One intervention per latent variable: generic identification of nonlinear causal representations on fixed-sign compact strata

CausalSmith*

September 12, 2026

Abstract

This paper studies nonlinear causal representation learning from an observational environment and one unknown-target perfect intervention for each scalar latent variable. In a positive C^3 compact-cube latent model with shared C^2 diffeomorphic mixing, causal minimality, and fixed own-coordinate derivative signs, likelihood ratios between interventional and observational laws define observable one-dimensional ratio coordinates. Comparing their laws across environments with Gaussian maximum mean discrepancy yields a directed ratio-discrepancy graph. The first result establishes that, in every nonempty fixed-sign stratum, the mechanisms for which these ratio laws separate all transported ancestral covers form an open dense set. On this cover-separated population class, a decoder recovers the transitive closure of the target-permuted latent DAG, constructs conditional-rank coordinates that agree with monotone transforms of the intervened latent variables, aligns intervention labels with coordinates, and prunes parents exactly by conditional independence under the observational law. Within the same structural class, equality of the observed environment-law family identifies the observed world up to componentwise C^2 coordinate changes and simultaneous relabeling of graph and targets. A sample-split procedure adds familywise-valid confidence edges for ancestral discoveries, conditional on a simultaneous first-stage L^1 likelihood-ratio error event. On the overlapping smooth compact-support faithful regime, the result gives an arbitrary-dimensional one-perfect-intervention-per-node counterpart to the bivariate law-separation route of von Kügelgen et al. [2023].

1 Introduction

High-dimensional measurements often arise from lower-dimensional causal variables observed through an unknown representation. Causal representation learning asks when distributional variation across environments identifies those variables, their causal graph, and the intervention labels that generated the variation [Schölkopf et al., 2021]. In nonlinear settings with invertible mixing, this question joins three traditions: graphical causal discovery from conditional independences [Pearl, 2009, Spirtes et al., 2000, Peters et al., 2017], identifiable nonlinear independent component analysis from distributional changes [Hyvärinen and Morioka, 2016, Hyvärinen et al.,

*Machine-generated by the CausalSmith research pipeline. Each displayed formal statement is either mapped to a Lean declaration or marked as a presentation-level definition, and each displayed proof renders a Lean proof, as recorded in the verification appendix (scope, toolchain, commit, and any statement conditional on a published input). Cited work otherwise supplies framing and positioning.

2019, Khemakhem et al., 2020, Hyvärinen et al., 2023], and interventional discovery when targets are known, unknown, or partially specified [Hauser and Bühlmann, 2012, Yang et al., 2018, Jaber et al., 2020, Dai et al., 2025, Zhang et al., 2026].

The setting here is an observational law and n interventional laws generated by one perfect intervention for each of n scalar latent coordinates. The target labels are permuted, so the analyst observes environment laws before recovering the alignment between environments and latent variables. The latent variables live on the compact cube, the observation map is shared and diffeomorphic across environments, the local mechanisms are positive and smooth, and each intervention likelihood ratio has a fixed own-coordinate derivative sign. Under these conditions, the likelihood ratio R_i, the density ratio for environment i against the observational law, is an observable scalar function whose latent form depends on the intervened coordinate and its parents.

The paper's central object is the law of one ratio coordinate under another environment. For each ordered pair (j, i), the Gaussian maximum mean discrepancy D_{ji}, the kernel discrepancy between the laws of R_i under environments 0 and j, measures whether intervention j changes the distribution of ratio coordinate i. Theorem 2 shows that, within every nonempty positive smooth causal-minimal fixed-sign stratum, the class on which these discrepancies are positive for all transported ancestral covers is open and dense. The proof uses explicit regular witnesses from Theorem 1 and analytic perturbations along normalized affine paths to move edge-specific second-moment contrasts away from cancellation.

On that generic population class, Theorem 3 gives an exact decoder. Starting from the environment laws, $\mathscr{D}$, the population decoder, forms the ratio-discrepancy graph, chooses a topological order, constructs conditional distribution transforms of log-ratios, aligns each environment with the coordinate recovered by its own rank, and prunes candidate predecessors using conditional independence under the observational law. The ratio graph recovers the transitive closure of G^π, the target-permuted DAG; the conditional ranks equal $Q_{\pi(i)}(V_{\pi(i)})$, the intervention distribution transform of the target latent coordinate, or its sign-reversed version; and causal minimality makes the true parent set the unique inclusion-minimal admissible set. Thus the output graph is exactly G^π, the latent DAG expressed on environment labels.

The result contributes to the unknown-intervention literature by using likelihood-ratio laws in place of scores, support geometry, quadratic precision structure, or supplied target metadata. The closest invertible-mixing comparator is von Kügelgen et al. [2023], who establish a bivariate one-intervention result under a continuous law-separation condition and an arbitrary-dimensional result with two paired perfect interventions per node. On the overlapping positive smooth compact-support faithful regime, Theorem 3 supplies an arbitrary-dimensional one-perfect-intervention-per-node population recovery statement and gives the Gaussian-MMD law-separation witness in the two-variable case. Relative to target-aligned approaches such as Wendong et al. [2023] and Yao et al. [2025], the ratio construction recovers the graph, order, and target alignment from the same observed law family.

The statistical layer translates population discrepancies into finite-sample edge statements. Theorem 4 studies a sample-split design: independent training folds produce fitted positive likelihood ratios, and independent evaluation folds compute empirical Gaussian-MMD discrepancies. Given a simultaneous first-stage L^1 ratio-error radius $a_N(\eta)$, the theorem provides a familywise event on which every selected edge in the confidence graph has an ancestral interpretation in the latent DAG. When every transported ancestral-cover discrepancy exceeds twice the simultaneous radius, the selected graph has the same transitive closure as G^π. This connects the ratio decoder to kernel two-sample concentration [Gretton et al., 2012, Schrab et al., 2023] while keeping the first-stage ratio-estimation requirement explicit.

The final section formulates the next coordinate-error problem in a bounded Hölder regime. Under smoothness, lower-envelope, own-derivative margin, shared-mixing geometry, observed log-ratio regularity, and conditional-design assumptions, Remark 2 states a generated-rank rate target for conditional-CDF estimation with fitted log-ratio responses and fitted conditioning covariates. The target combines second-stage smoothing bias, local empirical variation, and first-stage C^1 error propagated through bandwidth neighborhoods, drawing on conditional-CDF and generated-regressor theory [Xie, 2023, Mammen et al., 2012, Lee, 2018].

The whole argument rests on two observations that the rest of the paper makes exact. Under the stated positivity, smoothness, causal-minimality, shared-diffeomorphic-mixing, one-perfect-intervention, fixed-sign, and cover-separation conditions, the ratios R_i, the density ratios for interventional environments against the observational law, support ancestry recovery through population discrepancies, and the conditional-rank construction returns a monotone transform of the corresponding latent coordinate. Together with causal minimality, shared diffeomorphic mixing, and the one-perfect-intervention-per-node setup, positivity, smoothness, the sign condition, and cover separation make those two steps exact; the bookkeeping objects introduced below state these conditions precisely. A reader who wants the argument before its apparatus can read Algorithm 1 and Theorem 3 first and consult Section 3 for the objects they use. The formal layer accompanying the paper is machine-checked, with precise scope recorded in Section A.

Section 2 situates the paper in graphical causal inference, nonlinear ICA, causal representation learning, unknown-target interventions, and kernel testing. Section 3 introduces the positive compact-cube model, observed worlds, likelihood ratios, and ratio-law discrepancies. Section 4 proves generic ratio separation. Section 5 gives the population decoder and its exact recovery theorem. Section 6 establishes sample-split confidence edges, and Section 7 records the generated-rank frontier and further directions.

2 Related work

Graphical causal inference supplies the language for the paper's target object: a latent causal model on n scalar variables, indexed by the labeled vertex set $[n]$ and a finite DAG G. Classical constraint-based discovery identifies Markov equivalence classes from conditional independence information under faithfulness-type restrictions [Meek, 1995]. Recent work sharpens the role of typical or generic faithfulness in graphical models [Boeken et al., 2024]. The present analysis uses that tradition in a different observational object: ratios across environments generated by one intervention per latent coordinate. The resulting separation statements characterize when the ratio laws distinguish the transported graph, rather than when conditional independences alone determine a Markov equivalence class.

Identifiable nonlinear ICA provides a second point of contact. The modern line from temporal, auxiliary-variable, and exponential-family restrictions establishes identifiability of latent coordinates through variation in observed distributions [Hyvärinen et al., 2001, Hyvärinen and Morioka, 2016, Hyvärinen et al., 2019, Khemakhem et al., 2020, Hyvärinen et al., 2023]. Causal representation learning extends this idea by connecting identifiable coordinates to causal variables and interventions [Brehmer et al., 2022, Lippe et al., 2022, Squires et al., 2022, Zhang et al., 2023]. Here the source of variation is an observed environment family generated by a shared mixing map f, with observed laws $P^0, \ldots, P^n$ obtained from single-target interventions whose labels are permuted by π. The shared invertible observation map keeps the ratio law invariantly comparable across environments, while the unknown target alignment makes the graph recovery problem genuinely interventional.

The closest comparator in the invertible-mixing and unknown-target direction is von Kügelgen et al. [2023]. On the overlapping smooth, compact-support, faithful regime, both analyses exploit multiple environments whose targets are initially unlabeled. von Kügelgen et al. [2023, Theorem 3.2] settles the bivariate case from one perfect intervention per node under a continuous law-separation condition, and von Kügelgen et al. [2023, Theorem 3.4] reaches arbitrary dimension from two paired perfect interventions per node, with the one-intervention extension beyond two latent variables stated there as a conjecture. The present paper closes that count in a smaller model class rather than in theirs: on a positive C^3 compact latent cube with C^2 mixing, fixed own-derivative signs, and causal minimality, one perfect intervention per node suffices in every dimension, and the separation condition that makes this work is proved generic rather than assumed. The two results overlap rather than nest: their model class is the full Euclidean latent space at C^1 regularity, while the class used here is the compact cube at C^2, and the comparison therefore holds on the intersection. Inside that intersection, positive Gaussian-kernel discrepancy supplies the bivariate law-separation witness that von Kügelgen et al. [2023, Theorem 3.2] requires. The present paper characterizes a complementary ratio-law route: likelihood ratios R_i and log-ratios L_i produce observable discrepancies that separate transported ancestral covers and then support population decoding of the intervention labels and parent-pruned graph.

Several recent papers study interventional causal discovery with unknown or partially known targets. Hauser and Bühlmann [2012] analyzes interventional Markov equivalence under known targets, Yang et al. [2018] studies general interventional equivalence classes, and Jaber et al. [2020], Dai et al. [2025], and Zhang et al. [2026] develop graphical and algorithmic theory for target uncertainty. Related latent-variable and representation-learning approaches include Kivva et al. [2021], Xie et al. [2024], Jiang and Aragam [2023], Buchholz et al. [2023], Ahuja et al. [2023], Varıcı et al. [2024], Varıcı et al. [2025], Bing et al. [2024], and Ng et al. [2025]. Score-based analyses of the same design carry the closest published intervention counts. Varıcı et al. [2025] identify latent variables from score differences across interventional environments and show that one stochastic hard intervention per node suffices under linear transformations, while two per node suffice under general transformations; Varıcı et al. [2024] develop the general-identifiability side of that program. The present result keeps one perfect intervention per node under a nonlinear C^2 diffeomorphic transformation, at the price of the compact positive fixed-sign class and a genericity condition, and it works with likelihood-ratio laws and their kernel embeddings rather than with score functions. Within this literature, Wendong et al. [2023] and Yao et al. [2025] are especially close in their use of interventional or invariant structure for causal representations. The comparison turns on inputs: those approaches use graph, order, or target-alignment information appropriate to their identification strategies, whereas the ratio construction starts from the environment laws and recovers an observable directed discrepancy graph from pairwise ratio variation.

The statistical layer draws on kernel mean embeddings and maximum mean discrepancy. The Gaussian-kernel discrepancy construction uses the bounded kernel $k(a, b) = \exp(-(a-b)^2)$ to compare one-dimensional ratio laws, following the embedding and two-sample testing literature [Sriperumbudur et al., 2010, 2011, Gretton et al., 2007, Smola et al., 2007, Gretton et al., 2012, Schrab et al., 2023]. For the finite-sample confidence graph, sample splitting and independent environment samples align the empirical discrepancy analysis with standard concentration arguments for kernel averages. The generated-rank discussion uses tools for conditional distribution estimation and generated regressors, especially local conditional CDF theory and partial-mean arguments with first-stage inputs [Xie, 2023, Mammen et al., 2012, Lee, 2018]. Recent sample-complexity work for causal representation learning and few-environment recovery, including Acartürk et al. [2024] and Lee et al. [2026], provides the surrounding statistical context for translating population separation into finite-sample procedures. The confidence-edge guarantee proved here is a familywise statement about which edges are selected, valid on an event whose first-stage component is an assumed uniform ratio-error contract. Remark 2 states, as a target for later work, the rate a conditional-rank coordinate estimator would have to meet, expressed through an abstract first-stage C^1 rate rather than through any particular estimator.

3 Setup and ratio observables

Throughout the paper, n is the ambient number of scalar latent variables, $[n] = \{1, \dots, n\}$ is the labeled vertex set, and G is a finite labeled DAG. Vectors are written componentwise, with latent coordinates $V = (V_1, \dots, V_n)$ taking values in the compact cube $[0, 1]^n$. Generic constants may change from line to line, norms are taken in the space indicated by their subscripts, and all index ranges are over $[n]$ unless stated otherwise. The sign vector s records the prescribed own-coordinate derivative directions.

The structural model is specified by observational conditional densities p_i, parent-independent intervention densities q_i, a shared observation map f, and observed environment laws $P^0, \dots, P^n$ on the observed support $\mathcal{X}$. The first group of conditions fixes the smooth positive latent model and the common invertible observation system.

Assumption 1. [`ass:positive-normalized-smooth-mechanisms`] (Positive smooth mechanisms) *For every $i \in [n]$, the observational mechanism p_i and the intervention density q_i are strictly positive, normalized C^3 densities on their respective closed cubes.*

The smooth full-support compact-cube density condition in Assumption 1 places the observational and intervention mechanisms on a common regular support for the ratio construction in this paper. Positivity makes likelihood ratios well-defined on the latent cube, while normalization preserves the interpretation of each carrier as a conditional or marginal density.

Assumption 2. [`ass:shared-diffeomorphic-mixing`] (Shared diffeomorphic mixing) *The same map f is a C^2 diffeomorphism from $[0, 1]^n$ onto $\mathcal{X}$ in every environment.*

The shared diffeomorphic mixing condition in Assumption 2 is the standard invertible-observation restriction for this setting [von Kügelgen et al., 2023]. It ensures that every environment is observed through the same smooth coordinate change, so differences across $P^0, \dots, P^n$ come from the latent intervention laws rather than from environment-specific measurement maps.

We next make the graph notation and latent product law explicit.

Definition 1. [`synth_9`] (Parent set $\mathrm{pa}_G(i)$) *For a finite labeled directed acyclic graph G on $[n]$ and a vertex $i \in [n]$, $\mathrm{pa}_G(i)$ denotes the set of parents of i in G:*

$$\mathrm{pa}_G(i) = \{j \in [n] : j \to i \textit{ is an edge of } G\}.$$

For $A \subseteq [n]$, write $V_A = (V_j)_{j \in A}$.

The notation in Definition 1 is used throughout to express local Markov factorization through parent coordinates. In particular, p_i depends on its own coordinate and the parent set $\mathrm{pa}_G(i)$.

Definition 2. [synth_2] (Observational product density) *For a finite dimension n, a labeled DAG G on $[n]$, a mechanism θ with observational conditional-density carriers p_i, and a latent state $v = (v_1, \dots, v_n) \in \mathbb{R}^n$, define*

$$p^0(v) = \prod_{i=1}^{n} p_i\big(v_i \mid v_{\mathrm{pa}_G(i)}\big).$$

The product density p^0 in Definition 2 is the observational latent law associated with G and the local mechanisms. The target-i intervention law p^i replaces the ith observational mechanism by q_i, and this replacement is formalized in the observed laws below.

Assumption 3. [ass:one-perfect-intervention-per-node] (Perfect intervention laws) *For the mechanism θ, the shared mixing map f, the observed laws $P^0, \dots, P^n$, the target permutation π, and the supplied likelihood ratios $R_1, \dots, R_n$, the following conditions hold:*

- ***(Observational pushforward.)*** *The observational law satisfies*

$$P^0 = f_{\#}\big(p^0(v)\, dv\big), \qquad p^0(v) = \prod_{i=1}^{n} p_i(v_i \mid v_{\mathrm{pa}_G(i)}),$$

  with $v \in [0,1]^n$.

- ***(Single-target pushforwards.)*** *For every environment $e \in [n]$,*

$$P^e = f_{\#}\big(p^{\pi(e)}(v)\, dv\big), \qquad p^{\pi(e)}(v) = q_{\pi(e)}(v_{\pi(e)}) \prod_{\ell \neq \pi(e)} p_\ell(v_\ell \mid v_{\mathrm{pa}_G(\ell)}).$$

- ***(Radon–Nikodym ratios.)*** *For every $e \in [n]$,*

$$R_e(X) = \frac{dP^e}{dP^0}(X) \quad P^0\text{-almost surely.}$$

- ***(Latent ratio formula.)*** *For every $e \in [n]$ and every $v \in [0,1]^n$,*

$$R_e(f(v)) = \frac{q_{\pi(e)}\big(v_{\pi(e)}\big)}{p_{\pi(e)}\big(v_{\pi(e)} \mid v_{\mathrm{pa}_G(\pi(e))}\big)}.$$

The unknown-target stochastic perfect-intervention coverage condition in Assumption 3 is standard in this literature [von Kügelgen et al., 2023]. It ties each nonobservational environment to exactly one latent target through the permutation π, and it supplies the likelihood-ratio representation that turns observed laws into one-dimensional ratio coordinates. When $X = f(V)$, the ratio for an intervention targeting i depends on V_i and its parents through the displayed quotient.

Assumption 4. [ass:fixed-own-derivative-sign] (Fixed derivative sign) *For every $i \in [n]$,*

$$s_i\, \partial_{v_i} \log\{q_i(v_i)/p_i(v_i \mid v_{\mathrm{pa}_G(i)})\} > 0$$

throughout the closed cube.

Assumption 4 orients each ratio in its own coordinate through a compact-support fixed-sign derivative condition. It records that the log-ratio changes monotonically in the intervened coordinate after applying the prescribed sign s_i, which later supports the rank-based decoding step.

Assumption 5. [`ass:causal-minimality`] (Causal minimality) *For every edge $j \to i$ of G, the observational law p^0 satisfies*

$$V_i \not\perp\!\!\!\perp V_j \mid V_{\mathrm{pa}_G(i)\setminus\{j\}}.$$

The local edge form of causal minimality in Assumption 5 is the standard positivity-compatible requirement that each displayed parent contributes to the observational conditional law [Spirtes et al., 2000]. It anchors the graph G to conditional distributions under p^0, so the formal stratum below consists of mechanisms whose graph edges are statistically active.

Definition 3. [`def:model-stratum`] (Model stratum $\Theta_{G,s}$)

$$\Theta_{G,s} = \{((p_i, q_i))_{i=1}^n : \textit{Assumptions } 1, 4 \textit{and } 5 \textit{ hold for } G \textit{ and } s\}.$$

The set $\Theta_{G,s}$ is equipped with the relative product C^2 topology.

The model stratum $\Theta_{G,s}$ collects the latent mechanisms satisfying the positivity, minimality, and signed-ratio restrictions for the fixed graph and sign vector. The relative product C^2 topology is the topology used later for generic separation statements.

The remaining definitions describe the observed data object and the two population separation classes used by the ratio-decoding argument.

Definition 4. [`synth_19`] (Observed world data) *For a mechanism θ on the labeled DAG G with ambient vertex set $[n]$, an observed world is the following data:*

$$\big(f, \mathrm{unmix}, \pi, (P^e)_{e=0}^n, (R_i)_{i\in[n]}\big).$$

Here f and unmix *are maps from the latent state space $\mathbb{R}^{[n]}$ to itself, π is a permutation of $[n]$, P^e is a measure on $\mathbb{R}^{[n]}$ for each environment $e \in \{0, \dots, n\}$, and $R_i : \mathbb{R}^{[n]} \to \mathbb{R}$ is a real-valued ratio carrier for each target $i \in [n]$.*

Definition 4 packages the objects that are available to population procedures: common coordinates, environment laws, target labels, and ratio carriers. In the canonical observed world used below, the mixing and unmixing maps are identities, so the same definitions can be evaluated directly on latent laws.

Definition 5. [`def:edge-separated-set`] (Edge separation set) *For a target-label permutation π, define*

$$\mathcal{V}_{G,s,\pi} = \Big\{\theta \in \Theta_{G,s} : F^{\mathrm{can}}_{\pi^{-1}(j),\pi^{-1}(i)}(\theta;\pi) \neq 0 \textit{ for every edge } j \to i \textit{ of } G\Big\}.$$

Here $F^{\mathrm{can}}_{ab}(\theta;\pi)$ is the second-moment contrast in the canonical observed world generated by θ and π:

$$F^{\mathrm{can}}_{ab}(\theta;\pi) = \int \left(\frac{q_{\pi(b)}(v_{\pi(b)})}{p_{\pi(b)}(v_{\pi(b)} \mid v_{\mathrm{pa}_G(\pi(b))})}\right)^2 dP^a(v) - \int \left(\frac{q_{\pi(b)}(v_{\pi(b)})}{p_{\pi(b)}(v_{\pi(b)} \mid v_{\mathrm{pa}_G(\pi(b))})}\right)^2 dP^0(v).$$

The set $\mathcal{V}_{G,s,\pi}$ records mechanisms for which every direct edge has a nonzero canonical second-moment contrast. The contrast F_{ji} compares the squared ratio coordinate under the observational law and under another intervention environment after the target relabeling π.

Definition 6. [`synth_13`] (Population MMD discrepancy) *For a complete real Hilbert space H, let U be a feature map $r \mapsto \Phi(r) \in H$ with $\|\Phi(r)\|_H = 1$ and*

$$\langle \Phi(a), \Phi(b) \rangle_H = k(a,b) = \exp(-(a-b)^2).$$

For an observed world W with environment laws $P^0, \dots, P^n$ and ratio coordinates $R_1, \dots, R_n$, and for $j, i \in [n]$, define the population MMD discrepancy $D_{ji}(W)$ by

$$D_{ji}(W) = \left\| \int \Phi(r)\, d\big((R_i)_\# P^0\big)(r) - \int \Phi(r)\, d\big((R_i)_\# P^j\big)(r) \right\|_H.$$

When the observed world is fixed by the context we abbreviate $D_{ji}(W)$ to D_{ji}.

The discrepancy D_{ji} in Definition 6 is the maximum mean discrepancy between the one-dimensional laws of ratio coordinate i under environments 0 and j, using the bounded Gaussian kernel and its Hilbert-space feature map. This embeds the ratio-law comparison in the standard kernel two-sample framework [Gretton et al., 2012, Sriperumbudur et al., 2010, 2011].

Definition 7. [`def:cover-separated-set`] (Cover-separated mechanisms) *Let H be a complete real Hilbert space, and let U be a feature map $\phi : \mathbb{R} \to H$ satisfying*

$$\|\phi(r)\|_H = 1, \qquad \langle \phi(a), \phi(b) \rangle_H = k(a,b)$$

for all $r, a, b \in \mathbb{R}$. For a target permutation π, let $W^{\mathrm{can}}_{\theta,\pi}$ be the canonical observed world generated by θ and π: the mixing and unmixing maps are the identity, the target permutation is π, the observational environment has law induced by θ's observational law, intervention environment e has law induced by the intervention targeting $\pi(e)$, and its ratio carrier is

$$R_e(v) = \frac{q_{\pi(e)}\big(v_{\pi(e)}\big)}{p_{\pi(e)}\big(v_{\pi(e)} \mid v_{\mathrm{pa}_G(\pi(e))}\big)}.$$

The cover-separated set is

$$\mathcal{U}_{G,s,\pi} = \left\{ \theta \in \Theta_{G,s} : D_{ji}\big(W^{\mathrm{can}}_{\theta,\pi}\big) > 0 \text{ for every } j, i \in [n] \text{ with } \pi(j) \lessdot_G \pi(i) \right\},$$

where

$$D_{ji}\big(W^{\mathrm{can}}_{\theta,\pi}\big) = \left\| \int \phi(r)\, d\mathsf{P}^0_{\theta,\pi,i}(r) - \int \phi(r)\, d\mathsf{P}^j_{\theta,\pi,i}(r) \right\|_H,$$

$\mathsf{P}^0_{\theta,\pi,i}$ is the law of ratio i under the observational environment of $W^{\mathrm{can}}_{\theta,\pi}$, and $\mathsf{P}^j_{\theta,\pi,i}$ is the law of ratio i under intervention environment j of $W^{\mathrm{can}}_{\theta,\pi}$.

The cover relation $j \lessdot_G i$ selects ancestral covers in the latent DAG, and Definition 7 requires a positive ratio-law discrepancy on every such transported cover after applying π. Thus $\mathcal{U}_{G,s,\pi}$ is the population class on which the observable discrepancy graph contains the cover information needed by the decoder developed in the next section.

4 Generic ratio separation

The separation sets in Definitions 5 and 7 turn observable ratio laws into graph information. This section supplies the two ingredients behind that step. First, explicit three-node mechanisms show both a strict direct-edge discrepancy and an exact cancellation. Second, the genericity result shows that, within any nonempty fixed-sign stratum, the separating behavior is open and dense.

We begin with two concrete mechanisms on the graph $1 \to 2$, with node 3 isolated. The first mechanism is the separating endpoint used to certify that the relevant analytic contrast can be nonzero.

