:

$$\|\log \widetilde{r} - \log r\|_{C^2(\mathcal{U}_R)} \le \max_{j=0,1,2} \{\Lambda_{e,j} + \Lambda_{0,j}\}. \tag{176}$$

Proof. We first construct each inverse fibre map, then bound its Jacobian and the resulting marginal density. Fix $z \in \mathcal{U}_S$. For $y \in \mathcal{U}_R$, the map $x \mapsto y - (\mathsf{g}_z(x) - x)$ sends $\mathcal{U}_R^+$ into itself because $a \le \eta/2$, and it is a contraction with constant b. Banach's fixed-point theorem therefore produces a unique $x = \mathsf{G}_z(y) \in \mathcal{U}_R^+$ satisfying $\mathsf{g}_z(\mathsf{G}_z(y)) = y$. Moreover, for $x, x' \in \mathcal{U}_R^+$, integrating the Jacobian along their line segment gives $\|\mathsf{g}_z(x) - \mathsf{g}_z(x')\|_2 \ge (1-b)\|x - x'\|_2$, so the inverse inside $\mathcal{U}_R^+$ is single valued. Condition equation 168 rules out preimages outside this cube. Thus every target point has a unique support preimage. If the first-order bound extends to $\mathcal{D}_R$, the same line-segment calculation proves equation 168 directly.

Write $\mathbf{K}_z(y) := \mathrm{D}_y \mathsf{G}_z(y),\ J_z(y) := \det \mathbf{K}_z(y)$.

The Neumann-series bound, followed by the inverse identity, yields

$$\|\mathbf{K}_z\|_{\mathrm{op}} \le \kappa, \qquad \|\mathbf{K}_z - \mathbf{I}_{d_R}\|_{\mathrm{op}} \le b\kappa, \qquad \|\mathsf{G}_z(y) - y\|_2 \le a.$$

Differentiating $\mathsf{g}_z \circ \mathsf{G}_z = \mathrm{id}$ twice gives, for directions $v, w \in \mathbb{R}^{d_R}$,

$$\mathrm{D}_y^2 \mathsf{G}_z(y)[v, w] = -\mathbf{K}_z(y)\mathrm{D}_x^2 \mathsf{g}_z(\mathsf{G}_z(y))[\mathbf{K}_z(y)v, \mathbf{K}_z(y)w],$$

and hence

$$\|\mathrm{D}_y^2 \mathsf{G}_z\|_{\mathrm{op}} \le h\kappa^3. \tag{177}$$

To control the determinant derivatives, define $F_z(x) := \log\det \mathrm{D}_x \mathsf{g}_z(x)$.

The determinant is positive: the path $\mathbf{I}_{d_R} + t(\mathrm{D}_x \mathsf{g}_z - \mathbf{I}_{d_R})$ is nonsingular for $t \in [0, 1]$ and begins at the identity. Jacobi's formula gives

$$\mathrm{D}F_z(x)[v] = \mathrm{tr}\left[(\mathrm{D}_x \mathsf{g}_z(x))^{-1} \mathrm{D}_x^2 \mathsf{g}_z(x)[v, \cdot]\right], \tag{178}$$

so $\|\mathrm{D}F_z\|_{\mathrm{op}} \le d_R h\kappa$. Differentiating equation 178 and using $\mathrm{D}(\mathbf{A}^{-1})[\mathbf{H}] = -\mathbf{A}^{-1}\mathbf{H}\mathbf{A}^{-1}$ gives $\|\mathrm{D}^2 F_z\|_{\mathrm{op}} \le d_R(k\kappa + h^2\kappa^2)$. Since $\log J_z = -F_z \circ \mathsf{G}_z$, the first- and second-order chain rules, together with equation 177, yield

$$\|\mathrm{D}\log J_z\|_{\mathrm{op}} \le d_R h\kappa^2,$$
$$\|\mathrm{D}^2 \log J_z\|_{\mathrm{op}} \le d_R k\kappa^3 + 2d_R h^2\kappa^4.$$

The singular-value bounds imply $0 < J_z \le \kappa^{d_R}$ and $|J_z - 1| \le \delta_J$. Finally,

$$\mathrm{D}J_z = J_z \mathrm{D}\log J_z,$$
$$\mathrm{D}^2 J_z = J_z \left[\mathrm{D}^2 \log J_z + \mathrm{D}\log J_z \otimes \mathrm{D}\log J_z\right]$$

give $\|\mathrm{D}J_z\|_{\mathrm{op}} \le J_1,\ \|\mathrm{D}^2 J_z\|_{\mathrm{op}} \le J_2$.

Changing variables separately inside every fixed-z fibre gives the exact estimated-chart marginal density

$$\widetilde{p}_{c,R}(y) = \int_{\mathcal{U}_S} p_c(\mathsf{G}_z(y), z) J_z(y)\ \mathrm{d}z.$$

Unlike the block-local ratio, this ratio generally retains the Jacobian because $J_z(y)$ varies along the fibre. For the value difference, the identity

$$p_c(\mathsf{G}_z, z)J_z - p_c(y,z) = J_z[p_c(\mathsf{G}_z, z) - p_c(y,z)] + (J_z - 1)p_c(y,z)$$

and the mean-value theorem give equation 169 after integration.

For the first derivative,

$$\mathrm{D}_y[p_c(\mathsf{G}_z(y), z)J_z(y)] = J_z \mathbf{K}_z^\top \nabla_x p_c(\mathsf{G}_z, z) + p_c(\mathsf{G}_z, z)\mathrm{D}J_z. \tag{179}$$

Subtract $\nabla_x p_c(y,z)$ and split the first term as

$$J_z \mathbf{K}_z^\top [\nabla_x p_c(\mathsf{G}_z, z) - \nabla_x p_c(y,z)] + (J_z - 1)\nabla_x p_c(y,z) + J_z(\mathbf{K}_z^\top - \mathbf{I}_{d_R})\nabla_x p_c(y,z).$$

The three displayed terms are bounded respectively by $a\kappa^{d_R+1}M_{c,2}$, $\delta_J M_{c,1}$, and $b\kappa^{d_R+1}M_{c,1}$ after integration; the final term in equation 179 contributes $J_1 M_{c,0}$. This proves equation 170.

For the second derivative, the product and chain rules give

$$\begin{aligned}\mathrm{D}_y^2[p_c(\mathsf{G}_z, z)J_z] = {} & J_z \mathbf{K}_z^\top \mathrm{D}_x^2 p_c(\mathsf{G}_z, z)\mathbf{K}_z \\ & + J_z \nabla_x p_c(\mathsf{G}_z, z) \lrcorner \mathrm{D}_y^2 \mathsf{G}_z \\ & + \mathrm{D}_y[p_c(\mathsf{G}_z, z)] \otimes \mathrm{D}J_z + \mathrm{D}J_z \otimes \mathrm{D}_y[p_c(\mathsf{G}_z, z)] \\ & + p_c(\mathsf{G}_z, z)\mathrm{D}^2 J_z.\end{aligned} \tag{180}$$

The pulled-back Hessian minus $\mathrm{D}_x^2 p_c(y,z)$ is split into

$$\begin{aligned}& J_z \mathbf{K}_z^\top [\mathrm{D}_x^2 p_c(\mathsf{G}_z, z) - \mathrm{D}_x^2 p_c(y,z)]\mathbf{K}_z + (J_z - 1)\mathrm{D}_x^2 p_c(y,z) \\ & \quad + J_z(\mathbf{K}_z^\top - \mathbf{I}_{d_R})\mathrm{D}_x^2 p_c(y,z)\mathbf{K}_z + J_z \mathrm{D}_x^2 p_c(y,z)(\mathbf{K}_z - \mathbf{I}_{d_R}).\end{aligned}$$

After integration, these four terms contribute $a\kappa^{d_R+2}M_{c,3} + \left[\delta_J + b\kappa^{d_R+1}(\kappa+1)\right] M_{c,2}$.

The inverse-Hessian term in equation 180 contributes $h\kappa^{d_R+3}M_{c,1}$, the two cross terms contribute $2\kappa J_1 M_{c,1}$, and the last determinant term contributes $J_2 M_{c,0}$. Adding these quantities proves equation 171 and hence equation 172.

It remains to convert the density bounds into log-density bounds. Let $f = p_{c,R}$, $\widetilde{f} = \widetilde{p}_{c,R}$, $\delta_j = \|\mathrm{D}^j \widetilde{f} - \mathrm{D}^j f\|_{\mathrm{op},\infty}$, and $m^* = m_c - \delta_0$. The mean-value theorem for log gives

$$\|\log \widetilde{f} - \log f\|_\infty \le \delta_0 / m^*.$$

For the gradient, use

$$\frac{\nabla \widetilde{f}}{\widetilde{f}} - \frac{\nabla f}{f} = \frac{\nabla \widetilde{f} - \nabla f}{\widetilde{f}} + \nabla f \left(\frac{1}{\widetilde{f}} - \frac{1}{f}\right),$$

which yields equation 174. For the Hessian, write

$$\mathrm{D}^2 \log f = \frac{\mathrm{D}^2 f}{f} - \frac{\nabla f \otimes \nabla f}{f^2}.$$

The first quotient difference is bounded by

$$\frac{\delta_2}{m^*} + \frac{M_{c,2}\delta_0}{m_c m^*}.$$

For the outer-product quotient, use

$$\|\nabla\widetilde{f}\otimes\nabla\widetilde{f}-\nabla f\otimes\nabla f\|_{\mathrm{op}}\le\delta_1(2M_{c,1}+\delta_1),$$
$$\left|\frac{1}{\widetilde{f}^2}-\frac{1}{f^2}\right|\le\frac{\delta_0}{m_c m^*}\left(\frac{1}{m^*}+\frac{1}{m_c}\right).$$

Substituting $\delta_j\le\Delta_{c,j}$ proves equation 173–equation 175. Finally,

$$\log\widetilde{r}-\log r=(\log\widetilde{p}_{e,R}-\log p_{e,R})-(\log\widetilde{p}_{0,R}-\log p_{0,R}),$$

and the triangle inequality proves equation 176. □

Corollary L.21 (Linear bound for coupled chart error)**.** *Under Proposition L.20, let ρ collect the four transport errors. The dimension-dependent constants below turn the preceding bounds into a linear estimate:*

$$\rho:=\max\{a,b,h,k\}\le\frac{1}{2},\qquad D_*:=d_R2^{d_R+1},$$

and

$$A_1:=d_R2^{d_R+2},\qquad A_2:=d_R2^{d_R+3}+(d_R^2+2d_R)2^{d_R+4}.$$

For $c\in\{0,e\}$ define

$$C_{c,0}:=2^{d_R}M_{c,1}+D_*M_{c,0},\tag{181}$$
$$C_{c,1}:=2^{d_R+1}M_{c,2}+(D_*+2^{d_R+1})M_{c,1}+A_1M_{c,0},\tag{182}$$
$$C_{c,2}:=2^{d_R+2}M_{c,3}+(D_*+3\cdot2^{d_R+1})M_{c,2}+(2^{d_R+3}+d_R2^{d_R+4})M_{c,1}+A_2M_{c,0}.\tag{183}$$

Then $\Delta_{c,j}\le C_{c,j}\rho$. If additionally $C_{c,0}\rho\le m_c/2$ for $c=0,e$, define

$$K_{c,0}:=\frac{2C_{c,0}}{m_c},$$
$$K_{c,1}:=\frac{2C_{c,1}}{m_c}+\frac{2M_{c,1}C_{c,0}}{m_c^2},$$
$$K_{c,2}:=\frac{2C_{c,2}}{m_c}+\frac{2M_{c,2}C_{c,0}}{m_c^2}+\frac{4C_{c,1}(2M_{c,1}+C_{c,1})}{m_c^2}+\frac{6M_{c,1}^2C_{c,0}}{m_c^3}.$$

Consequently,

$$\|\log\widetilde{r}-\log r\|_{C^2(\mathcal{U}_R)}\le\rho\max_{j=0,1,2}\{K_{e,j}+K_{0,j}\}.\tag{184}$$

When the bounds from Corollary L.18 hold uniformly over these fibres, $\rho\le C_{\mathrm{tr}}(3)\epsilon_q(3)$. If also

$$\epsilon_q(3)\le\frac{1}{2C_{\mathrm{tr}}(3)},\qquad C_{c,0}C_{\mathrm{tr}}(3)\epsilon_q(3)\le\frac{m_c}{2},\quad c\in\{0,e\},$$

then Assumption L.16 holds on this active set with

$$C_{\mathrm{leak}}(R,3):=C_{\mathrm{tr}}(3)\max_{j=0,1,2}\{K_{e,j}+K_{0,j}\},\qquad\|\log\widetilde{r}-\log r\|_{C^2}\le C_{\mathrm{leak}}(R,3)\epsilon_q(3).$$

Proof. Since $b\le1/2$, $\kappa\le2$. The scalar mean-value theorem applied to $(1-b)^{-d_R}$ on $[0,1/2]$ gives $\delta_J\le D_*b\le D_*\rho$. Moreover, $J_1\le A_1\rho$, $J_2\le A_2\rho$, where $h^2\le\rho^2\le\rho$ is used in the second inequality. Substituting these three inequalities and $\kappa\le2$ into equation 169–equation 171 gives equation 181–equation 183 term by term.

The floor condition implies $m_c^*\ge m_c/2$. In equation 173–equation 175, substitute $\Delta_{c,j}\le C_{c,j}\rho$ and use $\rho\le1$. The four second-order contributions are bounded respectively by

$$\frac{2C_{c,2}}{m_c}\rho,\quad\frac{2M_{c,2}C_{c,0}}{m_c^2}\rho,\quad\frac{4C_{c,1}(2M_{c,1}+C_{c,1})}{m_c^2}\rho,\quad\frac{6M_{c,1}^2C_{c,0}}{m_c^3}\rho.$$

This proves the three $K_{c,j}$ bounds, and summing the factual and interventional log-density errors proves equation 184. □

L.4.1 BOUNDARY-TRUNCATED FULL CONDITIONAL CDF

Trimming keeps the chart derivatives controlled near the probit boundaries, but a population conditional CDF integrates over the full coordinate interval. The computed CDF uses the retained interval and renormalizes its mass. The following bound measures this change in both the CDF and its gradient. Calibration will also restrict where scores are evaluated; that query-domain restriction is separate from changing the CDF's integration interval.

Fix R, $j \in R$, and $\eta \in [0, 1/2)$. Write $w = u_{R\setminus\{j\}}$ and, for a positive C^1 ratio r, define

$$\mathcal{N}_j[r](w) := \int_0^1 r(w,t)\ \mathrm{d}t, \qquad \mathcal{A}_j[r](w,u_j) := \int_0^{u_j} r(w,t)\ \mathrm{d}t,$$
$$\mathcal{N}_{j,\eta}[r](w) := \int_\eta^{1-\eta} r(w,t)\ \mathrm{d}t, \qquad \mathcal{A}_{j,\eta}[r](w,u_j) := \int_\eta^{u_j} r(w,t)\ \mathrm{d}t.$$

The full and computed conditional CDF functionals are

$$\mathcal{C}_j[r] := \frac{\mathcal{A}_j[r]}{\mathcal{N}_j[r]}, \qquad \mathcal{C}_{j,\eta}[r] := \frac{\mathcal{A}_{j,\eta}[r]}{\mathcal{N}_{j,\eta}[r]}, \qquad u_R \in \mathcal{U}_R = [\eta, 1-\eta]^{d_R}.$$

At $\eta = 0$ the two functionals coincide.

Proposition L.22 (Explicit inner-fibre truncation error)**.** *For this fixed ratio and coordinate, abbreviate the full-fibre quantities as N, A, C and the truncated quantities as N_η, A_η, C_η. The CDF error norms below are evaluated on $\mathcal{U}_R$. Suppose on the full cube*

$$0 < r \le C_r, \qquad \|\nabla r\|_{2,\infty} \le R_1, \qquad N \ge n_- > 0.$$

Define the normalizer-tail and numerator-tail constants

$$T_0 := 2\eta C_r, \qquad t_0 := \eta C_r, \qquad T_1 := 2\eta R_1, \qquad t_1 := \eta R_1,$$

If $n_\eta := n_- - T_0 > 0$, then

$$\|C_\eta - C\|_\infty \le \frac{t_0 + T_0}{n_\eta} =: E_{\eta,0}, \tag{185}$$

$$\|\partial_{u_j} C_\eta - \partial_{u_j} C\|_\infty \le \frac{C_r T_0}{n_\eta n_-} =: E_{\eta,j}, \tag{186}$$

$$\|\nabla_w C_\eta - \nabla_w C\|_{2,\infty} \le \frac{t_1 + T_1}{n_\eta} + \frac{2R_1 T_0}{n_\eta n_-} + \frac{E_{\eta,0} R_1}{n_\eta} =: E_{\eta,\perp}. \tag{187}$$

Consequently,

$$\|C_\eta - C\|_{C^1(\mathcal{U}_R)} \le \max\left\{E_{\eta,0}, \sqrt{E_{\eta,j}^2 + E_{\eta,\perp}^2}\right\} =: \epsilon_{\mathrm{trunc}}(r,j,\eta). \tag{188}$$

Proof. The omitted normalizer comprises two intervals, each of length η, whereas the omitted numerator comprises only $[0,\eta]$. Hence

$$|N - N_\eta| \le T_0, \qquad |A - A_\eta| \le t_0,$$
$$\|\nabla_w N - \nabla_w N_\eta\|_2 \le T_1, \qquad \|\nabla_w A - \nabla_w A_\eta\|_2 \le t_1.$$

In particular, $N_\eta \ge N - T_0 \ge n_\eta$. Since $0 \le A/N \le 1$, adding and subtracting A/N_η gives

$$\left|\frac{A_\eta}{N_\eta} - \frac{A}{N}\right| \le \frac{|A_\eta - A|}{N_\eta} + \frac{A}{N}\frac{|N - N_\eta|}{N_\eta} \le \frac{t_0 + T_0}{n_\eta}.$$

This proves equation 185.

The fundamental theorem of calculus yields $\partial_{u_j} C = r/N$, $\partial_{u_j} C_\eta = r/N_\eta$. Their difference is at most $C_r T_0/(n_\eta n_-)$, proving equation 186.

For a transverse derivative, use the exact quotient identity

$$\nabla_w C = \frac{\nabla_w A}{N} - C\frac{\nabla_w N}{N},$$

and its inner-fibre analogue. The difference between the first quotient terms is bounded by

$$\frac{t_1}{n_\eta} + \frac{R_1 T_0}{n_\eta n_-}. \tag{189}$$

For the second quotient, add and subtract $C\nabla_w N_\eta/N_\eta$ and $C\nabla_w N/N_\eta$. Since both CDFs lie in $[0,1]$ and $\|\nabla_w N\|_2, \|\nabla_w N_\eta\|_2 \le R_1$, the three resulting terms are at most

$$\frac{E_{\eta,0} R_1}{n_\eta}, \qquad \frac{T_1}{n_\eta}, \qquad \frac{R_1 T_0}{n_\eta n_-}. \tag{190}$$

Adding equation 189 and equation 190 proves equation 187; combining the conditioned and transverse derivative bounds proves equation 188. □

L.5 From fitted ratios to certified sink scores

The sink decision compares conditional CDFs and their gradients. We first show how ratio error changes these comparisons, then bound the additional error from evaluating them numerically. The output of this section is a score-error bound that the next section combines with calibration uncertainty to decide whether a node is a sink.

L.5.1 From ratio error to sink-score error

The remaining step is deterministic: a small log-ratio error gives a small conditional-CDF error, which in turn gives a small score error. We derive these bounds once with generic envelopes, then specialize them to each active set. Numerical errors enter here as prescribed bounds; Section L.5.2 shows how to obtain them. A hat denotes an analytic fitted quantity and a check its numerical evaluation.

Proposition L.23 (Conditional CDF and score-integrand stability)**.** *Under Assumption L.14, suppose all fitted active log ratios obey*

$$\max_{R,e} \left\| \widehat{\ell}_e^R - \log r_e^R \right\|_{C^1} \le \epsilon_\ell,$$

on every inner cube used by the computation. Let $n_* > 0$ be a fixed lower bound satisfying

$$\mathcal{N}_{j,\eta}[r_e^R](w) \ge n_* \quad \text{for every } (j,e,R,w).$$

For example, Proposition L.22 licenses $n_* \le \min_R(\underline{n}_R - 2\eta C_{r,R})$. Suppose the ratio value error E_0 defined in equation 192 satisfies $E_0 < n_*/2$. Let q_C and q_K bound any analytic integration error in the CDF's C^1 norm and the density's value, respectively. Both are zero for the polynomial fits of Section L.3, evaluated by exact antiderivatives. On the finite query family $\mathcal{Q}_{\mathrm{cal}}$, let δ_C and δ_K bound the respective numerical discrepancies as in equation 211. Assume $0 \le q_K \le q_C$, $0 \le \delta_K \le \delta_C$. Then the analytic fitted object obeys the uniform bound

$$\max_{j,e,R} \left\| \widehat{C}_{j,e}^R - C_{j,e}^R \right\|_{C^1} \le \epsilon_C^{\mathrm{an}},$$

where ϵ_C^{an} includes fitted-ratio, boundary-truncation, and analytic integration terms as displayed in equation 201. At every $u \in \mathcal{Q}_{\mathrm{cal}}$,

$$\max\left\{ |\check{C}(u) - C(u)|, \|\nabla\check{C}(u) - \nabla C(u)\|_2 \right\} \le \epsilon_C^{\mathrm{qry}} := \epsilon_C^{\mathrm{an}} + \delta_C.$$

Assume $\epsilon_C^{\mathrm{an}} \le 1$ and let ϵ_K^{an} be the analytic conditional-density value error in equation 203. If $\epsilon_K^{\mathrm{an}} \le k_-/2$, then every identity or foliation population integrand h and its analytic fitted counterpart $\widehat{h}$ satisfy

$$\|\widehat{h} - h\|_\infty \le C_{\mathrm{score}} \epsilon_C^{\mathrm{an}}, \qquad 0 \le h, \widehat{h} \le B_{\mathrm{score}}. \tag{191}$$

The proof gives C_{score} and B_{score} explicitly in terms of the fixed envelopes.

Proof. Fix (e,R) and abbreviate $\ell = \log r$, $\widehat{r} = \exp(\widehat{\ell})$.