Definition 8. [`def:sparse-witness`] (Sparse witness θ^{sp}) *On the graph* $1 \to 2$ *with node* 3 *isolated, define* θ^{sp} *by*

$$h(z) = 2z - 1,$$

$$q_i(z) = \frac{4\exp(4z)}{\exp(4) - 1},$$

$$p_2(y \mid x) = 1 + 0.1h(x)h(y),$$

and

$$p_1 = p_3 = 1.$$

For a sign vector with a negative prescribed sign in coordinate i*, the corresponding coordinate is reflected by the map* $z \mapsto 1 - z$*.*

The affine dependence in Definition 8 creates a direct response of the ratio coordinate at node 2 to interventions at node 1. Reflections preserve the fixed-sign convention from Assumption 4, so the same construction serves every prescribed sign pattern.

The next mechanism has the same graph and the same intervention tilts, but it places the parent dependence in a form that exactly balances the ratio law across the relevant environments.

Definition 9. [`def:cancellation-witness`] (Cancellation witness θ^{can}) *On the graph* $1 \to 2$ *with node* 3 *isolated, define* θ^{can} *by*

$$q_i(z) = \frac{4\exp(4z)}{\exp(4) - 1}, \qquad p_1 = 1, \qquad p_3 = 1,$$

$$h(y) = 2y - 1, \qquad H(x) = \frac{\exp(4x) - 1}{\exp(4) - 1} - x, \qquad p_2(y \mid x) = 1 + 0.1H(x)h(y).$$

For any other sign vector, θ^{can} *is obtained from this mechanism by coordinate reflections.*

The contrast between Definitions 8 and 9 is useful because the ratio-law condition can hold strictly at one regular point and vanish at another regular point. Thus the generic result below is a topological statement about typical mechanisms in the stratum, with the cancellation example marking the boundary behavior that analytic perturbations must move away from.

For the witness calculation and the generic statement, it is helpful to name the observed law system induced by a mechanism and a shared coordinate map.

Definition 10. [synth_28] (Coherence of observed laws with an observed world) *Let W be an observed world generated by a mechanism $\theta = ((p_i, q_i))_{i\in[n]}$ on G, with shared mixing map $f : [0,1]^n \to \mathcal{X}$, inverse f^{-1}, and target permutation π. We say that observed laws $(P^0, \dots, P^n)$ are* coherent with W *when $P^0 = f_\# p^0$ and, for each interventional environment label $e \in [n]$, $P^e = f_\# p^{\pi(e)}$, where*

$$p^0(v) = \prod_{i=1}^{n} p_i(v_i \mid v_{\mathrm{pa}_G(i)})$$

and

$$p^a(v) = q_a(v_a) \prod_{\ell\neq a} p_\ell(v_\ell \mid v_{\mathrm{pa}_G(\ell)}) \qquad (a \in [n]).$$

Equivalently, under P^0 and P^e, the unmixed variable $V = f^{-1}(X)$ has densities p^0 and $p^{\pi(e)}$, respectively, on $[0,1]^n$.

Definition 10 records the population object available to the ratio analysis: the same observation map pushes forward the observational law and each single-target intervention law. In the explicit witness world, the shared map and target permutation are identities, so the latent and observed densities coincide.

Definition 11. [synth_27] (Sparse-witness environment laws) *Let W^{sp} be the observed world generated by the sparse witness mechanism θ^{sp} with identity mixing and identity target permutation on the three-node DAG $1 \to 2$ with node 3 isolated. For $e \in \{0,1,2,3\}$, P^e_{sp} denotes the environment-e law in this world. Thus P^0_{sp} is the observational law with density*

$$p^0_{\mathrm{sp}}(v) = p^{\mathrm{sp}}_1(v_1)\, p^{\mathrm{sp}}_2(v_2 \mid v_1)\, p^{\mathrm{sp}}_3(v_3),$$

and, for $e \in \{1,2,3\}$, P^e_{sp} is the intervention law at target e, with density

$$p^e_{\mathrm{sp}}(v) = q^{\mathrm{sp}}_e(v_e) \prod_{\ell\neq e} p^{\mathrm{sp}}_\ell(v_\ell \mid v_{\mathrm{pa}_G(\ell)}).$$

Because the mixing map is the identity, these latent densities are also the corresponding observed laws.

The graph-theoretic and probabilistic conventions used by the generic result are standard. We state them here to make the separation theorem self-contained.

Definition 12. [synth_24] (d-separation in a DAG) *Let G be a DAG on $[n]$, and let $A, B, C \subseteq [n]$ be pairwise disjoint. A path in the underlying undirected graph of G is C-active when every non-collider on the path is outside C, and every collider on the path belongs to C or has a directed path in G to some vertex of C. The sets A and B are d-separated by C in G when there is no C-active path in the underlying undirected graph of G with one endpoint in A and the other endpoint in B.*

Definition 13. [synth_10] (Faithfulness to G) *An observational law p^0 on $V = (V_1, \dots, V_n)$ is faithful to the directed acyclic graph G when, for all disjoint subsets $A, B, C \subseteq [n]$,*

$$V_A \perp\!\!\!\perp_{p^0} V_B \mid V_C \quad \Longleftrightarrow \quad A \text{ and } B \text{ are d-separated by } C \text{ in } G.$$

Equivalently, the conditional independence relations of p^0 are exactly the d-separation relations encoded by G.

Definition 14. [synth_23] (Ancestral order in a DAG) *For a DAG G on the finite labeled vertex set $[n]$ and vertices $j, i \in [n]$, write*

$$j <_G i$$

when j is a strict ancestor of i in G, equivalently when there exists a directed path in G of positive length from j to i.

Definition 15. [synth_4] (Ancestral cover relation) *For a DAG G on the finite labeled vertex set $[n]$, define $j \lessdot_G i$ to be the cover relation in the ancestor order of G:*

$$j \lessdot_G i \quad \Longleftrightarrow \quad j <_G i \text{ and } \{k \in [n] : j <_G k <_G i\} = \varnothing.$$

The cover relation in Definition 15 is the transitive-reduction analogue for the ancestor order. It selects the ancestral comparisons whose ratio-law discrepancies form the population input for the decoder in the next section.

We also record the ratio-law notation used in the witness certificate and in the open-dense theorem.

Definition 16. [synth_1] (Latent state space) *For $n \in \mathbb{N}$, the latent state $V = (V_1, \ldots, V_n)$ ranges over the real coordinate space*

$$\mathbb{R}^{[n]} = \{v : [n] \to \mathbb{R}\}.$$

Definition 17. [synth_20] (Observational, intervention, and ratio laws) *For a mechanism $\theta = ((p_i, q_i))_{i=1}^n \in \Theta_{G,s}$, let $\mathcal{L}(\theta)$ denote the observational law on the observed support $\mathcal{X}$ generated by*

$$p_\theta^0(v) = \prod_{\ell=1}^n p_\ell(v_\ell \mid v_{\mathrm{pa}_G(\ell)})$$

and the shared mixing map $X = f(V)$. For $i \in [n]$, let $\mathrm{do}(i)\theta$ denote the perfect-intervention law obtained from θ by replacing the i-th observational conditional mechanism $p_i(v_i \mid v_{\mathrm{pa}_G(i)})$ with the parent-independent intervention density $q_i(v_i)$, so its latent density is

$$p_\theta^i(v) = q_i(v_i) \prod_{\ell \neq i} p_\ell(v_\ell \mid v_{\mathrm{pa}_G(\ell)}),$$

again pushed forward by f to $\mathcal{X}$. Given a target permutation π, write $P_{\theta,\pi}^0 = \mathcal{L}(\theta)$ and, for $e \in [n]$, $P_{\theta,\pi}^e = \mathrm{do}(\pi(e))\theta$. For $i \in [n]$, let

$$R_i^{\theta,\pi} = \frac{dP_{\theta,\pi}^i}{dP_{\theta,\pi}^0}$$

be the observed density ratio for environment label i. For $e \in \{0, \ldots, n\}$, define

$$\mathcal{L}_e^{\theta,\pi,i} = (R_i^{\theta,\pi})_\# P_{\theta,\pi}^e,$$

the probability law on $\mathbb{R}$ obtained by pushing the canonical environment-e observed law forward through the ratio coordinate $R_i^{\theta,\pi}$.

Definition 18. [synth_3] (Intervention distribution function) *For a mechanism θ on a DAG G over $[n]$, with parent-independent intervention density q_i at coordinate i, define*

$$Q_i(v) = \int_0^v q_i(u)\, du$$

for $i \in [n]$ and $v \in \mathbb{R}$.

One further condition is recorded here for use by the explicit witness and by the comparison with prior work: the usual faithfulness condition, linking conditional independences in the observational density to d-separation in the graph. It is a strictly stronger requirement than the causal minimality used by Theorem 2 and Theorem 3, which quantify over the causal-minimal stratum; faithfulness enters only where it is cited by name.

Assumption 6. [ass:faithfulness] (Faithful observational law) *The observational latent density p^0 is faithful to the DAG G.*

Assumption 6 is the standard causal faithfulness condition [Spirtes et al., 2000]. It rules the observational conditional independences to be exactly those encoded by d-separation, so graphical ancestry and the statistical independences used by causal minimality are aligned.

The explicit witnesses now certify the two endpoint behaviors needed for the perturbation argument.

Theorem 1. [prop:sparse-witness-certificate] (Sparse witness certificate) *For every sign vector $s \in \{-1, 1\}^3$, let θ^{sp} be the construction in Definition 8 and let θ^{can} be the construction in Definition 9, both on the three-node DAG $1 \to 2$ with node 3 isolated. Let W^{sp} and W^{can} be the corresponding observed worlds with identity mixing and identity target permutation, so that environment 0 has the observational law, environment e has the intervention law at target e, and*

$$R_i(v) = \frac{q_i(v_i)}{p_i(v_i \mid v_{\mathrm{pa}_G(i)})}.$$

Then:

- ***(Sparse witness regularity.)*** *The mechanism θ^{sp} satisfies Assumption 1, Assumption 4 with sign vector s, and Assumption 6. Moreover, every observational conditional density p_i^{sp} is C^∞ on $[0,1]^3$, and every intervention density q_i^{sp} is C^∞ on $[0,1]$.*
- ***(Cancellation witness regularity.)*** *The mechanism θ^{can} satisfies Assumption 1, Assumption 4 with sign vector s, and Assumption 6. Moreover, every observational conditional density p_i^{can} is C^∞ on $[0,1]^3$, and every intervention density q_i^{can} is C^∞ on $[0,1]$.*
- ***(Sparse witness separation.)*** *The sparse witness has*

  $$\frac{3}{10000} < \int R_2(x)^2\, dP^0_{\mathrm{sp}}(x) - \int R_2(x)^2\, dP^1_{\mathrm{sp}}(x),$$

  and its Gaussian-kernel population discrepancy satisfies

  $$\frac{5}{100000000} < D_{12}(W^{\mathrm{sp}}).$$

- ***(Cancellation witness equality.)*** *For the cancellation witness, the law of R_2 under the observational environment equals the law of R_2 under environment* 1, *and*

$$D_{12}(W^{\mathrm{can}}) = 0.$$

Theorem 1 gives a regular separating point and a regular cancelling point inside the same elementary graph geometry. The strict lower bounds for the sparse witness show that the direct-edge second-moment contrast and the Gaussian-kernel discrepancy can both detect the edge $1 \to 2$. The cancellation equality shows that equality of ratio laws can occur at structured mechanisms satisfying the same smoothness, sign, and faithfulness requirements, which is exactly the situation addressed by a generic rather than universal separation theorem.

The main population result converts that endpoint calculation into a statement for every nonempty fixed-sign stratum. Its topology is the relative product C^2 topology built into Definition 3.

Theorem 2. [`thm:generic-cover-separation`] (Generic cover separation) *Let n be the ambient number of scalar latent variables. Assume:*

- ***(Finite labeled DAG.)*** *G is a finite labeled DAG on $[n]$.*
- ***(Sign stratum.)*** *s is a sign vector with coordinates in $\{-1, 1\}$, and the stratum $\Theta_{G,s}$ of Definition 3 is nonempty.*
- ***(Target permutation.)*** *π is a permutation of $[n]$.*

Let $\mathcal{V}_{G,s,\pi}$ be the direct-edge separation set of Definition 5, and let $\mathcal{U}_{G,s,\pi}$ be the ancestral-cover separation set of Definition 7 formed with the Gaussian feature map $\phi(a) = k(a, \cdot)$ for $k(a, b) = \exp\{-(a-b)^2\}$. Then $\mathcal{V}_{G,s,\pi}$ is open and dense in $\Theta_{G,s}$, and

$$\mathcal{V}_{G,s,\pi} \subseteq \mathcal{U}_{G,s,\pi}.$$

Moreover, $\mathcal{U}_{G,s,\pi}$ is open and dense in $\Theta_{G,s}$, and its complement in $\Theta_{G,s}$ is meagre, closed, and nowhere dense. If G has no directed edges, then

$$\mathcal{V}_{G,s,\pi} = \Theta_{G,s} \qquad \textit{and} \qquad \mathcal{U}_{G,s,\pi} = \Theta_{G,s}.$$

Theorem 2 establishes that ratio-law separation is typical in every nonempty stratum with fixed own-coordinate derivative signs. Openness says that a separated mechanism remains separated under sufficiently small perturbations of the normalized mechanisms. Density says that every admissible mechanism can be approximated by mechanisms whose direct-edge contrasts are nonzero and whose ancestral-cover discrepancies are positive. The meagreness and nowhere-density clauses give the corresponding topological description of the exceptional set.

The intuition is that the witness in Theorem 1 provides a local direction in which an edge-specific ratio contrast becomes nonzero. Along normalized analytic perturbation paths, a contrast that is nonzero at one endpoint can vanish only on a thin set of perturbation values unless it is identically zero. Applying this reasoning across the finitely many relevant edges and covers yields the open-dense separation class used by the population decoder in the next section.

5 Population decoding

The generic separation result in Theorem 2 supplies the observable input for a population construction. The construction is an oracle on the environment laws: its steps evaluate Radon–Nikodym derivatives, conditional-law versions, and conditional independences exactly, so it is a map from a law family to a graph and a set of coordinates rather than an estimator or an algorithm on data. Its order and pruning steps are well posed precisely on the cover-separated set, where Theorem 3 shows the ratio graph is acyclic and the admissible parent set is unique; on an arbitrary law family the ratio graph may contain cycles and the inclusion-minimal admissible set need not be unique, and the conclusions below are asserted only under the theorem's hypotheses. Starting from the environment laws, the decoder forms the directed ratio-discrepancy graph, uses its order information to compute conditional distribution transforms, aligns each intervention label with the coordinate recovered by its own rank, and then prunes the candidate predecessors through conditional independence under the observational law. The output is an environment-label graph, so we first record the relabeling convention that translates the latent DAG into intervention labels.

Definition 19. [`synth_17`] (Permuted environment graph) *For an observed world W with target permutation π, the environment-label DAG G^π is the directed graph on $[n]$ whose edge relation is*

$$j \to_{G^\pi} i \quad \Longleftrightarrow \quad G \text{ has the edge } \pi(j) \to \pi(i), \qquad j, i \in [n].$$

This graph is the structural object visible after intervention labels have been matched to latent coordinates. Because the target permutation is part of the observed-world description rather than known to the analyst, the decoder is formulated entirely in terms of laws and ratios.

The rank step requires a version of the conditional distribution function selected from observed laws. The next definition fixes that selection in a way that is compatible with regular conditional distributions and continuous support behavior.

Definition 20. [`synth_18`] (Selected conditional CDF) *For an observed probability-law family $P^0, \ldots, P^n$, an ordering, and a vertex i, let B_i be the predecessors of i in that ordering. The selected conditional distribution function*

$$C_i : \mathbb{R} \times \mathbb{R}^{B_i} \to [0, 1]$$

is defined as follows. If there exists a map $C : \mathbb{R} \times \mathbb{R}^{B_i} \to [0, 1]$ such that

- ***(Joint measurability.)*** *$C(t, \ell)$ is measurable as a function of (t, ℓ);*
- ***(Fixed-threshold version.)*** *for every threshold t, the function $\ell \mapsto C(t, \ell)$ agrees almost surely, under the observed target-i law of L_{B_i}, with the raw regular conditional CDF $\widetilde{C}_i(t, \ell)$ of L_i given $L_{B_i} = \ell$;*
- ***(Joint-law version.)*** *the function $(t, \ell) \mapsto C(t, \ell)$ agrees almost surely, under the observed target-i joint law of (L_i, L_{B_i}), with $(t, \ell) \mapsto \widetilde{C}_i(t, \ell)$;*
- ***(Support continuity.)*** *$(t, \ell) \mapsto C(t, \ell)$ is continuous on $\mathbb{R}$ times the support of the observed target-i law of L_{B_i},*

then C_i is one such map selected from the observed laws alone. In the complementary case,

$$C_i(t \mid L_{B_i} = \ell) = \begin{cases} \widetilde{C}_i(t, \ell), & \widetilde{C}_i(t, \ell) \in [0, 1], \\ 0, & \widetilde{C}_i(t, \ell) \notin [0, 1], \end{cases}$$

where $\widetilde{C}_i(t, \ell)$ is the raw regular conditional CDF of L_i given $L_{B_i} = \ell$ under the observed target-i law.

The continuity clause gives a population version of the probability integral transform along the conditional-ratio support. With that convention in place, the decoder $\mathscr{D}$ can be stated as an observable map from environment laws to a pruned environment-label DAG.

Algorithm 1 (Population decoder $\mathscr{D}$)

Given observed laws $(P^0, \ldots, P^n)$, the population decoder $\mathscr{D}$ is defined by the following steps.

1. Compute the likelihood ratios and log-ratios

$$R_i = \frac{dP^i}{dP^0}, \qquad L_i = \log R_i, \qquad i \in [n],$$

   and the Gaussian-MMD discrepancies D_{ji}.

2. Form the observable directed ratio-discrepancy graph

$$H_D = \{j \to i : j \neq i \text{ and } D_{ji} > 0\}.$$

3. Choose a topological ordering of H_D. For each i, let B_i be the predecessors of i in that ordering and define

$$U_i = C_i(L_i \mid L_{B_i}).$$

4. Align environment i with coordinate U_i.

5. For each i, range over $A \subseteq B_i$ and identify the unique inclusion-minimal subset satisfying

$$U_i \perp\!\!\!\perp_{P^0} U_{B_i \setminus A} \mid U_A.$$

   The output is the environment-label DAG

$$G^{\pi}_{\mathscr{D}} = \{a \to i : \ a \in A_i^{\min}\}, \qquad A_i^{\min} = \min_{\subseteq} \left\{ A \subseteq B_i : U_i \perp\!\!\!\perp_{P^0} U_{B_i \setminus A} \mid U_A \right\}.$$

The first two steps use the law discrepancies studied in Theorem 2; the third step converts log-ratios into coordinates through conditional ranks; the final step applies the Markov factorization logic under P^0, in the spirit of graphical-model recovery from conditional independences [Pearl, 2009, Spirtes et al., 2000]. The rank transform has a structural target: conditionally on the recovered predecessors, the log-ratio distribution is the intervention-target distribution induced by the latent mechanism.

Definition 21. [synth_26] (Structural conditional-ratio CDF) *For a mechanism* $\theta = (p_\ell, q_\ell)_{\ell\in[n]}$*, an environment label* i*, and target* $a = \pi(i)$*, write*

$$\lambda_{\theta,a}(w,z) = \log \frac{q_a(w)}{p_a(w \mid z)}, \qquad z \in [0,1]^{|\operatorname{pa}_G(a)|}.$$

The structural conditional-ratio CDF at target $a = \pi(i)$ *is*

$$C^{\mathrm{str}}_{\theta,\pi(i)}(t \mid z) = \int_0^1 \mathbf{1}\{\lambda_{\theta,\pi(i)}(w,z) \le t\}\, q_{\pi(i)}(w)\, dw.$$

Equivalently, $C^{\mathrm{str}}_{\theta,\pi(i)}(\cdot \mid z)$ *is the conditional distribution function of the target log-ratio* L_i *in the target-*$\pi(i)$ *intervention environment, conditional on the latent parent vector* $V_{\operatorname{pa}_G(\pi(i))} = z$*.*

The structural CDF describes the rank that would be computed with latent parents in hand. The theorem below states that the observed conditional CDF selected in Definition 20 agrees with this structural object on the relevant support, so the rank coordinates produced by Algorithm 1 coincide with monotone transforms of the intervention-target latent coordinates.

Representation recovery is stated up to componentwise coordinate changes and a simultaneous relabeling of the DAG and targets. The equivalence relation records exactly those transformations.

Definition 22. [synth_8] (Componentwise C^2 equivalence) *For* $m = 1, 2$*, let* W_m *be an observed world over* (G_m, θ_m) *with mixing map* f_m*, inverse-coordinate map* g_m*, target permutation* π_m*, and observed laws* $(P^{\mathrm{e}}_m)_{e=0}^n$*. With the same ambient number* n *of scalar latent variables, write*

$$\mathcal{X}_m = \{f_m(v) : v \in [0,1]^n\}, \qquad m = 1, 2.$$

The worlds W_1 *and* W_2 *are componentwise* C^2*-equivalent when the following conditions hold:*

- ***(Observed support.)*** *Their observed supports agree:*

$$\mathcal{X}_1 = \mathcal{X}_2.$$

- ***(DAG relabeling.)*** *There is a permutation* σ *of* $[n]$ *such that, for every* $j, i \in [n]$*,*

$$\sigma(j) \to_{G_2} \sigma(i) \iff j \to_{G_1} i.$$

- ***(Target alignment.)*** *For every* $e \in [n]$*,*

$$\pi_2(e) = \sigma(\pi_1(e)).$$

- ***(Coordinate changes.)*** *There are real functions* h_i *and* $\bar{h}_i$*, one pair for each* $i \in [n]$*, whose restrictions to* $[0,1]$ *are* C^2 *and map* $[0,1]$ *into* $[0,1]$*, such that, for every* $i \in [n]$ *and every* $z \in [0,1]$*,*

$$\bar{h}_i(h_i(z)) = z, \qquad h_i(\bar{h}_i(z)) = z.$$

- ***(Unmixing relation.)*** *For every* $x \in \mathcal{X}_1$ *and every* $i \in [n]$*,*

$$g_2(x)_{\sigma(i)} = h_i(g_1(x)_i)\,.$$

This is the usual observational equivalence scale for nonlinear component recovery: the coordinate system is identified up to separate invertible transformations of each latent coordinate, while the graph and target labels move together under the same permutation.

The ratio graph need not equal the parent graph before pruning, because a positive discrepancy can follow an ancestral path. The next definition fixes the graph operation used to state the exact order information recovered by the first stage.

Definition 23. [`synth_16`] (Transitive closure of a directed graph) *For a directed graph H on a vertex set V, $\mathrm{TC}(H)$ denotes its transitive closure: the directed graph on V with an arrow $a \to b$ exactly when there is a directed path of positive length from a to b in H.*

The headline result combines the generic ratio separation condition with conditional ranks and conditional-independence pruning. It gives the population decoding guarantee for the one-perfect-intervention-per-node design.

Theorem 3. [`thm:exact-ratio-decoder`] (Exact ratio decoding) *Fix n, a finite labelled DAG G on $[n]$, and a sign vector $s \in \{-1, +1\}^{[n]}$. Let $\theta = (p_i, q_i)_{i\in[n]}$ be a mechanism on G. Let* world *assign an observed world to every mechanism on G, and write $W = \mathsf{world}(\theta)$, with shared mixing map f, inverse map f^{-1}, target permutation π, observed laws $(P^0, \dots, P^n)$, likelihood ratios $R_i = dP^i/dP^0$, and log-ratios $L_i = \log R_i$. Assume:*

- ***(Smooth positive mechanisms.)*** *θ satisfies Assumption 1.*
- ***(Causal minimality.)*** *θ satisfies Assumption 5.*
- ***(Shared mixing.)*** *The maps f and f^{-1} in W satisfy Assumption 2.*
- ***(One intervention per node.)*** *The laws $(P^0, \dots, P^n)$, ratios $(R_i)_{i\in[n]}$, and target permutation π in W satisfy Assumption 3.*
- ***(Own-derivative signs.)*** *θ satisfies Assumption 4 with sign vector s.*
- ***(Gaussian cover separation.)*** *The stratum point $\theta \in \Theta_{G,s}$ belongs to the cover-separated set of Definition 7 with the Gaussian unit-norm feature map associated with $k(a,b) = \exp\{-(a-b)^2\}$ and target permutation π; equivalently, every transported ancestral cover has strictly positive population Gaussian-MMD discrepancy.*

Let $\mathscr{D}$ be the population decoder of Algorithm 1, let $H_D = \{j \to i : j \neq i \text{ and } D_{ji} > 0\}$, and let G^π be the environment-label DAG. Then

$$\mathrm{TC}(H_D) = \mathrm{TC}(G^\pi).$$

The ordering selected by $\mathscr{D}$ is a topological ordering of H_D. For every topological ordering of H_D, with predecessor set B_i for node i, and every $i \in [n]$, the selected conditional CDF $C_i(t \mid L_{B_i})$ is a continuous version of the conditional ratio CDF. For each such i, there exists a continuous version C such that, for every latent point v in the latent cube,

$$C(t \,|\, L_{B_i}(f(v))) = C^{\mathrm{str}}_{\theta,\pi(i)}(t \mid v).$$

Every continuous version agrees with the selected C_i on the observed conditional-ratio support, and the selected version satisfies the same structural identity at every mixed latent point $f(v)$.

Consequently, for every v in the latent cube,

$$U_i(f(v)) = \begin{cases} Q_{\pi(i)}(v_{\pi(i)}), & s_{\pi(i)} = 1, \\ 1 - Q_{\pi(i)}(v_{\pi(i)}), & s_{\pi(i)} = -1. \end{cases}$$

For every topological ordering of H_D and every $i \in [n]$,

$$\pi^{-1}\big(\mathrm{pa}_G(\pi(i))\big) \subseteq B_i.$$

Under P^0, for every such ordering, every $i \in [n]$, and every $A \subseteq B_i$,

$$U_i \perp\!\!\!\perp U_{B_i \setminus A} \mid U_A \quad \Longleftrightarrow \quad \pi^{-1}\big(\mathrm{pa}_G(\pi(i))\big) \subseteq A.$$

Thus, for every such ordering and every i, the set

$$\pi^{-1}\big(\mathrm{pa}_G(\pi(i))\big)$$

is admissible and is the unique inclusion-minimal admissible parent set. The observed laws are coherent with W, and the parent-pruned graph returned by $\mathscr{D}$ is exactly G^π.

For any DAG G' on $[n]$, mechanism θ', and observed world W' satisfying Assumption 1, Assumption 5, Assumption 4 with the same sign vector s, Assumption 2, and Assumption 3, equality between the observed environment-law family of W' and the observed environment-law family of W implies that W' is componentwise C^2-equivalent to W up to the simultaneous relabeling encoded by the target permutations. When $n = 2$, every edge $j \to i$ of G has strictly positive observed Gaussian-MMD discrepancy,

$$0 < D_{\pi^{-1}(j),\pi^{-1}(i)}.$$

Within the positive C^3 compact-cube, shared C^2-mixing, causal-minimal regime, $\mathscr{D}$ identifies the latent representation and structure from one perfect intervention per node, supplies the continuous law-separation witness used by von Kügelgen et al. [2023, Theorem 3.2] when $n = 2$, and recovers the graph, topological order, and intervention alignment relative to Wendong et al. [2023, Theorem 4.2] and Yao et al. [2025, Corollary D.1].

The theorem has three linked components. First, Gaussian cover separation makes the ratio-discrepancy graph rich enough to recover the transitive closure of the environment-label DAG. Second, every topological order of that graph supplies conditioning sets large enough for the selected conditional CDFs to agree with their structural counterparts, so the ranks are the coordinatewise transforms $Q_{\pi(i)}(V_{\pi(i)})$ or their sign-reversed versions. Third, causal minimality turns the observational conditional independences among these ranks into exact parent sets, which converts the order information into the graph G^π.

The final clauses state the representation comparison at the population level. Equality of the observed environment-law family pins down the recovered world up to the componentwise C^2 equivalence of Definition 22. In the two-variable case, the same conditions also yield strictly positive observed Gaussian-MMD discrepancy on every true edge, giving a compact-cube Gaussian-MMD counterpart to the continuous ratio-law witness used by von Kügelgen et al. [2023]. For the multi-node recovery task, the decoder delivers graph, order, and intervention alignment under the one-intervention-per-node design through ratio observables and conditional ranks.

6 Sample-split confidence edges

The population decoder in Algorithm 1 and Theorem 3 turns positive ratio-law discrepancies into ancestral order information. This section adds the finite-sample graph layer. The sampling scheme uses independent training folds to construct the fitted ratios $\widehat{R}_i$ and independent evaluation folds, with environment sample sizes N_e, to test the resulting one-dimensional ratio laws.

Definition 24. [`synth_5`] (Sample-split experiment world) *A sample-split experiment for a fixed observed environment family W, whose observations take values in $\mathbb{R}^n$, consists of a measurable sample space Ω together with:*

- ***(Experiment law.)*** *A measure on Ω.*
- ***(Evaluation fold.)*** *Environment sample sizes $N_e \in \mathbb{N}$, for $e = 0, \ldots, n$, and evaluation observations, each taking values in the observed state space,*

$$X^{\mathrm{eval}}_{e,m} : \Omega \to \mathbb{R}^n, \qquad m = 1, \ldots, N_e.$$

- ***(Training fold.)*** *Training sample sizes $N^{\mathrm{tr}}_e \in \mathbb{N}$, for $e = 0, \ldots, n$, and training observations, again valued in the observed state space,*

$$X^{\mathrm{tr}}_{e,m} : \Omega \to \mathbb{R}^n, \qquad m = 1, \ldots, N^{\mathrm{tr}}_e.$$

- ***(Fitted ratios.)*** *For each full training fold $T = \{T_{e,m} : e = 0, \ldots, n,\ m = 1, \ldots, N^{\mathrm{tr}}_e\}$, each $i \in [n]$, and each $x \in \mathbb{R}^n$, a fitted ratio value*

$$\widehat{R}_i(T, x) \in \mathbb{R}.$$

  For every i, the map $(T, x) \mapsto \widehat{R}_i(T, x)$ is jointly measurable; for every T and i, the map $x \mapsto \widehat{R}_i(T, x)$ is measurable; and

$$\widehat{R}_i(T, x) > 0$$

  for every T, i, and x.

- ***(First-stage radius.)*** *A first-stage radius $a_N(\eta) \in \mathbb{R}$, indexed by the first-stage failure level η, satisfying*

$$a_N(\eta) \geq 0$$

  for every $\eta \in \mathbb{R}$.

- ***(Error levels.)*** *A familywise evaluation error level $\alpha \in (0, 1)$ and a first-stage failure level $\eta \in (0, 1)$.*

Definition 24 isolates the objects needed for a sample-split analysis: a training sample, an evaluation sample, fitted positive ratios, a first-stage radius a_N, a familywise level α, and a first-stage failure level η. The radius $a_N(\eta)$ is an input to the analysis: the section takes as given a first-stage estimator delivering uniform L^1 ratio accuracy $a_N(\eta)$ with probability at least $1 - \eta$, and proves what follows from it. What the section delivers is therefore an inference interface for the ratio graph, valid for any first-stage estimator meeting that contract. The second-stage half of the radius, which comes from evaluation-fold kernel concentration, is explicit in the sample sizes and α alone. The next condition supplies the probabilistic independence that lets the evaluation fold be analyzed conditionally on the trained ratios.

Assumption 7. [`ass:independent-environment-sampling`] (Independent environment sampling) *The sample-split experiment is carried by a probability law, and each environment law P^e is a probability measure. It satisfies:*

- ***(Training fold.)*** *For each environment e, the training observations are measurable, independent within that environment, and each training observation in environment e has law P^e.*

- ***(Evaluation fold.)*** *For each environment e, the evaluation observations are measurable, independent within that environment, and each evaluation observation in environment e has law P^e.*

- ***(Fold independence.)*** *The complete training fold is independent of the complete array of evaluation observations across all environments.*

- ***(First-stage measurability.)*** *For each latent coordinate $i \in [n]$ and latent state x, the ratio estimate $\widehat{R}_i(x)$ is measurable with respect to the training fold, and the same estimate is jointly measurable in the training fold and x. The training-fold event*

$$\left\{\omega : \ \forall i \in [n], \ \forall e, \ \int \left|\widehat{R}_i(x) - R_i(x)\right| \, dP^e(x) \leq a_N(\eta) \text{ with finite integral}\right\}$$

  is measurable with respect to the training fold.

This is the standard independent multi-sample sampling with sample splitting condition used in kernel two-sample analysis [Gretton et al., 2012]. The within-environment independence clauses identify each empirical law with its environment distribution, while fold independence makes the evaluation concentration conditional on the realized training output. The measurability clause makes the simultaneous first-stage L^1 ratio-error event a genuine training-fold event.

The evaluation statistic compares the fitted-ratio distributions in an interventional environment and the observational environment through a Gaussian maximum mean discrepancy. The following definitions fix the evaluation notation, the Hilbert-space embedding, the empirical discrepancy $\widehat{D}_{ji}$, and the simultaneous radius $\varepsilon_N(\alpha, \eta)$.

Definition 25. [`synth_25`] (Evaluation-fold observations) *In the sample-split experiment, for each environment $e \in \{0, \dots, n\}$ let*

$$X_1^{(e)}, \dots, X_{N_e}^{(e)}$$

denote the independent evaluation-fold observations drawn from the observed law P^e, independently of the training folds used to construct the ratio estimators $\widehat{R}_i$. Thus $X_r^{(j)} = X_{j,r}^{\text{eval}}$ is the r-th evaluation observation from interventional environment j, and $X_r^{(0)} = X_{0,r}^{\text{eval}}$ is the r-th evaluation observation from the observational environment.

Definition 26. [`synth_21`] (Feature Hilbert space and kernel mean embedding) *Let $\mathcal{H}$ be the complete real Hilbert space in which the feature map $U = (\Phi, \|\Phi(\cdot)\|, \langle \Phi(\cdot), \Phi(\cdot)\rangle)$ takes its values. The feature map is normalized for the Gaussian kernel when*

$$\|\Phi(r)\|_{\mathcal{H}} = 1, \qquad \langle \Phi(a), \Phi(b)\rangle_{\mathcal{H}} = k(a, b), \qquad k(a, b) = \exp\{-(a - b)^2\}.$$

For any probability law ν on $\mathbb{R}$, define the kernel mean embedding

$$\mu_U(\nu) = \int_{\mathbb{R}} \Phi(r)\, d\nu(r) \in \mathcal{H},$$

with the integral understood as the Hilbert-valued Bochner integral.

Definition 27. [`synth_14`] (Empirical MMD discrepancy) *For $j, i \in [n]$, outcome ω, and a real Hilbert-space feature map $U = (\Phi, \dots)$ of unit norm satisfying $\langle \Phi(a), \Phi(b) \rangle = k(a, b) = \exp(-(a-b)^2)$, define*

$$\widehat{D}_{ji}(\omega; U) = \left\| \frac{1}{N_j} \sum_{r=1}^{N_j} \Phi\Big(\widehat{R}_i(X_r^{(j)}(\omega))\Big) - \frac{1}{N_0} \sum_{r=1}^{N_0} \Phi\Big(\widehat{R}_i(X_r^{(0)}(\omega))\Big) \right\|_{\mathcal{H}}.$$

Definition 28. [`synth_11`] (Simultaneous MMD radius) *In the sample-split experiment of Definition 24, assume:*

- ***(First-stage radius.)*** *$a_N(\eta) \in [0, \infty)$ is the deterministic radius in the simultaneous first-stage evaluation-law error event at failure level η.*
- ***(Evaluation sizes.)*** *N_e is the evaluation sample size in environment $e \in \{0, \dots, n\}$.*
- ***(Nondegeneracy.)*** *$n \geq 1$ and $\min_{0 \leq e \leq n} N_e \geq 1$.*
- ***(Levels.)*** *$\alpha \in (0, 1)$ and $\eta \in (0, 1)$.*

The simultaneous MMD confidence radius is

$$\varepsilon_N(\alpha, \eta) = 2\sqrt{2}\, a_N(\eta) + \frac{2\{1 + \sqrt{2 \log(n(n+1)/\alpha)}\}}{\sqrt{\min_{0 \leq e \leq n} N_e}}.$$

This radius is the threshold used to convert the empirical discrepancies $\widehat{D}_{ji}$ into simultaneous lower confidence bounds for the population discrepancies D_{ji}.

The first term in Definition 28 carries the training-fold ratio error into the MMD scale. The second term is the simultaneous evaluation fluctuation over all ordered pairs, using the unit-norm Gaussian feature representation in Definition 26; this is the same empirical-process layer underlying kernel two-sample testing and its aggregated variants [Gretton et al., 2012, Schrab et al., 2023].

The graph estimator thresholds empirical lower confidence bounds for the population discrepancies. Its output $\widehat{H}$ is an edge set on the environment labels.

Algorithm 2 (Sample-split confidence graph $\widehat{H}$)

Under Assumption 7, with training and evaluation folds and tuning levels α, η, the sample-split confidence graph $\widehat{H}$ is defined by the following steps.

1. On the training folds, fit likelihood-ratio estimators

$$\widehat{R}_i, \qquad i \in [n].$$

2. On the independent evaluation folds, compute the empirical Gaussian-MMD discrepancies

$$\widehat{D}_{ji}, \qquad j, i \in [n],\ j \neq i,$$

using $\widehat{R}_i$.

3. Output the directed graph

$$\widehat{H} = \Big\{ j \to i : \widehat{D}_{ji} - \varepsilon_N(\alpha, \eta) > 0 \Big\}.$$

The rule in Algorithm 2 selects an arrow exactly when the lower confidence bound for the corresponding population discrepancy is positive. The theorem below states the resulting simultaneous coverage and its implication for ancestral edge discoveries.

Theorem 4. [`thm:simultaneous-confidence-edges`] (Simultaneous confidence edges) *Fix $n \geq 1$, a labeled DAG G on $[n]$, a sign vector s, a mechanism θ, an observed world with target permutation π, and a sample-split experiment S. Suppose:*

- ***(Structural model.)*** *The mechanism, shared mixing representation, and one-intervention-per-node design satisfy Assumptions 1 to 3.*
- ***(Sample splitting.)*** *The training and evaluation folds satisfy Assumption 7, and each environment has evaluation sample size $N_e \geq 1$.*
- ***(Levels.)*** *The familywise evaluation level and first-stage failure level satisfy $\alpha, \eta \in (0, 1)$.*
- ***(First-stage event.)*** *With*

$$\mathcal{E}_{\mathrm{fs}} = \left\{ \omega : \forall i \in [n],\ \forall e \in \{0, \dots, n\},\ \int \left| \widehat{R}_i(x) - R_i(x) \right| dP^e(x) \leq a_N(\eta) \right\},$$

where the displayed integrals are finite, the training-fold event obeys

$$\Pr(\mathcal{E}_{\mathrm{fs}}) \geq 1 - \eta.$$

Let $\mathcal{E}_{\mathrm{MMD}}$ be the simultaneous evaluation event

$$\mathcal{E}_{\mathrm{MMD}} = \left\{ \omega : \forall j \neq i,\ \left| \widehat{D}_{ji}(\omega) - D_{ji} \right| \leq \varepsilon_N(\alpha, \eta) \right\}.$$

Then, conditionally on the training fold and on $\mathcal{E}_{\mathrm{fs}}$, this event has probability at least $1 - \alpha$: for every training-fold measurable event B,

$$(1 - \alpha) \Pr(\mathcal{E}_{\mathrm{fs}} \cap B) \leq \Pr(\mathcal{E}_{\mathrm{MMD}} \cap \mathcal{E}_{\mathrm{fs}} \cap B).$$

Consequently,

$$\Pr(\mathcal{E}_{\mathrm{MMD}}) \geq 1 - \alpha - \eta.$$

Moreover, for every $\omega \in \mathcal{E}_{\mathrm{MMD}}$, every selected arrow $j \to i$ of the sample-split confidence graph $\widehat{H}$ from Algorithm 2 satisfies that $\pi(j)$ is an ancestor of $\pi(i)$ in G. If, in addition,

$$2\varepsilon_N(\alpha, \eta) < D_{ji} \qquad \text{for every } j, i \text{ such that } \pi(j) \lessdot_G \pi(i),$$

then for every $\omega \in \mathcal{E}_{\mathrm{MMD}}$,

$$\mathrm{TC}\Big(\widehat{H}(\omega)\Big) = \mathrm{TC}(G^{\pi}).$$

The intuition is a conditioning argument. On the first-stage event, the fitted ratios are close enough to the population ratios in every environment law to perturb the population Gaussian-MMD discrepancies by at most the first part of $\varepsilon_N(\alpha, \eta)$. Given the training fold, the evaluation observations are independent across the multi-sample experiment, so the empirical embeddings concentrate simultaneously over ordered pairs. The selected arrows therefore have familywise-valid ancestral interpretation, and when every transported ancestral cover has population discrepancy larger than twice the radius, the selected graph has the same transitive closure as G^π.

7 Limitations and future work

The preceding sections establish population recovery under generic ratio separation and give sample-split confidence edges for ancestral discoveries. This section records the regularity regime in which a next statistical step can be formulated: coordinatewise error bars for the conditional-rank coordinates delivered by Algorithm 1. We fix a smoothness index β, lower-envelope constant c, radius bound M, compact evaluation set K, and latent boundary margin ρ; these parameters define the bounded subclass used to pose a generated-ratio, generated-rank problem.

Assumption 8. [`ass:bounded-holder-radius`] (Bounded Hölder radius) *Let $k = \lfloor \beta \rfloor$. The parameters satisfy $2 < \beta$ and $1 < M$. For every $i \in [n]$, the three functions*

$$p_i, \qquad q_i, \qquad v \mapsto L_i(f(v))$$

satisfy the common Hölder bound of order β and radius M on their respective domains: p_i and $v \mapsto L_i(f(v))$ on $[0,1]^n$, and q_i on $[0,1]$. That is, for each such function r with domain K, r is C^k on K,

$$\|D^m r(z)\| \leq M$$

for every $m \leq k$ and every $z \in K$, and

$$\|D^k r(z) - D^k r(z')\| \leq M \|z - z'\|^{\beta - k}$$

for every $z, z' \in K$.

Assumption 8 is the standard bounded Hölder ball condition used to obtain uniform local-polynomial control [Xie, 2023]. It places the observational mechanisms, intervention densities, and latent-coordinate log-ratios on a common smoothness scale.

Assumption 9. [`ass:density-lower-envelope`] (Density lower envelope) *For the bounded regime under consideration, the lower-envelope constant c satisfies:*

- ***(Range.)*** $0 < c < 1$.
- ***(Observational densities.)*** *For every $i \in [n]$ and every latent state $v \in [0,1]^n$,*

$$c \leq p_i(v_i \mid v_{\mathrm{pa}_G(i)}).$$

- ***(Intervention densities.)*** *For every $i \in [n]$ and every $z \in [0,1]$,*

$$c \leq q_i(z).$$

The lower-envelope restriction in Assumption 9 is the standard uniform full-support condition [Xie, 2023]. In the present model it gives a common lower bound for both observational conditionals and intervention marginals, keeping likelihood ratios well behaved on the closed latent cube.

Assumption 10. [`ass:own-derivative-margin`] (Own-derivative margin) *For every $i \in [n]$ and every $v \in [0,1]^n$,*

$$s_i\,\partial_{v_i} L_i(f(v)) \geq c.$$

Assumption 10 is specific to this analysis. It strengthens the fixed-sign condition from Assumption 4 into a uniform margin, so the own-coordinate monotonicity used by the rank construction remains quantitatively separated from zero.

Assumption 11. [`ass:common-interior-domain`] (Common interior domain) *The common interior-domain condition for K and radius ρ is the conjunction:*

- ***(Radius.)*** $0 < \rho < 1/2$.
- ***(Compactness.)*** *K is compact.*
- ***(Observed interior.)*** *$K \subseteq \operatorname{int}(\mathcal{X})$, where $\mathcal{X} = f([0,1]^n)$.*
- ***(Latent boundary margin.)*** *For every $v \in f^{-1}(K)$,*

$$\rho \leq \operatorname{dist}\bigl(v, \partial[0,1]^n\bigr).$$

The compact interior domain in Assumption 11 is the standard evaluation-region condition for uniform conditional-CDF estimation away from boundary effects [Xie, 2023]. Here it is imposed simultaneously in observed space and after pullback through the shared mixing map.

Assumption 12. [`ass:bounded-shared-mixing-geometry`] (Shared mixing geometry) *The shared mixing map satisfies*

$$\max\left\{\|f\|_{C^2([0,1]^n)}, \|f^{-1}\|_{C^2(\mathcal{X})}, \sup_{v\in[0,1]^n} |\det Df(v)|, \sup_{x\in\mathcal{X}} |\det Df^{-1}(x)|\right\} \leq M.$$

Assumption 12 is specific to this analysis. It gives a common geometric envelope for transporting smoothness, volume, and interior-domain statements between latent and observed coordinates.

Assumption 13. [`ass:observed-log-ratio-regularity`] (Observed log-ratio regularity) *The observed log-ratios satisfy*

$$\max_{i\in[n]} \|L_i\|_{C^2(K)} \leq M.$$

The generated regressors used in the second stage are the observed log-ratios, and Assumption 13 is the standard uniformly bounded smooth-generated-regressor condition on an interior compactum [Mammen et al., 2012]. It supplies the regularity needed to compare oracle and generated conditioning arguments.

The next objects describe the conditioning variables selected by the population decoder and the corresponding local design envelope.

Definition 29. [synth_12] (Conditioning-coordinate domain $\mathcal{Z}_i(K)$) *Fix the predecessor set B_i selected by the population decoder and write $L_{B_i}(x) = (L_b(x))_{b\in B_i}$. For a compact evaluation domain $K \subseteq \operatorname{int}(\mathcal{X})$, the conditioning-coordinate domain for node i is the image of K under these log-ratio coordinates:*

$$\mathcal{Z}_i(K) = \{L_{B_i}(x) : x \in K\} \subseteq \mathbb{R}^{|B_i|}.$$

When $B_i = \varnothing$, $\mathcal{Z}_i(K)$ is the singleton subset of $\mathbb{R}^0$.

Definition 29 turns the selected predecessor set B_i into the conditioning-coordinate support $\mathcal{Z}_i(K)$. When no predecessor is selected, the conditioning problem collapses to the unconditional one.

Assumption 14. [ass:conditional-design-envelope] (Conditional design envelope) *For every $i \in [n]$, the density g_i of $L_{B_i}(X)$ under P^i satisfies*

$$c \le g_i(z) \le M$$

for all $z \in \mathcal{Z}_i(K)$.

Assumption 14 is the standard interior local-polynomial design-density envelope condition [Xie, 2023]. The density g_i controls how much data are available in the conditioning neighborhoods indexed by Definition 29.

Definition 30. [def:bounded-holder-subclass] (Bounded Hölder subclass $\Theta^{\beta}_{G,s}(c, M, K, \rho)$) *The bounded Hölder subclass is*

$$\Theta^{\beta}_{G,s}(c, M, K, \rho) = \left\{ \theta \in \Theta_{G,s} : \begin{array}{l} \text{Assumption 8 holds for } n, G, s, \beta, c, M, K, \rho, \\ \text{Assumption 9 holds for } n, G, s, \beta, c, M, K, \rho, \\ \text{Assumption 10 holds for } n, G, s, \beta, c, M, K, \rho, \\ \text{Assumption 11 holds for } n, G, s, \beta, c, M, K, \rho, \\ \text{Assumption 12 holds for } n, G, s, \beta, c, M, K, \rho, \\ \text{Assumption 13 holds for } n, G, s, \beta, c, M, K, \rho, \\ \text{Assumption 14 holds for } n, G, s, \beta, c, M, K, \rho \end{array} \right\}.$$

Definition 30 collects the smoothness, support, margin, geometry, observed-regularity, and local-design requirements into the bounded subclass $\Theta^{\beta}_{G,s}(c, M, K, \rho)$. The definition gives a single parameter space over which uniform generated-rank questions can be posed.

The statistical program also requires a regime for sample sizes, first-stage ratio accuracy, and second-stage smoothing.

Definition 31. [synth_6] (Bounded regime parameters) *For an observed world W, a bounded generated-rank regime consists of:*

- ***(Order.)*** *An order map $o : [n] \to \mathbb{N}$.*
- ***(Environment sample sizes.)*** *A sample-size sequence $N_e(N) \in \mathbb{N}$ for each $N \in \mathbb{N}$ and each environment $e \in \{0, \dots, n\}$.*
- ***(First-stage ratio rate.)*** *A uniform first-stage C^1 rate b_N satisfying*

$$b_N > 0 \quad \text{for every } N, \qquad b_N \to 0.$$

- ***(First-stage failure rate.)*** *A first-stage failure sequence η_N satisfying*

$$\eta_N > 0 \quad \textit{for every } N, \qquad \eta_N \to 0.$$

- ***(Second-stage bandwidth.)*** *A second-stage bandwidth h_N satisfying*

$$h_N > 0 \quad \textit{for every } N, \qquad h_N \to 0.$$

In Definition 31, the rate b_N measures uniform first-stage C^1 log-ratio error, η_N records first-stage failure probability, and h_N is the second-stage bandwidth. These sequences separate the two statistical layers: estimating likelihood ratios and estimating conditional ranks from generated responses and covariates.

Definition 32. [`def:first-stage-ratio-contract`] (First-stage ratio contract) *For a generated observed-world assignment W_θ, a decoder-selected bounded-regime assignment R_θ, sample-split experiments $S_{\theta,N}$, a common environment sample-size sequence $N \mapsto (N_e(N))_{e=0}^n$, rate sequences b_N and η_N, and subclass parameters β, c, M, K, ρ, define*

$$\mathcal{E}_N^\infty(b_N) = \left\{ \omega : \begin{array}{l} \textit{for every } i \in [n],\ \log \widehat{R}_i(\omega,\cdot) \textit{ and } L_i \textit{ are } C^1 \textit{ on } \mathcal{X}, \\ \textit{and for every } x \in \mathcal{X},\ \left|\log \widehat{R}_i(\omega,x) - L_i(x)\right| \le b_N, \\ \left\|D_x \log \widehat{R}_i(\omega,x) - D_x L_i(x)\right\| \le b_N \end{array} \right\}.$$

The first-stage ratio contract holds when the following conditions are satisfied:

- ***(Vanishing positive rates.)*** *For every N, $b_N > 0$ and $\eta_N > 0$, and both sequences satisfy $b_N \to 0$ and $\eta_N \to 0$ as $N \to \infty$.*

- ***(Common selected-regime schedules.)*** *For every N and every $\theta \in \Theta^\beta_{G,s}(c, M, K, \rho)$ as in Definition 30, the selected bounded regime R_θ has environment sample-size sequence $N \mapsto (N_e(N))_{e=0}^n$, ratio-error rate b_N, and first-stage failure rate η_N.*

- ***(Sampling design.)*** *For every N and every $\theta \in \Theta^\beta_{G,s}(c, M, K, \rho)$, the experiment $S_{\theta,N}$ satisfies the independent-environment sample-splitting conditions in Assumption 7, and its environment sample sizes are exactly $(N_e(N))_{e=0}^n$.*

- ***(Measurable high-probability event.)*** *For every N and every $\theta \in \Theta^\beta_{G,s}(c, M, K, \rho)$, the event $\mathcal{E}_N^\infty(b_N)$ is measurable under $S_{\theta,N}$, and*

$$1 - \eta_N \le \Pr_{S_{\theta,N}} \{\mathcal{E}_N^\infty(b_N)\}.$$

Definition 32 defines the high-probability event $\mathcal{E}_N^\infty(b_N)$ on which all fitted log-ratios and their first derivatives are uniformly close to their population counterparts. Its C^1 form is deliberately matched to the generated-regressor literature, where first-stage perturbations enter through both fitted responses and fitted conditioning arguments [Pagan, 1984, Mammen et al., 2012, Lee, 2018].