Ratio values and gradients. Define the fitted-ratio envelope $\overline{C}_r$ and errors E_0, E_1 by

$$\overline{C}_r := C_r \exp(\epsilon_\ell), \qquad E_0 := C_r(\exp(\epsilon_\ell) - 1), \qquad E_1 := E_0(B_r + \epsilon_\ell) + C_r \epsilon_\ell. \tag{192}$$

Since $|\widehat{\ell} - \ell| \le \epsilon_\ell$ and $r \le C_r$,

$$|\widehat{r} - r| = r|\exp(\widehat{\ell} - \ell) - 1| \le E_0, \qquad \widehat{r} \le \overline{C}_r.$$

Differentiating $r = \exp(\ell)$ and adding and subtracting $r\nabla\widehat{\ell}$ gives

$$\begin{aligned}\|\nabla\widehat{r} - \nabla r\|_2 &\le |\widehat{r} - r|\,\|\nabla\widehat{\ell}\|_2 + r\|\nabla\widehat{\ell} - \nabla\ell\|_2 \\ &\le E_0(B_r + \epsilon_\ell) + C_r\epsilon_\ell = E_1. \end{aligned} \tag{193}$$

Conditional normalization. Write $Z = \mathcal{N}_{j,\eta}[r]$ and $\widehat{Z} = \mathcal{N}_{j,\eta}[\widehat{r}]$. The inner-fibre length is at most one, so

$$|\widehat{Z} - Z| \le E_0, \qquad \|\nabla_w\widehat{Z} - \nabla_w Z\|_2 \le E_1. \tag{194}$$

The stated condition gives $\widehat{Z} \ge n_*/2$. For the inner conditional densities $K_\eta = r/Z$ and $\widehat{K}_\eta = \widehat{r}/\widehat{Z}$,

$$|\widehat{K}_\eta - K_\eta| \le \frac{|\widehat{r} - r|}{\widehat{Z}} + r\frac{|\widehat{Z} - Z|}{\widehat{Z}Z} \le \frac{2E_0}{n_*} + \frac{2C_rE_0}{n_*^2} =: D_0. \tag{195}$$

For a nonconditioned coordinate k, the quotient expansion is

$$\partial_k\widehat{K}_\eta - \partial_k K_\eta = \frac{\partial_k\widehat{r} - \partial_k r}{\widehat{Z}} + \partial_k r\left(\frac{1}{\widehat{Z}} - \frac{1}{Z}\right) - \left(\frac{\widehat{r}\,\partial_k\widehat{Z}}{(\widehat{Z})^2} - \frac{r\,\partial_k Z}{Z^2}\right). \tag{196}$$

Expanding the last parenthesis gives the exact three-term identity

$$\frac{\widehat{r}\,\partial_k\widehat{Z}}{(\widehat{Z})^2} - \frac{r\,\partial_k Z}{Z^2} = \frac{(\widehat{r} - r)\,\partial_k\widehat{Z}}{(\widehat{Z})^2} + \frac{r(\partial_k\widehat{Z} - \partial_k Z)}{(\widehat{Z})^2} + r\,\partial_k Z\left((\widehat{Z})^{-2} - Z^{-2}\right). \tag{197}$$

The corresponding vector bounds use

$$R_1 := C_rB_r, \qquad \|\nabla r\|_{2,\infty}, \|\nabla_w Z\|_{2,\infty} \le R_1, \qquad \|\nabla_w\widehat{Z}\|_{2,\infty} \le R_1 + E_1. \tag{198}$$

Applying the expansion to the transverse gradient gives $\|\nabla_w\widehat{K}_\eta - \nabla_w K_\eta\|_2 \le D_1$, where

$$D_1 := \frac{2E_1}{n_*} + \frac{2R_1E_0}{n_*^2} + \frac{4E_0(R_1 + E_1)}{n_*^2} + \frac{4C_rE_1}{n_*^2} + \frac{8C_rR_1E_0\overline{C}_r}{n_*^4}. \tag{199}$$

The first two terms bound the first two quotients in equation 196; the remaining three bound equation 197. The final term uses $|(\widehat{Z})^{-2} - Z^{-2}| \le 4E_0(Z + \widehat{Z})/n_*^4$ and $Z + \widehat{Z} \le C_r + \overline{C}_r \le 2\overline{C}_r$.

CDF errors. The population truncated CDF $C_\eta[r]$ is the integral of $K_\eta[r]$ from η to u_j, and the exact analytic fitted CDF $\widehat{C}_\eta := C_\eta[\widehat{r}]$ is defined analogously. For $k \ne j$, differentiation under the integral gives

$$\partial_k C_\eta(u) = \int_\eta^{u_j} \partial_k K_\eta(w, t)\ \mathrm{d}t, \qquad \partial_j C_\eta(u) = K_\eta(u). \tag{200}$$

The jth gradient coordinate therefore has error at most D_0, and the transverse gradient has error at most D_1. Adding truncation and integration errors gives

$$\|\widehat{C}_\eta - C[r]\|_{C^1} \le \sqrt{D_0^2 + D_1^2} + \epsilon_{\mathrm{trunc}}(r_e^R, j, \eta) + q_C =: \epsilon_C^{\mathrm{an}}. \tag{201}$$

At a certified query q, the triangle inequality separates numerical evaluation, fitting and truncation:

$$\begin{aligned}\|\check{C}_\eta - C[r]\|_{C^1;q} &\le \|\check{C}_\eta - \widehat{C}_\eta\|_{C^1;q} + \|\widehat{C}_\eta - C_\eta[r]\|_{C^1} + \|C_\eta[r] - C[r]\|_{C^1} \\ &\le \delta_C + \epsilon_C^{\mathrm{an}} = \epsilon_C^{\mathrm{qry}}. \end{aligned} \tag{202}$$

Here $\|\cdot\|_{C^1;q}$ denotes the maximum value and Euclidean-gradient error at the single query. The corresponding total value error for the conditioned derivative $\partial_{u_j}C = K$ is

$$\epsilon_K^{\mathrm{an}} := D_0 + E_{\eta,j} + q_K \le \epsilon_C^{\mathrm{an}}, \qquad \epsilon_K^{\mathrm{qry}} := \epsilon_K^{\mathrm{an}} + \delta_K \le \epsilon_C^{\mathrm{qry}}. \tag{203}$$

Score integrands. Assume $\epsilon_C^{\mathrm{an}} \le 1$ and put $G_C := G_0 + 1$. Then all true and fitted CDF gradients are bounded by G_C. Set

$$\overline{k} := k_+ + \epsilon_K^{\mathrm{an}}, \qquad \rho_+ := \frac{2\overline{k}}{k_-}, \qquad c_\rho := \frac{2}{k_-} + \frac{2k_+}{k_-^2}.$$

The lower bound $\widehat{K}_f \ge k_-/2$ and the quotient identity imply

$$\left| \frac{\widehat{K}_e}{\widehat{K}_f} - \frac{K_e}{K_f} \right| \le c_\rho \epsilon_K^{\mathrm{an}}, \qquad \frac{\widehat{K}_e}{\widehat{K}_f}, \frac{K_e}{K_f} \le \rho_+.$$

With $B_{\mathrm{fol}} := G_C(1 + \rho_+)$, the change in the foliation numerator vector is at most

$$\epsilon_z := (1 + \rho_+)\epsilon_C^{\mathrm{an}} + G_C c_\rho \epsilon_K^{\mathrm{an}} \le (1 + \rho_+ + G_C c_\rho)\epsilon_C^{\mathrm{an}}.$$

Thus the identity and foliation integrands have the respective Lipschitz constants

$$\begin{aligned} C_{\mathrm{id}} &:= 2G_C + 4, \\ C_{\mathrm{fol}} &:= 2B_{\mathrm{fol}}(1 + \rho_+ + G_C c_\rho) + 2G_C B_{\mathrm{fol}}^2, \\ C_{\mathrm{score}} &:= \max\{C_{\mathrm{id}}, C_{\mathrm{fol}}\}. \end{aligned}$$

Indeed, for the identity integrand, the value-square difference is at most $3\epsilon_C^{\mathrm{an}}$: the true residual has magnitude at most one, while $\epsilon_C^{\mathrm{an}} \le 1$ makes the fitted residual at most two. The true CDF gradient is bounded by G_0, the fitted one by $G_C = G_0 + 1$, so the gradient-square difference is at most $(G_0 + 1 + G_C + 1)\epsilon_C^{\mathrm{an}} = (2G_C + 1)\epsilon_C^{\mathrm{an}}$. Their sum is $C_{\mathrm{id}}\epsilon_C^{\mathrm{an}}$. For the foliation integrand, the numerator-square difference is at most $2B_{\mathrm{fol}}\epsilon_z$, while changing the denominator $1 + \|\nabla C_e\|_2^2$ contributes at most $2G_C B_{\mathrm{fol}}^2 \epsilon_C^{\mathrm{an}}$; all denominators are at least one. Finally, both integrands are bounded by

$$B_{\mathrm{score}} := \max\{4 + (G_C + 1)^2, B_{\mathrm{fol}}^2\}.$$

This proves equation 191.

Numerical score evaluation. When comparing numerical and fitted quantities, both densities in the quotient may differ from their population values. The same calculation uses the following constants. Assume additionally $\epsilon_C^{\mathrm{qry}} \le 1$ and $\epsilon_K^{\mathrm{qry}} \le k_-/2$, and put

$$\begin{aligned} \overline{k}^{\mathrm{qry}} &:= k_+ + \epsilon_K^{\mathrm{qry}}, & \rho_+^{\mathrm{qry}} &:= \frac{2\overline{k}^{\mathrm{qry}}}{k_-}, \\ c_\rho^{\mathrm{num}} &:= \frac{2}{k_-} + \frac{4\overline{k}^{\mathrm{qry}}}{k_-^2}, & B_{\mathrm{fol}}^{\mathrm{num}} &:= G_C(1 + \rho_+^{\mathrm{qry}}), \end{aligned} \tag{204}$$

$$\begin{aligned} C_{\mathrm{id}}^{\mathrm{num}} &:= 2G_C + 6, & C_{\mathrm{fol}}^{\mathrm{num}} &:= 2B_{\mathrm{fol}}^{\mathrm{num}}(1 + \rho_+^{\mathrm{qry}} + G_C c_\rho^{\mathrm{num}}) + 2G_C (B_{\mathrm{fol}}^{\mathrm{num}})^2, \\ C_{\mathrm{score}}^{\mathrm{num}} &:= \max\{C_{\mathrm{id}}^{\mathrm{num}}, C_{\mathrm{fol}}^{\mathrm{num}}\}, & B_{\mathrm{score}}^{\mathrm{num}} &:= \max\{4 + (G_C + 1)^2, (B_{\mathrm{fol}}^{\mathrm{num}})^2\}. \end{aligned} \tag{205}$$

Indeed, the density-ratio quotient rule gives $\left| \check{K}_e/\check{K}_f - \widehat{K}_e/\widehat{K}_f \right| \le c_\rho^{\mathrm{num}} \delta_K$; the factor four comes from two denominators bounded below by $k_-/2$. Repeating the numerator-vector and denominator-square calculation with these constants, and using $\delta_K \le \delta_C$, gives

$$|\check{h}(q) - \widehat{h}(q)| \le C_{\mathrm{score}}^{\mathrm{num}} \delta_C. \tag{206}$$

This is a finite-query statement. Calibration sampling and query-law transport enter the simultaneous event over all active sets. □

Specializing to an active set. We now restore R and apply Proposition L.23 to the exact polynomial $\widehat{r}_e^{R,\sharp}$ induced by the represented coefficients. Its log-ratio error is bounded by $e_{R,e}^{\ell}$ in equation 222, including input and solve errors. Use the population envelopes

$$\begin{aligned} r_e^R \le C_{r,R}, \quad \|\nabla \log r_e^R\|_{2,\infty} \le B_{r,R}, \\ \mathcal{N}_j[r_e^R] \ge \underline{n}_R, \qquad k_{-,R} \le K_{j,e}^R \le k_{+,R}, \\ \|\nabla C_{j,e}^R\|_{2,\infty} \le G_{0,R}. \end{aligned}$$

uniformly over admissible j and e. Let $n_R^\eta := \underline{n}_R - 2\eta C_{r,R}$. Let $q_{C,R}$ and $q_{K,R}$ denote, respectively, the analytic quadrature errors for the conditional CDF in C^1 and its conditioned density in value, with $0 \le q_{K,R} \le q_{C,R}$. For any generic bound B in Proposition L.23, write $B|_R$ for its worst-case value after the substitutions

$$(\epsilon_\ell, C_r, B_r, n_*) \leftarrow (\max_{e\in[m]} e_{R,e}^\ell, C_{r,R}, B_{r,R}, n_R^\eta),$$
$$(k_-, k_+, G_0) \leftarrow (k_{-,R}, k_{+,R}, G_{0,R}),$$
$$(q_C, q_K) \leftarrow (q_{C,R}, q_{K,R}).$$

The same specialization takes the maximum over admissible (j, e) in the truncation bound of Proposition L.22, using $(C_r, R_1, n_-) = (C_{r,R}, C_{r,R}B_{r,R}, \underline{n}_R)$. Hence all formulas remain the explicit expressions in equation 192–equation 203 for every active set.

Only two propagated errors enter the peeling argument: the querywise C^1 error of the conditional CDF and the value error of its conditioned derivative $K = \partial_{u_j} C$. Let $\bar\epsilon_R^C$ and $\bar\epsilon_R^K$ be the fixed numerical budgets verified in Section L.5.2. Adding these to the analytic errors gives

$$e_R^C := \epsilon_C^{\mathrm{an}}|_R + \bar\epsilon_R^C, \tag{207}$$
$$e_R^K := \epsilon_K^{\mathrm{an}}|_R + \bar\epsilon_R^K. \tag{208}$$

The assumptions $n_R^\eta > 0$, $E_0|_R < n_R^\eta/2$, $e_R^C \le 1$, and $e_R^K \le k_{-,R}/2$ are exactly the positivity and stability conditions of the generic result. The quadrature and numerical density errors are bounded by their corresponding C^1 errors. These fixed bounds, and therefore e_R^C and e_R^K, are chosen before the calibration sample is observed.

The peeling theorem depends on the final score Lipschitz and envelope bounds. Specialize the numerical constants in equation 204– equation 205 by

$$(C_{\mathrm{score},R}, B_{\mathrm{score},R}) := (C_{\mathrm{score}}^{\mathrm{num}}, B_{\mathrm{score}}^{\mathrm{num}})|_R, \quad (G_0, k_-, k_+, \epsilon_K^{\mathrm{qry}}) \leftarrow (G_{0,R}, k_{-,R}, k_{+,R}, e_R^K).$$

Index the identity and foliation comparisons separately:

$$\mathcal{J}_R^{\mathrm{id}} := \{(\mathrm{id}, j, e) : j \in R,\ \sigma_{ej} = 0\},$$
$$\mathcal{J}_R^{\mathrm{fol}} := \{(\mathrm{fol}, j, e, f) : j \in R,\ 1 \le e < f \le m,\ \sigma_{ej} = \sigma_{fj} = 1\},$$

If $\mathcal{J}_R := \mathcal{J}_R^{\mathrm{id}} \dot\cup \mathcal{J}_R^{\mathrm{fol}}$ is empty, its empirical and population maximum score is defined as zero. The same $B_{\mathrm{score},R}$ and $C_{\mathrm{score},R}$ remain valid. Thus every quantity passed to the peeling analysis is determined by R and the stated inputs.

L.5.2 Numerical evaluation of fitted scores

The preceding bounds describe how an error in each fitted quantity affects the sink score. We now obtain the numerical error bounds used there. Interval arithmetic encloses an exact value between two outward-rounded endpoints, allowing each computation to be checked at finite precision. We first certify the polynomial fit, then its conditional CDFs, and finally the scores. Polynomial positivity is checked on the whole cube; CDF and score evaluation is checked at the calibration queries. These calculations supply the numerical contributions to the score radius, together with positivity and accepted-count checks.

Throughout the certification argument, fix an active set R and degree p. We omit these indices, and the environment index e when only one environment is involved. Thus $\boldsymbol{\psi}$, L, $\widehat{\mathbf{G}}$, $\widehat{\mathbf{b}}$ and $\widehat{\boldsymbol{\beta}}$ refer to this fit. Write $B_0 = B_{0,R,p}$ and $B_1 = B_{1,R,p}$ for the feature and derivative bounds. We use ξ for errors in the supplied inputs and δ for errors introduced by numerical computation; their subscripts identify the quantity being bounded. The fixed budgets at the end of the argument apply uniformly over the omitted indices.

Exact representation and outward rounding. Let $\mathbb{F}$ be the finite numbers in the chosen representation, and let $\iota(x)$ be the exact dyadic value of $x \in \mathbb{F}$:

$$\iota(x) = N_x 2^{-q_x} \in \mathbb{Q}, \qquad N_x \in \mathbb{Z}, \quad q_x \in \mathbb{N}_0, \quad q_x = 0 \text{ or } N_x \text{ is odd},$$

where for $x = 0$ we set $(N_x, q_x) := (0, 0)$; this reduced convention makes the pair unique. The map ι is applied componentwise to every supplied value. The estimator abstains if a supplied value is nonfinite or lies outside its fixed bound. A superscript $\sharp$ denotes the exact value represented by this embedding; acquisition and chart errors are bounded separately. For a finite interval enclosure $[x]$, define directed represented endpoints by

$$\mathrm{out}_+[x] := \min\{y \in \mathbb{F} : y \geq \sup[x]\}, \qquad \mathrm{out}_-[x] := \max\{y \in \mathbb{F} : y \leq \inf[x]\}.$$

Outward rounding preserves the enclosure: $\mathrm{out}_-[x] \leq \inf[x] \leq \sup[x] \leq \mathrm{out}_+[x]$.

From input errors to moment errors. The represented training values define the exact moments

$$\widehat{\mathbf{G}}^\sharp := \frac{1}{n_0}\sum_{s=1}^{n_0} \boldsymbol{\psi}(\widetilde{\mathbf{U}}^\sharp_{s,R})\boldsymbol{\psi}(\widetilde{\mathbf{U}}^\sharp_{s,R})^\top, \qquad \widehat{\mathbf{b}}^\sharp := \frac{1}{n_e}\sum_{s=1}^{n_e} \boldsymbol{\psi}(\widetilde{\mathbf{V}}^\sharp_{s,R}).$$

To compare them with the ideal empirical moments in equation 144, suppose the supplied chart values obey

$$\max_s \|\widetilde{\mathbf{U}}^\sharp_{s,R} - \widetilde{\mathbf{U}}_{s,R}\|_2 \leq \xi_U, \qquad \max_s \|\widetilde{\mathbf{V}}^\sharp_{s,R} - \widetilde{\mathbf{V}}_{s,R}\|_2 \leq \xi_V. \tag{209}$$

The mean-value theorem and equation 143 first give $\|\boldsymbol{\psi}(u_R^\sharp) - \boldsymbol{\psi}(u_R)\|_2 \leq B_1\|u_R^\sharp - u_R\|_2$.

Writing $\boldsymbol{\psi}_s^\sharp$ and $\boldsymbol{\psi}_s$ for the two feature vectors, expand each outer product before taking norms:

$$\begin{aligned}
\boldsymbol{\psi}_s^\sharp(\boldsymbol{\psi}_s^\sharp)^\top - \boldsymbol{\psi}_s\boldsymbol{\psi}_s^\top &= (\boldsymbol{\psi}_s^\sharp - \boldsymbol{\psi}_s)(\boldsymbol{\psi}_s^\sharp)^\top + \boldsymbol{\psi}_s(\boldsymbol{\psi}_s^\sharp - \boldsymbol{\psi}_s)^\top,\\
\|\widehat{\mathbf{G}}^\sharp - \widehat{\mathbf{G}}\|_{\mathrm{op}} &\leq 2B_0B_1\xi_U =: \xi_G,\\
\|\widehat{\mathbf{b}}^\sharp - \widehat{\mathbf{b}}\|_2 &\leq B_1\xi_V =: \xi_b.
\end{aligned}$$

These bounds account for acquisition and chart errors separately from the rounding errors in the calculation below.

Certified solve and coefficient errors. Outward evaluation of the normalized Legendre recurrence gives enclosures $[\widehat{\mathbf{G}}]$ and $[\widehat{\mathbf{b}}]$ containing these exact objects. To certify a requested empirical floor $\lambda_* > 0$, interval Cholesky is applied to $[\widehat{\mathbf{G}}] - \lambda_*\mathbf{I}$. In the fixed dictionary order, write $[\mathbf{L}]$ for the lower-triangular interval factor; its recursion is

$$[\mathbf{L}]_{kk} := \sqrt{[\widehat{\mathbf{G}}]_{kk} - \lambda_* - \sum_{\ell<k}[\mathbf{L}]_{k\ell}^2}, \qquad [\mathbf{L}]_{ik} := \frac{[\widehat{\mathbf{G}}]_{ik} - \sum_{\ell<k}[\mathbf{L}]_{i\ell}[\mathbf{L}]_{k\ell}}{[\mathbf{L}]_{kk}} \ \text{ for } \ i > k.$$

The condition $\inf([\widehat{\mathbf{G}}]_{kk} - \lambda_* - \sum_{\ell<k}[\mathbf{L}]_{k\ell}^2) > 0$ at every pivot implies, inductively, that the exact shifted Gram matrix has a real Cholesky factor, hence

$$\widehat{\mathbf{G}}^\sharp - \lambda_*\mathbf{I} \succ 0, \qquad \lambda_{\min}(\widehat{\mathbf{G}}^\sharp) > \lambda_*.$$

The interval solve then encloses the exact solution of the represented system:

$$[\widehat{\boldsymbol{\beta}}] \ni \widehat{\boldsymbol{\beta}}^{\mathrm{ex}} := (\widehat{\mathbf{G}}^\sharp)^{-1}\widehat{\mathbf{b}}^\sharp, \qquad \widehat{r}^{\mathrm{ex}} := (\widehat{\boldsymbol{\beta}}^{\mathrm{ex}})^\top\boldsymbol{\psi}.$$

Here "ex" distinguishes this solution from the ideal empirical coefficient $\widehat{\boldsymbol{\beta}}$. On the event of Theorem L.11, that ideal coefficient obeys $\|\widehat{\boldsymbol{\beta}}\|_2 \leq B_\beta + \epsilon_\beta =: B_{\widehat{\beta}}$, where $B_\beta = B_0/\lambda_{G,R,p}$ and $\epsilon_\beta = \epsilon_{\beta,R,e,p}$. Subtracting the two normal equations gives

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^{\mathrm{ex}} - \widehat{\boldsymbol{\beta}} &= (\widehat{\mathbf{G}}^\sharp)^{-1}\left[\widehat{\mathbf{b}}^\sharp - \widehat{\mathbf{b}} + (\widehat{\mathbf{G}} - \widehat{\mathbf{G}}^\sharp)\widehat{\boldsymbol{\beta}}\right],\\
\|\widehat{\boldsymbol{\beta}}^{\mathrm{ex}} - \widehat{\boldsymbol{\beta}}\|_2 &\leq \frac{\xi_b + \xi_G B_{\widehat{\beta}}}{\lambda_*} =: \xi_\beta.
\end{aligned}$$

Their fitted polynomials therefore differ by at most $\xi_0 := B_0\xi_\beta$, $\xi_1 := B_1\xi_\beta$ in value and C^1 norm, respectively.

The implementation returns coefficients with exact represented value $\widehat{\boldsymbol{\beta}}^\sharp$. Their distance from the enclosed solution, and the resulting value and C^1 errors, are bounded by

$$\delta_\beta := \mathrm{out}_+\left[\left(\sum_{a=1}^{L}\sup\left|[\widehat{\beta}_a] - \widehat{\beta}_a^\sharp\right|^2\right)^{1/2}\right], \qquad \delta_0 := B_0\delta_\beta, \qquad \delta_1 := B_1\delta_\beta.$$

Passing to log ratios. Suppose $\widehat{r}^{\mathrm{ex}}$ has floor $c_{\mathrm{ex}} > 0$ and gradient bound G_{ex}. If the coefficient error leaves a positive floor c_{fl}, the log-ratio error is at most δ_ℓ, where

$$c_{\mathrm{fl}} := c_{\mathrm{ex}} - \delta_0 > 0, \qquad \delta_\ell := \max\left\{\frac{\delta_0}{c_{\mathrm{fl}}}, \frac{\delta_1}{c_{\mathrm{fl}}} + \frac{G_{\mathrm{ex}}\delta_0}{c_{\mathrm{ex}}c_{\mathrm{fl}}}\right\}.$$

For the input error, use the ideal fit's floor and gradient bound:

$$\begin{aligned} c_{\mathrm{ideal}} &:= c^{R,e}_{G,r} - \epsilon^{R,e,p}_{G,r,0} > 0, \quad G_{\mathrm{ideal}} := B_1 B_{\widehat{\beta}}, \\ c_{\mathrm{io}} &:= \min\{c_{\mathrm{ideal}}, c_{\mathrm{ex}}\} > 0, \\ \xi_\ell &:= \max\left\{\frac{\xi_0}{c_{\mathrm{io}}}, \frac{\xi_1}{c_{\mathrm{io}}} + \frac{G_{\mathrm{ideal}}\xi_0}{c^2_{\mathrm{io}}}\right\}. \end{aligned}$$

The same decomposition as in equation 151 gives $\|\log \widehat{r}^{\mathrm{ex}} - \log \widehat{r}\|_{C^1} \le \xi_\ell$. Thus ξ_ℓ bounds the change from ideal to represented inputs, while δ_ℓ bounds the subsequent numerical solve. The statistical bounds allow the required floors and gradient envelope to be chosen as

$$\begin{aligned} 0 < \lambda_* \le \lambda_{G,R,p} - \epsilon_{G,R,p} - \xi_G, \\ 0 < c_{\mathrm{ex}} \le c_{\mathrm{ideal}} - \xi_0, \qquad G_{\mathrm{ex}} \ge G_{\mathrm{ideal}} + \xi_1. \end{aligned} \tag{210}$$

Both floors in equation 210 must be positive, and G_{ex} must be finite.

Full-cube positivity. Let $\widehat{r}^\sharp$ be the exact polynomial induced by the represented coefficients. Beginning with $[0,1]^{d_R}$, evaluate its natural interval extension $[\widehat{r}^\sharp](B)$ on a box B. A box is accepted when its outward lower endpoint is positive, rejected when its outward upper endpoint is nonpositive, and otherwise bisected through the midpoint of a widest coordinate. If every leaf is accepted before the fixed box and depth caps, finite covering and interval inclusion give

$$\widehat{r}^\sharp(u_R) \ge \min_{B \text{ accepted}} \mathrm{out}_-[\widehat{r}^\sharp](B) > 0 \quad (u_R \in [0,1]^{d_R}).$$

If a box remains unresolved at either cap, the estimator returns $\perp$.

Finite-query conditional CDF and density. Let $\mathcal{Q} = \{u^{(1),\sharp}, \dots, u^{(N),\sharp}\}$ be the accepted calibration queries, including multiplicities, with $N = N^\sharp_{\mathrm{cal}}$. Choose a dyadic trim level $\eta \in (0, 1/2)$ before sampling. The interval calculation uses η and $1-\eta$ exactly. For $\varphi_a(t) = \sqrt{2a+1}P_a(2t-1)$, interval arithmetic evaluates the polynomial antiderivative

$$J_0(t) = t, \qquad J_a(t) = \frac{P_{a+1}(2t-1) - P_{a-1}(2t-1)}{2\sqrt{2a+1}} \quad (a \ge 1),$$

which gives the integrals without a quadrature approximation. Within the next displays, $u = (w, u_j) \in \mathcal{Q}$ denotes an exact represented query; we omit its $\sharp$ superscript. For a fixed coordinate j and environment e, define the partial integral A and normalizer Z by

$$\begin{aligned} A &:= \int_\eta^{u_j} \widehat{r}^\sharp(w,t)\,\mathrm{d}t, & Z &:= \int_\eta^{1-\eta} \widehat{r}^\sharp(w,t)\,\mathrm{d}t, \\ C^\sharp_{j,e}(u) &:= A/Z, & K^\sharp_{j,e}(u) &:= \widehat{r}^\sharp(u)/Z. \end{aligned}$$

The interval lower bounds for Z and $K^\sharp_{j,e}$ must be positive. For $k \ne j$, the CDF derivatives follow from the quotient rule:

$$\partial_k C^\sharp_{j,e}(u) = \frac{(\partial_k A)Z - A(\partial_k Z)}{Z^2}, \qquad \partial_j C^\sharp_{j,e}(u) = K^\sharp_{j,e}(u).$$

Write $\check{C}_{j,e}$ and $\check{K}_{j,e}$ for the computed values, with $\nabla\check{C}_{j,e}$ denoting the returned gradient. Their discrepancies from these exact references are

$$\begin{aligned} \delta_C &:= \max_{j,e}\max_{u\in\mathcal{Q}} \max\left\{\begin{matrix} |\check{C}_{j,e}(u) - C^\sharp_{j,e}(u)|, \\ \|\nabla\check{C}_{j,e}(u) - \nabla C^\sharp_{j,e}(u)\|_2 \end{matrix}\right\}, \\ \delta_K &:= \max_{j,e}\max_{u\in\mathcal{Q}} |\check{K}_{j,e}(u) - K^\sharp_{j,e}(u)|. \end{aligned} \tag{211}$$

Both maxima range over $j \in R$ and $e \in [m]$. Interval evaluation gives outward upper bounds for δ_C and δ_K.

Identity and foliation scores. For each fixed j, let $c_e(u)$, $k_e(u)$ and $\mathbf{g}_e(u)$ be the exact dyadic values of the returned CDF, density and gradient. For $N > 0$, the exact empirical scores in equation 131–equation 132 are

$$\widehat{I}_{j,e} := \frac{1}{N} \sum_{u \in \mathcal{Q}} \left[(c_e(u) - u_j)^2 + \|\mathbf{g}_e(u) - \mathbf{e}_j\|_2^2 \right],$$
$$\widehat{F}_{j,ef} := \frac{1}{N} \sum_{u \in \mathcal{Q}} \frac{\|\mathbf{g}_e(u) - (k_e(u)/k_f(u))\mathbf{g}_f(u)\|_2^2}{1 + \|\mathbf{g}_e(u)\|_2^2}.$$

As in equation 132, the foliation score is normalized by the gradient of its first environment. Let $\check{I}_{j,e}$ and $\check{F}_{j,ef}$ be the returned scores. The additional error in evaluating these averages is

$$\delta_s := \max \left\{ \max_{\substack{j \in R \\ e \in [m]}} \left| \iota(\check{I}_{j,e}) - \widehat{I}_{j,e} \right|, \max_{\substack{j \in R \\ 1 \le e < f \le m}} \left| \iota(\check{F}_{j,ef}) - \widehat{F}_{j,ef} \right| \right\}. \tag{212}$$

Its outward interval bound is added to the propagated CDF and density errors.

To transfer the result to the ideal queries $u^{(s)}$, the represented and ideal trim decisions must agree:

$$B_s^\sharp = B_s \quad (s = 1, \dots, n_{\mathrm{cal}}), \qquad N = N_{\mathrm{cal}}, \tag{213}$$

Pair the accepted queries in sample order, and let $\mathcal{H}$ contain the exact identity and foliation integrands formed from $C^\sharp$ and $K^\sharp$. Their query-input error is

$$\xi_s := \max_{h \in \mathcal{H}} \frac{1}{N_{\mathrm{cal}}} \sum_{s: B_s = 1} |h(u_R^{(s),\sharp}) - h(u_R^{(s)})|, \tag{214}$$

when $N_{\mathrm{cal}} > 0$, and zero otherwise. Coordinate enclosures and a fixed Lipschitz bound for each h suffice to bound ξ_s.

Certifying the accepted count. If, conditionally on the fitted chart, the ideal accepted count satisfies $N_{\mathrm{cal}} \sim \mathrm{Bin}(n, \theta)$ with $\theta \ge \theta_0$, then for $1 \le k \le n$ monotonicity in θ and the beta identity give

$$\mathbb{P}(N_{\mathrm{cal}} < k) \le I_{1-\theta_0}(n - k + 1, k), \qquad \mathbb{P}(N_{\mathrm{cal}} \ge k) \ge I_{\theta_0}(k, n - k + 1), \tag{215}$$

where $I_x(a, b)$ is the regularized incomplete beta function. Interval arithmetic provides an outward upper endpoint for the first expression and an outward lower endpoint for the second. The cases $k = 0$, $k > n$, and $\theta_0 \in \{0, 1\}$ are evaluated directly.

Collecting the numerical conditions. Before inspecting the corresponding ratio or calibration sample, fix six active-set bounds that hold uniformly over every evaluated environment and basis index:

$$\bar{\epsilon}_R^{\ell,\mathrm{num}}, \quad \bar{\epsilon}_R^{\ell,\mathrm{io}}, \quad \bar{\epsilon}_R^{C}, \quad \bar{\epsilon}_R^{K}, \quad \bar{\epsilon}_R^{s,\mathrm{num}}, \quad \bar{\epsilon}_R^{s,\mathrm{io}}, \quad \bar{\epsilon}_R^{K} \le \bar{\epsilon}_R^{C}.$$

The numerical condition requires, for every evaluated index,

$$\delta_\ell \le \bar{\epsilon}_R^{\ell,\mathrm{num}}, \quad \delta_C \le \bar{\epsilon}_R^{C}, \quad \delta_K \le \bar{\epsilon}_R^{K}, \quad \delta_s \le \bar{\epsilon}_R^{s,\mathrm{num}}.$$

Define $\mathcal{E}_{\mathrm{io}}$ by equation 209, equation 213, and

$$\xi_\ell \le \bar{\epsilon}_R^{\ell,\mathrm{io}}, \quad \xi_s \le \bar{\epsilon}_R^{s,\mathrm{io}}.$$

Let $\mathcal{E}_{\mathrm{num}}$ collect all numerical validity conditions above, over every active set and query. The algorithm proceeds with an active-set update only after each required interval meets its bound; an unresolved comparison returns $\perp$.

Remark L.24 (Finite-query score enclosure). The interval checks certify the represented calculation at the finite query set $\mathcal{Q}$. The event $\mathcal{E}_{\mathrm{io}}$ connects this calculation to the ideal statistical estimator.

L.6 ADAPTIVE PEELING

We now turn the score bounds into a sequence of sink decisions. The active set is selected by earlier scores, using the same calibration sample. A bound for one fixed set would not justify applying it to a set selected from those data. We therefore control all relevant active sets simultaneously, so the bound covers whichever path is chosen. Deleting a sink leaves each remaining node together with all its ancestors; these are the ancestor-closed sets used below. For the true DAG, let $\mathrm{an}_{\mathcal{G}}(j)$ denote the strict ancestors of j and define

$$\mathcal{R}_{\mathcal{G}} := \left\{ R \subseteq [d] : R \neq \varnothing,\ \bigcup_{j \in R} \mathrm{an}_{\mathcal{G}}(j) \subseteq R \right\}, \qquad \mathcal{R}_{\mathrm{all}} := 2^{[d]} \setminus \{\varnothing\}.$$

We control the fitted ratios and scores simultaneously over $\mathcal{R}_{\mathcal{G}}$, using the shared ratio and calibration samples. The resulting event covers whichever peeling path is selected. Counting all nonempty subsets in $\mathcal{R}_{\mathrm{all}}$ gives bounds that do not require knowing the DAG.

For $R \in \mathcal{R}_{\mathcal{G}}$, write $L_{R,p} = \binom{d_R+p}{p}$. This active set contributes $L_{R,p}^2$ ordered Gram entries and $mL_{R,p}$ intervention-moment coordinates. Their DAG-independent upper counts are

$$\overline{M}_G := \sum_{r=1}^{d} \binom{d}{r} \binom{r+p}{p}^2, \qquad \overline{M}_b := m \sum_{r=1}^{d} \binom{d}{r} \binom{r+p}{p}.$$

These quantities count scalar concentration inequalities.

For calibration, let $m_{1j} := \sum_{e=1}^{m} \sigma_{ej},\ m_{0j} := m - m_{1j}$.

At node j, there are m_{0j} identity integrands and $\binom{m_{1j}}{2}$ unordered foliation integrands. Hence

$$M_{\mathrm{cal}} := \sum_{R \in \mathcal{R}_{\mathcal{G}}} \sum_{j \in R} \left[m_{0j} + \binom{m_{1j}}{2} \right], \qquad \overline{M}_{\mathrm{cal}} := 2^{d-1} \sum_{j=1}^{d} \left[m_{0j} + \binom{m_{1j}}{2} \right]. \tag{216}$$

Each node belongs to exactly 2^{d-1} nonempty subsets, which gives equation 216. Replacing the bracket by $m + \binom{m}{2}$ yields a signature-independent upper bound.

Definition L.25 (Failure-probability allocation)**.** Choose positive numbers

$$\delta_{\mathrm{ch}} + \delta_G + \delta_b + \delta_{\mathrm{cal}} + \delta_{\mathrm{gr}} \leq \delta < 1.$$

The five terms control target-signature recovery, factual Gram entries, intervention moments, sink-score calibration, and graph-score calibration, respectively.

Instantiate Theorem L.6 with $\delta = \delta_{\mathrm{ch}}$ and define the chart event:

$$\mathcal{E}_{\mathrm{ch}} := \left\{ \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \leq \epsilon_\mu,\ \|\widehat{\boldsymbol{\Sigma}}_X - \boldsymbol{\Sigma}_X\|_{\mathrm{op}} \leq \epsilon_\Sigma,\ \|\widehat{\mathbf{B}}_e - \mathbf{B}_e\|_{\mathrm{op}} \leq \epsilon_{B,e} \text{ for every } e \right\} \in \mathcal{F}_{\mathrm{ch}}.$$

Thus $\mathbb{P}(\mathcal{E}_{\mathrm{ch}}) \geq 1 - \delta_{\mathrm{ch}}$. On this event, equation 127 aligns the residual chart and target signatures up to one signed permutation. The population envelopes, query-density condition, and interval bounds are then verified directly; the algorithm proceeds exactly when they hold.

Proposition L.26 (Uniform ratio bound over active sets)**.** *Condition on the fitted chart and suppose $\mathcal{E}_{\mathrm{ch}}$ holds. Assume the Gram floors, projection tails, ratio-density floors, and chart-leakage bounds hold for every $(R, e) \in \mathcal{R}_{\mathcal{G}} \times [m]$. Let n_0^{rat} and n_e^{rat} be the independent factual and environment-e ratio-training counts. Define the population moments*

$$\mathbf{G}_R := \mathbb{E}_0[\boldsymbol{\psi}_{R,p}(\widetilde{\mathbf{U}}_R)\boldsymbol{\psi}_{R,p}(\widetilde{\mathbf{U}}_R)^\top], \qquad \mathbf{b}_e^R := \mathbb{E}_e[\boldsymbol{\psi}_{R,p}(\widetilde{\mathbf{V}}_R)],$$
$$\boldsymbol{\beta}_e^{R,\circ} := \mathbf{G}_R^{-1}\mathbf{b}_e^R, \qquad r_e^{R,\circ} := (\boldsymbol{\beta}_e^{R,\circ})^\top \boldsymbol{\psi}_{R,p}.$$

Assumption L.10 supplies the one-sided bounds

$$\lambda_{\min}(\mathbf{G}_R) \geq \underline{\lambda}_{G,R,p} > 0,$$
$$\|r_e^{R,\circ} - \widetilde{r}_e^R\|_{C^k} \leq \overline{A}_{G,k}^{R,e}(p), \qquad k \in \{0,1\},$$
$$\inf_{u_R} \widetilde{r}_e^R(u_R) \geq \underline{c}_{G,r}^{R,e} > 0.$$

For the fixed degree p, abbreviate the feature bounds, Gram floor, projection tails, and density floor as

$$B_{0,R} := B_{0,R,p}, \quad B_{1,R} := B_{1,R,p}, \quad \lambda_R := \underline{\lambda}_{G,R,p}, \quad A_{k,R,e} := \overline{A}^{R,e}_{G,k}(p), \quad a_{R,e} := \underline{c}^{R,e}_{G,r}.$$

The ratio error follows in three steps. First, concentration controls the Gram matrix and intervention moment:

$$g_R := B_{0,R}^2 \sqrt{\frac{2\log(2\overline{M}_G/\delta_G)}{n_0^{\mathrm{rat}}}}, \tag{217}$$

$$b_{R,e} := B_{0,R} \sqrt{\frac{2\log(2\overline{M}_b/\delta_b)}{n_e^{\mathrm{rat}}}}. \tag{218}$$

Second, if $g_R < \lambda_R$, the linear solve gives a coefficient radius, which yields a ratio-value radius and a remaining positive density floor:

$$r_{\beta,R,e} := \frac{b_{R,e} + g_R B_{0,R}/\lambda_R}{\lambda_R - g_R}, \tag{219}$$

$$r_{0,R,e} := A_{0,R,e} + B_{0,R} r_{\beta,R,e}, \tag{220}$$

$$c_{R,e} := a_{R,e} - r_{0,R,e}.$$

These bounds hold simultaneously with conditional probability at least $1 - \delta_G - \delta_b$ by Theorem L.11.

Third, for $c_{R,e} > 0$, pass to log ratios and add chart error:

$$\begin{aligned} e^{\mathrm{fit}}_{R,e} &:= \max\left\{ \frac{r_{0,R,e}}{c_{R,e}}, \quad \frac{A_{1,R,e} + B_{1,R} r_{\beta,R,e}}{c_{R,e}} + \frac{(A_{1,R,e} + B_{1,R}B_{0,R}/\lambda_R) r_{0,R,e}}{a_{R,e} c_{R,e}} \right\} \\ &\quad + C_{\mathrm{leak}}(R, k(R))\epsilon_q(k(R)), \\ e^{\ell}_{R,e} &:= e^{\mathrm{fit}}_{R,e} + \bar{\epsilon}^{\ell,\mathrm{io}}_R + \bar{\epsilon}^{\ell,\mathrm{num}}_R. \end{aligned} \tag{221}$$

The first radius includes sampling, projection, and chart error. The total radius also includes input representation and numerical evaluation. Let $\mathcal{E}_{\mathrm{rat}} \in \mathcal{F}_{\mathrm{fit}}$ collect these inequalities over every indexed active set:

$$\begin{aligned} \mathcal{E}_{G,R} &:= \left\{ \|\widehat{\mathbf{G}}_R - \mathbf{G}_R\|_{\mathrm{op}} \le g_R, \quad \lambda_{\min}(\widehat{\mathbf{G}}_R) \ge \underline{\lambda}_{G,R,p} - g_R \right\}, \qquad R \in \mathcal{R}_{\mathcal{G}}, \\ \mathcal{E}_{b,R,e} &:= \Big\{ \|\widehat{\mathbf{b}}^R_e - \mathbf{b}^R_e\|_2 \le b_{R,e}, \quad \|\widehat{\boldsymbol{\beta}}^R_e - \boldsymbol{\beta}^{R,\circ}_e\|_2 \le r_{\beta,R,e}, \\ &\qquad\qquad \|\log \widehat{r}^R_e - \log r^R_e\|_{C^1} \le e^{\mathrm{fit}}_{R,e} \Big\}, \quad (R,e) \in \mathcal{R}_{\mathcal{G}} \times [m], \\ \mathcal{E}_{\mathrm{rat}} &:= \bigcap_{R \in \mathcal{R}_{\mathcal{G}}} \mathcal{E}_{G,R} \cap \bigcap_{(R,e) \in \mathcal{R}_{\mathcal{G}} \times [m]} \mathcal{E}_{b,R,e}. \end{aligned}$$

Conditionally on $\mathcal{E}_{\mathrm{ch}}$, $\mathbb{P}(\mathcal{E}_{\mathrm{rat}} \mid \mathcal{F}_{\mathrm{ch}}) \ge 1 - \delta_G - \delta_b$. Section L.5.2 controls input representation through $\mathcal{E}_{\mathrm{io}}$ and interval arithmetic through $\mathcal{E}_{\mathrm{num}}$; together they extend $e^{\mathrm{fit}}_{R,e}$ to the total ratio error $e^{\ell}_{R,e}$.

Proof. For every active set and ordered Gram coordinate, apply the entrywise Hoeffding calculation in equation 148 with common log factor $\log(2\overline{M}_G/\delta_G)$. The number of entries is at most $\overline{M}_G$, so their union bound costs at most δ_G. Taking the Frobenius norm separately at each active set gives equation 217. Repeating the coordinatewise argument for all intervention moments costs at most δ_b and gives equation 218. Both union bounds remain valid under dependence among the reused factual Gram matrices and environment moments.

For each (R, e), Weyl's inequality gives the empirical Gram floor $\underline{\lambda}_{G,R,p} - g_R > 0$. The exact normal-equation subtraction equation 150 then gives equation 219. The feature envelopes and projection tails give the ratio-value radius in equation 220. Corollary L.12 gives equation 221. For a fixed (R, e), write $r = r^R_e$ and $\widetilde{r} = \widetilde{r}^R_e$ for the population ratios in the true and fitted charts. Let $\widehat{r}$ be the ideal fitted polynomial, r^{ex} the exact solve on represented moments, and $r^{\sharp}$ the represented polynomial. On $\mathcal{E}_{\mathrm{io}} \cap \mathcal{E}_{\mathrm{num}}$, the error is bounded by four terms:

$$\begin{aligned} \|\log r^{\sharp} - \log r\|_{C^1} &\le \|\log r^{\sharp} - \log r^{\mathrm{ex}}\|_{C^1} + \|\log r^{\mathrm{ex}} - \log \widehat{r}\|_{C^1} \\ &\quad + \|\log \widehat{r} - \log \widetilde{r}\|_{C^1} + \|\log \widetilde{r} - \log r\|_{C^1} \\ &\le \bar{\epsilon}^{\ell,\mathrm{num}}_R + \bar{\epsilon}^{\ell,\mathrm{io}}_R + e^{\mathrm{fit}}_{R,e} = e^{\ell}_{R,e}. \end{aligned} \tag{222}$$

The first two terms are controlled by the numerical and representation bounds in Section L.5.2. The last two combine the ratio fit with Assumption L.16. □

Peeling requires non-sink responses to remain visible on the queried inner cube. Let π_R^η be uniform probability measure on $\mathcal{U}_R$. We average the full population CDF scores under this law; the error from truncating CDF integration is already included in the approximation bound. Thus query restriction changes where a score is averaged, while CDF truncation changes the function being estimated. Define the population scores on this cube by

$$\begin{aligned} I_{j,e}^R(\eta) &:= \pi_R^\eta\left[(C_{j,e}^R - u_j)^2 + \|\nabla C_{j,e}^R - \mathbf{e}_j\|_2^2\right], \\ F_{j,ef}^R(\eta) &:= \pi_R^\eta\left[\frac{\|\nabla C_{j,e}^R - (K_{j,e}^R/K_{j,f}^R)\nabla C_{j,f}^R\|_2^2}{1+\|\nabla C_{j,e}^R\|_2^2}\right], \\ s_j^R(\eta) &:= \max\left\{\max_{e:\sigma_{ej}=0} I_{j,e}^R(\eta), \max_{e<f:\sigma_{ej}=\sigma_{fj}=1} F_{j,ef}^R(\eta)\right\}. \end{aligned} \tag{223}$$

Empty maxima equal zero. The inner-cube margin is

$$s_j^R(\eta) = 0 \quad \text{for every true sink,} \qquad s_j^R(\eta) \geq \Delta_{\mathrm{sink},R}^\eta > 0 \quad \text{for every non-sink.} \tag{224}$$

This separation holds uniformly at the selected trim level η.

Proposition L.27 (Uniform sink-score calibration)**.** *Condition on $\mathcal{F}_{\mathrm{fit}}$ and the preceding chart and ratio events. Let $\widetilde{\mathbf{U}}_1, \dots, \widetilde{\mathbf{U}}_{n_{\mathrm{cal}}}$ be independent factual queries in the fitted residual chart. A single full-dimensional acceptance rule supplies every active-set projection:*

$$B_s := \mathbf{1}\{\widetilde{\mathbf{U}}_s \in \mathcal{U}_{[d]}\}, \qquad N_{\mathrm{cal}} := \sum_{s=1}^{n_{\mathrm{cal}}} B_s, \qquad \widehat{\pi}_R := \frac{1}{N_{\mathrm{cal}}} \sum_{s:B_s=1} \delta_{\widetilde{\mathbf{U}}_{s,R}} \ (N_{\mathrm{cal}} > 0).$$

For $N_{\mathrm{cal}} = 0$, define the statistical event as the whole sample space because the algorithm abstains before scoring. On $\mathcal{E}_{\mathrm{io}}$, $N_{\mathrm{cal}}^\sharp = N_{\mathrm{cal}}$. Let $\widetilde{\pi}_{[d]}$ be the conditional law of $\widetilde{\mathbf{U}}$ given membership in $\mathcal{U}_{[d]}$, and define the active-set query law by projection,

$$\widetilde{\pi}_R := (\mathrm{pr}_R)_\# \widetilde{\pi}_{[d]}.$$

Normalized Lebesgue measure on $\mathcal{U}_{[d]}$ is a product measure, so its R-projection is exactly π_R^η.