Remark 1. [`def:generated-rank-handle`] (Generated-rank analysis handle) *$\mathfrak{H}_{\mathrm{rank}}$, the generated-rank proof-strategy handle, records a nonassertive construction for $\widehat{U}_i$, the proposed second-stage cross-fitted rank estimator. The estimator $\widehat{U}_i$ is formed by adapting Xie's local-linear conditional-CDF estimator to cross-fitted fitted log-ratio responses and cross-fitted conditioning covariates. The handle decomposes the second-stage rank error into the oracle conditional-CDF process, indicator boundary crossings from the generated threshold, and local-design perturbations from generated conditioning arguments; it then linearizes the local-design perturbation component using the generated-covariate expansion of Mammen, Rothe, and Schienle. The second-stage weight convention, bandwidth selector, influence representation, attainment claim, and sharpness claim are treated as open components outside the supplied handle.*

The handle $\mathfrak{H}_{\mathrm{rank}}$ organizes the proposed second-stage estimator $\widehat{U}_i$ around local-linear conditional-CDF estimation [Fan and Gijbels, 1996, Fan et al., 1996, Xie, 2023]. Its decomposition isolates oracle smoothing variation from the additional terms created by generated thresholds and generated conditioning coordinates.

Definition 33. [`synth_7`] (Maximum conditioning dimension) *The maximum conditioning dimension of the selected ordering is*

$$d_\star := \max_{i\in[n]} |B_i|,$$

where B_i is the set of labels preceding i in that ordering.

Definition 33 defines the effective nonparametric dimension $d_\star$. It is the dimension that appears in the local-neighborhood sample size for the conditional-CDF step.

Remark 2. [`oeq:generated-rank-frontier`] (Generated-rank frontier) *A natural next question is whether the handle $\mathfrak{H}_{\mathrm{rank}}$ yields uniform generated-rank estimators under the following conditions:*

- ***(First-stage contract.)***

$$\inf_{\theta\in\Theta^{\beta}_{G,s}(c,M,K,\rho)} \Pr_{\theta}\{\mathcal{E}^{\infty}_N(b_N)\} \geq 1-\eta_N.$$

- ***(Bandwidth scale.)*** *The bandwidth h_N satisfies $h_N \to 0$,*

$$b_N = o(h_N),$$

and

$$\frac{\underline{N} h_N^{d_\star}}{\log(\underline{N})} \to \infty.$$

The target is a construction of cross-fitted estimators $\widehat{U}_i$ such that

$$\max_{i\in[n]} \|\widehat{U}_i - U_i\|_{\infty,K} = O_{\Pr,\Theta^{\beta}_{G,s}(c,M,K,\rho)}(r_N),$$

where

$$r_N = h_N^2 + \sqrt{\frac{\log(\underline{N})}{\underline{N} h_N^{d_\star}}} + \frac{b_N}{h_N}.$$

When the first-stage log-ratio estimators additionally admit a uniform asymptotically linear representation on K, future work could derive the corresponding expansion for $\widehat{U}_i - U_i$, separating threshold crossings, generated-conditioning perturbations, oracle conditional-CDF variance, and smoothing bias, and could characterize the weakest sufficient first-stage remainder bound.

Remark 2 records the rate target in terms of the minimum environment sample size $\underline{N}$, that is $\underline{N} = \min_{0 \le e \le n} N_e$, and the proposed rate r_N. The three terms correspond to second-order smoothing bias, local empirical variation, and first-stage C^1 error propagated through bandwidth-scale neighborhoods. Meeting that target would supply a coordinate error analysis for the nonlinear representation delivered by Theorem 3, and would attach rank-coordinate uncertainty to the confidence-edge construction of Theorem 4. Whether a deployed pipeline such as ROPES [Kulkarni et al., 2025] satisfies the model assumed here is a separate question that this paper does not settle; the reference is to the kind of implementation such an error analysis would serve.

Several additional directions remain open. Exact-parent inference calls for conditional-dependence margins that distinguish direct parents from more distant ancestors after the population decoder has recovered the transitive structure. Broader support conditions would extend the compact-cube analysis to domains with boundary or tail behavior closer to applied continuous measurements. Vector-valued latent nodes would replace scalar monotone coordinates by multivariate intervention scores, changing both the separation argument and the conditional-rank step. Imperfect interventions would allow the intervention law to retain controlled parent dependence, connecting the one-intervention design here to general-environment formulations [Ng et al., 2025, Jin and Syrgkanis, 2024, Zhang et al., 2024, Ahuja et al., 2024] and to recent benchmark and application settings for interventional representation learning [Chen et al., 2025, Wang et al., 2024, Sun et al., 2024]. Noninvertible observation maps would move beyond the shared diffeomorphic mixing geometry in Assumption 2 and require an identification target defined on observational equivalence classes rather than on recovered latent coordinates.

Appendices

A Proofs and verification

This appendix collects the auxiliary statements that support the generic separation, population decoding, and sample-split confidence results. The first group records the explicit coordinate reflection and affine perturbation used to make a direct-edge ratio contrast nonzero inside a fixed-sign stratum.

Definition 34. [`synth_22`] (Coordinate reflection map) *For $\sigma, z \in \mathbb{R}$, define*

$$\mathrm{reflect}(\sigma, z) = \begin{cases} z, & \sigma = 1, \\ 1 - z, & \sigma \neq 1. \end{cases}$$

Definition 35. [`def:affine-path`] (Edge affine path) *For a finite labeled DAG G on $[n]$, a sign vector $s \in \{-1, 1\}^{[n]}$, a stratum point $\theta = ((p_\ell, q_\ell))_{\ell=1}^n \in \Theta_{G,s}$ as in Definition 3, an edge $j \to i$ in G, and $t \in [0, 1]$, define the reflected edge-specific sparse endpoint $\theta^{\mathrm{sp}} = ((p_\ell^{\mathrm{sp}}, q_\ell^{\mathrm{sp}}))_{\ell=1}^n$*

in G by

$$p_\ell^{\mathrm{sp}}(v) = \begin{cases} 1 + \dfrac{1}{10}\, h(\mathrm{reflect}(s_j, v_j))\, h(\mathrm{reflect}(s_i, v_i)), & \ell = i, \\ 1, & \ell \neq i, \end{cases} \qquad q_\ell^{\mathrm{sp}}(z) = \frac{4\exp(4\ \mathrm{reflect}(s_\ell, z))}{\exp(4) - 1},$$

where $h(x) = 2x - 1$. The affine path $\theta^t = ((p_\ell^t, q_\ell^t))_{\ell=1}^n$ is the mechanism on G with

$$p_\ell^t(v) = (1-t)p_\ell(v) + t\, p_\ell^{\mathrm{sp}}(v), \qquad q_\ell^t(z) = (1-t)q_\ell(z) + t\, q_\ell^{\mathrm{sp}}(z), \qquad \ell \in [n].$$

Definition 36. [`synth_15`] (Second-moment contrast) *For an observed world with environment laws $P^0, \ldots, P^n$ and likelihood-ratio carriers R_i, define, for $j, i \in [n]$,*

$$F_{ji} = \int_{\mathcal{X}} R_i(x)^2\, dP^j(x) - \int_{\mathcal{X}} R_i(x)^2\, dP^0(x).$$

Lemma 1. [`lem:analytic-edge-perturbation`] (Analytic edge perturbation) *After the harmless target relabeling $\pi = \mathrm{id}$, fix the ambient number n of scalar latent variables, a finite labeled DAG G on $[n]$, a sign vector $s \in \{-1, 1\}^{[n]}$, a mechanism $\theta \in \Theta_{G,s}$ as in Definition 3, and an edge $j \to i$ of G. Assume that the identity-target canonical observed world generated by θ satisfies Assumption 3. Let θ^t be the edge-specific affine path of Definition 35, and write $F_{ji}(\eta)$ for the direct-edge second-moment contrast computed in the identity-target canonical observed world generated by a mechanism η. Along this path,*

$$F_{ji}(\theta^t) = \int_{[0,1]^n} \frac{(q_i^t(v_i))^2 \big(q_j^t(v_j) - p_j^t(v_j \mid v_{\mathrm{pa}_G(j)})\big) \prod_{\ell \neq i,j} p_\ell^t(v_\ell \mid v_{\mathrm{pa}_G(\ell)})}{p_i^t(v_i \mid v_{\mathrm{pa}_G(i)})}\, dv.$$

There is an open set $O \subseteq \mathbb{R}$ with $[0,1] \subseteq O$ such that:

- ***(Integral identity.)*** *For every $t \in O$, the displayed integral equals $F_{ji}(\theta^t)$.*
- ***(Analyticity.)*** *The map $t \mapsto F_{ji}(\theta^t)$ is real analytic on a neighborhood of every point of O.*
- ***(Isolated zeros.)*** *If $t \in O$ and $F_{ji}(\theta^t) = 0$, then there is $\varepsilon > 0$ such that every $u \in (t - \varepsilon, t + \varepsilon)$ with $u \neq t$ satisfies $F_{ji}(\theta^u) \neq 0$.*

Moreover, $F_{ji}(\theta^1) \neq 0$. Finally, for every relative C^2 neighborhood N of θ, there exists $t \in (0,1)$ such that $\theta^t \in N$, $\theta^t \in \Theta_{G,s}$, and $F_{ji}(\theta^t) \neq 0$.

Proof of Lemma 1. Work after the stated relabeling, so the target permutation is the identity. Let $\theta^\star$ denote the edge-specific sparse endpoint used in Definition 35, and write the affine extension for real t as

$$p_\ell^t = (1-t)p_\ell + tp_\ell^\star, \qquad q_\ell^t = (1-t)q_\ell + tq_\ell^\star.$$

Step 1: the positive analytic parameter set. What the rest of the argument uses is an open convex set $O \supseteq [0,1]$ on which every path mechanism is positive and normalized, together with real analyticity on O of the parameter integral displayed in Step 2. We exhibit one such set. Strict positivity of $\theta \in \Theta_{G,s}$, strict positivity of the sparse endpoint, continuity on the compact cubes, and finiteness of the node set give a number $m > 0$ such that every $p_\ell, p_\ell^\star, q_\ell, q_\ell^\star$ is at least m on its domain; any positive lower bound serves equally well below. Define

$$O_p = \left\{ t : \ \frac{m}{2} < (1-t)p_\ell(v) + tp_\ell^\star(v) \text{ for all } \ell \text{ and all } v \in [0,1]^n \right\},$$

$$O_q = \left\{ t : \ \frac{m}{2} < (1-t)q_\ell(z) + tq_\ell^\star(z) \text{ for all } \ell \text{ and all } z \in [0,1] \right\}, \qquad O = O_p \cap O_q.$$

For fixed ℓ and cube point, the two displayed inequalities define open half-lines in the parameter t. Taking the infimum over the compact sets $[n] \times [0,1]^n$ and $[n] \times [0,1]$ gives continuous lower-envelope functions, so O_p and O_q are open; as intersections of affine strict-superlevel conditions, they are convex. If $0 \le t \le 1$, then for every admissible argument

$$(1-t)m + tm = m \le (1-t)p_\ell + tp_\ell^\star, \qquad (1-t)m + tm = m \le (1-t)q_\ell + tq_\ell^\star,$$

and hence $[0,1] \subseteq O$. On O, all p_ℓ^t and q_ℓ^t are positive. Normalization and C^3-smoothness are affine identities: for every real t, every $v \in [0,1]^n$, and every ℓ,

$$\int_0^1 p_\ell^t(z \mid v_{\mathrm{pa}_G(\ell)})\, dz = (1-t) + t = 1, \qquad \int_0^1 q_\ell^t(z)\, dz = (1-t) + t = 1,$$

and the derivatives up to order three are the same affine combinations of the endpoint derivatives. Moreover, for every $t \in O$ and $v \in [0,1]^n$,

$$\frac{m}{2} \le p_i^t(v_i \mid v_{\mathrm{pa}_G(i)}).$$

We now make the analytic-under-the-integral step explicit. Put

$$\mathsf{D}(t,v) = (1-t)p_i(v_i \mid v_{\mathrm{pa}_G(i)}) + tp_i^\star(v_i \mid v_{\mathrm{pa}_G(i)})$$

and

$$\mathsf{P}(t,v) = (q_i^t(v_i))^2 \big(q_j^t(v_j) - p_j^t(v_j \mid v_{\mathrm{pa}_G(j)})\big) \prod_{\ell \neq i,j} p_\ell^t(v_\ell \mid v_{\mathrm{pa}_G(\ell)}),$$

with the empty product equal to 1. Since the product has finitely many factors and every factor is affine in t, there are measurable coefficient functions $\mathsf{a}_0, \dots, \mathsf{a}_{n+3} : [0,1]^n \to \mathbb{R}$, continuous on the cube after choosing the standard continuous representatives on the ambient cube, such that

$$\mathsf{P}(t,v) = \sum_{r=0}^{n+3} \mathsf{a}_r(v) t^r, \qquad \sup_{v \in [0,1]^n} |\mathsf{a}_r(v)| < \infty \quad (0 \le r \le n+3).$$

Fix $t_0 \in O$. The denominator satisfies $\mathsf{D}(t_0, v) \ge m/2$ on the compact latent cube, while

$$\mathsf{D}(t,v) = \mathsf{D}(t_0,v) + (t - t_0)\{p_i^\star(v_i \mid v_{\mathrm{pa}_G(i)}) - p_i(v_i \mid v_{\mathrm{pa}_G(i)})\}.$$

By compactness, choose $\rho_{t_0} > 0$ so that for $|t - t_0| < \rho_{t_0}$,

$$\left| (t - t_0) \frac{p_i^\star(v_i \mid v_{\mathrm{pa}_G(i)}) - p_i(v_i \mid v_{\mathrm{pa}_G(i)})}{\mathsf{D}(t_0,v)} \right| \le \frac{1}{2} \quad \text{for all } v \in [0,1]^n.$$

Then

$$\frac{1}{\mathsf{D}(t,v)} = \frac{1}{\mathsf{D}(t_0,v)} \sum_{r=0}^{\infty} \left[-(t-t_0) \frac{p_i^\star(v_i \mid v_{\mathrm{pa}_G(i)}) - p_i(v_i \mid v_{\mathrm{pa}_G(i)})}{\mathsf{D}(t_0,v)} \right]^r$$

converges uniformly in v. Multiplying by the polynomial $\mathsf{P}(t,v)$, the coefficient bounds and the finite volume of $[0,1]^n$ give a summable dominating majorant on a smaller neighborhood of t_0. Termwise integration therefore gives a power series for

$$t \longmapsto \int_{[0,1]^n} \frac{\mathsf{P}(t,v)}{\mathsf{D}(t,v)}\,dv$$

near t_0. Thus the parameter integral is real analytic on a neighborhood of every point of O.

Step 2: the integral identity. Since $j \to i$ is an edge in a DAG, $j \neq i$. For each $t \in O$, the construction of O supplies positive normalized smooth mechanisms. In the identity-target canonical observed world generated by θ^t, the latent observational density, target-j density, and target-i ratio are

$$p^{0,t}(v) = \prod_{\ell=1}^{n} p_\ell^t(v_\ell \mid v_{\mathrm{pa}_G(\ell)}), \qquad p^{j,t}(v) = q_j^t(v_j)\prod_{\ell\neq j} p_\ell^t(v_\ell \mid v_{\mathrm{pa}_G(\ell)}),$$

and

$$R_i^t(v) = \frac{q_i^t(v_i)}{p_i^t(v_i \mid v_{\mathrm{pa}_G(i)})}.$$

The lower bound on p_i^t makes the ratio continuous and integrable on the compact latent cube. Therefore, with the empty product interpreted as 1,

$$\begin{aligned} F_{ji}(\theta^t) &= \int_{[0,1]^n} (R_i^t(v))^2\{p^{j,t}(v) - p^{0,t}(v)\}\,dv \\ &= \int_{[0,1]^n} \frac{(q_i^t(v_i))^2\big(q_j^t(v_j) - p_j^t(v_j \mid v_{\mathrm{pa}_G(j)})\big)\prod_{\ell\neq i,j} p_\ell^t(v_\ell \mid v_{\mathrm{pa}_G(\ell)})}{p_i^t(v_i \mid v_{\mathrm{pa}_G(i)})}\,dv. \end{aligned}$$

This proves the displayed identity for every $t \in O$. The right-hand side is the analytic parameter integral constructed above, so the same is true of $t \mapsto F_{ji}(\theta^t)$.

Step 3: the endpoint is nonzero. At $t = 1$, the path is the embedded sparse endpoint. Write

$$h(z) = 2z - 1, \qquad \mathfrak{q}(z) = \frac{4\exp(4z)}{\exp(4) - 1}, \qquad \rho_\ell(z) = \mathrm{reflect}(s_\ell, z),$$

where ρ_ℓ is the coordinate reflection appearing in Definition 35. At the sparse endpoint,

$$p_i^1(v) = 1 + \frac{1}{10}h(\rho_j(v_j))h(\rho_i(v_i)), \qquad p_j^1(v) = 1, \qquad q_\ell^1(z) = \mathfrak{q}(\rho_\ell(z)),$$

and $p_\ell^1(v) = 1$ for $\ell \neq i$. Since the edge $j \to i$ gives $j \neq i$, the two selected coordinates form a two-element set. The integrand in the identity from Step 2 is therefore

$$\Phi(v) = \frac{\mathfrak{q}(\rho_i(v_i))^2\big(\mathfrak{q}(\rho_j(v_j)) - 1\big)}{1 + \frac{1}{10}h(\rho_j(v_j))h(\rho_i(v_i))},$$

which depends only on (v_j, v_i); the finite product over $\ell \neq i,j$ is 1, with the empty-product convention covering the case of no unused coordinates. Product integration on $[0,1]^n$ then gives

$$F_{ji}(\theta^1) = \int_0^1\int_0^1 \frac{\mathfrak{q}(\rho_i(y))^2\big(\mathfrak{q}(\rho_j(x)) - 1\big)}{1 + \frac{1}{10}h(\rho_j(x))h(\rho_i(y))}\,dy\,dx.$$

Each ρ_ℓ is either the identity map or $z \mapsto 1-z$, hence preserves Lebesgue measure on $[0,1]$. Define, for $x \in [0,1]$,

$$\mathcal{A}(x) = \int_0^1 \frac{\mathfrak{q}(y)^2}{1+\frac{1}{10}h(x)h(y)}\,dy.$$

Changing variables first in y and then in x gives

$$F_{ji}(\theta^1) = \int_0^1 \big(\mathfrak{q}(x)-1\big)\mathcal{A}(x)\,dx.$$

It remains to establish the scalar gap for this last integral. Introduce the auxiliary primitive

$$H(x) = \frac{\exp(4x)-1}{\exp(4)-1} - x,$$

and, for $a \in [-1/10, 1/10]$, set

$$\omega_a(y) = \frac{\mathfrak{q}(y)^2}{(1+ah(y))^2}, \qquad y \in [0,1].$$

The identities $H(0)=H(1)=0$ and $H'(x)=\mathfrak{q}(x)-1$ imply, by integration by parts,

$$\int_0^1 \big(1-\mathfrak{q}(x)\big)\mathcal{A}(x)\,dx = \int_0^1 H(x)\mathcal{A}'(x)\,dx.$$

For $x \in [0,1]$, differentiating the defining integral for $\mathcal{A}$ under the integral sign gives

$$\mathcal{A}'(x) = -\frac{1}{5}\int_0^1 h(y)\omega_{h(x)/10}(y)\,dy.$$

The differentiation is justified because $|h(y)| \le 1$ and $h(x)/10 \in [-1/10, 1/10]$, so the denominator is bounded away from zero uniformly on the square.

For every $a \in [-1/10, 1/10]$ and $y \in [0,1]$, the same bounds give $1+ah(y) \in [9/10, 11/10]$. Together with $13 < \exp(4) < 55$, this yields

$$\omega_a(y) > \left(\frac{4}{729}\right)\left(\frac{100}{121}\right), \qquad 8 - \frac{4a}{1+ah(y)} \ge \frac{68}{9}.$$

Since

$$\omega_a'(y) = \omega_a(y)\left(8 - \frac{4a}{1+ah(y)}\right),$$

we have the uniform derivative bound

$$\left(\frac{68}{9}\right)\left(\frac{4}{729}\right)\left(\frac{100}{121}\right) \le \omega_a'(y).$$

Consequently, for $0 \le z \le y \le 1$,

$$\left(\frac{68}{9}\right)\left(\frac{4}{729}\right)\left(\frac{100}{121}\right)(y-z) \le \omega_a(y) - \omega_a(z).$$

Applying this secant estimate on the intervals from $1/2$ to y and from y to $1/2$ gives, for every $y \in [0,1]$,

$$\left(\frac{68}{9}\right)\left(\frac{4}{729}\right)\left(\frac{100}{121}\right) 2\left(y-\frac{1}{2}\right)^2 \le h(y)\big(\omega_a(y)-\omega_a(1/2)\big).$$

Because $\int_0^1 h(y)\,dy = 0$ and $\int_0^1 2(y-1/2)^2\,dy = 1/6$, integration gives

$$\frac{1}{6}\left(\frac{68}{9}\right)\left(\frac{4}{729}\right)\left(\frac{100}{121}\right) \le \int_0^1 h(y)\omega_a(y)\,dy.$$

Hence, for $x \in [0,1]$,

$$\mathcal{A}'(x) \le -\frac{1}{30}\left(\frac{68}{9}\right)\left(\frac{4}{729}\right)\left(\frac{100}{121}\right).$$

The exponential distribution function in H lies in $[0,1]$, and convexity of the exponential on $[0,4]$ gives $(\exp(4x)-1)/(\exp(4)-1) \le x$; therefore $-1 \le H(x) \le 0$ on $[0,1]$. Multiplying the last derivative bound by the nonpositive function H and integrating yields

$$\int_0^1 \big(1-\mathfrak{q}(x)\big)\mathcal{A}(x)\,dx \ge \frac{1}{30}\left(\frac{68}{9}\right)\left(\frac{4}{729}\right)\left(\frac{100}{121}\right)\int_0^1 -H(x)\,dx.$$

A direct integration of the exponential density gives

$$\int_0^1 x\mathfrak{q}(x)\,dx = \frac{1}{1-\exp(-4)} - \frac{1}{4},$$

so integration by parts gives

$$\int_0^1 -H(x)\,dx = \frac{1}{1-\exp(-4)} - \frac{3}{4} > \frac{29}{108}.$$

Combining the rational factors,

$$\frac{1}{30}\left(\frac{68}{9}\right)\left(\frac{4}{729}\right)\left(\frac{100}{121}\right)\left(\frac{29}{108}\right) = \frac{19720}{64304361} > \frac{3}{10000},$$

and hence

$$\frac{3}{10000} < \int_0^1 (1-\mathfrak{q}(x))\mathcal{A}(x)\,dx.$$

Thus

$$F_{ji}(\theta^1) = -\int_0^1 \big(1-\mathfrak{q}(x)\big)\mathcal{A}(x)\,dx < -\frac{3}{10000},$$

and in particular

$$F_{ji}(\theta^1) \neq 0.$$

Step 4: isolated zeros. Let $t_0 \in O$ and suppose $F_{ji}(\theta^{t_0}) = 0$. The set O is convex, hence preconnected, and $1 \in O$ with $F_{ji}(\theta^1) \neq 0$. For a real analytic function on a preconnected set, the local dichotomy at t_0 is: either the function vanishes on a neighborhood of t_0, or it is nonzero on a punctured neighborhood of t_0. The first alternative would force the function to vanish throughout O, contradicting the nonzero value at 1. Thus there is $\varepsilon > 0$ such that

$$u \in (t_0-\varepsilon, t_0+\varepsilon), \quad u \neq t_0 \quad \Longrightarrow \quad F_{ji}(\theta^u) \neq 0.$$

Step 5: small stratum-preserving nonzero perturbations. Let N be any neighborhood of θ in the topology of Definition 3; in particular this covers every relative C^2 neighborhood in the statement. The affine path satisfies

$$\|p_\ell^t - p_\ell\|_{C^2} = t\|p_\ell^\star - p_\ell\|_{C^2}, \qquad \|q_\ell^t - q_\ell\|_{C^2} = t\|q_\ell^\star - q_\ell\|_{C^2},$$

so $\theta^t \to \theta$ in the topology of Definition 3. Hence there is $\delta_N > 0$ such that $0 < t < \delta_N$ implies $\theta^t \in N$.

The conditions defining $\Theta_{G,s}$ persist on a sufficiently short initial segment. Positivity, normalization, and smoothness are handled first. For every $t \in [0,1]$, convexity of the positive endpoint densities gives positivity of every p_ℓ^t and q_ℓ^t, and the affine identities

$$\int_0^1 p_\ell^t(z \mid v_{\mathrm{pa}_G(\ell)})\,dz = (1-t) + t = 1, \qquad \int_0^1 q_\ell^t(z)\,dz = (1-t) + t = 1$$

show normalization; C^3-smoothness follows by applying the same affine combination to derivatives up to order three.

For the fixed-sign part, define for $\ell \in [n]$, $t \in \mathbb{R}$, and $v \in [0,1]^n$,

$$\Sigma_\ell(t,v) = s_\ell\Bigg(\frac{(1-t)\partial_z q_\ell(v_\ell) + t\partial_z q_\ell^\star(v_\ell)}{(1-t)q_\ell(v_\ell) + tq_\ell^\star(v_\ell)} - \frac{(1-t)\partial_{v_\ell} p_\ell(v) + t\partial_{v_\ell} p_\ell^\star(v)}{(1-t)p_\ell(v) + tp_\ell^\star(v)}\Bigg),$$

where ∂_z is the one-dimensional derivative on $[0,1]$ and ∂_{v_ℓ} is the own-coordinate derivative on the latent cube. Here $p_\ell(v)$ abbreviates $p_\ell(v_\ell \mid v_{\mathrm{pa}_G(\ell)})$ regarded as a function on the whole latent cube, and likewise $p_\ell^\star(v)$. At $t = 0$, this is exactly the fixed own-coordinate log-ratio derivative required in Assumption 4. The numerator and denominator terms in Σ_ℓ are continuous in (t,v), and the denominators are bounded away from zero at $t = 0$. Since $[0,1]^n$ is compact and there are finitely many ℓ, the strict inequalities $\Sigma_\ell(0,v) > 0$ persist uniformly for all $v \in [0,1]^n$ and all sufficiently small $t \ge 0$.