Suppose the full estimated-chart factual density obeys

$$\|\widetilde{p}_{0,[d]} - 1\|_{\infty;\mathcal{U}_{[d]}} \leq \delta_{\pi,d} < 1. \tag{225}$$

Then total variation contracts under every coordinate projection, and hence

$$\mathrm{TV}(\widetilde{\pi}_R, \pi_R^\eta) \leq \mathrm{TV}(\widetilde{\pi}_{[d]}, \pi_{[d]}^\eta) \leq \frac{\delta_{\pi,d}}{1-\delta_{\pi,d}} =: \omega_d \tag{226}$$

simultaneously for every $R \subseteq [d]$. For coupled transport one may take $\delta_{\pi,d} = \Delta_{0,0}$ from equation 172; for block-local transport one may use the corresponding Jacobian-determinant error on the full chart. Here the leakage bound is instantiated at $R = [d]$.

Let $B_{\mathrm{score},R}$ and $C_{\mathrm{score},R}$ be the constants from Proposition L.23; e_R^C and e_R^K are the two propagated errors in equation 207–equation 208. Assume for every $R \in \mathcal{R}_{\mathcal{G}}$ that

$$n_R^\eta > 0, \qquad E_0|_R < \frac{n_R^\eta}{2}, \qquad e_R^C \leq 1, \qquad e_R^K \leq \frac{k_{-,R}}{2},$$

with all numerical bounds fixed before $\mathcal{D}_{\mathrm{cal}}$ is observed. For $N_{\mathrm{cal}} > 0$, group the approximation, representation, calibration, and query-transport contributions as

$$\begin{aligned} r_R^{\mathrm{det}} &:= C_{\mathrm{score},R} e_R^C + \bar{\epsilon}_R^{\mathrm{s,num}} + \bar{\epsilon}_R^{\mathrm{s,io}} + B_{\mathrm{score},R}\omega_d, \\ r_R^{\mathrm{cal}} &:= B_{\mathrm{score},R}\sqrt{\frac{\log(2\max\{1, \overline{M}_{\mathrm{cal}}\}/\delta_{\mathrm{cal}})}{2N_{\mathrm{cal}}}}, \\ e_R^{\mathrm{s}} &:= r_R^{\mathrm{det}} + r_R^{\mathrm{cal}}. \end{aligned} \tag{227}$$

For $a \in \mathcal{J}_R$, let $h^{\mathrm{fit}}_{R,a}$ be the corresponding exact identity or foliation integrand formed from the represented polynomial $\widehat{r}^{R,\sharp}_e$ through its exact truncated CDF and density. The chart and ratio samples determine these functions and their input and numerical bounds before calibration. Conditional on $\mathcal{F}_{\mathrm{fit}}$, the fits satisfying those bounds form a fixed family. The calibration event requires each sample average to be close to its expectation:

$$\begin{aligned}\mathcal{E}_{\mathrm{cal}} := \{N_{\mathrm{cal}} = 0\} \\ \cup \left\{ N_{\mathrm{cal}} > 0, \quad \left|(\widehat{\pi}_R - \widetilde{\pi}_R)[h^{\mathrm{fit}}_{R,a}]\right| \le r^{\mathrm{cal}}_R \quad \substack{\text{for every } R\in\mathcal{R}_{\mathcal{G}} \text{ satisfying the ratio-stage} \\ \text{input and numerical bounds, and } a\in\mathcal{J}_R} \right\}.\end{aligned} \tag{228}$$

Then $\mathbb{P}(\mathcal{E}_{\mathrm{cal}} \mid \mathcal{F}_{\mathrm{fit}}) \ge 1 - \delta_{\mathrm{cal}}$. Moreover, on $\mathcal{E}_{\mathrm{cal}}$, whenever both the representation and numerical conditions hold for a required active set R, simultaneously for every $j \in R$,

$$\left|\check{s}^R_j - s^R_j(\eta)\right| \le e^s_R. \tag{229}$$

Proof. Put $v_d := (1-2\eta)^d$ and let $p_{\mathrm{acc},d} := \int_{\mathcal{U}_{[d]}} \widetilde{p}_{0,[d]}(u)\, \mathrm{d}u$. Equation 225 gives

$$|p_{\mathrm{acc},d} - v_d| \le v_d \delta_{\pi,d}, \qquad p_{\mathrm{acc},d} \ge v_d(1-\delta_{\pi,d}) > 0.$$

The corresponding conditional density is $\widetilde{p}_{0,[d]}/p_{\mathrm{acc},d}$, whereas the full inner-cube uniform density is $1/v_d$. Therefore,

$$\begin{aligned}\int_{\mathcal{U}_{[d]}} \left| \frac{\widetilde{p}_{0,[d]}}{p_{\mathrm{acc},d}} - \frac{1}{v_d} \right| \mathrm{d}u &= \frac{1}{v_d p_{\mathrm{acc},d}} \int_{\mathcal{U}_{[d]}} |v_d \widetilde{p}_{0,[d]} - p_{\mathrm{acc},d}|\, \mathrm{d}u \\ &\le \frac{v_d^2 \delta_{\pi,d} + v_d |v_d - p_{\mathrm{acc},d}|}{v_d p_{\mathrm{acc},d}} \le \frac{2\delta_{\pi,d}}{1-\delta_{\pi,d}}.\end{aligned}$$

Dividing by two proves the full-dimensional total-variation bound. Applying the data processing inequality to the measurable map pr_R gives the first inequality in equation 226. Since every score integrand lies in $[0, B_{\mathrm{score},R}]$, changing the projected query law costs at most $B_{\mathrm{score},R}\omega_d$.

Conditionally on $N_{\mathrm{cal}} = n > 0$, the n accepted full vectors are i.i.d. from $\widetilde{\pi}_{[d]}$; their R-projections are therefore i.i.d. from $\widetilde{\pi}_R$. Hoeffding's inequality applies with the observed denominator n. Union-bound it over the deterministic family of at most M_{cal} scalar integrands and then average over the binomial count N_{cal}. On $N_{\mathrm{cal}} = 0$ the event holds by definition and the estimator abstains. This proves the radius in equation 227 and the probability claim for equation 228 while retaining the randomness of the accepted count.

On $\mathcal{E}_{\mathrm{cal}}$, fix an active set whose representation and numerical bounds hold. For one component score, let h be its population integrand and $\widehat{h} = h^{\mathrm{fit}}_{R,a}$ the exact integrand from the represented polynomial. Substituting the returned CDF, density, and gradient values into the score formula gives the exact integrand $\check{h}$. Its exact empirical average is $\widehat{s}$, and its numerical evaluation is the returned score $\check{s}$. For this calculation only, use $P, \widetilde{P}, P_n, P^{\sharp}_n$ for the uniform, learned, empirical, and represented query measures, respectively. With $\widehat{s} = P^{\sharp}_n[\widehat{h}]$, adding and subtracting the corresponding averages gives

$$\begin{aligned}|\check{s} - P[h]| &\le |\check{s} - \widehat{s}| + P^{\sharp}_n[|\check{h} - \widehat{h}|] + |(P^{\sharp}_n - P_n)[\widehat{h}]| \\ &\quad + |(P_n - \widetilde{P})[\widehat{h}]| + |\widetilde{P}[\widehat{h} - h]| + |(\widetilde{P} - P)[h]| \\ &\le \bar{\epsilon}^{s,\mathrm{num}}_R + \bar{\epsilon}^{s,\mathrm{io}}_R + r^{\mathrm{cal}}_R + C_{\mathrm{score},R} e^C_R + B_{\mathrm{score},R}\omega_d = e^s_R.\end{aligned}$$

The first two terms cover score evaluation and numerical CDF error; the third covers input representation. Their bounds are given in equation 212, equation 206, and equation 214. Conditional Hoeffding controls the fourth term. Score stability equation 191 and the query-law bound equation 226 control the last two. Finally, $|\max_a x_a - \max_a y_a| \le \max_a |x_a - y_a|$ transfers this componentwise bound to equation 229. □

Combine the three random events with the two verified error conditions:

$$\mathcal{E}_{\mathrm{stat}} := \mathcal{E}_{\mathrm{ch}} \cap \mathcal{E}_{\mathrm{rat}} \cap \mathcal{E}_{\mathrm{cal}}, \qquad \mathcal{E}_{\mathrm{peel}} := \mathcal{E}_{\mathrm{stat}} \cap \mathcal{E}_{\mathrm{io}} \cap \mathcal{E}_{\mathrm{num}}.$$

The probability budget applies to $\mathcal{E}_{\mathrm{stat}}$. $\mathcal{E}_{\mathrm{io}}$ and $\mathcal{E}_{\mathrm{num}}$ verify the stated representation and numerical bounds before a score is used.

Theorem L.28 (Exact sink decisions under simultaneous control)**.** *Suppose the chart and target-signature conditions recover the residual chart up to one signed permutation. On $\mathcal{E}_{\mathrm{peel}}$, $N_{\mathrm{cal}}^{\sharp} = N_{\mathrm{cal}}$. A zero count leads directly to abstention. For a positive count and active set R, let $t_R \geq 0$ be the sink tolerance and let $\Delta_{\mathrm{rej},R} \geq 0$ be the rejection gap. Define the upper- and lower-confidence-bound sets*

$$\widehat{\mathcal{S}}_R^{\mathrm{pass}} := \left\{ j \in R : \check{s}_j^R + e_R^s \leq t_R \right\},$$
$$\widehat{\mathcal{S}}_R^{\mathrm{reject}} := \left\{ j \in R : \max\{0, \check{s}_j^R - e_R^s\} > t_R + \Delta_{\mathrm{rej},R} \right\}.$$

Coordinates outside both sets are unresolved. Assume, for every active set $R \in \mathcal{R}_{\mathcal{G}}$,

$$2e_R^s \leq t_R, \qquad t_R + \Delta_{\mathrm{rej},R} + 2e_R^s < \Delta_{\mathrm{sink},R}^{\eta}. \tag{230}$$

Then the pass set is the set of true sinks of $\mathcal{G}[R]$, and the reject set is its complement; together they partition R.

Starting from $R_0 = [d]$, the recursion updates the active set only after the sample-count, error, and separation conditions pass; otherwise it abstains. Each update removes one true sink, so d updates complete the peeling order.

Proof. At an iteration whose representation and numerical conditions hold, $\mathcal{E}_{\mathrm{stat}}$ and equation 229 imply that a true sink has population score zero, so $\check{s}_j^R + e_R^s \leq 2e_R^s \leq t_R$. Thus every true sink passes. For a non-sink,

$$\max\{0, \check{s}_j^R - e_R^s\} \geq \Delta_{\mathrm{sink},R}^{\eta} - 2e_R^s > t_R + \Delta_{\mathrm{rej},R},$$

so every non-sink is rejected. This proves the exact partition.

It remains to check that the adaptive path stays within the sets covered by the simultaneous event. Initially, $R_0 = [d]$ is ancestor-closed. At step t, the exact partition selects a true sink j_t of $\mathcal{G}[R_t]$. Deleting this sink preserves ancestor-closure, so the same event already covers $R_{t+1} = R_t \setminus \{j_t\}$ whenever it is nonempty. The chart alignment stays fixed, and each selected node has a targeting environment because its signature is nonzero. Induction therefore gives d correct deletions and $R_d = \varnothing$. If a required condition fails earlier, the procedure abstains before deletion. □

Corollary L.29 (Peeling error probability and calibration count)**.** *Under the independent channel splits and stated failure-probability allocation,*

$$\mathbb{P}(\textit{the algorithm returns an incorrect peeling decision}) \leq \delta_{\mathrm{ch}} + \delta_G + \delta_b + \delta_{\mathrm{cal}}. \tag{231}$$

Let

$$L_{\mathrm{cal}} := \log\left(\frac{2\max\{1, \overline{M}_{\mathrm{cal}}\}}{\delta_{\mathrm{cal}}}\right), \qquad \tau_R := \min\left\{\frac{t_R}{2}, \frac{\Delta_{\mathrm{sink},R}^{\eta} - t_R - \Delta_{\mathrm{rej},R}}{2}\right\} - r_R^{\mathrm{det}}.$$

The model class, sample sizes, probability allocation, and fixed numerical bounds determine r_R^{det}, τ_R, and all thresholds before sampling. The observed accepted count enters only through r_R^{cal}. For $\tau_R > 0$, define the active-set threshold and its DAG-independent upper bound:

$$n_\star^{\mathcal{G}} := 1 + \left\lfloor \max_{R \in \mathcal{R}_{\mathcal{G}}} \frac{B_{\mathrm{score},R}^2 L_{\mathrm{cal}}}{2\tau_R^2} \right\rfloor, \qquad \overline{n}_\star := 1 + \left\lfloor \max_{R \in \mathcal{R}_{\mathrm{all}}} \frac{B_{\mathrm{score},R}^2 L_{\mathrm{cal}}}{2\tau_R^2} \right\rfloor,$$
$$\underline{p}_{\mathrm{acc}} := v_d(1 - \delta_{\pi,d}), \qquad n_\star^{\mathcal{G}} \leq \overline{n}_\star.$$

The first maximum ranges over active sets; the computable second maximum ranges over all nonempty subsets. If the stated error conditions hold on $\mathcal{E}_{\mathrm{stat}} \cap \{N_{\mathrm{cal}} \geq \overline{n}_\star\}$, then correct complete peeling has probability at least

$$1 - \delta_{\mathrm{ch}} - \delta_G - \delta_b - \delta_{\mathrm{cal}} - \overline{T}_{\mathrm{bin}}, \tag{232}$$

where the binomial lower-tail bound is

$$\overline{T}_{\mathrm{bin}} := \mathbf{1}\{\overline{n}_\star > n_{\mathrm{cal}}\} + \mathbf{1}\{1 \leq \overline{n}_\star \leq n_{\mathrm{cal}}\} \sum_{k=0}^{\overline{n}_\star - 1} \binom{n_{\mathrm{cal}}}{k} \underline{p}_{\mathrm{acc}}^k (1 - \underline{p}_{\mathrm{acc}})^{n_{\mathrm{cal}} - k}.$$

Its certified value is the outward upper interval endpoint in equation 215.

Proof. The event definitions and the tower property give

$$\begin{aligned}\mathbb{P}(\mathcal{E}_{\mathrm{stat}}^{\mathrm{c}}) &\le \mathbb{P}(\mathcal{E}_{\mathrm{ch}}^{\mathrm{c}}) + \mathbb{E}[\mathbf{1}_{\mathcal{E}_{\mathrm{ch}}}\mathbb{P}(\mathcal{E}_{\mathrm{rat}}^{\mathrm{c}} \mid \mathcal{F}_{\mathrm{ch}})] + \mathbb{E}[\mathbf{1}_{\mathcal{E}_{\mathrm{ch}}\cap\mathcal{E}_{\mathrm{rat}}}\mathbb{P}(\mathcal{E}_{\mathrm{cal}}^{\mathrm{c}} \mid \mathcal{F}_{\mathrm{fit}})] \\ &\le \delta_{\mathrm{ch}} + \delta_G + \delta_b + \delta_{\mathrm{cal}}. \end{aligned} \tag{233}$$

On $\mathcal{E}_{\mathrm{stat}}$, the rule verifies the representation and numerical conditions before deletion. Therefore an incorrect returned deletion lies in $\mathcal{E}_{\mathrm{stat}}^{\mathrm{c}}$, proving equation 231.

If $N_{\mathrm{cal}} \ge \overline{n}_\star \ge n_\star^{\mathcal{G}}$, then for every active set R,

$$\begin{aligned} r_R^{\mathrm{cal}} &= B_{\mathrm{score},R}\sqrt{\frac{L_{\mathrm{cal}}}{2N_{\mathrm{cal}}}} < \tau_R, \\ e_R^{\mathrm{s}} &= r_R^{\mathrm{det}} + r_R^{\mathrm{cal}} < \min\left\{\frac{t_R}{2}, \frac{\Delta_{\mathrm{sink},R}^{\eta} - t_R - \Delta_{\mathrm{rej},R}}{2}\right\}. \end{aligned}$$

This is stronger than both conditions in equation 230.

Conditionally on $\mathcal{F}_{\mathrm{fit}}$, $N_{\mathrm{cal}} \sim \mathrm{Bin}(n_{\mathrm{cal}}, p_{\mathrm{acc},d})$. On the chart density condition, $p_{\mathrm{acc},d} \ge \underline{p}_{\mathrm{acc}}$, and binomial lower tails are nonincreasing in their success probability. Hence, almost surely, $\mathbf{1}_{\mathcal{E}_{\mathrm{ch}}}\mathbb{P}(N_{\mathrm{cal}} < \overline{n}_\star \mid \mathcal{F}_{\mathrm{fit}}) \le \mathbf{1}_{\mathcal{E}_{\mathrm{ch}}}\overline{T}_{\mathrm{bin}}$. Taking expectations and combining this with equation 233 proves the unconditional complete-peeling bound equation 232. □

L.7 DAG RECOVERY FROM FRESH SAMPLES

Peeling recovers a coordinate map Y for the causal variables, up to componentwise monotone transformations. Its inverse $H = Y^{-1}$ reconstructs each node's noise from its own value and the values of its direct parents. As shown in equation 13, the off-diagonal derivative $\partial_{y_j} H_i$ vanishes for a non-parent and is nonzero somewhere for a direct parent. Squaring and averaging this derivative avoids cancellation between positive and negative values; the edge margin makes its magnitude detectable under the query law. The next proposition bounds the effect of estimating Y on this inverse derivative, after which fresh queries estimate the edge scores.

Proposition L.30 (Deterministic inverse-Jacobian perturbation)**.** *Let Ω_Y be the fixed common latent domain and suppose $Y, \widehat{Y} : \Omega_Y \to \mathbb{R}^d$ are injective C^1 maps onto their images. Suppose the assembled Jacobian obeys*

$$\inf_{u\in\Omega_Y} \sigma_{\min}(\mathrm{D}Y(u)) \ge \mu_Y > 0, \qquad \sup_{u\in\Omega_Y} \left\|\mathrm{D}\widehat{Y}(u) - \mathrm{D}Y(u)\right\|_{\mathrm{op}} \le \epsilon_{\mathrm{Jac}} < \mu_Y/2. \tag{234}$$

Then $\mathrm{D}\widehat{Y}(u)$ is invertible for every $u \in \Omega_Y$. Writing $H := Y^{-1}$ and $\widehat{H} := \widehat{Y}^{-1}$ on the respective images, the inverse function theorem gives

$$\mathrm{D}H(Y(u)) = [\mathrm{D}Y(u)]^{-1}, \qquad \mathrm{D}\widehat{H}(\widehat{Y}(u)) = [\mathrm{D}\widehat{Y}(u)]^{-1}.$$

Moreover,

$$\sup_{u\in\Omega_Y} \left\|\mathrm{D}\widehat{H}(\widehat{Y}(u)) - \mathrm{D}H(Y(u))\right\|_{\mathrm{op}} \le \frac{\epsilon_{\mathrm{Jac}}}{\mu_Y(\mu_Y - \epsilon_{\mathrm{Jac}})} =: \epsilon_{\mathrm{inv}}. \tag{235}$$

For any probability law ν on Ω_Y, define $\Delta_{\mathrm{edge}}(\nu) := \min_{j\to i} \|[\mathrm{D}H(Y(\cdot))]_{ij}\|_{L^2(\nu)}$.

If an estimated root-mean-square edge score approximates all corresponding $L^2(\nu)$ scores within ϵ_{gr} and $3\epsilon_{\mathrm{gr}} < \Delta_{\mathrm{edge}}(\nu)$, thresholding at $2\epsilon_{\mathrm{gr}}$ recovers the exact DAG.

Proof. Weyl's inequality gives $\sigma_{\min}(\mathrm{D}\widehat{Y}) \ge \mu_Y - \epsilon_{\mathrm{Jac}} > 0$. At each fixed u, the resolvent identity

$$(\mathrm{D}\widehat{Y})^{-1} - (\mathrm{D}Y)^{-1} = (\mathrm{D}\widehat{Y})^{-1}(\mathrm{D}Y - \mathrm{D}\widehat{Y})(\mathrm{D}Y)^{-1}$$

and submultiplicativity give

$$\begin{aligned}\|[\mathrm{D}\widehat{Y}(u)]^{-1} - [\mathrm{D}Y(u)]^{-1}\|_{\mathrm{op}} &\le \|[\mathrm{D}\widehat{Y}(u)]^{-1}\|_{\mathrm{op}}\|\mathrm{D}\widehat{Y}(u) - \mathrm{D}Y(u)\|_{\mathrm{op}}\|[\mathrm{D}Y(u)]^{-1}\|_{\mathrm{op}} \\ &\le (\mu_Y - \epsilon_{\mathrm{Jac}})^{-1}\epsilon_{\mathrm{Jac}}\mu_Y^{-1}. \end{aligned}$$

Taking the supremum proves equation 235. Nonedges have population score zero and estimated score at most ϵ_{gr}; edges have estimated score at least $\Delta_{\mathrm{edge}}(\nu) - \epsilon_{\mathrm{gr}} > 2\epsilon_{\mathrm{gr}}$. □

The general estimator completes the inverse-Jacobian construction with fresh graph queries. Appendix O defines the atomic estimator using direct moments, and Appendix O.7 reports its numerical evaluation. Both estimators first recover a sink order and then identify direct edges.

After peeling fixes the coordinate map, the graph step draws fresh queries from a finite product grid. Its error has three sources: map estimation, numerical evaluation, and finite query sampling. The graph step uses the remaining failure budget δ_{gr}.

Let $\mathcal{F}_{\mathrm{pre}}$ denote the sigma-field generated by all chart, ratio, calibration, and peeling quantities. Conditional on $\mathcal{F}_{\mathrm{pre}}$, the population map Y, its fitted counterpart $\widehat{Y}$, and the dyadic product-grid query design are fixed. Write $H = Y^{-1}$ and $\widehat{H} = \widehat{Y}^{-1}$ on their respective images. For $i \neq j$, define the population and fitted inverse-Jacobian fields

$$g_{ij}(u) := \big|[\mathrm{D}H(Y(u))]_{ij}\big|, \qquad \widehat{g}_{ij}(u) := \big|[\mathrm{D}\widehat{H}(\widehat{Y}(u))]_{ij}\big|,$$

Assumption L.31 (Dyadic product-grid approximation)**.** For every correct state after peeling, conditionally on $\mathcal{F}_{\mathrm{pre}}$,

$$\mathcal{Q}_{\mathrm{gr}} = \prod_{k=1}^{d} \mathcal{Q}_{\mathrm{gr},k} \subseteq \Omega_Y, \qquad \pi_{\mathrm{gr}} = \bigotimes_{k=1}^{d} \mathrm{Unif}(\mathcal{Q}_{\mathrm{gr},k}),$$

where every grid point is dyadic and the grid is $\mathcal{F}_{\mathrm{pre}}$-measurable. Uniformly over the probit–linear class and all correct states after peeling,

$$\max_{i\neq j} \big|\pi_{\mathrm{gr}}[g_{ij}^2] - \pi_Y[g_{ij}^2]\big| \leq \tau_{\mathrm{gr}} < \big(\Delta^{\pi}_{\mathrm{edge}}\big)^2.$$

Define $\Delta^{\mathrm{gr}}_{\mathrm{edge}} := \sqrt{\big(\Delta^{\pi}_{\mathrm{edge}}\big)^2 - \tau_{\mathrm{gr}}} > 0$. Graph recovery uses fresh fair bits independent of all fitted samples. For a coordinate grid of size L_k, it draws $b_k = \lceil \log_2 L_k \rceil$ bits, interprets them as an integer, rejects an integer at least L_k, and repeats; accepted coordinate is therefore exactly uniform on $\mathcal{Q}_{\mathrm{gr},k}$.

Assumption L.32 (Inverse-Jacobian graph stability)**.** For every correct state after peeling, Y and $\widehat{Y}$ are injective on Ω_Y and satisfy equation 234 with $\epsilon_{\mathrm{Jac}} < \mu_Y/2$. Let ϵ_{inv} be the resulting inverse-Jacobian error in equation 235. Outward evaluation produces values $\widehat{g}^{\mathrm{num}}_{ij}$ satisfying

$$\max_{i\neq j} \sup_{u\in\mathcal{Q}_{\mathrm{gr}}} \big|\widehat{g}^{\mathrm{num}}_{ij}(u) - \widehat{g}_{ij}(u)\big| \leq \epsilon_{\mathrm{gr,num}}.$$

A fixed constant B_{gr} bounds both g_{ij} and $\widehat{g}^{\mathrm{num}}_{ij}$ on the query grid. Set $\epsilon_{\mathrm{det}} := \epsilon_{\mathrm{inv}} + \epsilon_{\mathrm{gr,num}}$ and $s_{ij} := \|g_{ij}\|_{L^2(\pi_{\mathrm{gr}})}$. Assumptions L.15 and L.31 imply $s_{ij} = 0$ for nonedges and $s_{ij} \geq \Delta^{\mathrm{gr}}_{\mathrm{edge}}$ for edges.

Lemma L.33 (Independent graph-score event)**.** *Let $Q_1, \ldots, Q_{n_{\mathrm{gr}}}$ be conditionally independent draws from π_{gr}, generated after peeling fixes the coordinate map, and set*

$$\widehat{s}_{ij} := \left\{ \frac{1}{n_{\mathrm{gr}}} \sum_{t=1}^{n_{\mathrm{gr}}} \widehat{g}^{\mathrm{num}}_{ij}(Q_t)^2 \right\}^{1/2}.$$

Under Assumption L.32, the deterministic inverse-map and outward-evaluation bounds give

$$\max_{i\neq j} \sup_{u\in\mathcal{Q}_{\mathrm{gr}}} |\widehat{g}^{\mathrm{num}}_{ij}(u) - g_{ij}(u)| \leq \epsilon_{\mathrm{det}}. \tag{236}$$

With $M_{\mathrm{gr}} := \max\{d(d-1), 1\}$, define $r_{\mathrm{gr}} := B^2_{\mathrm{gr}} \sqrt{\log(2M_{\mathrm{gr}}/\delta_{\mathrm{gr}})/2n_{\mathrm{gr}}}$, $\epsilon_{\mathrm{gr}} := \epsilon_{\mathrm{det}} + \sqrt{r_{\mathrm{gr}}}$.

Then, conditionally on $\mathcal{F}_{\mathrm{pre}}$,

$$\mathcal{E}_{\mathrm{gr}} := \left\{ \max_{i\neq j} |\widehat{s}_{ij} - s_{ij}| \leq \epsilon_{\mathrm{gr}} \right\}, \qquad \mathbb{P}\big(\mathcal{E}^c_{\mathrm{gr}} \mid \mathcal{F}_{\mathrm{pre}}\big) \leq \delta_{\mathrm{gr}}.$$

If $3\epsilon_{\mathrm{gr}} < \Delta^{\mathrm{gr}}_{\mathrm{edge}}$, thresholding $\widehat{s}_{ij}$ at $2\epsilon_{\mathrm{gr}}$ identifies $j \to i$, hence entry (j,i) of the adjacency matrix, on this event.

Proof. Conditional Hoeffding and a union bound over the M_{gr} ordered pairs give, with conditional probability at least $1-\delta_{\mathrm{gr}}$,

$$\max_{i\neq j}\left|\frac{1}{n_{\mathrm{gr}}}\sum_{t=1}^{n_{\mathrm{gr}}} g_{ij}(Q_t)^2 - s_{ij}^2\right| \le r_{\mathrm{gr}}.$$

The inequality $|\sqrt{x}-\sqrt{y}| \le \sqrt{|x-y|}$ converts this to an L^2-score error of at most $\sqrt{r_{\mathrm{gr}}}$. The reverse triangle inequality for empirical L^2 norms and equation 236 add at most ϵ_{det}. Nonedges are therefore below ϵ_{gr}, whereas every edge is above $\Delta_{\mathrm{edge}}^{\mathrm{gr}} - \epsilon_{\mathrm{gr}} > 2\epsilon_{\mathrm{gr}}$. □

Define the selective graph output $\widehat{\mathbf{A}}_{\mathrm{out}} \in \{\perp\} \sqcup \{0,1\}^{d\times d}$ as follows. Any abstention in the chart, ratio, calibration, or peeling steps, or any failed graph-query conditioning or numerical condition, returns $\perp$. Otherwise threshold $\widehat{s}_{ij}$ at $2\epsilon_{\mathrm{gr}}$ and place that decision in $\widehat{\mathbf{A}}_{\mathrm{out},ji}$; return the resulting adjacency when it is acyclic and $\perp$ otherwise. Let $\mathcal{E}_{\mathrm{complete}}$ be the intersection of $\mathcal{E}_{\mathrm{stat}}$ and all sample-count, envelope, representation, numerical, and separation conditions in Theorem L.28.

Theorem L.34 (Finite-sample graph recovery with abstention)**.** *Let $\mathfrak{m}_\star$ satisfy the following four groups of conditions:*

(i) population identification and signal regularity in Assumptions K.1, L.1, L.14, and L.15;

(ii) approximation and estimation control in Assumptions L.5, L.10, and L.16;

(iii) the graph-query conditions in Assumptions L.31 and L.32, using the independent samples in equation 128; and

(iv) the chart, score, peeling, and edge separations in equation 127, equation 224, equation 230, and $3\epsilon_{\mathrm{gr}} < \Delta_{\mathrm{edge}}^{\mathrm{gr}}$, together with the numerical enclosures in Section L.5.2.

Let $\mathbf{A}_\star$ be the DAG of $\mathfrak{m}_\star$. Then the total output $\widehat{\mathbf{A}}_{\mathrm{out}} \in \{\perp\} \sqcup \{0,1\}^{d\times d}$ satisfies

$$\begin{aligned}\mathbb{P}\Big(\widehat{\mathbf{A}}_{\mathrm{out}} \in \{\mathbf{A}_\star, \perp\}\Big) &\ge 1 - (\delta_{\mathrm{ch}} + \delta_G + \delta_b + \delta_{\mathrm{cal}} + \delta_{\mathrm{gr}}) \\ &\ge 1 - \delta. \end{aligned} \tag{237}$$

On $\mathcal{E}_{\mathrm{complete}} \cap \mathcal{E}_{\mathrm{gr}}$, the output is $\mathbf{A}_\star$.

Proof. The chart, ratio, and calibration sigma-fields are nested in split order. Proposition L.26, Proposition L.27, and the indicator-weighted tower argument in Corollary L.29 bound $\mathcal{E}_{\mathrm{stat}}^{\mathrm{c}}$ by the first four allocations. On $\mathcal{E}_{\mathrm{stat}}$, a failed representation or numerical condition returns $\perp$; whenever the procedure continues, Theorem L.28 permits only true-sink deletions and determines the correct coordinate map after peeling.

Conditional on that state, the fair-bit sampler in Assumption L.31 gives the exact product-grid law. Proposition L.30 and Lemma L.33 recover every directed edge on $\mathcal{E}_{\mathrm{gr}}$, whose conditional failure probability is at most δ_{gr}. Thus a wrong returned graph requires either failure of $\mathcal{E}_{\mathrm{stat}}$ or failure of the graph event after a correct peeling state. Conditional expectation bounds their combined probability by the five failure allocations, proving equation 237. On $\mathcal{E}_{\mathrm{complete}} \cap \mathcal{E}_{\mathrm{gr}}$, all required checks pass and the output is $\mathbf{A}_\star$. □

M WEAK-RESPONSE LOWER BOUND

The upper guarantees require the response margin to dominate statistical and numerical uncertainty. The following two-point construction proves that this dependence is intrinsic: it satisfies the probit–linear model with atomic target signatures, fixed support, smooth mechanisms, uniform replacements, and a known probit residual chart, so the lower bound holds even with oracle access to the chart, signatures, and nuisance parameters, leaving response strength as the sole unknown.

The design fixes sample sizes, environment labels, and data splits before any observation is read. Chart-environment rows are atomic factual-counterfactual pairs; ratio rows are the corresponding

marginal counterfactuals. Let $\mathcal{D}_{\mathrm{rep}}$ denote the represented split-labelled data. An allowed estimator is any Markov kernel

$$K : \mathcal{D}_{\mathrm{rep}} \rightsquigarrow \{\perp\} \sqcup \{0,1\}^{2\times 2}, \tag{238}$$

including all parameter-independent algorithmic randomness. Its input is exactly $\mathcal{D}_{\mathrm{rep}}$; the response strength, population conditional CDFs, population Jacobians, and graph remain latent. The number of distinct informative environment-two units is

$$N_2 := n_2^{\mathrm{ch}} + n_2^{\mathrm{rat}}. \tag{239}$$

Each physical unit contributes one independent likelihood factor.

Let $d = m = 2$, $S_1 = \{1\}$, and $S_2 = \{2\}$. For $|\upsilon| \le 1/4$, define

$$F_1^{\upsilon}(u) = u_1, \qquad F_2^{\upsilon}(u) = u_2 + \upsilon(2u_1 - 1)(u_2^2 - u_2).$$

Use uniform replacements on $(0,1)$ and decoder

$$g_{\upsilon}(\mathbf{z}) = \Phi^{-1}((F^{\upsilon})^{-1}(\mathbf{z})). \tag{240}$$

Then the factual observation is $\mathbf{X} = \Phi^{-1}(\mathbf{U})$ for every υ, and its residual chart is $q(\mathbf{x}) = \Phi(\mathbf{x})$, the probit-linear model with $\mathbf{A}_{\mathrm{mix}} = \mathbf{I}_2$ and $\boldsymbol{\mu} = \mathbf{0}$.

Theorem M.1 (Selective Le Cam tradeoff)**.** *Fix $0 < \upsilon \le 1/4$ and let*

$$\mathbf{A}_0 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{A}_{\upsilon} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

For $b \in \{0, \upsilon\}$, define the disjoint abstention, correct-return and wrong-return events

$$\mathcal{B}_b := \{\widehat{\mathbf{A}}_{\mathrm{out}} = \perp\}, \quad \mathcal{C}_b := \{\widehat{\mathbf{A}}_{\mathrm{out}} = \mathbf{A}_b\}, \quad \mathcal{W}_b := \{\widehat{\mathbf{A}}_{\mathrm{out}} \notin \{\mathbf{A}_b, \perp\}\}.$$

For every estimator of the form given in equation 238, put

$$V_{N_2,\upsilon} := \frac{\upsilon\sqrt{N_2}}{6\sqrt{1-\upsilon^2}}. \tag{241}$$

Then

$$\sum_{b\in\{0,\upsilon\}} [\mathbb{P}_b(\mathcal{B}_b) + \mathbb{P}_b(\mathcal{W}_b)] \ge 1 - V_{N_2,\upsilon}, \tag{242}$$

and in particular

$$\max_{b\in\{0,\upsilon\}} \mathbb{P}_b(\widehat{\mathbf{A}}_{\mathrm{out}} \ne \mathbf{A}_b) \ge \frac{1 - V_{N_2,\upsilon}}{2}. \tag{243}$$

If $\max_b \mathbb{P}_b(\mathcal{W}_b) \le \delta_{\mathrm{s}}$, then

$$\max_b \mathbb{P}_b(\mathcal{B}_b) \ge \left[\frac{1 - V_{N_2,\upsilon}}{2} - \delta_{\mathrm{s}}\right]_+. \tag{244}$$

Conversely, if $\max_b \mathbb{P}_b(\mathcal{B}_b) \le \alpha$, then

$$\max_b \mathbb{P}_b(\mathcal{W}_b) \ge \left[\frac{1 - V_{N_2,\upsilon}}{2} - \alpha\right]_+. \tag{245}$$

All distinguishability, and hence the bound, depends exclusively on the N_2 environment-two units. Selective abstention is therefore the mechanism that reconciles sound output with weak response.

Proof. The derivative in the own-noise coordinate is $\partial_{u_2} F_2^{\upsilon} = 1 + \upsilon(2u_1 - 1)(2u_2 - 1)$. Because $|(2u_1 - 1)(2u_2 - 1)| < 1$ and $|\upsilon| \le 1/4$, $3/4 < \partial_{u_2} F_2^{\upsilon} < 5/4$. Also $F_2^{\upsilon}(u_1, 0+) = 0$ and $F_2^{\upsilon}(u_1, 1-) = 1$, so the scalar map is an increasing diffeomorphism of $(0,1)$ onto itself. Its parent derivative is $\partial_{u_1} F_2^{\upsilon} = 2\upsilon(u_2^2 - u_2)$, which is identically zero only at $\upsilon = 0$. Thus the two graphs differ.

The decoder in equation 240 cancels the structural map in the factual channel: if $\mathbf{Z} = F^{\upsilon}(\mathbf{U})$, then

$$\mathbf{X} = g_{\upsilon}(\mathbf{Z}) = \Phi^{-1}(\mathbf{U}), \qquad q(\mathbf{X}) = \Phi(\mathbf{X}) = \mathbf{U},$$

so the factual law P_0 is the same for both parameters. In residual coordinates, an environment-one atomic pair is $(U_1, U_2; W_1, U_2)$ and is parameter free. An environment-two atomic pair is $(U_1, U_2; U_1, V_2)$, where $F_2^{\upsilon}(U_1, V_2) = W_2$.

For environment 2, the mixed map from residuals to intrinsic coordinates is

$$\mathsf{T}_2^{\upsilon}(u_1, u_2) = (u_1, F_2^{\upsilon}(u_1, u_2)), \qquad \det \mathrm{D}\mathsf{T}_2^{\upsilon} = \partial_{u_2} F_2^{\upsilon}.$$

The source density in mixed coordinates is uniform because the replacement is uniform. The change-of-variables formula therefore gives

$$\rho_{\upsilon}(u_1, u_2, v_2) = \det \mathrm{D}\mathsf{T}_2^{\upsilon}(u_1, v_2) = 1 + \upsilon(2u_1 - 1)(2v_2 - 1),$$

which is positive and integrates to one; the factual u_2 coordinate is an independent uniform ancillary. Thus this is the likelihood ratio of the full atomic pair. One complete pair contributes this single likelihood ratio. Integrating out u_2 gives the same ratio for the environment-two ratio marginal.

For the exact information calculation, put $x = 2u_1 - 1$, $y = 2v_2 - 1$, and $t = \upsilon xy$. Under one null environment-two pair, x and y are independent uniform variables on $(-1, 1)$ and the likelihood ratio is $1 + t$. Absolute convergence for $|t| \le 1/4$ and vanishing odd moments give

$$D_{\upsilon} := D_{\mathrm{KL}}(P_{2,\upsilon}^{\mathrm{pair}} \| P_{2,0}^{\mathrm{pair}}) = \sum_{k=1}^{\infty} \frac{\upsilon^{2k}}{(2k)(2k-1)(2k+1)^2}.$$

Indeed, the even coefficient of order $2k$ in $(1+t)\log(1+t) - t$ is $1/((2k)(2k-1))$, and $\mathbb{E}[(xy)^{2k}] = (2k+1)^{-2}$. Hence

$$\frac{\upsilon^2}{18} \le D_{\upsilon} \le \frac{\upsilon^2}{18(1-\upsilon^2)} \le \frac{8\upsilon^2}{135}.$$

Integrating out the ancillary factual u_2 leaves the same likelihood ratio, so one environment-two ratio marginal also contributes D_{υ}.

The sample blocks form an independent product subexperiment. Exactly the N_2 distinct chart-pair or ratio units in equation 239 contribute information; all other block laws agree. Product additivity, followed by finite represented acquisition and every parameter-independent randomized post-processing step, therefore gives by data processing and Pinsker

$$\mathrm{TV}(\mathbb{P}_{\upsilon}, \mathbb{P}_0) \le \sqrt{\frac{N_2 D_{\upsilon}}{2}} \le \frac{\upsilon\sqrt{N_2}}{6\sqrt{1-\upsilon^2}} = V_{N_2,\upsilon}.$$

Because $\mathbf{A}_0 \neq \mathbf{A}_{\upsilon}$, the correct-return events $\mathcal{C}_0$ and $\mathcal{C}_{\upsilon}$ are disjoint. The variational definition of total variation yields $\mathbb{P}_0(\mathcal{C}_0) \le \mathbb{P}_{\upsilon}(\mathcal{C}_0) + V_{N_2,\upsilon}$. Adding $\mathbb{P}_{\upsilon}(\mathcal{C}_{\upsilon})$ and using disjointness under $\mathbb{P}_{\upsilon}$ bounds the sum of the two correct-return probabilities by $1 + V_{N_2,\upsilon}$. Taking complements proves equation 242. Averaging proves equation 243; subtracting the two wrong-return probabilities or the two abstention probabilities and averaging proves equation 244 and equation 245, respectively. □

Corollary M.2 (Value-only CFP margin consequence)**.** *For $\upsilon \neq 0$, node 1 has child 2, while environment 2 targets node 2. The conditional CDF of node 1 is*

$$C_{1,2}(u_1, u_2) = u_1 + \upsilon(u_1^2 - u_1)(2u_2 - 1).$$

Consequently,

$$\kappa_{\upsilon} := \|C_{1,2} - u_1\|_{L^2((0,1)^2)} = \frac{|\upsilon|}{\sqrt{90}}. \tag{246}$$

Thus any uniformly reliable exact-DAG theorem requires $N_2\kappa_{\upsilon}^2$ to be bounded away from zero; this is the value-only counterpart of the value-plus-gradient score.

Proof. The normalizer along u_1 is

$$\mathcal{N}_{1,2}(u_2) = \int_0^1 [1 + \upsilon(2t-1)(2u_2-1)]\,\mathrm{d}t = 1 + \upsilon(2u_2-1)\int_0^1 (2t-1)\,\mathrm{d}t = 1.$$

Thus $K_{1,2} = \rho_\upsilon$ and integrating the conditional density from zero to u_1 gives

$$C_{1,2}(u_1, u_2) = \int_0^{u_1} [1 + \upsilon(2t-1)(2u_2-1)]\,\mathrm{d}t = u_1 + \upsilon(u_1^2 - u_1)(2u_2 - 1).$$

The squared L^2 deviation factorizes:

$$\|C_{1,2} - u_1\|_{L^2}^2 = \upsilon^2 \left[\int_0^1 (u_1^2 - u_1)^2\,\mathrm{d}u_1\right]\left[\int_0^1 (2u_2-1)^2\,\mathrm{d}u_2\right] = \upsilon^2\left(\frac{1}{30}\right)\left(\frac{1}{3}\right) = \frac{\upsilon^2}{90}.$$

Taking the square root yields equation 246; substituting $\upsilon^2 = 90\kappa_\upsilon^2$ into equation 241 makes the required $N_2\kappa_\upsilon^2$ scaling explicit. □

Proposition M.3 (Sink-score identity)**.** *Let* $U_\eta \sim \mathrm{Unif}([\eta, 1-\eta])$ *and define*

$$\begin{aligned} M_{x,\eta} &:= \mathbb{E}[(2U_\eta - 1)^2], \\ M_{a,\eta} &:= \mathbb{E}[(U_\eta^2 - U_\eta)^2], \\ \Gamma_\eta &:= M_{a,\eta}M_{x,\eta} + (M_{x,\eta})^2 + 4M_{a,\eta} > 0. \end{aligned}$$

At $R = \{1, 2\}$*, the population score in equation 223 is exactly* $\Delta_\upsilon^{\mathrm{sink}} := s_1^R(\eta) = \upsilon^2\Gamma_\eta$.

Consequently,

$$V_{N_2,\upsilon} = \sqrt{\frac{N_2\Delta_\upsilon^{\mathrm{sink}}}{36\Gamma_\eta(1 - \Delta_\upsilon^{\mathrm{sink}}/\Gamma_\eta)}}.$$

Uniform nontrivial recovery on this score requires $N_2\Delta_\upsilon^{\mathrm{sink}} = \Omega(1)$.

Proof. The identity residual and gradient residual are

$$\begin{aligned} C_{1,2} - u_1 &= \upsilon(u_1^2 - u_1)(2u_2 - 1), \\ \nabla C_{1,2} - \mathbf{e}_1 &= \upsilon\big((2u_1 - 1)(2u_2 - 1), 2(u_1^2 - u_1)\big). \end{aligned}$$

Integrating their squared value and squared Euclidean norm under two independent copies of U_η yields the three terms in Γ_η. Node 1 has only this non-target identity requirement, so its target-only foliation maximum is empty. Substituting $\upsilon^2 = \Delta_\upsilon^{\mathrm{sink}}/\Gamma_\eta$ into equation 241 proves the displayed identity. □

Remark M.4 (Response-margin scaling). Theorem M.1 gives the value-only scaling $1/\kappa_\upsilon^2$ and Proposition M.3 gives the corresponding score requirement $N_2\Delta_\upsilon^{\mathrm{sink}} = \Omega(1)$.

N Robust selective recovery

A second, distributional selective estimator forms an MMD plausible set around the observed channel laws, so a single channelwise radius can absorb both distribution shift and noisy pairs. Let $\widetilde{P}_c$ be the actual observed channel law and suppose a fixed ideal model $\mathfrak{m}_\star \in \mathfrak{M}_B$ satisfies a known channel budget

$$\left\|\mu_c^\varphi(\widetilde{P}_c) - \mu_c^\varphi(P_c^{\mathfrak{m}_\star})\right\|_{\mathcal{H}_c} \le b_c. \tag{247}$$

Let $a_c(\delta)$ be the sampling radius in equation 109 associated with channel c, and set $r_c = b_c + a_c(\delta)$.

Definition N.1 (Plausible model set and selective output)**.** Define

$$\widehat{\mathfrak{M}} = \left\{\mathfrak{m} \in \mathfrak{M}_B : \left\|\mu_c^\varphi(P_c^{\mathfrak{m}}) - \widehat{\mu^\varphi}_c\right\|_{\mathcal{H}_c} \le r_c \text{ for every } c\right\}.$$

For a discrete target $\vartheta_{\mathrm{disc}}$ such as the DAG, output

$$\widehat{\vartheta}_{\mathrm{FC}} = \begin{cases} v, & \{\vartheta_{\mathrm{disc}}(\mathfrak{m}) : \mathfrak{m} \in \widehat{\mathfrak{M}}\} = \{v\}, \\ \perp, & \text{otherwise.} \end{cases} \tag{248}$$

An empty plausible set therefore outputs $\perp$. For a continuous quotient target, replace singleton agreement by a fixed maximum diameter and output the entire target confidence set.

Theorem N.2 (Selective soundness and robust completeness). *Suppose samples are i.i.d. from $\widetilde{P}_c$ in every channel and equation 247 holds for $\mathfrak{m}_\star$. Then, with probability at least $1-\delta$:*

(i) $\mathfrak{m}_\star \in \widehat{\mathfrak{M}}$;

(ii) every non-$\perp$ output in equation 248 equals $\vartheta_{\mathrm{disc}}(\mathfrak{m}_\star)$;

(iii) if every $\mathfrak{m}$ with $\vartheta_{\mathrm{disc}}(\mathfrak{m}) \neq \vartheta_{\mathrm{disc}}(\mathfrak{m}_\star)$ has at least one channel c satisfying

$$\|\mu_c^\varphi(P_c^{\mathfrak{m}}) - \mu_c^\varphi(P_c^{\mathfrak{m}_\star})\|_{\mathcal{H}_c} > 2r_c, \tag{249}$$

then the procedure returns the exact target.

For an outer approximation $\widetilde{\mathfrak{M}} \supseteq \widehat{\mathfrak{M}}$, soundness remains valid when a return requires target unanimity over $\widetilde{\mathfrak{M}}$. The corresponding completeness guarantee follows when the outer approximation satisfies the two-sided sandwich and witness conditions in Theorem N.4.