It remains to record the local witness used for causal minimality. For a positive normalized smooth mechanism ϑ, let $\mathcal{F}_\vartheta$ be the compact product factorization on $\prod_{r\in[n]}[0,1]$ with local factor

$$\varphi_{\vartheta,r}(\xi) = p_{r,\vartheta}(\xi_r \mid \xi_{\mathrm{pa}_G(r)}), \qquad \xi \in [0,1]^n.$$

For an edge $j_0 \to i_0$, a base point $\xi \in [0,1]^n$, child values $y_0, y_1 \in [0,1]$, and parent values $x_0, x_1 \in [0,1]$, define the local determinant

$$\mathcal{L}^\vartheta_{i_0 j_0}(\xi; y_0, y_1, x_0, x_1) = \varphi_{\vartheta,i_0}(\xi[i_0 := y_0, j_0 := x_0])\varphi_{\vartheta,i_0}(\xi[i_0 := y_1, j_0 := x_1]) - \varphi_{\vartheta,i_0}(\xi[i_0 := y_0, j_0 := x_1])\varphi_{\vartheta,i_0}(\xi[i_0 := y_1, j_0 := x_0]),$$

where $\xi[i_0 := y, j_0 := x]$ is obtained from ξ by replacing the i_0-coordinate by y and the j_0-coordinate by x. In a positive compact product factorization,

$$V_{i_0} \perp\!\!\!\perp V_{j_0} \mid V_{\mathrm{pa}_G(i_0)\setminus\{j_0\}} \quad \Longleftrightarrow \quad \mathcal{L}^\vartheta_{i_0 j_0}(\xi; y_0, y_1, x_0, x_1) = 0$$

for all such ξ, y_0, y_1, x_0, x_1. Indeed, after conditioning on $V_{\mathrm{pa}_G(i_0)\setminus\{j_0\}}$, positivity lets one cancel the remaining factors, and conditional independence is exactly the rank-one condition that all two-by-two determinants of the child factor as a function of (v_{i_0}, v_{j_0}) vanish. The compact-coordinate statement is the same as the coordinate conditional-independence statement in Assumption 5, because both use the same cube-supported observational density.

Since $\theta \in \Theta_{G,s}$, for every edge $j_0 \to i_0$ there is a witness

$$(\xi_{i_0 j_0}, y^0_{i_0 j_0}, y^1_{i_0 j_0}, x^0_{i_0 j_0}, x^1_{i_0 j_0})$$

with

$$\mathcal{L}^{\theta}_{i_0 j_0}(\xi_{i_0 j_0}; y^0_{i_0 j_0}, y^1_{i_0 j_0}, x^0_{i_0 j_0}, x^1_{i_0 j_0}) \neq 0.$$

For each fixed witness, the map $(z_1, z_2, z_3, z_4) \mapsto z_1 z_2 - z_3 z_4$ is continuous, so the witness remains nonzero when the corresponding local factors are uniformly close to those of θ. Taking the minimum over the finite edge set gives $\varepsilon_{\mathrm{fac}} > 0$ such that every positive compact factorization satisfying

$$\max_{r\in[n]} \sup_{\xi\in[0,1]^n} \left|\varphi_{\vartheta,r}(\xi) - \varphi_{\theta,r}(\xi)\right| < \varepsilon_{\mathrm{fac}}$$

has the same nonzero witnesses and hence satisfies causal minimality. If the edge set is empty, the condition is vacuous.

Along the affine path,

$$\max_{r\in[n]} \sup_{\xi\in[0,1]^n} \left|p^t_r(\xi) - p_r(\xi)\right| \le t \max_{r\in[n]} \sup_{\xi\in[0,1]^n} \left|p^\star_r(\xi) - p_r(\xi)\right|,$$

so the factorization neighborhood above contains θ^t for all sufficiently small $t \ge 0$. Combining the closed-segment positive normalized smoothness with the local fixed-sign and causal-minimality neighborhoods, there is $\delta_s > 0$ such that $0 \le t \le 1$ and $t < \delta_s$ imply $\theta^t \in \Theta_{G,s}$.

It remains to choose t avoiding the zero set. If $F_{ji}(\theta) \neq 0$, analyticity gives continuity at 0, so $F_{ji}(\theta^u) \neq 0$ for all sufficiently small u. If $F_{ji}(\theta) = 0$, the isolated-zero conclusion at $0 \in O$ gives the same conclusion for all sufficiently small nonzero u. Hence there is $\delta_z > 0$ such that

$$0 < |u| < \delta_z \quad\Longrightarrow\quad F_{ji}(\theta^u) \neq 0.$$

Set

$$r = \min\{1, \delta_s, \delta_N, \delta_z\}, \qquad t = \frac{r}{2}.$$

Then $t \in (0,1)$, $\theta^t \in N$, $\theta^t \in \Theta_{G,s}$, and $F_{ji}(\theta^t) \neq 0$. □

The perturbation statement gives the local ingredient behind the density argument in Theorem 2: the affine path stays in the same fixed-sign stratum near the starting mechanism and produces a nonzero direct-edge second-moment contrast. The companion zero statement records the population implication for labels whose transported targets have no ancestral relation, using the Gaussian mean-embedding discrepancy that defines the ratio graph in Algorithm 1.

Lemma 2. [`lem:ratio-nonancestor-zero`] (Nonancestor ratio discrepancy vanishes) *Fix an ambient number n of scalar latent variables, a finite labeled DAG G on $[n]$, a mechanism θ, and an observed world W with environment laws $P^0, \dots, P^n$, likelihood ratios $R_i = dP^i/dP^0$, and target permutation π. Assume Assumptions 1 to 3. Let $j, i \in [n]$ satisfy $j \neq i$, and suppose that $\pi(j)$ is not an ancestor of $\pi(i)$ in G. Then the Gaussian-kernel discrepancy of ratio i between the observational environment and environment j vanishes:*

$$D_{ji} = \left\| \int \phi(r)\, d\big((R_i)_\# P^0\big)(r) - \int \phi(r)\, d\big((R_i)_\# P^j\big)(r) \right\|_{\mathcal{H}_k} = 0,$$

where $k(a,b) = \exp(-(a-b)^2)$ is the Gaussian kernel and $\phi(a) = k(a,\cdot)$ is its feature map.

Proof of Lemma 2. 1. Let μ^0 be the latent observational law with density $p^0(v)$ on $[0,1]^n$, and let $\mu^{\pi(j)}$ be the latent law for the single intervention at $\pi(j)$, with density $p^{\pi(j)}(v)$ on $[0,1]^n$. On the cube define

$$\rho_i(v) = \frac{q_{\pi(i)}(v_{\pi(i)})}{p_{\pi(i)}(v_{\pi(i)} \mid v_{\mathrm{pa}_G(\pi(i))})}, \qquad v \in [0,1]^n.$$

The denominator is positive on $[0,1]^n$ by Assumption 1. By Assumption 3, the observed laws are the f-pushforwards of the latent laws, and the likelihood-ratio formula stated there identifies the two ratio laws as

$$(R_i)_\# P^0 = (\rho_i)_\# \mu^0, \qquad (R_i)_\# P^j = (\rho_i)_\# \mu^{\pi(j)}.$$

2. Since $j \neq i$ and π is injective, $\pi(j) \neq \pi(i)$. Together with the hypothesis that $\pi(j)$ is not an ancestor of $\pi(i)$, this gives

$$\pi(j) \notin \mathcal{A}_i, \qquad \mathcal{A}_i = \{\pi(i)\} \cup \{\mathfrak{u} \in [n] : \mathfrak{u} \text{ is an ancestor of } \pi(i) \text{ in } G\}.$$

The set $\mathcal{A}_i$ is parent-closed: if $\mathfrak{u} \in \mathcal{A}_i$ and $\mathfrak{v} \in \mathrm{pa}_G(\mathfrak{u})$, then $\mathfrak{v} \in \mathcal{A}_i$, because a parent of $\pi(i)$ is an ancestor of $\pi(i)$, and a parent of an ancestor of $\pi(i)$ is again an ancestor of $\pi(i)$. Let $\mathrm{pr}_{\mathcal{A}_i} : [0,1]^n \to [0,1]^{\mathcal{A}_i}$ be the coordinate projection. For $\mathfrak{a} \in [0,1]^{\mathcal{A}_i}$ and $\mathfrak{b} \in [0,1]^{[n]\setminus\mathcal{A}_i}$, write $\mathfrak{a} \oplus \mathfrak{b}$ for the point of $[0,1]^n$ whose $\mathcal{A}_i$-coordinates are $\mathfrak{a}$ and whose complementary coordinates are $\mathfrak{b}$. For such $(\mathfrak{a},\mathfrak{b})$, set

$$\Gamma_i^0(\mathfrak{a},\mathfrak{b}) = \prod_{\ell \in [n]\setminus\mathcal{A}_i} p_\ell\big((\mathfrak{a}\oplus\mathfrak{b})_\ell \mid (\mathfrak{a}\oplus\mathfrak{b})_{\mathrm{pa}_G(\ell)}\big)$$

and

$$\Gamma_i^j(\mathfrak{a},\mathfrak{b}) = q_{\pi(j)}(\mathfrak{b}_{\pi(j)}) \prod_{\substack{\ell \in [n]\setminus\mathcal{A}_i \\ \ell \neq \pi(j)}} p_\ell\big((\mathfrak{a}\oplus\mathfrak{b})_\ell \mid (\mathfrak{a}\oplus\mathfrak{b})_{\mathrm{pa}_G(\ell)}\big),$$

where the coordinate $\mathfrak{b}_{\pi(j)}$ is defined because $\pi(j) \notin \mathcal{A}_i$. Parent-closedness implies that, for every $\ell \in \mathcal{A}_i$, the factor p_ℓ evaluated at $\mathfrak{a}\oplus\mathfrak{b}$ depends only on $\mathfrak{a}$. Hence

$$p^0(\mathfrak{a}\oplus\mathfrak{b}) = \left\{\prod_{\ell\in\mathcal{A}_i} p_\ell\big(\mathfrak{a}_\ell \mid \mathfrak{a}_{\mathrm{pa}_G(\ell)}\big)\right\}\Gamma_i^0(\mathfrak{a},\mathfrak{b})$$

and

$$p^{\pi(j)}(\mathfrak{a}\oplus\mathfrak{b}) = \left\{\prod_{\ell\in\mathcal{A}_i} p_\ell\big(\mathfrak{a}_\ell \mid \mathfrak{a}_{\mathrm{pa}_G(\ell)}\big)\right\}\Gamma_i^j(\mathfrak{a},\mathfrak{b}).$$

The complement factors have unit iterated integral. Indeed, if a nonempty collection of complementary vertices remains to be integrated, choose one that is maximal in a topological ordering of the induced remaining set. Parent-closedness prevents this vertex from being a parent of any retained factor, and maximality prevents it from being a parent of any other remaining factor. Its coordinate therefore appears in the current product only through its own factor; integrating that factor gives one by the normalization of p_ℓ, and, when the vertex is $\pi(j)$, by the normalization of $q_{\pi(j)}$. Iterating this deletion, with the empty remaining collection as the base case, gives

$$\int_{[0,1]^{[n]\setminus\mathcal{A}_i}} \Gamma_i^0(\mathfrak{a},\mathfrak{b})\,d\mathfrak{b} = 1, \qquad \int_{[0,1]^{[n]\setminus\mathcal{A}_i}} \Gamma_i^j(\mathfrak{a},\mathfrak{b})\,d\mathfrak{b} = 1.$$

Consequently the two complement-marginal densities on $[0,1]^{\mathcal{A}_i}$ are the same:

$$\int_{[0,1]^{[n]\setminus\mathcal{A}_i}} p^0(\mathfrak{a}\oplus\mathfrak{b})\,d\mathfrak{b} = \prod_{\ell\in\mathcal{A}_i} p_\ell\big(\mathfrak{a}_\ell \mid \mathfrak{a}_{\mathrm{pa}_G(\ell)}\big) = \int_{[0,1]^{[n]\setminus\mathcal{A}_i}} p^{\pi(j)}(\mathfrak{a}\oplus\mathfrak{b})\,d\mathfrak{b}.$$

Thus the projected latent laws agree:

$$(\mathrm{pr}_{\mathcal{A}_i})_\#\mu^0 = (\mathrm{pr}_{\mathcal{A}_i})_\#\mu^{\pi(j)}.$$

Since $\{\pi(i)\}\cup\mathrm{pa}_G(\pi(i)) \subseteq \mathcal{A}_i$, the ratio map ρ_i factors through this projection. Explicitly, the map $\widetilde{\rho}_i : [0,1]^{\mathcal{A}_i} \to \mathbb{R}$ defined by

$$\widetilde{\rho}_i(\mathfrak{a}) = \frac{q_{\pi(i)}\big(\mathfrak{a}_{\pi(i)}\big)}{p_{\pi(i)}\big(\mathfrak{a}_{\pi(i)} \mid \mathfrak{a}_{\mathrm{pa}_G(\pi(i))}\big)}$$

satisfies $\rho_i = \widetilde{\rho}_i \circ \mathrm{pr}_{\mathcal{A}_i}$. Pushing the projected-law identity forward by $\widetilde{\rho}_i$ gives

$$(\rho_i)_\#\mu^0 = (\rho_i)_\#\mu^{\pi(j)},$$

and hence, by the ratio-law identifications in the preceding step,

$$(R_i)_\#P^0 = (R_i)_\#P^j.$$

3. The population discrepancy is the norm of the difference of the two Gaussian-kernel mean embeddings of these ratio laws. Using the equality just proved,

$$\begin{aligned} D_{ji} &= \left\| \int \phi(r)\,d\big((R_i)_\#P^0\big)(r) - \int \phi(r)\,d\big((R_i)_\#P^j\big)(r) \right\|_{\mathcal{H}_k} \\ &= \left\| \int \phi(r)\,d\big((R_i)_\#P^0\big)(r) - \int \phi(r)\,d\big((R_i)_\#P^0\big)(r) \right\|_{\mathcal{H}_k} = 0. \end{aligned}$$

□

The next statements isolate the triangular change of variables used by the conditional-rank decoder. They express predecessor log-ratios as coordinates for the already ordered predecessors and then factor the target intervention law conditional on those scores.

Definition 37. [`synth_29`] (Triangular predecessor-score range and inverse) *Fix an observed world W, a mechanism θ, and an injective topological order $\tau : [n] \to \mathbb{N}$ for the observed ratio graph. For $i \in [n]$, write*

$$B_i^\tau = \{j \in [n] : \tau(j) < \tau(i)\}$$

for the predecessor labels of i in this order. The predecessor-score space is $\mathbb{R}^{B_i^\tau}$, with coordinates indexed by $j \in B_i^\tau$, and the predecessor latent cube is $[0,1]^{B_i^\tau}$.

For $z \in [0,1]^{B_i^\tau}$*, let* $\tilde z \in [0,1]^n$ *be the latent vector that places* z_j *in coordinate* $\pi(j)$ *for* $j \in B_i^\tau$ *and fills all remaining coordinates by* 0*. The triangular predecessor-score map is*

$$\Lambda_i^\tau : [0,1]^{B_i^\tau} \longrightarrow \mathbb{R}^{B_i^\tau}, \qquad \big(\Lambda_i^\tau(z)\big)_j = \log \frac{q_{\pi(j)}(\tilde z_{\pi(j)})}{p_{\pi(j)}(\tilde z)} \quad (j \in B_i^\tau).$$

Equivalently, when $v \in [0,1]^n$ *and* $z_j = v_{\pi(j)}$ *for* $j \in B_i^\tau$*,*

$$\Lambda_i^\tau(z) = \big(L_j(f(v))\big)_{j \in B_i^\tau}.$$

The realized predecessor-score range is the image

$$\mathcal{R}_i^\tau = \Lambda_i^\tau\big([0,1]^{B_i^\tau}\big) \subseteq \mathbb{R}^{B_i^\tau}.$$

Under the positive smooth mechanism, shared mixing, one-intervention-per-node, and fixed own-derivative-sign conditions used in Lemma 6, Λ_i^τ *is injective and hence is a homeomorphism from* $[0,1]^{B_i^\tau}$ *onto* $\mathcal{R}_i^\tau$*. Its inverse is written*

$$(\Lambda_i^\tau)^{-1} : \mathcal{R}_i^\tau \longrightarrow [0,1]^{B_i^\tau},$$

so that for $\ell \in \mathcal{R}_i^\tau$*,* $(\Lambda_i^\tau)^{-1}(\ell)$ *is the unique predecessor latent vector* $z \in [0,1]^{B_i^\tau}$ *with* $\Lambda_i^\tau(z) = \ell$*, and*

$$\Lambda_i^\tau \circ (\Lambda_i^\tau)^{-1} = \mathrm{id}_{\mathcal{R}_i^\tau}, \qquad (\Lambda_i^\tau)^{-1} \circ \Lambda_i^\tau = \mathrm{id}_{[0,1]^{B_i^\tau}}.$$

The range $\mathcal{R}_i^\tau$ *is the continuous image of a compact cube, hence a nonempty compact Borel subset of* $\mathbb{R}^{B_i^\tau}$*. Formulas indexed by an arbitrary score vector are totalized through a fixed Borel retraction*

$$\rho_i^\tau : \mathbb{R}^{B_i^\tau} \longrightarrow \mathcal{R}_i^\tau, \qquad \rho_i^\tau(\ell) = \ell \quad (\ell \in \mathcal{R}_i^\tau),$$

which sends every $\ell \notin \mathcal{R}_i^\tau$ *to one fixed point of* $\mathcal{R}_i^\tau$*. Writing* ρ_i^τ *for this retraction and* $(\Lambda_i^\tau)^{-1}$ *for the inverse keeps the two maps distinct: the composition* $(\Lambda_i^\tau)^{-1} \circ \rho_i^\tau$ *is defined on all of* $\mathbb{R}^{B_i^\tau}$*, and off* $\mathcal{R}_i^\tau$ *the convention carries no content, since every continuity and support claim is asserted on* $\mathcal{R}_i^\tau$*.*

Lemma 3. [`lem:predecessor-score-homeomorphism`] (Predecessor score homeomorphism) *Let* G *be a DAG on* $[n]$*, let* s *be a sign vector on* $[n]$*, let* θ *be a mechanism on* G*, and let* W *be an observed world with observed laws* $P^0, \ldots, P^n$*, likelihood ratios* R_i*, shared observation map* f*, and target permutation* π*. Assume:*

- ***(Regular mechanism and observation model.)*** *The hypotheses of Assumptions 1 to 4 hold for* G, θ, W*, with prescribed sign vector* s*.*
- ***(Observed ratio ordering.)*** *The map* $\tau : [n] \to \mathbb{N}$ *is injective and is a topological ordering of the observable ratio-discrepancy graph* $H_D = \{j \to k : D_{jk} > 0\}$ *constructed from the observed laws with the Gaussian feature map in Algorithm 1: whenever* $D_{jk} > 0$ *and* $j \neq k$*,* $\tau(j) < \tau(k)$*.*
- ***(Transported latent ordering.)*** *The same order also respects the transported latent DAG: whenever* $\pi(j) \to \pi(k)$ *is an edge of* G*,* $\tau(j) < \tau(k)$*.*

Fix $i \in [n]$, set $B_i^\tau = \{b : \tau(b) < \tau(i)\}$, and let

$$\mathcal{Z}_i^\tau = \prod_{b \in B_i^\tau} [0,1].$$

For $\chi \in \mathcal{Z}_i^\tau$, let $E_i^\tau(\chi) \in [0,1]^n$ be the zero-filled latent point with $E_i^\tau(\chi)_{\pi(b)} = \chi_b$ for $b \in B_i^\tau$ and all other coordinates equal to zero. Define $\Lambda_i^\tau : \mathcal{Z}_i^\tau \to \mathbb{R}^{B_i^\tau}$ by

$$\Lambda_i^\tau(\chi)_b = \log \frac{q_{\pi(b)}\big(E_i^\tau(\chi)_{\pi(b)}\big)}{p_{\pi(b)}\big(E_i^\tau(\chi)_{\pi(b)} \mid E_i^\tau(\chi)_{\mathrm{pa}_G(\pi(b))}\big)}, \qquad b \in B_i^\tau,$$

and write $\mathcal{R}_i^\tau = \Lambda_i^\tau(\mathcal{Z}_i^\tau)$. Then Λ_i^τ is a homeomorphism from $\mathcal{Z}_i^\tau$ onto $\mathcal{R}_i^\tau$. In particular, the inverse map $(\Lambda_i^\tau)^{-1} : \mathcal{R}_i^\tau \to \mathcal{Z}_i^\tau$ is continuous.

Proof of Lemma 3. Write $B = B_i^\tau$, $\mathcal{Z} = \mathcal{Z}_i^\tau$, and $E = E_i^\tau$. For $\chi \in \mathcal{Z}$, set $v_\chi = E(\chi)$, so

$$(v_\chi)_a = \begin{cases} \chi_b, & a = \pi(b) \text{ for some } b \in B, \\ 0, & a \notin \pi(B). \end{cases}$$

This point lies in $[0,1]^n$, since each displayed predecessor coordinate lies in $[0,1]$ and every filled coordinate is 0.

1. The map Λ_i^τ is continuous. Indeed, $E : \mathcal{Z} \to [0,1]^n$ is continuous coordinatewise: each coordinate is either one of the product-coordinate projections $\chi \mapsto \chi_b$ or the constant 0. For each $b \in B$, positivity gives

$$p_{\pi(b)}\big((v_\chi)_{\pi(b)} \mid (v_\chi)_{\mathrm{pa}_G(\pi(b))}\big) > 0, \qquad q_{\pi(b)}\big((v_\chi)_{\pi(b)}\big) > 0,$$

   and the C^3 regularity in the hypotheses gives continuity of both factors on the relevant closed cubes. Hence the quotient is a continuous positive function of χ, and applying log preserves continuity. Since this holds for every coordinate $b \in B$, the product map $\Lambda_i^\tau : \mathcal{Z} \to \mathbb{R}^B$ is continuous.

2. The map Λ_i^τ is injective. Let $\chi, \chi' \in \mathcal{Z}$ satisfy

$$\Lambda_i^\tau(\chi) = \Lambda_i^\tau(\chi'),$$

   and put $v = v_\chi$, $w = v_{\chi'}$. With $L_b = \log R_b$, the latent ratio identity in the perfect-intervention hypotheses gives, for every $b \in B$,

$$L_b(f(v)) = \log \frac{q_{\pi(b)}(v_{\pi(b)})}{p_{\pi(b)}(v_{\pi(b)} \mid v_{\mathrm{pa}_G(\pi(b))})}, \qquad L_b(f(w)) = \log \frac{q_{\pi(b)}(w_{\pi(b)})}{p_{\pi(b)}(w_{\pi(b)} \mid w_{\mathrm{pa}_G(\pi(b))})}.$$

   Thus equality of the score vectors is exactly equality of the predecessor log-ratio projections:

$$\big(L_b(f(v))\big)_{b \in B} = \big(L_b(f(w))\big)_{b \in B}.$$

   We prove by strong induction on $m \in \mathbb{N}$ that

$$\forall b \in B, \quad \tau(b) = m \Longrightarrow v_{\pi(b)} = w_{\pi(b)}.$$

Fix $b \in B$ with $\tau(b) = m$, and assume the claim for all smaller order values. If $a \in \mathrm{pa}_G(\pi(b))$, let $c = \pi^{-1}(a)$. Then $\pi(c) \to \pi(b)$ is an edge of G, so the transported latent ordering gives

$$\tau(c) < \tau(b) = m.$$

Since $b \in B$, also $\tau(b) < \tau(i)$, hence $\tau(c) < \tau(i)$ and $c \in B$. The induction hypothesis gives

$$v_a = v_{\pi(c)} = w_{\pi(c)} = w_a.$$

Therefore

$$v_a = w_a \qquad \text{for all } a \in \mathrm{pa}_G(\pi(b)).$$

For $\zeta \in [0,1]$, let $v^{[b:\zeta]} \in [0,1]^n$ be the point obtained from v by replacing the coordinate $\pi(b)$ by ζ:

$$\left(v^{[b:\zeta]}\right)_a = \begin{cases} \zeta, & a = \pi(b), \\ v_a, & a \neq \pi(b). \end{cases}$$

Define, on $[0,1]$, the one-coordinate score

$$\psi_b(\zeta) = \log \frac{q_{\pi(b)}(\zeta)}{p_{\pi(b)}\big(\zeta \mid v_{\mathrm{pa}_G(\pi(b))}\big)}.$$

Strict positivity and C^3 regularity make ψ_b continuous on $[0,1]$ and differentiable on $(0,1)$. For $0 < \zeta < 1$, the point $v^{[b:\zeta]}$ lies in the latent cube, so the fixed own-derivative sign hypothesis gives

$$0 < s_{\pi(b)} \frac{d}{d\zeta} \log \frac{q_{\pi(b)}(\zeta)}{p_{\pi(b)}\big(\zeta \mid v_{\mathrm{pa}_G(\pi(b))}\big)} = s_{\pi(b)}\, \psi_b'(\zeta).$$

Since $s_{\pi(b)} \in \{-1, 1\}$, the mean-value theorem yields, for any $0 \le \zeta_1 < \zeta_2 \le 1$,

$$\psi_b(\zeta_2) - \psi_b(\zeta_1) = \psi_b'(\zeta_\circ)(\zeta_2 - \zeta_1)$$

for some $\zeta_\circ \in (\zeta_1, \zeta_2)$, with positive right-hand side when $s_{\pi(b)} = 1$ and negative right-hand side when $s_{\pi(b)} = -1$. Thus ψ_b is strictly monotone on $[0,1]$.