Proof. Apply Lemma K.5 to the observed laws $\widetilde{P}_c$. On its simultaneous event,

$$\begin{aligned}\left\|\widehat{\mu^\varphi}_c - \mu_c^\varphi(P_c^{\mathfrak{m}_\star})\right\|_{\mathcal{H}_c} &\le \left\|\widehat{\mu^\varphi}_c - \mu_c^\varphi(\widetilde{P}_c)\right\|_{\mathcal{H}_c} + \left\|\mu_c^\varphi(\widetilde{P}_c) - \mu_c^\varphi(P_c^{\mathfrak{m}_\star})\right\|_{\mathcal{H}_c} \\ &\le a_c(\delta) + b_c = r_c.\end{aligned}$$

Thus $\mathfrak{m}_\star \in \widehat{\mathfrak{M}}$, proving (i). If the plausible target set is a singleton, it contains the true target and therefore equals it, proving (ii). For any $\mathfrak{m} \in \widehat{\mathfrak{M}}$, the triangle inequality gives

$$\|\mu_c^\varphi(P_c^{\mathfrak{m}}) - \mu_c^\varphi(P_c^{\mathfrak{m}_\star})\|_{\mathcal{H}_c} \le r_c + r_c = 2r_c$$

for every channel. Condition equation 249 therefore excludes all alternate targets, proving (iii). Replacing the exact set by a superset preserves soundness; completeness follows when the separation condition excludes every alternate target. □

N.1 Robustness across sample groups

The distributional estimator also covers the sample groups used by the constructive procedure and permits dependence between groups. Let $\mathcal{C}_{\mathrm{data}}$ index one factual chart group, two paired chart groups, three ratio groups, and one calibration group. Conditional on the sampling design, let $Y_{c,1}, \dots, Y_{c,n_c}$ be i.i.d. from Q_c within group c; the groups may be dependent. Let $\widehat{\mu^\varphi}_c^{\mathrm{obs}}$ be the mean embedding computed from the observed rows. For the true model $\mathfrak{m}_\star \in \mathfrak{M}_B$, assume that a fixed budget b_c covers both law shift and the discrepancy introduced by row corruption, acquisition, and numerical evaluation:

$$\|\mu_c^\varphi(Q_c) - \mu_c^\varphi(P_c^{\mathfrak{m}_\star})\|_{\mathcal{H}_c} + \left\|\widehat{\mu^\varphi}_c^{\mathrm{obs}} - \frac{1}{n_c}\sum_i \varphi_c(Y_{c,i})\right\|_{\mathcal{H}_c} \le b_c. \tag{250}$$

Proposition N.5 gives standard choices for the two terms.

For kernels with $\sup_x k_c(x,x) \le K_c$, choose $\delta_c \in (0,1)$ with $\sum_c \delta_c \le \delta$ and define

$$a_c := \sqrt{\frac{K_c}{n_c}} + \sqrt{\frac{2K_c \log(1/\delta_c)}{n_c}}, \qquad r_c := a_c + b_c. \tag{251}$$

Lemma N.3 (Embedding error across sample groups). *There is a simultaneous event $\mathcal{E}_{\mathrm{samp}}$ with $\mathbb{P}(\mathcal{E}_{\mathrm{samp}}) \ge 1-\delta$ on which*

$$\left\|\widehat{\mu^\varphi}_c^{\mathrm{obs}} - \mu_c^\varphi(P_c^{\mathfrak{m}_\star})\right\|_{\mathcal{H}_c} \le r_c \quad \textit{for every } c \in \mathcal{C}_{\mathrm{data}}. \tag{252}$$

Proof. For each group, bounded-kernel concentration gives empirical-embedding error at most a_c with probability $1-\delta_c$. A union bound permits arbitrary dependence between groups. Adding equation 250 gives equation 252. □

For a nonnegative vector $\mathbf{t}$, define

$$\widehat{\mathfrak{M}}(\mathbf{t}) := \left\{ \mathfrak{m} \in \mathfrak{M}_B : \left\| \mu_c^{\varphi}(P_c^{\mathfrak{m}}) - \widehat{\mu^{\varphi}}_c^{\mathrm{obs}} \right\|_{\mathcal{H}_c} \le t_c \text{ for every } c \right\}.$$

Fix an inflation vector $\boldsymbol{\rho} \ge \mathbf{0}$. Let $\mathcal{E}_{\mathrm{out}}$ denote successful construction of an outer set $\widetilde{\mathfrak{M}}$ satisfying

$$\widehat{\mathfrak{M}}(\mathbf{r}) \subseteq \widetilde{\mathfrak{M}} \subseteq \widehat{\mathfrak{M}}(\mathbf{r} + \boldsymbol{\rho}). \tag{253}$$

The selective output is

$$\widehat{\mathbf{A}}_{\mathrm{rob}} := \begin{cases} \mathbf{A}, & \mathcal{E}_{\mathrm{out}} \text{ holds and } \{\mathbf{A}(\mathfrak{m}) : \mathfrak{m} \in \widetilde{\mathfrak{M}}\} = \{\mathbf{A}\}, \\ \perp, & \text{otherwise.} \end{cases}$$

Theorem N.4 (Robust graph recovery across sample groups)**.** *Let $\mathbf{A}_\star$ be the adjacency matrix of $\mathfrak{m}_\star$. Then*

$$\mathbb{P}\{\widehat{\mathbf{A}}_{\mathrm{rob}} \in \{\mathbf{A}_\star, \perp\}\} \ge 1 - \delta. \tag{254}$$

If every alternate graph is separated in at least one sample group,

$$\mathbf{A}(\mathfrak{m}) \ne \mathbf{A}_\star \Longrightarrow \exists c : \quad \|\mu_c^{\varphi}(P_c^{\mathfrak{m}}) - \mu_c^{\varphi}(P_c^{\mathfrak{m}_\star})\|_{\mathcal{H}_c} > 2r_c + \rho_c, \tag{255}$$

then $\widehat{\mathbf{A}}_{\mathrm{rob}} = \mathbf{A}_\star$ on $\mathcal{E}_{\mathrm{samp}} \cap \mathcal{E}_{\mathrm{out}}$. Consequently,

$$\mathbb{P}(\widehat{\mathbf{A}}_{\mathrm{rob}} = \mathbf{A}_\star) \ge 1 - \delta - \mathbb{P}(\mathcal{E}_{\mathrm{samp}} \cap \mathcal{E}_{\mathrm{out}}^c).$$

Proof. On $\mathcal{E}_{\mathrm{samp}}$, Lemma N.3 places $\mathfrak{m}_\star$ in $\widehat{\mathfrak{M}}(\mathbf{r})$. The left inclusion in equation 253 therefore places its graph in the outer set, so every returned label equals $\mathbf{A}_\star$. This proves equation 254.

For $\mathfrak{m} \in \widetilde{\mathfrak{M}}$, the right inclusion and triangle inequality give

$$\|\mu_c^{\varphi}(P_c^{\mathfrak{m}}) - \mu_c^{\varphi}(P_c^{\mathfrak{m}_\star})\|_{\mathcal{H}_c} \le 2r_c + \rho_c$$

for every sample group. Condition equation 255 makes $\mathbf{A}_\star$ the unique graph label on $\mathcal{E}_{\mathrm{samp}} \cap \mathcal{E}_{\mathrm{out}}$, and the probability bound follows. □

Proposition N.5 (Useful corruption budgets)**.** *Under equation 107, write $\mu_\varphi(P) := \mathbb{E}_P[\varphi(X)] \in \mathcal{H}$ for the generic single-channel kernel mean embedding. Then:*

(a) If $Q = (1-\epsilon)P + \epsilon R$ is Huber contamination, then

$$\|\mu_\varphi(Q) - \mu_\varphi(P)\|_{\mathcal{H}} \le 2\epsilon\sqrt{K}.$$

(b) If at most k of n represented rows are replaced or swapped after the center sample is drawn, then the pathwise empirical-embedding displacement is at most

$$\frac{2k\sqrt{K}}{n}.$$

An atomic chart pair counts as one joint row.

(c) If $(X, \widetilde{X})$ is a coupling of P, Q and the feature map is L_φ-Lipschitz, then

$$\|\mu_\varphi(Q) - \mu_\varphi(P)\|_{\mathcal{H}} \le L_\varphi \mathbb{E}[d_{\mathcal{X}}(X, \widetilde{X})].$$

The bounds apply channelwise to pair swaps, measurement perturbations, or other fixed corruptions whenever their coupling or contamination fraction is known.

Proof. For (a), $\mu_\varphi(Q) - \mu_\varphi(P) = \epsilon(\mu_\varphi(R) - \mu_\varphi(P))$ and each mean embedding has norm at most $\sqrt{K}$. For (b), each replaced row changes the empirical embedding by at most $2\sqrt{K}/n$. For (c), Jensen's inequality gives

$$\left\| \mathbb{E}[\varphi(\widetilde{X}) - \varphi(X)] \right\|_{\mathcal{H}} \le \mathbb{E} \left\| \varphi(\widetilde{X}) - \varphi(X) \right\|_{\mathcal{H}} \le L_\varphi \mathbb{E}[d_{\mathcal{X}}(X, \widetilde{X})].$$

□

Corollary N.6 (Detection of sufficiently separated misspecification)**.** *On the sampling event, if for every $\mathfrak{m} \in \mathfrak{M}_B$ there is a channel c with*

$$\left\|\mu_c^{\varphi}(\widetilde{P}_c) - \mu_c^{\varphi}(P_c^{\mathfrak{m}})\right\|_{\mathcal{H}_c} > b_c + 2a_c(\delta), \tag{256}$$

then $\widehat{\mathfrak{M}} = \varnothing$ and the rule outputs $\perp$.

Proof. For the channel supplied by equation 256,

$$\begin{aligned}\left\|\widehat{\mu^{\varphi}}_c - \mu_c^{\varphi}(P_c^{\mathfrak{m}})\right\|_{\mathcal{H}_c} &\geq \left\|\mu_c^{\varphi}(\widetilde{P}_c) - \mu_c^{\varphi}(P_c^{\mathfrak{m}})\right\|_{\mathcal{H}_c} - \left\|\widehat{\mu^{\varphi}}_c - \mu_c^{\varphi}(\widetilde{P}_c)\right\|_{\mathcal{H}_c} \\ &> b_c + a_c(\delta) = r_c.\end{aligned}$$

Thus $\widehat{\mathfrak{M}} = \varnothing$. □

Proposition N.7 (Undetectable misspecification forces abstention)**.** *Let $\mathfrak{m}_0, \mathfrak{m}_1$ have different discrete targets. Suppose one full law Q_{data} of the complete represented data is admissible in both fixed corruption worlds, with group marginals $(\widetilde{P}_c)_c$ inside both budget balls:*

$$\left\|\mu_c^{\varphi}(\widetilde{P}_c) - \mu_c^{\varphi}(P_c^{\mathfrak{m}_b})\right\|_{\mathcal{H}_c} \leq b_c, \qquad b \in \{0,1\}, \quad c = 0, \ldots, m.$$

The conditional Markov kernel for all algorithmic randomness is shared by the two worlds. Uniform soundness therefore requires abstention with high probability under their common data law. More precisely, if a randomized procedure has probability at most δ of a wrong non-$\perp$ output under each admissible world, then

$$\mathbb{P}_{Q_{\mathrm{data}}}(\textit{output } \perp) \geq 1 - 2\delta. \tag{257}$$

Proof. The complete input law and the conditional algorithmic kernel are the same in both worlds. Let A_b be the event that the procedure outputs the target of $\mathfrak{m}_b$, and let A_o be the event of any other non-$\perp$ output. Soundness in world 0 gives $\mathbb{P}_{Q_{\mathrm{data}}}(A_1 \cup A_o) \leq \delta$; soundness in world 1 gives $\mathbb{P}_{Q_{\mathrm{data}}}(A_0 \cup A_o) \leq \delta$. Hence the total probability of returning a target is at most 2δ, proving equation 257. □

When sample groups may be dependent, a common full-law construction requires joint compatibility in addition to equality of the group marginals. Under a product sampling design, a common tuple of group laws induces the required Q_{data}.

Corollary N.6 and Proposition N.7 together characterize the robustness boundary: separation from the model class is detectable, whereas a corruption compatible with two legal targets requires selective abstention.

Two-node response-margin example. This two-node family makes the response threshold and robust radius concrete. Define

$$\mathcal{V} = \{0\} \cup [\upsilon_{\min}, \upsilon_{\max}], \qquad \mathbf{A}(0) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{A}(\upsilon) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \quad (\upsilon > 0). \tag{258}$$

The seven sample groups comprise one factual chart group, two paired chart groups, three ratio groups, and one calibration group. Each chart pair is one joint observation. Two groups have interaction mean $\upsilon/9$; the other groups have mean zero. For group c, use the kernel

$$k_c(x, y) = \psi_c(x)\psi_c(y) + \lambda \prod_{j=1}^{d_c} \frac{1 + \min(x_j, y_j)}{2}, \qquad K_c \geq 1 + \lambda,$$

where the interaction indices are fixed by the group definition. The strictly positive product-Brownian term is characteristic, while the scalar component retains the response margin. Use the radius $r_c = a_c + b_c$ from equation 251; Proposition N.5 supplies b_c for Huber contamination, row replacement, and bounded displacement.

The outer-set rule covers the continuous class in equation 258. Outward distance intervals certify exclusion from $\widehat{\mathfrak{M}}(\mathbf{r})$ when their lower endpoint exceeds r_c^2, and inclusion in $\widehat{\mathfrak{M}}(\mathbf{r} + \boldsymbol{\rho})$ when their

upper endpoint is below $(r_c + \rho_c)^2$. A unique graph label yields a graph; ambiguity yields $\perp$. The response parameter remains latent.

Theorem N.4 gives $\mathbb{P}\{\widehat{\mathbf{A}}_{\mathrm{rob}} \in \{\mathbf{A}_\star, \perp\}\} \geq 1 - \delta$; with the response separation in equation 255, it returns the true graph on $\mathcal{E}_{\mathrm{samp}} \cap \mathcal{E}_{\mathrm{out}}$.

O ATOMIC FINITE-SAMPLE RECOVERY

This appendix turns population CFP into a sound graph-or-abstention rule. Chart recovery, simultaneous response intervals, sink peeling, and an independent direct-parent query are covered by one common event; on that event, every returned DAG is the true DAG. Atomic interventions and an orthogonal chart make each comparison a bounded-moment decision while preserving the essential separation between marginal response and direct adjacency. The general fixed-support construction appears in Section L.

O.1 ATOMIC FINITE-SAMPLE MODEL AND SELECTIVE RULE

Let $\mathbf{U} = (U_1, \ldots, U_d)^\top$ have independent $\mathrm{Unif}(0,1)$ coordinates and let $\mathbf{Z} = F(\mathbf{U})$ be generated in a topological order by

$$Z_j = U_j + U_j(1 - U_j) \sum_{p \in \mathrm{pa}(j)} a_{pj}(2Z_p - 1). \tag{259}$$

The observation map is

$$\mathbf{X} = \mathbf{M}\Phi^{-1}(\mathbf{U}), \qquad \mathbf{M}^\top \mathbf{M} = \mathbf{I}_d, \tag{260}$$

where Φ is the standard-normal CDF applied coordinatewise and $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_d]$. Environment e_j replaces Z_j by an independent $\mathrm{Unif}(0,1)$ draw and leaves the other structural assignments unchanged. Every environment is labeled, and a pair shares the same factual residual draw before the replacement is made.

Six independent splits align the statistical conditioning with the three decisions: chart recovery, sink-order recovery, and direct-edge recovery. Write

$$\mathcal{D} = \big(\mathcal{D}_{\mathrm{ch},1}, \mathcal{D}_{\mathrm{ch},2}, \mathcal{D}_{\mathrm{ch},3}, \mathcal{D}_{\mathrm{rat}}, \mathcal{D}_{\mathrm{cal}}, \mathcal{D}_{\mathrm{gr}}\big). \tag{261}$$

Each split contains N factual observations and, for every atomic environment, N factual–interventional pairs; hence N denotes the sample size per split and channel. The three chart splits fit the residual directions, confirm them on new pairs, and certify held-out factual moments. The ratio and calibration splits provide independent response intervals, while the graph split is queried after the sink order is fixed. Let $\mathcal{F}_{\mathrm{ch}}$ be the sigma-field generated by the three chart samples.

Let P_0 be the factual residual law and P_{e_j} the law under the atomic replacement of node j. Set

$$W_k := 2U_k - 1, \qquad \widetilde{Z}_k := 2Z_k - 1, \qquad \gamma_{kj} := 3\mathbb{E}_{e_j}[W_k W_j] \quad (k \neq j). \tag{262}$$

The pairwise Legendre moments in $\mathbf{\Gamma} = (\gamma_{kj})$ summarize marginal response along directed paths. They are estimated once and restricted to each active set during recursion. Direct adjacency is instead encoded by the structural coefficients a_{pj} in equation 259 and recovered from the independent graph split.

Assumption O.1 (Finite-sample model)**.** The following conditions hold.

(i) The data satisfy equation 259– equation 260; the d labeled environments are atomic, the pairs are valid, and every replacement has nonzero excitation almost surely.

(ii) The DAG has maximum indegree two. In the canonical residual orientation, its direct coefficients satisfy $a_{pj} \in [a_{\min}, a_{\max}]$ and $\sum_{p \in \mathrm{pa}(j)} a_{pj} < 1$, where $0 < a_{\min} \leq a_{\max}$. Coordinate reflections may change coefficient signs but preserve their magnitudes and all graph decisions. Accordingly, the direct-coefficient set is $\mathcal{B}_a := \{0\} \cup [-a_{\max}, -a_{\min}] \cup [a_{\min}, a_{\max}]$.

(iii) The pair moments in equation 262 satisfy $|\gamma_{kj}| \leq \beta_{\max}$, with $\beta_{\min} := a_{\min}/3$. The estimator uses the separated response set

$$\mathcal{B}_\gamma := \{0\} \cup [-\beta_{\max}, -\beta_{\min}] \cup [\beta_{\min}, \beta_{\max}], \qquad 0 < \beta_{\min} \leq \beta_{\max}. \tag{263}$$

Strict-ancestor responses may satisfy $0 < |\gamma_{kj}| < \beta_{\min}$; adjacency is determined later by the direct moment system.

(iv) The six samples in equation 261 are independent. All model constants, interval allocations, and tie-breaking rules are fixed before these samples are read.

(v) For every node, the principal system indexed by its true parent set satisfies the fixed inverse-stability bound. The outward linear solve contains the true coefficient vector and its zero full-predecessor residual whenever the input moment intervals contain their population values.

Lemma O.2 (Pair-moment and direct-system identities)**.** *Define the factual cross-moment matrix and the atomic right-hand side by*

$$\mathbf{G}[q,p] := 3\mathbb{E}_0[W_q \widetilde{Z}_p], \qquad \mathbf{b}_j[q] := -9\mathbb{E}_{e_j}[W_q W_j].$$

In a source-to-sink order, $\mathbf{G}[p,p] = 1$, $\mathbf{G}[q,p] = 0$ *for* $q \notin \mathrm{an}(p) \cup \{p\}$, *and*

$$\mathbf{G}[q,p] = \frac{1}{3} \sum_{h \in \mathrm{pa}(p)} \mathbf{G}[q,h] a_{hp} \qquad (q \neq p). \tag{264}$$

Writing $\mathbf{G}_j := \mathbf{G}[\mathrm{Pred}(j), \mathrm{Pred}(j)]$, $\mathbf{b}_j := \mathbf{b}_j[\mathrm{Pred}(j)]$, *and* $\mathbf{a}_j := (a_{pj})_{p \in \mathrm{Pred}(j)}$ *gives*

$$\mathbf{G}_j \mathbf{a}_j = \mathbf{b}_j, \qquad \gamma_{qj} = -\frac{1}{3}\mathbf{b}_j[q] = -\frac{1}{3} \sum_{p \in \mathrm{pa}(j)} \mathbf{G}[q,p] a_{pj}. \tag{265}$$

If the recovered chart reflects coordinates by $\mathbf{D}_s = \mathrm{diag}(s_1, \ldots, s_d)$, *let* $\mathbf{D}_{s,j} := \mathrm{diag}(s_p : p \in \mathrm{Pred}(j))$ *and define*

$$\widetilde{\mathbf{G}} = \mathbf{D}_s \mathbf{G} \mathbf{D}_s, \qquad \widetilde{\mathbf{G}}_j = \mathbf{D}_{s,j} \mathbf{G}_j \mathbf{D}_{s,j}, \qquad \widetilde{\mathbf{b}}_j = s_j \mathbf{D}_{s,j} \mathbf{b}_j, \qquad \widetilde{\mathbf{a}}_j = s_j \mathbf{D}_{s,j} \mathbf{a}_j. \tag{266}$$

Then $\widetilde{\mathbf{G}}_j \widetilde{\mathbf{a}}_j = \widetilde{\mathbf{b}}_j$ *and* $\widetilde{\gamma}_{qj} = s_q s_j \gamma_{qj}$. *Consequently,* $\gamma_{qj} = 0$ *for* $q \notin \mathrm{an}(j)$, *while every direct edge* $q \to j$ *satisfies* $|\gamma_{qj}| \geq |a_{qj}|/3 \geq \beta_{\min}$.

Proof. Under P_{e_j}, the residual density relative to the factual uniform law is

$$r_{e_j}(u) = 1 - W_j \sum_{p \in \mathrm{pa}(j)} a_{pj} \widetilde{Z}_p.$$

Hence independence of W_j under P_0 and $\mathbb{E}_0 W_j^2 = 1/3$ give

$$\mathbb{E}_{e_j}[W_q W_j] = -\frac{1}{3} \sum_{p \in \mathrm{pa}(j)} a_{pj} \mathbb{E}_0[W_q \widetilde{Z}_p],$$

which proves equation 265. Rewriting equation 259 as $\widetilde{Z}_p = W_p + 2U_p(1-U_p) \sum_{h \in \mathrm{pa}(p)} a_{hp} \widetilde{Z}_h$ yields $\mathbf{G}[p,p] = 1$ and equation 264. In the canonical orientation, the recursion and the positive coefficients make every ancestor entry of $\mathbf{G}$ nonnegative. Only ancestors contribute directed path terms, so every nonancestor has zero cross-moment. For a direct parent q of j, the $p = q$ summand in equation 265 equals a_{qj} and all other summands are nonnegative. Reflections change signs while preserving the zero pattern and magnitude bound. Finally, the reflected centered coordinates transform by the same diagonal signs. Hence their cross moments are $\mathbf{D}_s \mathbf{G} \mathbf{D}_s$, while the atomic right-hand side is $s_j \mathbf{D}_{s,j} \mathbf{b}_j$. Multiplying $\mathbf{G}_j \mathbf{a}_j = \mathbf{b}_j$ by these sign matrices proves equation 266 and the transformed system identity. □

Algorithm 3 realizes selective graph recovery for the atomic estimator $\widehat{\mathbf{A}}_N$. It certifies the residual chart, classifies every ordered-pair response, peels certified sinks, and resolves each parent set from the independent graph split. A complete sequence of unique decisions returns a DAG; any unresolved decision returns $\perp$. The binary64 study evaluates the same rule with the outward linear-system bounds from Section O.5.

Algorithm 3 Atomic Interval Rule

Require: independent splits $\mathcal{D}_{\mathrm{ch},1:3}$, $\mathcal{D}_{\mathrm{rat}}$, $\mathcal{D}_{\mathrm{cal}}$, $\mathcal{D}_{\mathrm{gr}}$; fixed model constants
Ensure: adjacency matrix $\widehat{\mathbf{A}}_N$ or $\perp$

1. Certify the residual chart
Fit and certify the atomic residual chart from $\mathcal{D}_{\mathrm{ch},1:3}$
if the chart is unresolved **then**
  return $\perp$
end if
2. Classify ordered-pair responses
for each ordered pair (k, j) with $k \neq j$ **do**
  Form $I_{kj}^{(\mathrm{rat})}$ and $I_{kj}^{(\mathrm{cal})}$ from $\mathcal{D}_{\mathrm{rat}}$ and $\mathcal{D}_{\mathrm{cal}}$
  $I_{kj} \leftarrow I_{kj}^{(\mathrm{rat})} \cap I_{kj}^{(\mathrm{cal})}$
  if I_{kj} is unresolved in $\mathcal{B}_\gamma$ **then**
    return $\perp$
  end if
  Project I_{kj} onto its resolved component to obtain J_{kj}
end for
Form $\widehat{S}_{k,R}$ for all $\varnothing \neq R \subseteq [d]$, $k \in R$ by equation 274
3. Peel certified sinks
$R \leftarrow [d]$, $\pi \leftarrow ()$ ▷ active nodes and sink order
while $R \neq \varnothing$ **do**
  $\widehat{\mathcal{S}}_R \leftarrow \{k \in R : \sup \widehat{S}_{k,R} = 0\}$
  if $\widehat{\mathcal{S}}_R = \varnothing$ **then**
    return $\perp$
  end if
  $k \leftarrow \min \widehat{\mathcal{S}}_R$; $\pi \leftarrow (\pi, k)$
  $R \leftarrow R \setminus \{k\}$ ▷ peel the certified sink
end while
4. Recover direct parents
for each j in the reverse sink order π **do**
  Estimate $\mathbf{b}_j$ from $\mathcal{D}_{\mathrm{gr}}$ and form the current $\mathbf{G}_j$
  Resolve $\widehat{\mathrm{pa}}(j)$ over $S \subseteq \mathrm{Pred}(j)$, $|S| \leq 2$, using $\mathbf{G}_j \mathbf{a}_j = \mathbf{b}_j$
  if $\widehat{\mathrm{pa}}(j)$ is unresolved **then**
    return $\perp$
  end if
  Update $\mathbf{G}$ by equation 264
end for
Assemble $\widehat{\mathbf{A}}_N$ from $\{\widehat{\mathrm{pa}}(j) : j \in [d]\}$
return $\widehat{\mathbf{A}}_N$ ▷ recovered adjacency matrix

O.2 Residual-chart recovery from atomic pairs

Same-unit atomic pairs make the residual chart directly observable after orthogonal mixing: each paired difference lies on one labeled residual axis. This rank-one signal is the atomic counterpart of minimum-code chart recovery.

Let $\mathbf{m}_j$ be column j of $\mathbf{M}$. For paired unit s in environment e_j, set $\Delta\mathbf{X}_s^{(j)} = \mathbf{X}_s^{(j)} - \mathbf{X}_s$. Valid pairing gives

$$\Delta\mathbf{X}_s^{(j)} = \xi_{sj}\mathbf{m}_j, \qquad \xi_{sj} = \Phi^{-1}(U_{sj}^{(j)}) - \Phi^{-1}(U_{sj}). \tag{267}$$

The fitted axis is the leading unit eigenvector of

$$\widehat{\mathbf{H}}_j^{\mathrm{ch}} := \sum_{s=1}^{N} \Delta\mathbf{X}_s^{(j)} \Delta\mathbf{X}_s^{(j)\top}. \tag{268}$$

Its sign is fixed by requiring the largest-magnitude entry to be positive. The same axes must explain the paired differences in the second chart sample. The third chart sample supplies the held-out factual

moments. If $\mathbf{V} = [\widehat{\mathbf{m}}_1, \dots, \widehat{\mathbf{m}}_d]$, the encoder is $\mathbf{L} = \mathbf{V}^{-1}$ and the recovered residuals are

$$\widehat{\mathbf{U}} = \Phi(\mathbf{L}\mathbf{X}).$$

Lemma O.3 (Exact atomic chart)**.** *Under Assumption O.1, if every atomic sample has nonzero excitation, then $\widehat{\mathbf{m}}_j = s_j \mathbf{m}_j$ for signs $s_j \in \{-1, 1\}$. Consequently*

$$\widehat{U}_j = \begin{cases} U_j, & s_j = 1, \\ 1 - U_j, & s_j = -1, \end{cases} \tag{269}$$

and the labeled target signatures are recovered exactly. The event has probability one for a continuous nondegenerate replacement.

Proof. Substituting equation 267 into equation 268 gives

$$\widehat{\mathbf{H}}_j^{\mathrm{ch}} = \left(\sum_{s=1}^{N} \xi_{sj}^2 \right) \mathbf{m}_j \mathbf{m}_j^\top.$$

The multiplier is positive under nonzero excitation, so the leading eigenspace is $\mathrm{span}(\mathbf{m}_j)$. Orthogonality of $\mathbf{M}$ gives $\mathbf{V} = \mathbf{M}\,\mathrm{diag}(s_1, \dots, s_d)$ and hence $\mathbf{L} = \mathrm{diag}(s_1, \dots, s_d)\mathbf{M}^\top$. Applying Φ to the signed normal scores proves equation 269. Reflections preserve the zero/nonzero coefficient classes and squared sink scores. □

Held-out moments certify that the recovered coordinates have the factual uniform-product structure required downstream. With $\widetilde{W} = 2\widehat{\mathbf{U}} - \mathbf{1}$, the targets are

$$\mathbb{E}\widehat{U}_k = \frac{1}{2}, \qquad \mathbb{E}\widetilde{W}_k^2 = \frac{1}{3}, \qquad \mathbb{E}\widetilde{W}_k \widetilde{W}_\ell = 0 \quad (k < \ell). \tag{270}$$

There are $2d + \binom{d}{2}$ scalar comparisons. The estimator accepts the chart when all comparisons pass.

For an interval calculation, if $T_1, \dots, T_N$ have range width w, empirical variance $\widehat{v}$, and scalar allocation α, define

$$r_{\mathrm{EB}}(T; \alpha, w) := \sqrt{\frac{2\widehat{v}\log(2/\alpha)}{N}} + \frac{7w\log(2/\alpha)}{3(N-1)}. \tag{271}$$

The empirical Bernstein inequality bounds the deviation of the sample mean by equation 271 with probability at least $1 - \alpha$. Equal allocation across equation 270 gives total chart probability δ_{ch}.

Remark O.4 (Chart uncertainty). By Lemma O.3, the population chart has zero transport error. The finite-precision rule enlarges downstream intervals by a nonnegative term e_{ch}, computed from rank-one residuals and axis orthogonality across the three chart samples. This enlargement preserves coverage under Assumption O.1.

Conditioning on the recovered signs in Lemma O.3 reduces all subsequent decisions to reflected coordinates. The corresponding objects are $(\widetilde{W}, \widetilde{\mathbf{Z}}, \widetilde{\mathbf{G}}, \widetilde{\mathbf{b}}, \widetilde{\mathbf{a}}, \widetilde{\mathbf{\Gamma}})$ in equation 266. Because reflection preserves every zero pattern and coefficient magnitude, we henceforth write them as $(W, \widetilde{\mathbf{Z}}, \mathbf{G}, \mathbf{b}, \mathbf{a}, \mathbf{\Gamma})$. The encoded sample moments then target these population quantities directly.

O.3 Simultaneous response intervals and sink peeling

Population sink score. The atomic pair moment preserves exactly the sink set characterized by the population conditional-CDF foliation. In the positive-coefficient model, equation 262 therefore supplies a bounded score on every active set reached by recursive sink deletion.

For $k \in R$, define

$$s_{k,R} := \sum_{j \in R \setminus \{k\}} (\gamma_{kj})^2.$$

Lemma O.5 (Population sink score)**.** *Under Assumption O.1, let R be any set reached from $[d]$ by deleting sinks. Then $s_{k,R} = 0$ if and only if k is a sink of the active DAG on R. Every non-sink satisfies $s_{k,R} \geq \beta_{\min}^2$.*

Proof. Every set reached in this way is ancestor closed. If k is a sink of the induced graph on R, it is not an ancestor of any $j \in R \setminus \{k\}$. Lemma O.2 therefore gives $\gamma_{kj} = 0$ for every retained j, and hence $s_{k,R} = 0$.

If k is not a sink, it has a direct child j that remains in R. The direct-edge witness in Lemma O.2 gives $|\gamma_{kj}| \geq \beta_{\min}$, so the corresponding summand proves $s_{k,R} \geq \beta_{\min}^2$. The same pair moments are used at every active set; recursion only restricts their index set to R. □

Marginal-response intervals. Two independently estimated interval families certify every marginal response before the data-dependent peeling path is chosen. Their intersection must identify a unique response component for each ordered pair.

For split $q \in \{\mathrm{rat}, \mathrm{cal}\}$, atomic environment j, and $k \neq j$, define

$$\widehat{\gamma}_{kj}^{(q)} := \frac{3}{N} \sum_{s=1}^{N} W_{sk}^{(q,j)} W_{sj}^{(q,j)}. \tag{272}$$

The summand lies in $[-3, 3]$. Let $\alpha_q \leq \delta_q/[d(d-1)]$ and set

$$r_{kj}^{(q)} := r_{\mathrm{EB}}\Big(3W_k^{(q,j)} W_j^{(q,j)}; \alpha_q, 6\Big) + 12e_{\mathrm{ch}}.$$

The raw interval is $I_{kj}^{(q)} = [\widehat{\gamma}_{kj}^{(q)} - r_{kj}^{(q)}, \widehat{\gamma}_{kj}^{(q)} + r_{kj}^{(q)}]$.

The two samples are combined by intersection:

$$I_{kj} = I_{kj}^{(\mathrm{rat})} \cap I_{kj}^{(\mathrm{cal})}. \tag{273}$$

The interval is resolved when it meets exactly one connected component of $\mathcal{B}_\gamma$ in equation 263. Its projection onto that component is denoted $J_{kj} = [\ell_{kj}, u_{kj}]$ and is labeled zero or large. Empty intersections and intersections spanning multiple components remain unresolved. A covered strict-ancestor response below $\beta_{\min}$ may receive either label or remain unresolved, depending on which components its interval meets. Either resolved label preserves the sink set: a retained strict descendant certifies a non-sink, and every non-sink also has a large parent–child witness.

Lemma O.6 (Sound response classification)**.** *Let $\mathcal{E}_{\mathrm{rat}}$ and $\mathcal{E}_{\mathrm{cal}}$ be the events that all $d(d-1)$ intervals from the respective samples contain their population coefficients. Then*

$$\mathbb{P}(\mathcal{E}_q^c \mid \mathcal{F}_{\mathrm{ch}}) \leq \delta_q, \qquad q \in \{\mathrm{rat}, \mathrm{cal}\}.$$

On $\mathcal{E}_{\mathrm{rat}} \cap \mathcal{E}_{\mathrm{cal}}$, a population zero receives the zero label, and a parent–child coefficient in the separated part of $\mathcal{B}_\gamma$ receives the large label. Consequently, if every ordered pair is resolved, then at every reachable active set the projected score is zero exactly at its sinks. Strict-ancestor responses with $0 < |\gamma_{kj}| < \beta_{\min}$ enter only the ordering step and may receive either label when resolved.

Proof. Conditional on the independently fitted chart, each summand in equation 272 is bounded and has expectation γ_{kj}. Apply equation 271 and a union bound over the $d(d-1)$ ordered pairs. On the common event, the intersection equation 273 contains the true coefficient. Coverage places a population zero in the zero component and each parent–child witness in its separated component. Every outgoing response of a sink is zero, while each non-sink has a retained child with a large response. Once every pair meets a unique component, the projected score separates sinks regardless of the labels assigned to small strict-ancestor responses. □

Recursive sink peeling. Simultaneous response coverage converts the population sink score into a sound recursive decision. Squaring and summing the projected intervals preserves the zero-versus-positive separation on every active set.

For every nonempty $R \subseteq [d]$ and $k \in R$, square the projected intervals exactly. If $J = [\ell, u]$, write

$$J^{\langle 2 \rangle} = \begin{cases} [0, \max\{\ell^2, u^2\}], & 0 \in J, \\ [\min\{\ell^2, u^2\}, \max\{\ell^2, u^2\}], & 0 \notin J. \end{cases}$$

The score interval is

$$\widehat{S}_{k,R} := \sum_{j \in R \setminus \{k\}} J_{kj}^{\langle 2 \rangle}. \tag{274}$$

The rule forms the $2^d - 1$ active-set families before selecting a recursive path. At a visited set R, a coordinate is eligible precisely when the upper endpoint of $\widehat{S}_{k,R}$ is zero, and the smallest eligible index is deleted. A score interval with lower endpoint zero and positive upper endpoint remains unresolved. The rule returns $\perp$ when a visited set contains an unresolved score or lacks an eligible coordinate.

Lemma O.7 (Sink peeling under simultaneous coverage)**.** *On $\mathcal{E}_{\mathrm{rat}} \cap \mathcal{E}_{\mathrm{cal}}$, if every ordered pair is resolved, the recursion starting from $R = [d]$ deletes a true sink at every step and returns a complete sink-to-source order. The same simultaneous event covers the data-dependent path.*

Proof. The initial set $[d]$ is ancestor closed, and deleting a sink preserves this property. Hence every set visited by the recursion satisfies the conditions of Lemma O.6. At such a set, a true sink has score interval $[0, 0]$, whereas every non-sink has a retained parent–child witness with a strictly positive squared lower endpoint. The eligible coordinates are therefore precisely the sinks of the active DAG. Induction gives the complete order. The common event was indexed before the path was chosen, so its probability also covers the selected active-set family. □

O.4 Direct-edge recovery

Direct-edge recovery separates the ordering signal from the adjacency target. The peeling coefficient $\gamma_{qj} = -\mathbf{b}_j[q]/3$ aggregates response along directed paths and determines the sink order, whereas the solution of $\mathbf{G}_j\mathbf{a}_j = \mathbf{b}_j$ isolates the direct coefficients. Thus adjacency is read from $\mathbf{a}_j$, even when $\mathbf{b}_j$ contains indirect-path contributions.

The graph sample estimates one bounded scalar moment for each ordered pair $q \prec j$ in the recovered source order:

$$\widehat{\mathbf{b}}_j[q] = -\frac{9}{N}\sum_{s=1}^{N} W_{sq}^{(\mathrm{gr},j)} W_{sj}^{(\mathrm{gr},j)}.$$

There are $\sum_{j=1}^{d}(j-1) = d(d-1)/2$ such moments. Equal allocation of δ_{gr} and equation 271 with range width two produce simultaneous radii

$$r_{b,jq} = 9r_{\mathrm{EB}}\left(W_q^{(\mathrm{gr},j)} W_j^{(\mathrm{gr},j)}; \alpha_{\mathrm{gr}}, 2\right) + 36e_{\mathrm{ch}}. \tag{275}$$

The entries of $\mathbf{G}$ are propagated from previously accepted coefficient intervals using equation 264; conditional on those intervals, this propagation is deterministic.

For node j, let $\mathrm{Pred}(j)$ be its predecessor set in the recovered order and enumerate $\mathfrak{P}_j := \{S \subseteq \mathrm{Pred}(j) : |S| \le 2\}$. For each $S \in \mathfrak{P}_j$, solve the restricted system $\mathbf{G}_j[S, S]\mathbf{a}_{j,S} = \mathbf{b}_j[S]$ with interval-enclosed inputs. A candidate is compatible when all of the following hold: the restricted system is well-conditioned; every selected coefficient interval meets one nonzero component of $\mathcal{B}_a$; and the full predecessor residual

$$\mathbf{b}_j - \mathbf{G}_j[:, S]\mathbf{a}_{j,S} \tag{276}$$

contains zero coordinatewise. Node j is resolved when exactly one candidate is compatible and each of its selected coefficients lies in a unique nonzero component. Otherwise the estimator returns $\perp$.

Lemma O.8 (Unique compatible parent set)**.** *Let $\mathcal{E}_{\mathrm{gr}}$ be the event that every interval in equation 275 contains its population moment. Conditional on the chart, the ratio and calibration samples, and the resulting peeling order,*

$$\mathbb{P}(\mathcal{E}_{\mathrm{gr}}^{c} \mid \mathcal{D}_{\mathrm{ch},1:3}, \mathcal{D}_{\mathrm{rat}}, \mathcal{D}_{\mathrm{cal}}) \le \delta_{\mathrm{gr}}. \tag{277}$$

On $\mathcal{E}_{\mathrm{gr}}$ and a correct peeling order, Assumption O.1(v) makes the true parent set compatible. Hence any resolved node has the true parent set, and if every node is resolved the returned DAG equals $\mathbf{A}_\star$.

Proof. The graph sample is independent of the sigma-field in equation 277. Conditional on that sigma-field, the order and all queried moment coordinates are fixed. Empirical Bernstein and a union bound over $d(d-1)/2$ bounded products prove the probability statement.

For the true parent set, Lemma O.2 makes the population residual equation 276 equal to zero. Simultaneous interval containment and Assumption O.1(v) preserve the true coefficient vector and zero residual through the inverse enclosure; the direct coefficients lie in the nonzero part of $\mathcal{B}_a$. The

true set is therefore among the compatible candidates. If exactly one candidate remains, it must be the true set. Induction along the source order preserves the propagated cross-moment intervals and proves the graph statement. □

O.5 Finite-precision soundness

Outward rounding carries the selective rule to binary64 arithmetic. Each restricted linear solve receives an outward radius derived from bounds on the input matrix, right-hand side, and rounding error.

Lemma O.9 (Residual-enclosed linear solve). *Let $\mathbf{G}$ be an invertible matrix, let $\widetilde{\mathbf{G}}$ and $\widetilde{\mathbf{b}}$ be computed centers, and suppose componentwise deterministic bounds $|\mathbf{G}-\widetilde{\mathbf{G}}| \le R_G$ and $|\mathbf{b}-\widetilde{\mathbf{b}}| \le r_b$ are available, and let e_{fp} bound the componentwise floating-point error in each matrix calculation below. Let $\mathbf{B}$ be a computed inverse center and*

$$\tau := \|I - \mathbf{B}\widetilde{\mathbf{G}}\|_\infty < 1. \tag{278}$$

Then

$$\|\widetilde{\mathbf{G}}^{-1} - \mathbf{B}\|_\infty \le \frac{\|\mathbf{B}\|_\infty \tau}{1-\tau}. \tag{279}$$

If $\widetilde{\mathbf{a}}$ is the computed solution center, a valid componentwise coefficient radius is

$$r_a = \left(|\mathbf{B}| + \frac{\|\mathbf{B}\|_\infty \tau}{1-\tau}\mathbf{1}\mathbf{1}^\top\right)\left(r_b + R_G a_{\max}\mathbf{1} + |\widetilde{\mathbf{G}}\widetilde{\mathbf{a}} - \widetilde{\mathbf{b}}| + e_{\mathrm{fp}}\right), \tag{280}$$

with endpoints rounded outward. The full residual radius is

$$r_{\mathrm{res}} = r_b + R_G a_{\max}\mathbf{1} + |\widetilde{\mathbf{G}}| r_a + e_{\mathrm{fp}}. \tag{281}$$

Proof. Write $\mathbf{B}\widetilde{\mathbf{G}} = I - E$ with $\|E\|_\infty = \tau < 1$. The Neumann series gives $(I-E)^{-1} = \sum_{k\ge 0} E^k$ and $\|(I-E)^{-1} - I\|_\infty \le \tau/(1-\tau)$. Since $\widetilde{\mathbf{G}}^{-1} = (I-E)^{-1}\mathbf{B}$, this proves equation 279. The identity $\widetilde{\mathbf{G}}(\mathbf{a}-\widetilde{\mathbf{a}}) = (\widetilde{\mathbf{G}}-\mathbf{G})\mathbf{a} + (\mathbf{b}-\widetilde{\mathbf{b}}) + (\widetilde{\mathbf{b}} - \widetilde{\mathbf{G}}\widetilde{\mathbf{a}})$ and the triangle inequality give equation 280. Substitution into $\mathbf{b} - \mathbf{G}_S\mathbf{a}_S$ gives equation 281. All bounds are componentwise, with e_{fp} covering the finite-precision error in both calculations. □

The direct-parent calculation rounds every endpoint outward. With $u_{\mathrm{fp}} = 2^{-52}$ and $\mathrm{up}_{\mathrm{fp}}(x)$ denoting upward rounding, the graph-moment radius is

$$\mathrm{up}_{\mathrm{fp}}\left[\sqrt{\frac{2\widehat{v}_{\mathrm{fl}}\log(2/\alpha)}{N}} + \frac{14\log(2/\alpha)}{3(N-1)} + 64u_{\mathrm{fp}}(N+1)\right].$$

For candidate row q, the common arithmetic term in the solve and residual enclosures is the qth component of e_{fp}:

$$[e_{\mathrm{fp}}]_q = 64u_{\mathrm{fp}}\left(1 + |\widetilde{\mathbf{b}}_q| + a_{\max}\sum_p |\widetilde{\mathbf{G}}_{qp}|\right).$$

For a floating-point center c and radius r, the interval endpoints are its represented rational bounds rounded outward. Direct-parent recovery uses these enclosures; chart fitting, marginal coefficients, intersections, and scores use standard binary64 arithmetic.

O.6 Selective finite-sample theorem

Let $\mathcal{F}_{\mathrm{pre}}$ be the field obtained from $\mathcal{F}_{\mathrm{ch}}$ after adding the ratio and calibration samples. Set

$$\delta := \delta_{\mathrm{ch}} + \delta_{\mathrm{rat}} + \delta_{\mathrm{cal}} + \delta_{\mathrm{gr}}.$$

Theorem O.10 (Selective graph recovery). *Under Assumption O.1, the estimator defined in Section O.1 satisfies*

$$\mathbb{P}\Big\{\widehat{\mathbf{A}}_N \in \{\mathbf{A}_\star, \perp\}\Big\} \ge 1 - \delta. \tag{282}$$

The bound holds for every integer $N \ge 2$. On the common event, the estimator returns the true graph whenever the following finite collection of separation conditions holds.

Suppose in addition that the following separations hold:

(i) every coefficient intersection, including a small strict-ancestor response, lies in one component of $\mathcal{B}_\gamma$;

(ii) every restricted true-parent system satisfies equation 278;

(iii) each wrong parent set is excluded by a full-residual interval; and

(iv) every enlarged true direct coefficient lies in a nonzero component.

Then, on the simultaneous event,

$$\widehat{\mathbf{A}}_N = \mathbf{A}_\star. \tag{283}$$

For population zeros and parent–child witnesses, radii below $\beta_{\min}/2$ are sufficient for their respective zero and nonzero separations. A strict-ancestor response in the gap additionally requires its realized intersection to meet a single component. Analogous direct-edge and residual separations are read from equation 280–equation 281.

Proof. Let $\mathcal{E}_{\mathrm{ch}}$ be simultaneous coverage of the held-out chart moments, and use $\mathcal{E}_{\mathrm{rat}}$, $\mathcal{E}_{\mathrm{cal}}$, and $\mathcal{E}_{\mathrm{gr}}$ from Lemmas O.6 and O.8. The exact atomic chart lemma holds almost surely; the held-out moment event has failure probability at most δ_{ch}. Conditional on $\mathcal{F}_{\mathrm{ch}}$, the two coefficient families satisfy

$$\mathbb{P}(\mathcal{E}_{\mathrm{rat}}^c \mid \mathcal{F}_{\mathrm{ch}}) \le \delta_{\mathrm{rat}}, \qquad \mathbb{P}(\mathcal{E}_{\mathrm{cal}}^c \mid \mathcal{F}_{\mathrm{ch}}) \le \delta_{\mathrm{cal}}.$$

Conditional on $\mathcal{F}_{\mathrm{pre}}$, the independent graph sample satisfies equation 277. The tower property and a union bound therefore give

$$\mathbb{P}[(\mathcal{E}_{\mathrm{ch}} \cap \mathcal{E}_{\mathrm{rat}} \cap \mathcal{E}_{\mathrm{cal}} \cap \mathcal{E}_{\mathrm{gr}})^c] \le \delta.$$

On the common event, population zeros retain the zero label and every retained parent–child witness retains a nonzero label by Lemma O.6; reachable sink scores absorb the labels of small strict-ancestor responses. Every deletion is a true sink by Lemma O.7, and every resolved parent set is true by Lemma O.8. Hence every returned graph equals $\mathbf{A}_\star$, proving equation 282. Under the additional separation conditions, all ordered pairs, active sets, and parent sets are resolved, and equation 283 follows. □

The numerical rule allocates its simultaneous error probability as $\delta_{\mathrm{ch}} = \delta_{\mathrm{rat}} = \delta_{\mathrm{cal}} = \delta_{\mathrm{gr}} = 0.008$, $\delta = 0.032$. Each family divides its allocation equally among its bounded scalar events, rounding downward when necessary.