Moreover,

$$\psi_b(v_{\pi(b)}) = \log \frac{q_{\pi(b)}\big(v_{\pi(b)}\big)}{p_{\pi(b)}\big(v_{\pi(b)} \mid v_{\mathrm{pa}_G(\pi(b))}\big)},$$

and parent locality, together with $v_a = w_a$ for all $a \in \mathrm{pa}_G(\pi(b))$, gives

$$p_{\pi(b)}\big(w_{\pi(b)} \mid v_{\mathrm{pa}_G(\pi(b))}\big) = p_{\pi(b)}\big(w_{\pi(b)} \mid w_{\mathrm{pa}_G(\pi(b))}\big),$$

so

$$\psi_b(w_{\pi(b)}) = \log \frac{q_{\pi(b)}\big(w_{\pi(b)}\big)}{p_{\pi(b)}\big(w_{\pi(b)} \mid w_{\mathrm{pa}_G(\pi(b))}\big)}.$$

The equality $L_b(f(v)) = L_b(f(w))$ therefore gives

$$\psi_b(v_{\pi(b)}) = \psi_b(w_{\pi(b)}).$$

By strict monotonicity of ψ_b,

$$v_{\pi(b)} = w_{\pi(b)}.$$

The induction is complete. Finally, for each $b \in B$,

$$\chi_b = v_{\pi(b)} = w_{\pi(b)} = \chi'_b,$$

so $\chi = \chi'$ in the product cube. Thus Λ_i^τ is injective.

3. Since B is finite, $\mathcal{Z} = \prod_{b\in B}[0,1]$ is compact, and $\mathbb{R}^B$ is Hausdorff. Every closed subset of $\mathcal{Z}$ is compact, so its image under the continuous map Λ_i^τ is compact and hence closed in $\mathbb{R}^B$. Together with injectivity, this makes $\Lambda_i^\tau : \mathcal{Z} \to \mathbb{R}^B$ a topological embedding with closed image. Therefore the induced map

$$\Lambda_i^\tau : \mathcal{Z} \longrightarrow \Lambda_i^\tau(\mathcal{Z}) = \mathcal{R}_i^\tau$$

is a homeomorphism onto its range. Consequently $(\Lambda_i^\tau)^{-1} : \mathcal{R}_i^\tau \to \mathcal{Z}_i^\tau$ is continuous. The same argument covers the case $B = \varnothing$, where both products are singleton spaces.

□

The homeomorphism clause supplies the support-level continuity needed for conditional distribution transforms: predecessor scores identify predecessor latent coordinates on the realized range. The following two marginal identities then remove coordinates outside the retained ancestral set while preserving the factors that enter the target log-ratio.

Lemma 4. [`lem:equation-ten-intervention-marginal`] (Intervention marginal identity) *Fix $i \in [n]$. Suppose that:*

- ***(Regular mechanisms.)*** *The mechanism satisfies Assumption 1.*
- ***(Ratio-graph order.)*** *The map $\tau : [n] \to \mathbb{N}$ is injective and is a topological ordering of the observable ratio-discrepancy graph H_D in Algorithm 1: whenever $D_{ji} > 0$, $\tau(j) < \tau(i)$.*
- ***(Transported-DAG order.)*** *The same order respects the transported latent DAG: whenever $\pi(j) \to \pi(i)$ is an edge of G, $\tau(j) < \tau(i)$.*

Let

$$B_i^\tau = \{j \in [n] : \tau(j) < \tau(i)\}, \qquad A_i^\tau = \{\pi(i)\} \cup \pi(B_i^\tau).$$

For $v \in [0,1]^n$ and $r \in [0,1]^{[n]\setminus A_i^\tau}$, let $u^{v,\tau}(r) \in [0,1]^n$ be the point with

$$u^{v,\tau}(r)_a = v_a \quad (a \in A_i^\tau), \qquad u^{v,\tau}(r)_a = r_a \quad (a \notin A_i^\tau).$$

Then, for every $v \in [0,1]^n$,

$$\int_{[0,1]^{[n]\setminus A_i^\tau}} q_{\pi(i)}\big(u^{v,\tau}(r)_{\pi(i)}\big) \prod_{\ell \neq \pi(i)} p_\ell\big(u^{v,\tau}(r)_\ell \mid u^{v,\tau}(r)_{\mathrm{pa}_G(\ell)}\big)\, dr = q_{\pi(i)}(v_{\pi(i)}) \prod_{\ell \in A_i^\tau \setminus \{\pi(i)\}} p_\ell\big(v_\ell \mid v_{\mathrm{pa}_G(\ell)}\big)\,.$$

Proof of Lemma 4. Put $t_i = \pi(i) \in [n]$ and $\mathcal{A}_i = A_i^\tau \subseteq [n]$.

1. First, $\mathcal{A}_i$ is parent-closed:
$$\lambda \in \mathcal{A}_i, \qquad \chi \in \mathrm{pa}_G(\lambda) \quad \Longrightarrow \quad \chi \in \mathcal{A}_i.$$
Indeed, if $\lambda = t_i = \pi(i)$, then $\pi^{-1}(\chi) \to i$ in the transported environment-label graph, so the transported-DAG order gives
$$\tau(\pi^{-1}(\chi)) < \tau(i).$$
Thus $\pi^{-1}(\chi) \in B_i^\tau$, hence $\chi \in \pi(B_i^\tau) \subseteq \mathcal{A}_i$. If instead $\lambda = \pi(\xi)$ for some $\xi \in B_i^\tau$, then $\pi^{-1}(\chi) \to \xi$ in the transported graph, and the same order gives
$$\tau(\pi^{-1}(\chi)) < \tau(\xi) < \tau(i).$$
Thus again $\pi^{-1}(\chi) \in B_i^\tau$, hence $\chi \in \mathcal{A}_i$.

2. Define, for $\omega \in [0,1]^n$ and $\lambda \in [n]$,
$$\widetilde{p}_\lambda(\omega) = \begin{cases} q_{t_i}(\omega_{t_i}), & \lambda = t_i, \\ p_\lambda\big(\omega_\lambda \mid \omega_{\mathrm{pa}_G(\lambda)}\big), & \lambda \neq t_i. \end{cases}$$
The intervention density at target t_i is therefore
$$\prod_{\lambda \in [n]} \widetilde{p}_\lambda(\omega) = q_{t_i}(\omega_{t_i}) \prod_{\lambda \neq t_i} p_\lambda\big(\omega_\lambda \mid \omega_{\mathrm{pa}_G(\lambda)}\big).$$
For $v \in [0,1]^n$ and $r \in [0,1]^{[n]\setminus\mathcal{A}_i}$, write $u^{v,\tau}(r)$ for the point satisfying
$$u^{v,\tau}(r)_a = \begin{cases} v_a, & a \in \mathcal{A}_i, \\ r_a, & a \notin \mathcal{A}_i. \end{cases}$$
Since $t_i \in \mathcal{A}_i$, the preceding display is the ordinary product density after replacing the t_i-factor by q_{t_i}.

   For a parent-closed subset $\mathcal{A} \subseteq [n]$, an index $\jmath \in \mathcal{A}$, and a normalized replacement density $q_\jmath$, define
$$\widehat{p}_\lambda(\omega) = \begin{cases} q_\jmath(\omega_\jmath), & \lambda = \jmath, \\ p_\lambda\big(\omega_\lambda \mid \omega_{\mathrm{pa}_G(\lambda)}\big), & \lambda \neq \jmath, \end{cases}$$
and let
$$\widehat{p}^{\mathrm{obs}}(\omega) = \prod_{\lambda \in [n]} p_\lambda\big(\omega_\lambda \mid \omega_{\mathrm{pa}_G(\lambda)}\big), \qquad \widehat{p}^{\mathrm{int}}(\omega) = \prod_{\lambda \in [n]} \widehat{p}_\lambda(\omega).$$
The replacement construction gives
$$\widehat{p}^{\mathrm{int}}(\omega) = q_\jmath(\omega_\jmath) \prod_{\lambda \neq \jmath} p_\lambda\big(\omega_\lambda \mid \omega_{\mathrm{pa}_G(\lambda)}\big),$$
that is, the observational product factorization with the single $\jmath$-factor replaced. For this replaced factorization, parent-closed marginalization gives
$$\int_{[0,1]^{[n]\setminus\mathcal{A}}} \widehat{p}^{\mathrm{int}}(u)\, du_{[n]\setminus\mathcal{A}} = \prod_{\lambda \in \mathcal{A}} \widehat{p}_\lambda(\omega),$$

where the integrated point u agrees with ω on $\mathcal{A}$. This is the usual finite product-density marginal identity for a parent-closed coordinate set: after the single-factor replacement, the factors again form a normalized directed factorization, and parent-closedness ensures that retaining $\mathcal{A}$ leaves exactly the product of the retained factors. The sigma-finiteness and finite-coordinate integration bookkeeping are supplied by the unit-interval reference measure, while Assumption 1 supplies the normalized measurable factors and the normalized replacement density.

Apply this identity with $\mathcal{A} = \mathcal{A}_i$ and $\jmath = t_i$. Unfolding the retained product gives

$$\prod_{\lambda \in \mathcal{A}_i} \widetilde{p}_\lambda(v) = q_{t_i}(v_{t_i}) \prod_{\lambda \in \mathcal{A}_i \setminus \{t_i\}} p_\lambda\big(v_\lambda \mid v_{\mathrm{pa}_G(\lambda)}\big) ,$$

because the erased product contains no t_i-factor. Therefore

$$\int_{[0,1]^{[n]\setminus \mathcal{A}_i}} \prod_{\lambda \in [n]} \widetilde{p}_\lambda(u^{v,\tau}(r))\, dr = q_{t_i}(v_{t_i}) \prod_{\lambda \in \mathcal{A}_i \setminus \{t_i\}} p_\lambda\big(v_\lambda \mid v_{\mathrm{pa}_G(\lambda)}\big) .$$

3. Substituting $t_i = \pi(i)$ and $\mathcal{A}_i = A_i^\tau$ in the preceding identity gives, for every $v \in [0,1]^n$,

$$\int_{[0,1]^{[n]\setminus A_i^\tau}} q_{\pi(i)}\big(u^{v,\tau}(r)_{\pi(i)}\big) \prod_{\ell \neq \pi(i)} p_\ell\big(u^{v,\tau}(r)_\ell \mid u^{v,\tau}(r)_{\mathrm{pa}_G(\ell)}\big)\, dr = q_{\pi(i)}(v_{\pi(i)}) \prod_{\ell \in A_i^\tau \setminus \{\pi(i)\}} p_\ell\big(v_\ell \mid v_{\mathrm{pa}_G(\ell)}\big) ,$$

which is the asserted identity.

□

Lemma 5. [`lem:equation-ten-fiber-product`] (Fiber product factorization) *In the setting of Assumption 1, fix an observed world with target permutation π. Let τ be a numerical order on $[n]$, and fix $i \in [n]$. Assume:*

- ***(Ratio-graph order.)*** *The order τ is injective and topological for the observable ratio graph $H_D = \{j \to i : j \neq i,\ D_{ji} > 0\}$ of Algorithm 1: whenever $D_{ji} > 0$, one has $\tau(j) < \tau(i)$.*
- ***(Transported-DAG order.)*** *The same order respects every transported latent edge: whenever $\pi(j) \to \pi(i)$ is an edge of G, one has $\tau(j) < \tau(i)$.*

Write

$$B_i^\tau = \{j : \tau(j) < \tau(i)\}, \qquad A_i^\tau = \{\pi(i)\} \cup \pi(B_i^\tau).$$

For $x \in \mathbb{R}^n$ and $r \in [0,1]^{[n]\setminus A_i^\tau}$, let $u^{x,\tau}(r)$ be the latent point whose coordinates in A_i^τ agree with x and whose remaining coordinates are filled by r. For every $v \in \mathbb{R}^n$ and every $w \in \mathbb{R}$, if $v^{(w)}$ denotes v with its $\pi(i)$-coordinate replaced by w, then

$$\int_{[0,1]^{[n]\setminus A_i^\tau}} q_{\pi(i)}\Big(u^{v^{(w)},\tau}(r)_{\pi(i)}\Big) \prod_{\ell \neq \pi(i)} p_\ell\Big(u^{v^{(w)},\tau}(r)_\ell \mid u^{v^{(w)},\tau}(r)_{\mathrm{pa}_G(\ell)}\Big)\, dr = q_{\pi(i)}(w) \prod_{\ell \in A_i^\tau \setminus \{\pi(i)\}} p_\ell\big(v_\ell \mid v_{\mathrm{pa}_G(\ell)}\big) .$$

Proof of Lemma 5. Let $i_\pi = \pi(i)$, and put $A = A_i^\tau$. For $a \in [n]$, write

$$v_a^{(w)} = \begin{cases} w, & a = i_\pi, \\ v_a, & a \neq i_\pi. \end{cases}$$

1. The intervention marginal factorization supplied by Lemma 4 is used here in its ambient density-carrier form: for the arbitrary point $v^{(w)} \in \mathbb{R}^n$,

$$\int_{[0,1]^{[n]\setminus A}} q_{i_\pi}\Big(u^{v^{(w)},\tau}(r)_{i_\pi}\Big) \prod_{\ell \neq i_\pi} p_\ell\Big(u^{v^{(w)},\tau}(r)_\ell \mid u^{v^{(w)},\tau}(r)_{\mathrm{pa}_G(\ell)}\Big)\, dr$$
$$= q_{i_\pi}\Big(v^{(w)}_{i_\pi}\Big) \prod_{\ell \in A\setminus\{i_\pi\}} p_\ell\Big(v^{(w)}_\ell \mid v^{(w)}_{\mathrm{pa}_G(\ell)}\Big).$$

The density carriers p_ℓ and q_{i_π} are evaluated as their ambient functions on $\mathbb{R}^n$ and $\mathbb{R}$, respectively, so the coordinate replacement is admissible for every $v \in \mathbb{R}^n$ and $w \in \mathbb{R}$. Since $v^{(w)}_{i_\pi} = w$, the scalar factor on the right is $q_{i_\pi}(w)$.

2. It remains to identify the retained observational product after the coordinate replacement. We claim that

$$\prod_{\ell \in A\setminus\{i_\pi\}} p_\ell\Big(v^{(w)}_\ell \mid v^{(w)}_{\mathrm{pa}_G(\ell)}\Big) = \prod_{\ell \in A\setminus\{i_\pi\}} p_\ell\big(v_\ell \mid v_{\mathrm{pa}_G(\ell)}\big).$$

Fix $\ell \in A \setminus \{i_\pi\}$. By the definition

$$A = \{i_\pi\} \cup \pi(B_i^\tau), \qquad B_i^\tau = \{j : \tau(j) < \tau(i)\},$$

there is $j \in B_i^\tau$ with $\ell = \pi(j)$. The own coordinate satisfies

$$\ell \neq i_\pi,$$

because ℓ lies in $A \setminus \{i_\pi\}$. Now let $a \in \mathrm{pa}_G(\ell)$. If $a = i_\pi$, then $i_\pi \to \ell = \pi(j)$ is an edge of G, so the transported-DAG order gives

$$\tau(i) < \tau(j).$$

But $j \in B_i^\tau$ gives $\tau(j) < \tau(i)$, a contradiction. Hence

$$a \neq i_\pi \qquad \text{for every } a \in \mathrm{pa}_G(\ell).$$

Thus $v^{(w)}$ and v agree on $\{\ell\} \cup \mathrm{pa}_G(\ell)$, and consequently

$$p_\ell\Big(v^{(w)}_\ell \mid v^{(w)}_{\mathrm{pa}_G(\ell)}\Big) = p_\ell\big(v_\ell \mid v_{\mathrm{pa}_G(\ell)}\big).$$

Multiplying this equality over $\ell \in A \setminus \{i_\pi\}$ proves the claimed product identity; when the index set is empty, both products are the empty product 1.

3. Substituting the product identity from the previous step into the marginal identity from the first step yields

$$\int_{[0,1]^{[n]\setminus A_i^\tau}} q_{\pi(i)}\Big(u^{v^{(w)},\tau}(r)_{\pi(i)}\Big) \prod_{\ell \neq \pi(i)} p_\ell\Big(u^{v^{(w)},\tau}(r)_\ell \mid u^{v^{(w)},\tau}(r)_{\mathrm{pa}_G(\ell)}\Big)\, dr$$
$$= q_{\pi(i)}(w) \prod_{\ell \in A_i^\tau\setminus\{\pi(i)\}} p_\ell\big(v_\ell \mid v_{\mathrm{pa}_G(\ell)}\big),$$

which is the asserted fiber product factorization.

□

Lemma 6. [lem:equation-eleven-kernel-factorization] (Kernel factorization of scores) *Let n be fixed, let G be a finite labeled DAG on $[n]$, let s be a sign vector with entries in $\{-1, 1\}$, and let W be an observed world generated by a mechanism θ. Assume:*

- ***(Structural regularity.)*** *The mechanism and observed world satisfy Assumptions 1 to 4 with sign vector s.*
- ***(Observed ratio order.)*** *The map $\tau : [n] \to \mathbb{N}$ is injective and is a topological ordering of the observed Gaussian-MMD ratio graph H_D of Algorithm 1: whenever H_D contains $j \to k$, $\tau(j) < \tau(k)$.*
- ***(Transported graph order.)*** *The same order respects the transported latent DAG: whenever $\pi(j) \to \pi(k)$ is an edge of G, $\tau(j) < \tau(k)$.*

Fix $i \in [n]$, and put

$$B_i^\tau = \{j \in [n] : \tau(j) < \tau(i)\}.$$

For each predecessor log-ratio vector $\ell \in \mathbb{R}^{B_i^\tau}$, let $\kappa_i^\tau(\ell, \cdot)$ be the law on $\mathbb{R}$ of

$$\log \frac{q_{\pi(i)}(W_i)}{p_{\pi(i)}(\widetilde{v}_i^\tau(z_i^\tau(\ell), W_i))}, \qquad W_i \sim q_{\pi(i)}(w)\, dw \text{ on } [0, 1],$$

where $z_i^\tau = (\Lambda_i^\tau)^{-1} \circ \rho_i^\tau$ is the triangular predecessor-score inverse composed with the retraction onto the realized score range $\mathcal{R}_i^\tau$, and $\widetilde{v}_i^\tau(z, w)$ is the latent point whose $\pi(j)$-coordinate is z_j for $j \in B_i^\tau$, whose $\pi(i)$-coordinate is w, and whose remaining coordinates are zero. Then, under the target-$\pi(i)$ observed environment law P^i, the joint law of the predecessor log-ratios and the target log-ratio factorizes as

$$(L_{B_i^\tau}, L_i)_\# P^i = (L_{B_i^\tau})_\# P^i \otimes \kappa_i^\tau.$$

Proof of Lemma 6.

1. Let μ_i be the latent target-$\pi(i)$ intervention law on $[0, 1]^n$,

$$\mu_i(dv) = q_{\pi(i)}(v_{\pi(i)}) \prod_{\ell \neq \pi(i)} p_\ell(v_\ell \mid v_{\mathrm{pa}_G(\ell)})\, dv,$$

and let

$$\nu_i(dw) = q_{\pi(i)}(w) \mathbf{1}_{[0,1]}(w)\, dw.$$

By Assumption 1, ν_i is a probability law. Define the total clamped predecessor-coordinate map

$$\zeta_i^\tau : \mathbb{R}^n \longrightarrow [0, 1]^{B_i^\tau}, \qquad \left(\zeta_i^\tau(v)\right)_j = \max\{0, \min\{1, v_{\pi(j)}\}\}, \quad j \in B_i^\tau.$$

Thus, for $v \in [0, 1]^n$,

$$\zeta_i^\tau(v) = (v_{\pi(j)})_{j \in B_i^\tau}.$$

For $\xi \in [0, 1]^{B_i^\tau}$, let $\widehat{\xi} \in [0, 1]^n$ be the latent point with

$$\widehat{\xi}_{\pi(j)} = \xi_j \quad (j \in B_i^\tau), \qquad \widehat{\xi}_a = 0 \quad (a \notin \pi(B_i^\tau)).$$

The triangular predecessor-score map is

$$\Lambda_i^\tau : [0,1]^{B_i^\tau} \to \mathbb{R}^{B_i^\tau}, \qquad \left(\Lambda_i^\tau(\xi)\right)_j = \log \frac{q_{\pi(j)}(\xi_j)}{p_{\pi(j)}\left(\widehat{\xi}_{\pi(j)} \mid \widehat{\xi}_{\mathrm{pa}_G(\pi(j))}\right)}, \quad j \in B_i^\tau.$$

The assumptions in the statement make Λ_i^τ one-to-one on the compact predecessor cube, with realized range $\mathcal{R}_i^\tau$, as recorded in Definition 37. For $\xi \in [0,1]^{B_i^\tau}$ and $w \in \mathbb{R}$, put

$$\bar{w} = \max\{0, \min\{1, w\}\}, \qquad \Gamma_i^\tau(\xi, w) = \log \frac{q_{\pi(i)}(\bar{w})}{p_{\pi(i)}(\widetilde{v}_i^\tau(\xi, \bar{w}))}.$$

The clamping makes Γ_i^τ a Borel map on the full domain $[0,1]^{B_i^\tau} \times \mathbb{R}$, while on $w \in [0,1]$ it is exactly the target score in the statement. Since the retraction onto $\mathcal{R}_i^\tau$ fixes every realized value and the inverse of Λ_i^τ sends $\Lambda_i^\tau(\xi)$ back to ξ, the kernel satisfies the fiber identity

$$\kappa_i^\tau(\Lambda_i^\tau(\xi), \cdot) = \left(\Gamma_i^\tau(\xi, \cdot)\right)_{\#}\nu_i, \qquad \xi \in [0,1]^{B_i^\tau}.$$

When $B_i^\tau = \varnothing$, the predecessor cube is the singleton product and the same formula is the corresponding singleton-index identity.

2. The latent predecessor coordinates and the target coordinate factor under μ_i:

$$(\zeta_i^\tau, v_{\pi(i)})_{\#}\mu_i = (\zeta_i^\tau)_{\#}\mu_i \otimes \nu_i.$$

Indeed, Lemma 5 integrates the target-$\pi(i)$ intervention density over all coordinates outside

$$A_i^\tau = \{\pi(i)\} \cup \pi(B_i^\tau)$$

and leaves $q_{\pi(i)}(w)$ as the target-coordinate factor, multiplying a factor depending only on the retained predecessor coordinates. The normalization in Assumption 1 identifies this target factor with ν_i.

3. Define the latent joint score map

$$\mathcal{J}_i^\tau(v) = \left(\Lambda_i^\tau(\zeta_i^\tau(v)), \Gamma_i^\tau(\zeta_i^\tau(v), v_{\pi(i)})\right).$$

For a measure α on $\mathbb{R}^{B_i^\tau}$ and a Markov kernel K from $\mathbb{R}^{B_i^\tau}$ to $\mathbb{R}$, write $\alpha \otimes_m K$ for the measure on $\mathbb{R}^{B_i^\tau} \times \mathbb{R}$ defined by

$$(\alpha \otimes_m K)(E) = \int K(a, \{y : (a,y) \in E\})\, d\alpha(a)$$

for measurable $E \subseteq \mathbb{R}^{B_i^\tau} \times \mathbb{R}$. This is the product denoted by $\alpha \otimes K$ in the displayed statement. Mapping the product identity in the previous step through

$$(\xi, w) \longmapsto \left(\Lambda_i^\tau(\xi), \Gamma_i^\tau(\xi, w)\right)$$

and using the fiber identity for κ_i^τ gives the measure-kernel product factorization

$$(\mathcal{J}_i^\tau)_{\#}\mu_i = (\Lambda_i^\tau \circ \zeta_i^\tau)_{\#}\mu_i \otimes_m \kappa_i^\tau.$$

Indeed, for every measurable $E \subseteq \mathbb{R}^{B_i^\tau} \times \mathbb{R}$,

$$\int \nu_i\{w : (\Lambda_i^\tau(\xi), \Gamma_i^\tau(\xi, w)) \in E\}\, d((\zeta_i^\tau)_{\#}\mu_i)(\xi) = \int \kappa_i^\tau(a, \{y : (a,y) \in E\})\, d((\Lambda_i^\tau \circ \zeta_i^\tau)_{\#}\mu_i)(a),$$

which is exactly the displayed measure-kernel product.

4. For μ_i-almost every v, one has $v \in [0,1]^n$, because the support of μ_i is the closed latent cube. On this event, clamping is inactive. Fix $j \in B_i^\tau$. The point $\widehat{\zeta_i^\tau(v)}$ agrees with v in coordinate $\pi(j)$. It also agrees on each parent of $\pi(j)$: if $u \to \pi(j)$ is an edge of G, write $a = \pi^{-1}(u)$. Then the transported edge is $\pi(a) \to \pi(j)$, and the transported graph order gives $\tau(a) < \tau(j)$. Since $j \in B_i^\tau$, $\tau(j) < \tau(i)$, so $\tau(a) < \tau(i)$ and hence $a \in B_i^\tau$. Thus the retained predecessor vector supplies the coordinate $u = \pi(a)$. By the parent-locality of $p_{\pi(j)}$ and the latent ratio formula in Assumption 3,

$$\big(\Lambda_i^\tau(\zeta_i^\tau(v))\big)_j = \log \frac{q_{\pi(j)}(v_{\pi(j)})}{p_{\pi(j)}(v_{\pi(j)} \mid v_{\mathrm{pa}_G(\pi(j))})} = L_j(f(v)).$$

Thus

$$\Lambda_i^\tau(\zeta_i^\tau(v)) = L_{B_i^\tau}(f(v)) \qquad \text{for } \mu_i\text{-almost every } v.$$

5. The same comparison gives the target coordinate. If $u \to \pi(i)$ is a parent edge in G, write $a = \pi^{-1}(u)$. Then $a \neq i$, and the transported graph order applied to $\pi(a) \to \pi(i)$ gives $\tau(a) < \tau(i)$, so $a \in B_i^\tau$. Therefore $\widetilde{v}_i^\tau(\zeta_i^\tau(v), v_{\pi(i)})$ agrees with v on $\pi(i)$ and on every parent coordinate $u = \pi(a)$ of $\pi(i)$. Hence, again by Assumption 3,

$$\Gamma_i^\tau(\zeta_i^\tau(v), v_{\pi(i)}) = \log \frac{q_{\pi(i)}(v_{\pi(i)})}{p_{\pi(i)}(v_{\pi(i)} \mid v_{\mathrm{pa}_G(\pi(i))})} = L_i(f(v))$$

for μ_i-almost every v. Combining this with the preceding step,

$$\mathcal{J}_i^\tau(v) = \big(L_{B_i^\tau}(f(v)), L_i(f(v))\big) \qquad \text{for } \mu_i\text{-almost every } v.$$

6. By Assumption 3, the observed target-$\pi(i)$ law satisfies

$$P^i = f_{\#}\mu_i.$$

Therefore the almost-sure identity in the previous step yields

$$(L_{B_i^\tau}, L_i)_{\#} P^i = (\mathcal{J}_i^\tau)_{\#}\mu_i.$$

For the marginal predecessor pushforward, the almost-sure identity from step 4 gives

$$\Lambda_i^\tau(\zeta_i^\tau(v)) = L_{B_i^\tau}(f(v)) \qquad \text{for } \mu_i\text{-almost every } v,$$

and the same pushforward calculation gives

$$(L_{B_i^\tau})_{\#} P^i = (\Lambda_i^\tau \circ \zeta_i^\tau)_{\#}\mu_i.$$

Substituting these two pushforward identities into the latent factorization from step 3 gives

$$(L_{B_i^\tau}, L_i)_{\#} P^i = (L_{B_i^\tau})_{\#} P^i \otimes_m \kappa_i^\tau,$$

which is the asserted factorization, with the statement's product understood as this measure-kernel product.