O.7 Finite-sample results

The recovery rule uses sampled pairs and constants fixed before sampling; graph labels enter only when its outputs are scored. The experiment measures how often increasing sample size converts abstention into a correct graph. It varies dimension, graph structure, response separation, and sample size across 160 independent replicate series.

The interval rule exhibits the predicted abstention-to-return transition. Its correct-return rate over the 160 evaluations at each sample size is $0\%, 0\%, 10\%, 70\%$, and 100%. Across all 800 evaluations, the rule returns 288 graphs and abstains 512 times; all 288 returned graphs have structural Hamming distance zero.

The decision counts identify how additional data unlock a return. Chart recovery succeeds throughout; response classification then certifies the sink order, and the independent graph moments certify the parent sets. At $N = 16{,}384$, all three decisions resolve all 160 instances.

The mechanism grid confirms that ordering from marginal response remains distinct from direct-edge recovery. Among the 160 data-generating instances, 72 contain 484 nonzero strict-ancestor responses below $\beta_{\min}$, with magnitudes from 1.60×10^{-8} to 0.02321. The direct parent–child witnesses instead range from 0.077 to 0.116025, all at or above their regime-specific $\beta_{\min}$. The grid therefore exercises indirect-path responses while retaining the separated witness required by peeling.

The selective transition persists across every dimension–response stratum. Near-boundary strata require larger N as dimension grows, yet every stratum reaches $20/20$ correct outputs at $N = 16{,}384$ (Table 2).

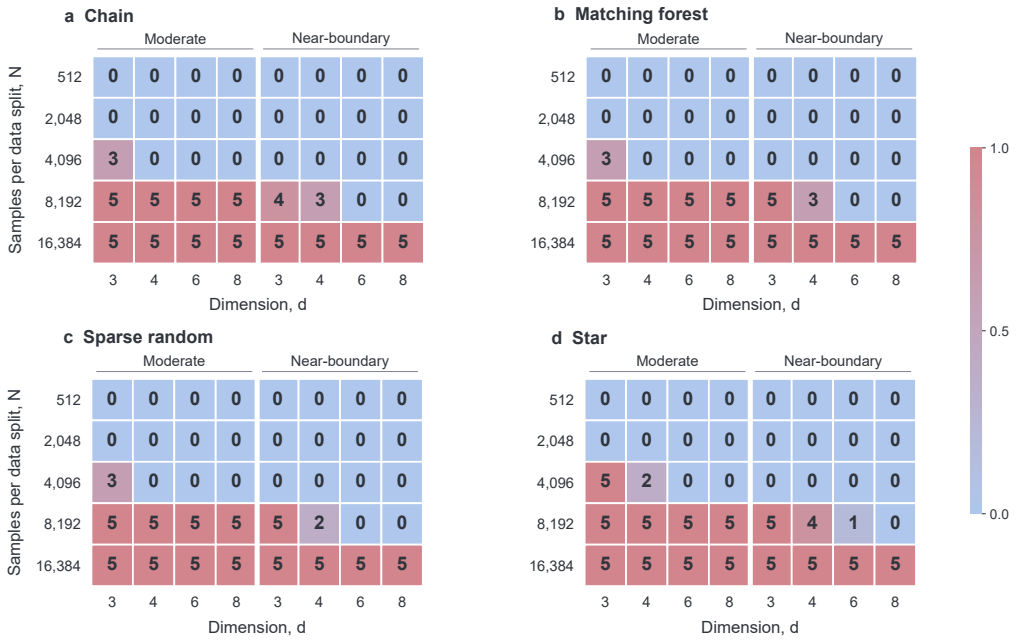

Figure 4: **Correct graph recovery spans four graph families. a–d,** Chains, matching forests, sparse random DAGs, and stars, respectively. Each cell reports correct returns among five replicate series. Every return is correct, and every cell reaches $5/5$ at $N = 16{,}384$.

Table 2: Correct outputs by dimension, response regime, and sample size per data split. Each entry is out of 20 evaluations: four graph families and five independently seeded series per family. The remaining evaluations abstain, and every graph return is correct.

Regime	Dimension	512	2,048	4,096	8,192	16,384
Moderate	3	0	0	14	20	20
Moderate	4	0	0	2	20	20
Moderate	6	0	0	0	20	20
Moderate	8	0	0	0	20	20
Near-boundary	3	0	0	0	19	20
Near-boundary	4	0	0	0	12	20
Near-boundary	6	0	0	0	1	20
Near-boundary	8	0	0	0	0	20
All strata	–	0	0	16	112	160

Within each replicate series, shrinking response intervals produce a monotone selective transition: the largest interval radius never increases with N, and every series continues to return after its first graph return. The five evaluations share the generating model and split seeds, with larger factual blocks extending smaller ones and paired blocks using N-dependent offsets. The series-level diagnostic therefore treats each complete series as one independent unit.

Series-level error diagnostic. The 160 independent series contain zero observed series-level errors over their fixed five-point sample-size grids. We use these independent units to estimate a one-sided upper bound on the mean series-level error probability; the five evaluations within each series share the generating model and seeds. Let E_i be the event that series i contains at least one incorrect graph output on the five-point N grid, and let $p_i = \mathbb{P}(E_i)$. The series are independent and may have heterogeneous p_i. If $\bar{p} = 160^{-1} \sum_i p_i$, then the arithmetic–geometric mean inequality gives

$$\mathbb{P}\left(\bigcap_{i=1}^{160} E_i^c\right) = \prod_{i=1}^{160}(1 - p_i) \le (1 - \bar{p})^{160}. \tag{284}$$

All 160 observed series satisfy E_i^c. Inverting equation 284 at level 0.05 gives the one-sided 95% upper bound

$$\bar{p} \le 1 - 0.05^{1/160} = 0.0185491341. \tag{285}$$

This 1.855% bound summarizes the mean probability of an incorrect output over the fixed five-point grid; the selective guarantee is Theorem O.10. Exact return counts appear in Table 3.

Table 3: Decision counts out of 160 evaluations at each sample size per data split. "Marginal order" requires complete coefficient classification and peeling; "direct graph" resolves the independent graph moments; "correct output" is a graph return from the full interval rule.

N	Chart	Marginal order	Direct graph	Correct output
512	160	0	0	0
2,048	160	0	0	0
4,096	160	29	18	16
8,192	160	135	115	112
16,384	160	160	160	160

O.8 Same-unit pairing supplies the chart signal

This experiment tests whether same-unit alignment supplies the chart signal predicted by the population theory. It compares valid pairs with matched perturbations that either break unit correspondence or remove one environment; Theorem L.6 gives the corresponding population analysis.

For factual observations with covariance $\mathbf{\Sigma}_X$, let $\mathbf{W}_\Sigma = \mathbf{\Sigma}_X^{-1/2}$ be the symmetric whitener. Holding this whitener fixed, form the sample paired-difference moment in environment e:

$$\widehat{\mathbf{B}}_e = \frac{1}{N}\sum_{s=1}^{N}\left[\mathbf{W}_\Sigma(\mathbf{X}_s^{(e)} - \mathbf{X}_s)\right]\left[\mathbf{W}_\Sigma(\mathbf{X}_s^{(e)} - \mathbf{X}_s)\right]^\top.$$

The eigenspace of $\widehat{\mathbf{B}}_e$ above a fixed relative threshold gives the sample projector $\widehat{\mathbf{P}}_e$. Form

$$\widehat{\mathbf{C}}_{\text{sig}} = \sum_{e=1}^{m} 2^{e-1}\widehat{\mathbf{P}}_e.$$

Rounding its eigenvalues gives candidate binary target-signature IDs. The signatures are accepted when their IDs are distinct, nonzero, and separated from the rounding boundary. The population operator explains this construction.

Proposition O.11 (Population signature operator)**.** *Suppose the factual covariance is nonsingular, the paired differences in environment e span exactly the residual coordinates in its intervention signature, and the binary signature columns are distinct and nonzero. Let*

$$\mathbf{B}_e = \mathbb{E}\left[\mathbf{W}_\Sigma(\mathbf{X}^{(e)} - \mathbf{X})(\mathbf{X}^{(e)} - \mathbf{X})^\top\mathbf{W}_\Sigma^\top\right].$$

Let $\mathbf{P}_e$ be the orthogonal projector onto its range. Then the projectors commute, and the eigendecomposition of $\mathbf{C}_{\text{sig}} = \sum_{e=1}^{m} 2^{e-1}\mathbf{P}_e$ recovers the residual directions and exact target signatures up to a common coordinate permutation and sign reflections.

Proof. Write $\mathbf{O} = \mathbf{W}_\Sigma\mathbf{A}_{\text{mix}}$ for the orthogonal whitened mixing matrix in equation 113. In this basis, the paired-difference moment has range spanned by the targeted coordinate axes, so

$$\mathbf{P}_e = \mathbf{O}\,\mathrm{diag}(\sigma_{e1},\dots,\sigma_{ed})\mathbf{O}^\top,$$

$$\mathbf{C}_{\text{sig}} = \mathbf{O}\,\mathrm{diag}(c_1,\dots,c_d)\mathbf{O}^\top, \qquad c_i = \sum_{e=1}^{m} 2^{e-1}\sigma_{ei}.$$

Distinct IDs separate the one-dimensional common eigenspaces and identify $\mathbf{O}$ up to signed permutation. The residual encoder $\mathbf{L} = \mathbf{O}^\top\mathbf{W}_\Sigma = \mathbf{A}_{\text{mix}}^{-1}$ then gives the chart; signs before Φ become coordinate reflections. □

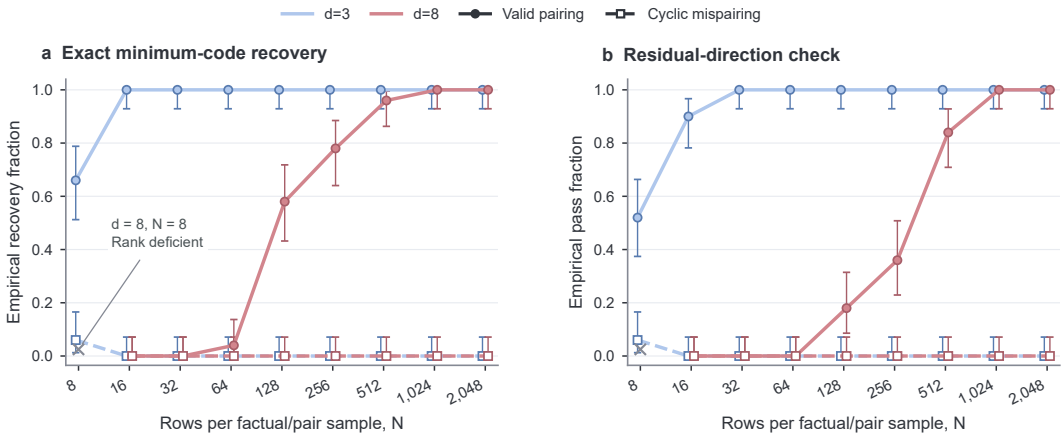

Figure 5: **Same-unit pairs identify residual directions and target signatures. a,** Exact signature recovery up to a common coordinate permutation. Valid pairing reaches $50/50$ from $N = 16$ for $d = 3$ and from $N = 1{,}024$ for $d = 8$; cyclic mispairing yields $3/900$ recoveries. **b,** Fraction of trials passing the residual-direction criterion, with pointwise 95% Clopper–Pearson intervals over 50 replicates per cell. The two conditions share each generated sample; crosses mark the rank-deficient $d = 8, N = 8$ cell.

The matched design changes pairing information while holding the underlying mechanism and generated observations fixed within each cell. It uses two generator templates with mixing and data redrawn across repetitions, $d \in \{3, 8\}$, nine sample sizes $N \in \{8, 16, 32, 64, 128, 256, 512, 1024, 2048\}$, three pairing conditions, and 50 repetitions per cell, for $2 \times 9 \times 3 \times 50 = 2700$ evaluations. Each evaluation contains N factual rows and N pairs per retained environment. The three conditions are:

(i) valid factual–interventional pairs;

(ii) a deterministic within-environment cyclic shift of the interventional rows by one position; and

(iii) removal of one environment, leaving too few binary signatures for the dimension.

The same generator and data seed define all three conditions within each cell.

The matched contrast confirms the predicted chart signal under same-unit alignment. With valid pairing, exact target-signature recovery is $33/50$ at $d = 3, N = 8$ and $50/50$ for every $d = 3$ cell from $N = 16$ onward. For $d = 8$, the counts across $N = 16, 32, 64, 128, 256, 512, 1024, 2048$ are $0, 0, 2, 29, 39, 48, 50, 50$ out of 50. Cyclic shifting yields $3/900$ recoveries, isolating loss of same-unit signal. Removing one environment yields $0/900$, matching the $2^m - 1$ signature-capacity bound.

O.9 Implementation details

Every quantity that enters the selective decision is fixed before sampling. The design uses orthogonal mixing, identity atomic signatures, uniform replacements, maximum indegree two, and the structural map equation 259. Moderate and near-boundary response regimes multiply the base marginal coefficient 0.07 by 1.5 and 1.1, respectively. Each replicate series fixes its graph structure, mixing, and random draws; the chain, star, and matching-forest families share their structural template, while the sparse-random graph is drawn independently for each replicate. The corresponding graph supplies the recovery target.

Each of the 160 replicate series fixes its graph instance, mixing, and random draws across the five values of N. Complete split-specific samples within a series share seeds but use different paired offsets; distinct series use independent random draws. The latter define the independent units in equation 285.

In the pairing experiment, repetitions use independent mixing and data draws. Within each dimension–sample-size–repetition cell, all three conditions begin from the same generated observations before re-pairing or environment removal. This matched construction attributes the recovery contrast to pairing information while retaining sampling variation across repetitions.

Table 4: Implementation details for the finite-sample study. The values in this table determine the return rule; recovery metrics are evaluated against the corresponding graph.

Quantity	**Value**
Model and data splits	Orthogonal mixing; labeled uniform atomic replacements; maximum indegree two; six independent splits; each split has N factual rows and N pairs per environment
Sample-size grid	$N \in \{512, 2048, 4096, 8192, 16384\}$
Marginal-response class	$\beta_{\min} = 0.105$ (moderate) or 0.077 (near-boundary), $\beta_{\max} = 0.15$
Direct-coefficient class	Generated magnitude $3\beta_{\min}$; nonzero range $[3\beta_{\min}, 0.45]$; scale from pair response to direct coefficient equal to 3
Chart conditions	Rank-one residual at most 10^{-9} on each chart split; $\|\mathbf{V}^\top\mathbf{V} - \mathbf{I}\|_2 \le 10^{-9}$; $\kappa_2(\mathbf{V}) \le 1.000001$; trim $\eta = 1/16$; C^3 radius at most 10^{-7}; +ve Jacobian floor
Direct-system conditions	$\kappa_2(\mathbf{G}_j[S,S]) \le 10^6$; inverse residual below $1/2$; inactive-pair tolerance 10^{-9}; parent sets of size at most two
Failure probabilities	$\delta_{\mathrm{ch}} = \delta_{\mathrm{rat}} = \delta_{\mathrm{cal}} = \delta_{\mathrm{gr}} = 0.008$, so $\delta = 0.032$

P Response faithfulness and conditional-law regularity

The following constructions delimit the response condition and the conditional-law formulations used in the identification argument.

P.1 Terminal-response faithfulness

Path cancellation can preserve causal minimality while removing every terminal response to a non-sink. The following example isolates why compressed designs use TRF.

Example P.1 (Causal minimality and TRF are distinct). Let $a, b \neq 0$, define $\varepsilon_i(u) := \Phi^{-1}(u)$ and $E_i := \varepsilon_i(U_i)$, and consider

$$Z_1 = E_1, \qquad Z_2 = aZ_1 + E_2, \qquad Z_3 = bZ_2 - abZ_1 + E_3.$$

Every displayed parent coefficient is nonzero, and every noise coordinate maps $(0,1)$ diffeomorphically onto the fixed conditional support $\mathbb{R}$. Indeed, $\varepsilon_i'(u_i) = 1/\phi(\Phi^{-1}(u_i)) > 0$, where ϕ is the standard Gaussian density. Thus the SCM satisfies scalar noise solvability and causal minimality. Substituting the second equation into the third gives

$$Z_3 = b(aZ_1 + E_2) - abZ_1 + E_3 = bE_2 + E_3.$$

Consequently,

$$\partial_{u_1} F_3 = \partial_{u_1}\big(b\varepsilon_2(u_2) + \varepsilon_3(u_3)\big) = 0, \tag{286}$$

while $\partial_{u_2} F_3 = b\,\varepsilon_2'(u_2) \neq 0$. The displayed DAG contains $1 \to 2$, $2 \to 3$, and $1 \to 3$. In the full active set, node 1 is a non-sink and node 3 is its only sink descendant. Equation 286 therefore violates the exact quantifier in Assumption 2.1. The cancellation separates structural minimality from TRF. Compressed multi-target designs therefore use an explicit response condition such as TRF. Under atomic coverage, Proposition 4.5 supplies direct nodewise recovery.

P.2 Conditional-law regularity

Exceptional fibers. The following smooth fixed-support SCM has a divergent pointwise conditional normalizer on a null fiber, making the regular-conditional formulation in Section 3.3 essential. Fix $\varepsilon \in (0,1)$ and, for $s, t \in (0,1)$, define

$$A(s) := \int_0^1 \exp\left(-\frac{(z-\frac{1}{2})^2}{s^4}\right) \mathrm{d}z,$$
$$\kappa_s(t) := \frac{1}{A(s)} \exp\left(-\frac{(t-\frac{1}{2})^2}{s^4}\right),$$
$$\rho(s,t) := (1-\varepsilon) + \varepsilon\kappa_s(t).$$

The function ρ is positive and smooth on the open square, and

$$\int_0^1 \rho(s,t)\ \mathrm{d}t = 1 \qquad \text{for every } s \in (0,1). \tag{287}$$

The dependence on s is nontrivial. Differentiation under the integral sign is valid locally on $s \in (0,1)$ because the integrand and its s-derivative are uniformly bounded on every compact s-subinterval. It gives

$$A'(s) = \int_0^1 \frac{4(z-\frac{1}{2})^2}{s^5} \exp\left(-\frac{(z-\frac{1}{2})^2}{s^4}\right) \mathrm{d}z > 0.$$

At the center of the spike,

$$\partial_s \rho(s, \tfrac{1}{2}) = -\varepsilon \frac{A'(s)}{A(s)^2} < 0, \tag{288}$$

so both ρ and $\log \rho$ vary with s.

Consider the two-node SCM

$$F_1(s) = s, \qquad F_2(s,u) = \int_0^u \rho(s,t)\ \mathrm{d}t$$

with graph $1 \to 2$ and identity decoder. Positivity and equation 287 make $u \mapsto F_2(s,u)$ a smooth increasing diffeomorphism from $(0,1)$ onto the fixed interval $(0,1)$ for every s. The triangular map $F = (F_1, F_2)$ is therefore a smooth global diffeomorphism of the open square: its inverse is $(z_1, z_2) \mapsto (z_1, \psi(z_1, z_2))$, where ψ is the unique solution of $F_2(z_1, \psi) = z_2$ and is smooth by the implicit function theorem. To verify that the edge is active, suppose for contradiction that $\partial_s F_2 \equiv 0$. Smoothness permits differentiation with respect to u, which would give

$$0 = \partial_u \partial_s F_2(s,u) = \partial_s \partial_u F_2(s,u) = \partial_s \rho(s,u) \qquad \text{for every } (s,u).$$

This contradicts equation 288. Hence $\partial_s F_2 \not\equiv 0$; the edge $1 \to 2$ is structurally active, and the same nonzero response verifies TRF for the unique possible active non-sink.

Use two labeled environments that target nodes 1 and 2 separately, with uniform replacement densities. Their signatures are distinct and nonzero, and the induced paired densities are positive on their complete support relations. For the node-1 environment, let ξ_1 be the uniform replacement. The counterfactual endogenous variables and their factual residual inverse satisfy

$$Z_1^{(1)} = \xi_1, \quad Z_2^{(1)} = F_2(\xi_1, U_2), \quad V_1^{(1)} = \xi_1, \quad V_2^{(1)} = U_2.$$

The last equality follows from injectivity of $F_2(\xi_1, \cdot)$. Thus the source coordinates (U_1, U_2, ξ_1) have unit density and parametrize the whole relation $V_2^{(1)} = U_2$, proving positive intrinsic density for this environment.

For the node-2 environment, let ξ_2 be its uniform replacement. Here $V_1^{(2)} = U_1$ and the target residual is the unique solution of $F_2(U_1, V_2^{(2)}) = \xi_2$. Changing variables from ξ_2 to v_2 has positive derivative $\partial_{v_2} F_2(U_1, v_2) = \rho(U_1, v_2)$. Hence the intrinsic density in the free support coordinate v_2 is $\rho > 0$. In this node-2 environment, the residual density ratio is exactly

$$r_{e_2}(s,t) = \rho(s,t).$$

On the transverse fiber $t = 1/2$,

$$A(s) = \int_0^1 \exp\left(-\frac{(z-\frac{1}{2})^2}{s^4}\right) \mathrm{d}z \overset{x=z-\frac{1}{2}}{=\!=\!=} \int_{-1/2}^{1/2} e^{-x^2/s^4}\ \mathrm{d}x \le \int_{\mathbb{R}} e^{-x^2/s^4}\ \mathrm{d}x = \sqrt{\pi}\, s^2.$$

and hence

$$\int_0^1 r_{e_2}(s, \tfrac{1}{2})\ \mathrm{d}s = \int_0^1 \rho(s, \tfrac{1}{2})\ \mathrm{d}s \ge \frac{\varepsilon}{\sqrt{\pi}} \int_0^1 \frac{1}{s^2}\ \mathrm{d}s = \infty.$$

Thus the pointwise formula for the conditional density of u_1 given $u_2 = \frac{1}{2}$ is undefined even though every primitive model assumption is satisfied. Meanwhile,

$$\int_0^1 \int_0^1 \rho(s,t)\ \mathrm{d}t\ \mathrm{d}s = 1,$$

so Lemma E.1 gives a finite conditional normalizer for almost every u_2. The exceptional fiber has zero measure, and the conditional-law class, a.e. identity test, and a.e. foliation test remain well defined. The construction separates joint integrability from pointwise fiber normalizability. Moreover, equation 288 gives $\partial_s \log \rho \not\equiv 0$, so Lemma E.2 makes node 1 fail the non-target identity, whereas the true sink 2 has the continuous target representative $C_{2,e_2} = F_2$.

Disconnected conditional support. Connected conditional-CDF fibers turn the differential wedge test into a global statement. On the disconnected domain

$$\mathcal{D} = ((-2,-1) \cup (1,2)) \times (0,1),$$

let $D(x,u) = u$, and set $C(x,u) = u$ on the left component and $C(x,u) = u^2$ on the right component. Both functions have positive u-derivative in the interior and the conditional-CDF boundary values. On the left component, $\mathrm{d}C = \mathrm{d}u,\ \mathrm{d}D = \mathrm{d}u$, whereas on the right component, $\mathrm{d}C = 2u\ \mathrm{d}u,\ \mathrm{d}D = \mathrm{d}u$.

Thus, on both components, $\mathrm{d}C \wedge \mathrm{d}D = 0$. Now take $x_L = (-3/2, 1/2), x_R = (3/2, 1/2)$.

They lie in different connected components and satisfy

$$D(x_L) = D(x_R) = \frac{1}{2}, \qquad C(x_L) = \frac{1}{2}, \qquad C(x_R) = \frac{1}{4}.$$

If a global scalar Γ with $C = \Gamma \circ D$ existed, it would have to satisfy simultaneously $\Gamma(1/2) = 1/2$ and $\Gamma(1/2) = 1/4$, a contradiction. Thus the differential criterion requires the connected-fiber hypothesis to imply a global composition; the main a.e. foliation definition tests that global composition directly.