□

Together, these identities turn the target intervention environment into a product between predecessor scores and a one-dimensional kernel indexed by those scores. The continuous-rank clauses below state the corresponding conditional-CDF representation and the resulting rank coordinate used by the population decoder.

Lemma 7. [`lem:exact-ratio-decoder-continuous-rank-clauses`] (Continuous rank decoder clauses) *Let n be the ambient number of scalar latent variables, let G be the ambient finite labeled DAG on $[n]$, let θ be the mechanism, let W be the observed world with shared observation map f, observed laws $P^0, \dots, P^n$, likelihood ratios $R_i = dP^i/dP^0$, log ratios $L_i = \log R_i$, and target permutation π, and let s be a sign vector with $s_i \in \{-1, 1\}$ for every i. Assume:*

- ***(Mechanism regularity.)*** *The mechanism satisfies Assumption 1.*
- ***(Shared mixing.)*** *The observed world satisfies Assumption 2.*
- ***(Single-target interventions.)*** *The observed laws and likelihood ratios satisfy Assumption 3.*
- ***(Own-derivative signs.)*** *The mechanism satisfies Assumption 4 with sign vector s.*
- ***(Ratio-order recovery.)*** *For the observable directed ratio-discrepancy graph H_D of Algorithm 1, built from the Gaussian kernel $k(a, b) = \exp(-(a - b)^2)$, the transitive closures agree:*

$$\mathrm{TC}(H_D) = \mathrm{TC}(G^\pi).$$

For any injective topological ordering τ of H_D, any $i \in [n]$, and $B_i^\tau = \{j : \tau(j) < \tau(i)\}$, call a function C a continuous conditional-CDF version of L_i given $L_{B_i^\tau}$ when C is jointly Borel measurable, for every fixed $t \in \mathbb{R}$ it represents the regular conditional probability $P^i(L_i \le t \mid L_{B_i^\tau} = \ell)$ for $(L_{B_i^\tau})_\# P^i$-almost every ℓ, the same identity holds for $(L_i, L_{B_i^\tau})_\# P^i$-almost every (t, ℓ), and C is continuous on the observed conditional-ratio support

$$\mathcal{S}_i^\tau = \mathbb{R} \times \operatorname{supp}\big((L_{B_i^\tau})_\# P^i\big).$$

Then, for every such τ and every $i \in [n]$, with C_i denoting the law-selected conditional CDF used by Algorithm 1, the following hold:

- *C_i is a continuous conditional-CDF version of L_i given $L_{B_i^\tau}$.*
- *There exists a continuous conditional-CDF version C such that, for all $t \in \mathbb{R}$ and all $v \in [0, 1]^n$,*

$$C\big(t \mid L_{B_i^\tau}(f(v))\big) = \int_0^1 \mathbf{1}\left\{\log \frac{q_{\pi(i)}(w)}{p_{\pi(i)}\big(v^{(w)}\big)} \le t\right\} q_{\pi(i)}(w)\, dw,$$

  where $v^{(w)}$ is the latent point obtained from v by replacing coordinate $\pi(i)$ by w.

- *Every continuous conditional-CDF version C agrees with C_i on $\mathcal{S}_i^\tau$:*

$$C(z_1 \mid z_2) = C_i(z_1 \mid z_2) \qquad \textit{for every } (z_1, z_2) \in \mathcal{S}_i^\tau.$$

- *For all $t \in \mathbb{R}$ and all $v \in [0, 1]^n$,*

$$C_i\big(t \mid L_{B_i^\tau}(f(v))\big) = \int_0^1 \mathbf{1}\left\{\log \frac{q_{\pi(i)}(w)}{p_{\pi(i)}\big(v^{(w)}\big)} \le t\right\} q_{\pi(i)}(w)\, dw.$$

- *The rank coordinate* $U_i = C_i(L_i \mid L_{B_i^\tau})$ *returned by* Algorithm 1 *satisfies, for every* $v \in [0,1]^n$,
$$U_i(f(v)) = \begin{cases} Q_{\pi(i)}(v_{\pi(i)}), & s_{\pi(i)} = 1, \\ 1 - Q_{\pi(i)}(v_{\pi(i)}), & s_{\pi(i)} = -1, \end{cases}$$
where $Q_{\pi(i)}(x) = \int_0^x q_{\pi(i)}(u)\,du$.

Proof of Lemma 7. Fix an injective topological ordering τ of H_D, and fix $i \in [n]$. Write

$$B_i^\tau = \{j \in [n] : \tau(j) < \tau(i)\}, \qquad L_{B_i^\tau}(x) = (L_j(x))_{j \in B_i^\tau}.$$

1. First record the order-recovery consequence that is passed to the ordered decoder argument. If G^π contains $j \to k$, then the one-edge path $j \to k$ is an element of $\mathrm{TC}(G^\pi)$. The assumed equality
$$\mathrm{TC}(H_D) = \mathrm{TC}(G^\pi)$$
therefore gives a directed H_D-path from j to k. Along each edge $a \to b$ of this path, topologicality of τ for H_D gives $\tau(a) < \tau(b)$; induction over the path and transitivity of $<$ yield
$$G^\pi : j \to k \quad \Longrightarrow \quad \tau(j) < \tau(k).$$
Thus τ is also topological for the transported graph G^π, and in particular every G^π-parent of k lies in B_k^τ.

2. The supplied observed laws are probability laws. The observational law is the pushforward by f of the normalized observational latent density, and each target-intervention law is the pushforward by the same f of the corresponding normalized intervention density. Hence, for every $e = 0, \ldots, n$,
$$P^e(\mathbb{R}^n) = 1.$$
Consequently the decoder is computed from these laws themselves: writing P for the law family it uses,
$$\mathsf{P}_e = P^e, \qquad e = 0, \ldots, n.$$
In each subsequent conditional-CDF statement, the selected family is rewritten by this identity to the stated observed laws.

3. Let $\kappa_i^\tau(\ell, \cdot)$ be the probability kernel from Lemma 6, extended to all $\ell \in \mathbb{R}^{B_i^\tau}$ by first retracting ℓ onto the realized predecessor-score range. Thus, for every $\ell \in \mathbb{R}^{B_i^\tau}$, $\kappa_i^\tau(\ell, \cdot)$ is the law of
$$\log \frac{q_{\pi(i)}(W_i)}{p_{\pi(i)}(\widetilde{v}_i^\tau(z_i^\tau(\ell), W_i))}, \qquad W_i \sim q_{\pi(i)}(w)\,dw \text{ on } [0,1],$$
where the retraction in z_i^τ supplies the off-range values. Define the total candidate
$$C_i^\star(t \mid \ell) = \kappa_i^\tau\big(\ell, (-\infty, t]\big), \qquad (t, \ell) \in \mathbb{R} \times \mathbb{R}^{B_i^\tau}.$$
The same formula applies when $B_i^\tau = \varnothing$, with $\mathbb{R}^{B_i^\tau}$ the one-point product space. Put
$$\mu_i^\tau = (L_{B_i^\tau})_\# P^i, \qquad \zeta_i^\tau = (L_{B_i^\tau}, L_i)_\# P^i.$$
By Lemma 6,
$$\zeta_i^\tau = \mu_i^\tau \otimes \kappa_i^\tau.$$

The candidate $C_i^\star$ has the four properties required of a continuous conditional-CDF version. First, $(t,\ell) \mapsto C_i^\star(t \mid \ell)$ is Borel measurable because it is the lower-half-line evaluation of a measurable probability kernel. Second, uniqueness of regular conditional distributions in the displayed product disintegration gives, for every fixed $t \in \mathbb{R}$,

$$C_i^\star(t \mid \ell) = P^i(L_i \le t \mid L_{B_i^\tau} = \ell) \quad \text{for } \mu_i^\tau\text{-a.e. } \ell.$$

The same disintegration, pulled back along the argument map $x \mapsto (L_i(x), L_{B_i^\tau}(x))$, gives the joint almost-everywhere version

$$C_i^\star(r \mid \ell) = P^i(L_i \le r \mid L_{B_i^\tau} = \ell) \quad \text{for } (L_i, L_{B_i^\tau})_\# P^i\text{-a.e. } (r,\ell).$$

For continuity, put

$$\mathcal{R}_i^\tau = \{L_{B_i^\tau}(f(u)) : u \in [0,1]^n\}.$$

On $\mathbb{R} \times \mathcal{R}_i^\tau$, the retraction in the definition of κ_i^τ is the identity, and the predecessor-score homeomorphism rewrites $C_i^\star$ as the lower-level integral of the one-dimensional score

$$w \longmapsto \log \frac{q_{\pi(i)}(w)}{p_{\pi(i)}(\widetilde{v}_i^\tau(z,w))}, \qquad z = (\Lambda_i^\tau)^{-1}(\ell).$$

The positivity and smoothness in Assumption 1, together with the fixed own-derivative sign in Assumption 4, make this score continuous in (z,w) and strictly monotone in w on $[0,1]$; hence its lower-level integral varies continuously in (t,ℓ) on $\mathbb{R} \times \mathcal{R}_i^\tau$. The observed conditional-ratio support satisfies

$$\mathcal{S}_i^\tau \subseteq \mathbb{R} \times \mathcal{R}_i^\tau,$$

because under P^i the predecessor projection is supported on realized predecessor scores. Thus $C_i^\star$ is continuous on $\mathcal{S}_i^\tau$, and so it is a continuous conditional-CDF version of L_i given $L_{B_i^\tau}$.

4. Since the class of continuous conditional-CDF versions is nonempty, the law-selected conditional CDF C_i used by Algorithm 1 is itself one of these versions. This proves the first asserted clause:

$$C_i \quad \text{is a continuous conditional-CDF version of } L_i \text{ given } L_{B_i^\tau}.$$

Taking $C = C_i^\star$ proves the asserted existence of a continuous version once its pointwise latent formula is verified in the next step.

5. Fix $v \in [0,1]^n$, and set

$$\ell_v = L_{B_i^\tau}(f(v)).$$

The realized score ℓ_v belongs to $\mathcal{R}_i^\tau$, so the retraction used to define κ_i^τ fixes it. The score-image kernel then evaluates at the predecessor coordinates reconstructed from ℓ_v. For each $w \in [0,1]$, let $\bar{v}_i^\tau(\ell_v, w)$ be this reconstructed point with target coordinate $\pi(i)$ set to w. The order-recovery step places every transported parent of $\pi(i)$ in B_i^τ, and the inverse predecessor-score map reconstructs those predecessor coordinates from ℓ_v. Hence

$$\bar{v}_i^\tau(\ell_v, w)_a = v_a^{(w)} \qquad \text{for every } a \in \mathrm{pa}_G(\pi(i)),$$

where $v^{(w)}$ is obtained from v by replacing coordinate $\pi(i)$ by w. Parent locality of the mechanism gives

$$p_{\pi(i)}\big(\bar{v}_i^\tau(\ell_v, w)\big) = p_{\pi(i)}\big(v^{(w)}\big).$$

Substituting this identity into the lower-half-line evaluation of κ_i^τ yields, for every $t \in \mathbb{R}$,

$$C_i^\star\big(t \mid L_{B_i^\tau}(f(v))\big) = \int_0^1 \mathbf{1}\left\{\log \frac{q_{\pi(i)}(w)}{p_{\pi(i)}\big(v^{(w)}\big)} \le t\right\} q_{\pi(i)}(w)\, dw.$$

This is the second asserted clause, with $C = C_i^\star$.

6. Any two continuous conditional-CDF versions agree on the observed conditional-ratio support. To see this, fix another continuous version C. For each fixed threshold t, the defining conditional-CDF property gives

$$C(t \mid \ell) = C_i(t \mid \ell) \quad \text{for } \mu_i^\tau\text{-a.e. } \ell.$$

Both sides are continuous in ℓ on $\operatorname{supp} \mu_i^\tau$. Therefore, for every $t \in \mathbb{R}$ and every $\ell \in \operatorname{supp} \mu_i^\tau$, that is, for every $(t, \ell) \in \mathcal{S}_i^\tau = \mathbb{R} \times \operatorname{supp} \mu_i^\tau$,

$$C(t \mid \ell) = C_i(t \mid \ell).$$

Equivalently,

$$C(z_1 \mid z_2) = C_i(z_1 \mid z_2) \qquad \text{for every } (z_1, z_2) \in \mathcal{S}_i^\tau.$$

It remains to place realized predecessor scores in this support. For every $v \in [0,1]^n$,

$$L_{B_i^\tau}(f(v)) \in \operatorname{supp} \mu_i^\tau.$$

Indeed, P^i is the pushforward by f of the target-$\pi(i)$ intervention law; by strict positivity of the latent density on the cube and the shared diffeomorphic mixing, $f(v)$ lies in the support of P^i. The predecessor-log-ratio projection is continuous on the observed support, so every neighborhood of $L_{B_i^\tau}(f(v))$ has preimage with positive P^i-mass. Applying the support uniqueness just proved to $C_i^\star$, and then using the previous step, gives

$$C_i\big(t \mid L_{B_i^\tau}(f(v))\big) = \int_0^1 \mathbf{1}\left\{\log \frac{q_{\pi(i)}(w)}{p_{\pi(i)}\big(v^{(w)}\big)} \le t\right\} q_{\pi(i)}(w)\, dw.$$

This proves the third and fourth asserted clauses.

7. It remains to evaluate the rank at the realized target log-ratio. Fix $v \in [0,1]^n$. The latent ratio formula in Assumption 3, together with the selected log-ratio version used by the decoder, gives the pointwise log-ratio identity

$$L_i(f(v)) = \log \frac{q_{\pi(i)}(v_{\pi(i)})}{p_{\pi(i)}(v)}.$$

By definition of the rank coordinate and by the fourth clause already proved,

$$U_i(f(v)) = \int_0^1 \mathbf{1}\left\{\log \frac{q_{\pi(i)}(w)}{p_{\pi(i)}\big(v^{(w)}\big)} \le \log \frac{q_{\pi(i)}(v_{\pi(i)})}{p_{\pi(i)}(v)}\right\} q_{\pi(i)}(w)\, dw.$$

Define, for this fixed $v \in [0,1]^n$,

$$\Gamma_{i,v}(w) = \log \frac{q_{\pi(i)}(w)}{p_{\pi(i)}(v^{(w)})}, \qquad w \in [0,1].$$

For every $w \in [0,1]$, the point $v^{(w)}$ remains in $[0,1]^n$. Positivity and smoothness from Assumption 1 make $\Gamma_{i,v}$ continuous on $[0,1]$ and differentiable in the interior, and its derivative is the own-coordinate log-ratio derivative at $v^{(w)}$:

$$\Gamma'_{i,v}(w) = \partial_u \log \frac{q_{\pi(i)}(u)}{p_{\pi(i)}((v^{(w)})^{(u)})}\bigg|_{u=w}.$$

The fixed sign condition in Assumption 4 gives

$$s_{\pi(i)}\,\Gamma'_{i,v}(w) > 0, \qquad 0 < w < 1.$$

Thus $\Gamma_{i,v}$ is strictly increasing on $[0,1]$ when $s_{\pi(i)} = 1$, and strictly decreasing on $[0,1]$ when $s_{\pi(i)} = -1$. Since

$$\log \frac{q_{\pi(i)}(v_{\pi(i)})}{p_{\pi(i)}(v)} = \Gamma_{i,v}(v_{\pi(i)}),$$

we have

$$U_i(f(v)) = \int_0^1 \mathbf{1}\{\Gamma_{i,v}(w) \le \Gamma_{i,v}(v_{\pi(i)})\} q_{\pi(i)}(w)\,dw.$$

If $s_{\pi(i)} = 1$, strict increase makes the lower level set $[0, v_{\pi(i)}]$, so

$$U_i(f(v)) = \int_0^{v_{\pi(i)}} q_{\pi(i)}(w)\,dw = Q_{\pi(i)}(v_{\pi(i)}).$$

If $s_{\pi(i)} = -1$, strict decrease makes the lower level set $[v_{\pi(i)}, 1]$, and normalization of $q_{\pi(i)}$ gives

$$U_i(f(v)) = \int_{v_{\pi(i)}}^1 q_{\pi(i)}(w)\,dw = 1 - Q_{\pi(i)}(v_{\pi(i)}).$$

Thus, for every $v \in [0,1]^n$,

$$U_i(f(v)) = \begin{cases} Q_{\pi(i)}(v_{\pi(i)}), & s_{\pi(i)} = 1, \\ 1 - Q_{\pi(i)}(v_{\pi(i)}), & s_{\pi(i)} = -1. \end{cases}$$

This proves the rank-coordinate clause.

8. The argument was for an arbitrary injective topological ordering τ of H_D and an arbitrary $i \in [n]$. The five asserted clauses therefore hold for every such τ and every i.

□

The last auxiliary bound supports the simultaneous confidence graph in Theorem 4. It is a Hilbert-space empirical-mean concentration statement for bounded Gaussian-kernel feature maps, matching the kernel mean-embedding discrepancy used in maximum mean discrepancy testing [Gretton et al., 2012, Schrab et al., 2023].

Lemma 8. [`lem:bounded-rkhs-empirical-mean`] (Conditional RKHS mean bound) *Fix $n \geq 1$, a labeled DAG G on $[n]$, a mechanism θ, an observed world with environment laws $P^0, \dots, P^n$, and a sample-split experiment S. Let $\mathcal{H}_k$ be a complete real Hilbert space equipped with a measurable unit-norm feature map ϕ for the Gaussian kernel $k(a,b) = \exp(-(a-b)^2)$.*

Assume:

- ***(Conditional evaluation sampling.)*** *The experiment satisfies the conditional evaluation-sampling conditions: the ambient and environment laws are probability measures; the complete training fold is measurable; the evaluation observations are measurable, independent within each environment, distributed according to their corresponding P^e, and independent of the training fold; and the fitted ratio map is jointly measurable in the training fold and the evaluation point.*
- ***(Evaluation sizes.)*** *Each environment has at least one evaluation observation, $N_e \geq 1$ for every $e \in \{0, \dots, n\}$. Write*
$$\underline{N} = \min_{0 \leq e \leq n} N_e.$$
- ***(Familywise level.)*** *The error level satisfies $\alpha \in (0,1)$.*

For each training realization, let $\widehat{R}_i$ denote the fitted ratio obtained from the complete training fold, and define

$$\widehat{\mu}_{ei} = \frac{1}{N_e} \sum_{r=1}^{N_e} \phi\Big(\widehat{R}_i(X_{er})\Big), \qquad \mu_{ei} = \mathbb{E}_{X \sim P^e}\Big[\phi\Big(\widehat{R}_i(X)\Big)\Big].$$

Then, conditionally on the complete training fold, the simultaneous event

$$\forall i \in [n]\ \forall e \in \{0, \dots, n\}, \qquad \|\widehat{\mu}_{ei} - \mu_{ei}\|_{\mathcal{H}_k} \leq \frac{1 + \sqrt{2\log(n(n+1)/\alpha)}}{\sqrt{\underline{N}}}$$

has probability at least $1 - \alpha$. Equivalently, for every event B measurable with respect to the complete training fold,

$$(1-\alpha)\mathbb{P}(B) \leq \mathbb{P}\left(B \cap \left\{\forall i \in [n]\ \forall e \in \{0, \dots, n\},\ \|\widehat{\mu}_{ei} - \mu_{ei}\|_{\mathcal{H}_k} \leq \frac{1 + \sqrt{2\log(n(n+1)/\alpha)}}{\sqrt{\underline{N}}}\right\}\right).$$

Proof of Lemma 8. Fix an event B measurable with respect to the complete training fold. It is enough to prove the displayed lower bound for this B.

1. Set
$$\Lambda = n(n+1), \qquad \delta = \frac{\alpha}{\Lambda}.$$
Since $n \geq 1$, $\Lambda \geq 1$. Hence $0 < \delta < 1$, and
$$\frac{1}{\delta} = \frac{n(n+1)}{\alpha}.$$
For every environment e, $\underline{N} \leq N_e$ and $\underline{N} \geq 1$. Therefore
$$\begin{aligned}\frac{1}{\sqrt{N_e}} + \sqrt{\frac{2\log(1/\delta)}{N_e}} &= \frac{1 + \sqrt{2\log\{n(n+1)/\alpha\}}}{\sqrt{N_e}} \\ &\leq \frac{1 + \sqrt{2\log\{n(n+1)/\alpha\}}}{\sqrt{\underline{N}}},\end{aligned}$$

where the equality uses $1/\delta = n(n+1)/\alpha$, and the inequality uses monotonicity of the square root together with the nonnegativity of the numerator.

2. Fix $i \in [n]$, an environment $e \in \{0, \dots, n\}$, and a training realization ω_{tr}. Define

$$\psi_{i,\omega_{\mathrm{tr}}}(x) = \phi\Big(\widehat{R}_i^{\omega_{\mathrm{tr}}}(x)\Big), \qquad T_{i,e,\omega_{\mathrm{tr}}}(z) = \left\| \frac{1}{N_e} \sum_{r=1}^{N_e} \psi_{i,\omega_{\mathrm{tr}}}(z_r) - \int \psi_{i,\omega_{\mathrm{tr}}}(x)\, dP^e(x) \right\|_{\mathcal{H}_k},$$

for $z = (z_1, \dots, z_{N_e})$. The unit-norm feature-map assumption gives

$$\|\psi_{i,\omega_{\mathrm{tr}}}(x)\|_{\mathcal{H}_k} \le 1 \quad \text{for all } x.$$

Under the product law $(P^e)^{\otimes N_e}$,

$$\mathbb{E} T_{i,e,\omega_{\mathrm{tr}}} \le \big(\mathbb{E} T_{i,e,\omega_{\mathrm{tr}}}^2\big)^{1/2} \le \frac{1}{\sqrt{N_e}}.$$

Indeed, the squared centered empirical mean expands into diagonal and off-diagonal Hilbert inner products; the off-diagonal terms vanish by product independence after centering, while the diagonal terms are bounded by the unit second moment. If $z^{(r,x')}$ is obtained from z by replacing only coordinate r by x', then

$$\Big| T_{i,e,\omega_{\mathrm{tr}}}(z) - T_{i,e,\omega_{\mathrm{tr}}}(z^{(r,x')}) \Big| \le \frac{1}{N_e} \big\| \psi_{i,\omega_{\mathrm{tr}}}(z_r) - \psi_{i,\omega_{\mathrm{tr}}}(x') \big\|_{\mathcal{H}_k} \le \frac{2}{N_e}.$$

McDiarmid's bounded-difference inequality therefore yields, for

$$s_e = \sqrt{\frac{2\log(1/\delta)}{N_e}},$$

the bound

$$(P^e)^{\otimes N_e} \left\{ z : \ \frac{1}{\sqrt{N_e}} + s_e < T_{i,e,\omega_{\mathrm{tr}}}(z) \right\} \le \exp\!\left(-\frac{N_e s_e^2}{2} \right) = \delta.$$

3. Let

$$\mathcal{A}_{i,e}(\omega_{\mathrm{tr}}) = \left\{ z : \ \frac{1}{\sqrt{N_e}} + \sqrt{\frac{2\log(1/\delta)}{N_e}} < T_{i,e,\omega_{\mathrm{tr}}}(z) \right\}.$$

The preceding step gives

$$(P^e)^{\otimes N_e}\big(\mathcal{A}_{i,e}(\omega_{\mathrm{tr}})\big) \le \delta \quad \text{for every training realization } \omega_{\mathrm{tr}}.$$

The conditional evaluation-sampling assumptions give measurability of these sections, independence of the complete training fold from the evaluation vector

$$Y_e = (X_{e1}, \dots, X_{eN_e}),$$

and the law $Y_e \sim (P^e)^{\otimes N_e}$. Hence, for every training-fold event B,

$$\mathbb{P}(\{Y_e \in \mathcal{A}_{i,e}\} \cap B) = \int_B (P^e)^{\otimes N_e}\big(\mathcal{A}_{i,e}(\omega_{\mathrm{tr}})\big)\, d\mathbb{P}(\omega_{\mathrm{tr}}) \le \delta\, \mathbb{P}(B).$$

By the definitions of the empirical and fitted mean embeddings, together with the change-of-variables identity for the fitted ratio map,

$$\{Y_e \in \mathcal{A}_{i,e}\} = \left\{ \omega : \ \frac{1}{\sqrt{N_e}} + \sqrt{\frac{2\log(1/\delta)}{N_e}} < \|\widehat{\mu}_{ei} - \mu_{ei}\|_{\mathcal{H}_k} \right\}.$$

Thus

$$\mathbb{P}\left(B \cap \left\{ \frac{1}{\sqrt{N_e}} + \sqrt{\frac{2\log(1/\delta)}{N_e}} < \|\widehat{\mu}_{ei} - \mu_{ei}\|_{\mathcal{H}_k} \right\} \right) \le \delta\, \mathbb{P}(B).$$

4. Define the pairwise bad events

$$\mathsf{Bad}_{i,e} = \left\{ \omega : \ \frac{1}{\sqrt{N_e}} + \sqrt{\frac{2\log(1/\delta)}{N_e}} < \|\widehat{\mu}_{ei} - \mu_{ei}\|_{\mathcal{H}_k} \right\},$$

and set

$$\mathsf{Bad}_{\mathrm{all}} = \bigcup_{i \in [n]} \bigcup_{e=0}^{n} \mathsf{Bad}_{i,e}.$$

The union bound and the preceding pairwise estimate give

$$\mathbb{P}\big(B \cap \mathsf{Bad}_{\mathrm{all}}\big) \le \sum_{i \in [n]} \sum_{e=0}^{n} \mathbb{P}(B \cap \mathsf{Bad}_{i,e}) \le n(n+1)\delta\, \mathbb{P}(B) = \alpha\, \mathbb{P}(B).$$

5. Let

$$\mathsf{Good} = \left\{ \omega : \ \forall i \in [n]\ \forall e \in \{0, \dots, n\}, \ \|\widehat{\mu}_{ei} - \mu_{ei}\|_{\mathcal{H}_k} \le \frac{1 + \sqrt{2\log\{n(n+1)/\alpha\}}}{\sqrt{\underline{N}}} \right\}.$$

If $\omega \in B \setminus \mathsf{Bad}_{\mathrm{all}}$, then each pair (i, e) avoids $\mathsf{Bad}_{i,e}$, so the pairwise inequality from Step 1 implies $\omega \in \mathsf{Good}$. Consequently,

$$B \setminus \mathsf{Bad}_{\mathrm{all}} \subseteq B \cap \mathsf{Good}.$$

Since all measures involved are finite,

$$\begin{aligned}
\mathbb{P}(B \cap \mathsf{Good}) &\ge \mathbb{P}(B \setminus \mathsf{Bad}_{\mathrm{all}}) \\
&= \mathbb{P}(B) - \mathbb{P}(B \cap \mathsf{Bad}_{\mathrm{all}}) \\
&\ge (1 - \alpha)\mathbb{P}(B).
\end{aligned}$$

Because B was an arbitrary complete-training-fold measurable event, this is the asserted conditional probability lower bound.

□

Verification record

The formal layer accompanying this paper is a Lean 4 development, built with toolchain `leanprover/lean4:v4.33.0` against the Mathlib revision pinned in that repository. Its sources are the module tree `CausalSmith/ExactID/EID_CrlCoverratioMmdGenericity_Research/`. The theorem statements and proofs corresponding to the declaration-naming formal objects in this paper are machine-checked there, including Theorems 1 to 4 and the auxiliary statements displayed in this appendix. The certificate covers the frozen assumptions, definitions, auxiliary lemmas, and theorems recorded in the paper crosswalk; external citations and Remark 2 enter through the stated theorem-local comparisons, assumptions, and labelled future-work objects.

B Proofs of the main results

Proof of Theorem 1. Fix $s \in \{-1, 1\}^3$, and write

$$T_i^s(z) = \begin{cases} z, & s_i = 1, \\ 1 - z, & s_i = -1. \end{cases}$$

The map T_i^s is a unit-Jacobian bijection of $[0, 1]$. Put

$$h(z) = 2z - 1, \qquad \mathfrak{q}(z) = \frac{4\exp(4z)}{\exp(4) - 1}, \qquad H(z) = \frac{\exp(4z) - 1}{\exp(4) - 1} - z.$$

1. First consider the structural assertions. Since

$$\int_0^1 \mathfrak{q}(z)\, dz = 1, \qquad \int_0^1 h(z)\, dz = 0,$$

the intervention densities and the displayed child conditionals in Definitions 8 and 9 are normalized. Convexity gives

$$0 \le \frac{\exp(4x) - 1}{\exp(4) - 1} \le x, \qquad x \in [0, 1],$$

so $H(x) \in [-1, 0]$. Hence, for both $K = h$ and $K = H$,

$$\frac{9}{10} \le 1 + \frac{1}{10} K(x)h(y) \le \frac{11}{10}, \qquad x, y \in [0, 1].$$

The reflected mechanisms therefore have strictly positive normalized factors. Their factors are restrictions of smooth elementary functions, so every p_i is C^∞ on $[0, 1]^3$ and every q_i is C^∞ on $[0, 1]$.

For the unreflected child factor,

$$\partial_y \log \frac{\mathfrak{q}(y)}{1 + \frac{1}{10} K(x)h(y)} = 4 - \frac{\frac{1}{5} K(x)}{1 + \frac{1}{10} K(x)h(y)} \ge 4 - \frac{2}{9} > 0,$$

and at nodes 1 and 3 the same own-coordinate derivative equals 4. Applying T_i^s multiplies the i-th own derivative by s_i, so $s_i \partial_{v_i} \log R_i > 0$ throughout the closed cube. This proves Assumption 1 and Assumption 4 for both witnesses.

It remains to check faithfulness for the two observational laws. Define, for $x, y \in [0,1]$,

$$\chi_1(x) = h(T_1^s(x)), \qquad \chi_2(y) = h(T_2^s(y)), \qquad \kappa_1(x) = H(T_1^s(x)).$$

Reflection preserves Lebesgue measure and sends h to h or $-h$, so

$$\int_0^1 \chi_1(x)\,dx = \int_0^1 \chi_2(y)\,dy = 0, \qquad \int_0^1 \chi_1(x)^2\,dx = \int_0^1 \chi_2(y)^2\,dy = \frac{1}{3}.$$

For the sparse law, the child density is $1 + \frac{1}{10}\chi_1(x)\chi_2(y)$. Hence, for each fixed x,

$$\int_0^1 \left(1 + \frac{1}{10}\chi_1(x)\chi_2(y)\right) \chi_1(x)\,dy = \chi_1(x),$$

$$\int_0^1 \left(1 + \frac{1}{10}\chi_1(x)\chi_2(y)\right) \chi_2(y)\,dy = \frac{1}{30}\chi_1(x),$$

and

$$\int_0^1 \left(1 + \frac{1}{10}\chi_1(x)\chi_2(y)\right) \chi_1(x)\chi_2(y)\,dy = \frac{1}{30}\chi_1(x)^2.$$

After integrating over x and over the independent third coordinate, these identities give

$$\int \chi_1(v_1)\,dP^0_{\mathrm{sp}}(v) = 0, \qquad \int \chi_2(v_2)\,dP^0_{\mathrm{sp}}(v) = 0,$$

and

$$\int \chi_1(v_1)\chi_2(v_2)\,dP^0_{\mathrm{sp}}(v) = \frac{1}{90}.$$

If the first two coordinates were independent under P^0_{sp}, the last integral would be the product of the preceding two means, equal to zero; this contradicts $1/90 > 0$. Thus the edge endpoints are dependent in the sparse observational law.

For the cancellation law, put

$$\bar\kappa = \int_0^1 \kappa_1(x)\,dx, \qquad \overline{\kappa^2} = \int_0^1 \kappa_1(x)^2\,dx.$$

The function κ_1 is continuous and nonconstant: indeed $H(0) = H(1) = 0$, while

$$H(1/2) = \frac{\exp(2) - 1}{\exp(4) - 1} - \frac{1}{2} = \frac{1}{\exp(2) + 1} - \frac{1}{2} < 0.$$

Therefore

$$0 < \int_0^1 \{\kappa_1(x) - \bar\kappa\}^2\,dx = \overline{\kappa^2} - \bar\kappa^2.$$

The cancellation child density is $1 + \frac{1}{10}\kappa_1(x)\chi_2(y)$, and the same centered-coordinate calculation gives

$$\int \kappa_1(v_1)\,dP^0_{\mathrm{can}}(v) = \bar\kappa, \qquad \int \chi_2(v_2)\,dP^0_{\mathrm{can}}(v) = \frac{1}{30}\bar\kappa,$$

$$\int \kappa_1(v_1)\chi_2(v_2)\,dP^0_{\mathrm{can}}(v) = \frac{1}{30}\overline{\kappa^2}.$$

Independence of the first two coordinates under P^0_{can} would force

$$\frac{1}{30}\overline{\kappa^2} = \bar{\kappa}\cdot\frac{1}{30}\bar{\kappa},$$

contradicting $\overline{\kappa^2} - \bar{\kappa}^2 > 0$. Hence the edge endpoints are also dependent in the cancellation observational law. Finally, the factorization over $1 \to 2$ with node 3 isolated gives the local-Markov independence of nodes 1 and 3. In this three-node graph, any failed d-separation between disjoint coordinate blocks places the endpoints of the unique edge in opposite blocks; if the conditioning block contains node 3, contraction with the independence of nodes 1 and 3 reduces it to the unconditional edge independence just ruled out. Both observational laws are therefore faithful to the three-node DAG, giving Assumption 6.

2. We next prove the sparse second-moment separation. In the unreflected sparse construction define

$$A_{\mathrm{sp}}(x) = \int_0^1 \frac{\mathfrak{q}(y)^2}{1+\frac{1}{10}h(x)h(y)}\,dy.$$

The product-density calculation gives

$$\int R_2^2\,dP^0_{\mathrm{sp}} - \int R_2^2\,dP^1_{\mathrm{sp}} = \int_0^1 \{1-\mathfrak{q}(x)\}A_{\mathrm{sp}}(x)\,dx,$$

and reflection changes variables in the same integral, so the identity holds for every sign vector.

For $\tau \in [-1/10, 1/10]$, set

$$\Omega_\tau(y) = \frac{\mathfrak{q}(y)^2}{(1+\tau h(y))^2}.$$

The elementary exponential bounds

$$13 < \exp(4) < 55$$

imply

$$\frac{2}{27} < \frac{4}{\exp(4)-1}, \qquad \Omega_\tau(y) > \frac{4}{729}\frac{100}{121}.$$

Moreover

$$\frac{d}{dy}\log\Omega_\tau(y) = 8 - \frac{4\tau}{1+\tau h(y)} \geq \frac{68}{9},$$

so

$$\frac{68}{9}\frac{4}{729}\frac{100}{121} \leq \Omega_\tau'(y).$$

Integrating this derivative bound from $1/2$ to y, with the direction reversed when $y < 1/2$, gives

$$\frac{68}{9}\frac{4}{729}\frac{100}{121}\cdot 2\Big(y-\frac{1}{2}\Big)^2 \leq h(y)\{\Omega_\tau(y) - \Omega_\tau(1/2)\}.$$

After integration over $y \in [0,1]$, using $\int_0^1 h(y)\,dy = 0$ and

$$\int_0^1 2\Big(y-\frac{1}{2}\Big)^2\,dy = \frac{1}{6},$$

we obtain

$$\frac{1}{6}\frac{68}{9}\frac{4}{729}\frac{100}{121} \le \int_0^1 h(y)\Omega_\tau(y)\,dy.$$

Since

$$A'_{\rm sp}(x) = -\frac{1}{5}\int_0^1 h(y)\Omega_{h(x)/10}(y)\,dy,$$

and $H'(x) = \mathfrak{q}(x) - 1$ with $H(0) = H(1) = 0$, integration by parts yields

$$\int_0^1 \{1-\mathfrak{q}(x)\}A_{\rm sp}(x)\,dx = \int_0^1 H(x)A'_{\rm sp}(x)\,dx \ge \frac{1}{30}\frac{68}{9}\frac{4}{729}\frac{100}{121}\int_0^1 \{-H(x)\}\,dx.$$

The last integral is

$$\int_0^1 \{-H(x)\}\,dx = \frac{1}{1-\exp(-4)} - \frac{3}{4} > \frac{29}{108}.$$

Therefore

$$\int R_2^2\,dP^0_{\rm sp} - \int R_2^2\,dP^1_{\rm sp} > \frac{1}{30}\frac{68}{9}\frac{4}{729}\frac{100}{121}\frac{29}{108} = \frac{19720}{64304361} > \frac{3}{10000}.$$

3. Let $\mu_{\rm sp}$ and $\nu_{\rm sp}$ be the laws of R_2 under $P^0_{\rm sp}$ and $P^1_{\rm sp}$. The bounds above also give

$$0 < R_2 < \frac{13/3}{9/10} < 5,$$

so both laws are supported on $[0,5]$. The preceding item gives

$$\frac{3}{10000} < \int r^2\,d\mu_{\rm sp}(r) - \int r^2\,d\nu_{\rm sp}(r).$$

We now derive the quantitative Gaussian recovery estimate used on this common support. Put

$$\gamma_m = \left(\frac{2^m}{m!}\right)^{1/2}, \qquad \Gamma_m(r) = \gamma_m \exp(-r^2)r^m, \qquad \Gamma(r) = (\Gamma_m(r))_{m\ge 0}.$$

The power series for the exponential gives

$$\langle \Gamma(a), \Gamma(b)\rangle = \exp(-a^2-b^2)\sum_{m=0}^{\infty}\frac{(2ab)^m}{m!} = \exp(-(a-b)^2),$$

so Γ is the Gaussian feature map. For probability laws μ, ν supported on $[0,5]$, write

$$\delta(\mu,\nu) = \left\|\int \Gamma(r)\,d\mu(r) - \int \Gamma(r)\,d\nu(r)\right\|_{\mathcal{H}_k}.$$

Reading the m-th coordinate and bounding it by the Hilbert norm yields

$$\left|\int \exp(-r^2)r^m\,d\mu(r) - \int \exp(-r^2)r^m\,d\nu(r)\right| \le \gamma_m^{-1}\delta(\mu,\nu) = \left(\frac{m!}{2^m}\right)^{1/2}\delta(\mu,\nu).$$

Define the degree-202 truncation

$$\mathrm{Tr}_{100}(r) = \sum_{\ell=0}^{100} \frac{r^{2\ell+2}}{\ell!}.$$

For $0 \le r \le 5$, the tail of the exponential series satisfies

$$0 \le \exp(r^2) - \sum_{\ell=0}^{100} \frac{(r^2)^\ell}{\ell!} \le \frac{(r^2)^{101}}{101!} \frac{102}{102 - r^2},$$

because after the first omitted term the successive ratios are bounded by $r^2/102$. Since $r^2 \le 25$ and $102 - r^2 \ge 77$, multiplication by $r^2 \exp(-r^2)$ gives the uniform approximation bound

$$\left| r^2 - \exp(-r^2)\,\mathrm{Tr}_{100}(r) \right| \le 25 \frac{25^{101}}{101!} \frac{102}{77} =: \mathrm{rem}_{100}\,.$$

Consequently, for every such pair μ, ν,

$$\begin{aligned}\left| \int r^2\, d\mu(r) - \int r^2\, d\nu(r) \right| &\le \left| \int \exp(-r^2)\,\mathrm{Tr}_{100}(r)\, d\mu(r) - \int \exp(-r^2)\,\mathrm{Tr}_{100}(r)\, d\nu(r) \right| + 2\,\mathrm{rem}_{100} \\ &\le \delta(\mu,\nu) \sum_{\ell=0}^{100} \frac{1}{\ell!} \left(\frac{(2\ell+2)!}{2^{2\ell+2}} \right)^{1/2} + 2\,\mathrm{rem}_{100}\,.\end{aligned}$$

Since

$$(2\ell+2)! = \binom{2\ell+2}{\ell+1} ((\ell+1)!)^2,$$

the central-binomial estimate $\binom{2\ell+2}{\ell+1} \le 4^{\ell+1}$ implies

$$\frac{1}{\ell!} \left(\frac{(2\ell+2)!}{2^{2\ell+2}} \right)^{1/2} \le \frac{(\ell+1)!}{\ell!} = \ell + 1.$$

Thus

$$\sum_{\ell=0}^{100} \frac{1}{\ell!} \left(\frac{(2\ell+2)!}{2^{2\ell+2}} \right)^{1/2} \le \sum_{\ell=0}^{100} (\ell+1) = 5151.$$

The remaining rational tail comparison is

$$2\,\mathrm{rem}_{100} = 50 \frac{25^{101}}{101!} \frac{102}{77} \le \frac{1}{10^{15}}.$$

Combining the last three displays gives, for laws supported on $[0,5]$,

$$\left| \int r^2\, d\mu(r) - \int r^2\, d\nu(r) \right| \le 5151\, \delta(\mu,\nu) + \frac{1}{10^{15}}.$$

Applying this estimate to $\mu_{\mathrm{sp}}, \nu_{\mathrm{sp}}$, where $\delta(\mu_{\mathrm{sp}}, \nu_{\mathrm{sp}}) = D_{12}(W^{\mathrm{sp}})$, and using the strict moment gap above, we obtain

$$D_{12}(W^{\mathrm{sp}}) > \frac{\frac{3}{10000} - 10^{-15}}{5151} > \frac{5}{100000000}.$$

4. It remains to identify the cancellation ratio law. Let ψ be bounded and continuous. After expanding the observational and parent-interventional densities for θ^{can}, the difference of the two test-function expectations equals

$$\int_0^1 \int_0^1 \{\mathfrak{q}(x) - 1\} \left(1 + \frac{1}{10} H(x)h(y)\right) \psi\left(\frac{\mathfrak{q}(y)}{1 + \frac{1}{10}H(x)h(y)}\right) dx\, dy,$$

with the same value after the sign reflections by the changes of variables $x \mapsto T_1^s(x)$ and $y \mapsto T_2^s(y)$. For fixed y, define

$$B_y(u) = \left(1 + \frac{1}{10} uh(y)\right) \psi\left(\frac{\mathfrak{q}(y)}{1 + \frac{1}{10}uh(y)}\right), \qquad u \in [-1, 0].$$

Since $1 + \frac{1}{10}uh(y) \geq 9/10$, B_y is continuous and has an antiderivative $\mathcal{B}_y$. Using $H'(x) = \mathfrak{q}(x) - 1$ and $H(0) = H(1) = 0$,

$$\int_0^1 \{\mathfrak{q}(x) - 1\} B_y(H(x))\, dx = \mathcal{B}_y(H(1)) - \mathcal{B}_y(H(0)) = 0.$$

Fubini's theorem applies on the compact square, hence the two expectations agree for every bounded continuous ψ. Thus the law of R_2 under the observational environment equals its law under environment 1 for W^{can}.

Since D_{12} is the Hilbert norm of the difference of the Gaussian mean embeddings of these two ratio laws, equality of the laws gives

$$D_{12}(W^{\mathrm{can}}) = 0.$$

Combining the four numbered assertions gives exactly the four clauses of the theorem. □

Proof of Theorem 2. 1. First, $\mathcal{V}_{G,s,\pi} \subseteq \mathcal{U}_{G,s,\pi}$. If $\pi(j) \lessdot_G \pi(i)$, then the cover relation in the ancestral order is a direct edge:

$$\pi(j) \lessdot_G \pi(i) \quad \Longrightarrow \quad G.\mathrm{edge}(\pi(j), \pi(i)).$$

Indeed, an ancestral relation in a DAG is either a child edge or factors through an intermediate ancestor; the cover property excludes the latter. Hence, for $\theta \in \mathcal{V}_{G,s,\pi}$, Definition 5 gives

$$F_{ji}^{\mathrm{can}}(\theta; \pi) \neq 0.$$

Let $\mu^0_{\theta,\pi,i}$ and $\mu^j_{\theta,\pi,i}$ be the two ratio laws appearing in Definition 7:

$$\mu^0_{\theta,\pi,i} = \mathsf{P}^0_{\theta,\pi,i}, \qquad \mu^j_{\theta,\pi,i} = \mathsf{P}^j_{\theta,\pi,i}.$$

By Definition 3, the ratio

$$r_i(v) = \frac{q_{\pi(i)}(v_{\pi(i)})}{p_{\pi(i)}(v_{\pi(i)} \mid v_{\mathrm{pa}_G(\pi(i))})}$$

is continuous and bounded on the compact latent cube. Hence there is $\mathfrak{b}_{\theta,\pi,i} \geq 0$ such that, with

$$I_{\theta,\pi,i} = [-\mathfrak{b}_{\theta,\pi,i}, \mathfrak{b}_{\theta,\pi,i}],$$

the two ratio laws satisfy

$$\mu^0_{\theta,\pi,i}(I^c_{\theta,\pi,i}) = 0, \qquad \mu^j_{\theta,\pi,i}(I^c_{\theta,\pi,i}) = 0.$$

We use the following bounded-support Gaussian recovery fact. If finite real laws $\bar{\mu}$ and $\bar{\nu}$ are supported on the same interval $[-\mathfrak{c}, \mathfrak{c}]$, with $\mathfrak{c} \geq 0$, then

$$\int r^2\, d\bar{\mu}(r) \neq \int r^2\, d\bar{\nu}(r) \quad \Longrightarrow \quad 0 < \left\| \int \phi(r)\, d\bar{\mu}(r) - \int \phi(r)\, d\bar{\nu}(r) \right\|_H .$$

To prove the fact, use the explicit Gaussian feature coordinates

$$\psi_m(r) = \left(\frac{2^m}{m!}\right)^{1/2} e^{-r^2} r^m, \qquad m = 0, 1, 2, \ldots.$$

Equality of the two mean embeddings makes the integrals of $e^{-r^2} r^m$ equal for every m, since the coefficient multiplying this weighted monomial is strictly positive. Hence the integrals of $e^{-r^2} P(r)$ agree for every polynomial P. For any continuous test function ζ_{test} on $[-\mathfrak{c}, \mathfrak{c}]$ and any $\varepsilon > 0$, choose P_ε so that

$$\sup_{|r| \leq \mathfrak{c}} \left| \zeta_{\text{test}}(r) e^{r^2} - P_\varepsilon(r) \right| \leq \frac{\varepsilon}{\bar{\mu}(\mathbb{R}) + \bar{\nu}(\mathbb{R}) + 1}.$$

Then $e^{-r^2} P_\varepsilon(r)$ has equal integrals under $\bar{\mu}$ and $\bar{\nu}$, and the common support gives

$$\left| \int \zeta_{\text{test}}\, d\bar{\mu} - \int \zeta_{\text{test}}\, d\bar{\nu} \right| \leq \varepsilon.$$

Thus all continuous test integrals agree, so $\bar{\mu} = \bar{\nu}$, and their second moments agree. The displayed implication is the contrapositive.

For the canonical observed world generated by θ and π, the observational and interventional law formulas and the ratio formula of Assumption 3 hold directly from the product observational density, the target-$\pi(e)$ intervention density, and the identity mixing map. Therefore the two second moments of the ratio laws are

$$\int r^2\, d\mu^0_{\theta,\pi,i}(r) = \int r_i(v)^2\, dP^0_{\theta,\pi}(v), \qquad \int r^2\, d\mu^j_{\theta,\pi,i}(r) = \int r_i(v)^2\, dP^j_{\theta,\pi}(v),$$

and the canonical second-moment contrast is

$$F^{\text{can}}_{ji}(\theta; \pi) = \int r^2\, d\mu^j_{\theta,\pi,i}(r) - \int r^2\, d\mu^0_{\theta,\pi,i}(r).$$

Since this contrast is nonzero, the two second moments differ. The bounded-support recovery fact applied to $\mu^0_{\theta,\pi,i}$ and $\mu^j_{\theta,\pi,i}$ gives

$$0 < \left\| \int \phi(r)\, d\mu^0_{\theta,\pi,i}(r) - \int \phi(r)\, d\mu^j_{\theta,\pi,i}(r) \right\|_H = D_{ji}\big(W^{\text{can}}_{\theta,\pi}\big).$$

Thus $D_{ji} > 0$ for every transported ancestral cover, and $\theta \in \mathcal{U}_{G,s,\pi}$.

2. The direct-edge separated set is open and dense, and the cover-separated set is open. For a directed edge $a \to b$, put

$$\mathcal{O}^{\pi}_{ab} = \left\{ \theta \in \Theta_{G,s} : F^{\mathrm{can}}_{\pi^{-1}(a),\pi^{-1}(b)}(\theta;\pi) \neq 0 \right\}.$$

For this paragraph write $F^{\mathrm{id}}_{ab}(\theta)$ for the identity-target contrast obtained by intervening on latent coordinate a and measuring the ratio coordinate b. The canonical relabeling identity is

$$F^{\mathrm{can}}_{\pi^{-1}(a),\pi^{-1}(b)}(\theta;\pi) = F^{\mathrm{id}}_{ab}(\theta).$$

This identity-target contrast is the cube integral

$$F^{\mathrm{id}}_{ab}(\theta) = \int_{[0,1]^n} \frac{q_b(v_b)^2 \big(q_a(v_a) - p_a(v_a \mid v_{\mathrm{pa}_G(a)})\big) \prod_{\ell \neq a,b} p_\ell(v_\ell \mid v_{\mathrm{pa}_G(\ell)})}{p_b(v_b \mid v_{\mathrm{pa}_G(b)})}\, dv.$$

The relative product C^2 topology of Definition 3, strict positivity of the denominators, and compactness of the cube make this contrast continuous on $\Theta_{G,s}$; hence $\mathcal{O}^{\pi}_{ab}$ is open. For a fixed edge $a \to b$, density is obtained from the edge-specific affine path. Let $\mathcal{N}_\theta$ be a relative neighborhood of $\theta \in \Theta_{G,s}$. Since $\theta \in \Theta_{G,s}$, the positivity and smoothness conditions in Definition 3 give strictly positive normalized densities on the latent cube. The identity-target canonical observed world is formed by taking the mixing map to be the identity, taking P^0 to be the law with density p^0_θ, and taking environment e to be the law with density p^e_θ. With these canonical choices, the observational pushforward, the single-target pushforwards, the Radon–Nikodym identities, and the latent ratio formula listed in Assumption 3 are verified directly; in particular

$$R_e(v) = \frac{q_e(v_e)}{p_e(v_e \mid v_{\mathrm{pa}_G(e)})}.$$

Thus the hypotheses of Lemma 1 are met for $a \to b$. The lemma gives a parameter value on the affine path, arbitrarily close to θ in the relative C^2 topology, for which the identity-target contrast F^{id}_{ab} is nonzero. Relabeling the canonical intervention environments back by π, every $\mathcal{N}_\theta$ contains a stratum point θ' satisfying

$$\theta' \in \mathcal{N}_\theta, \qquad F^{\mathrm{can}}_{\pi^{-1}(a),\pi^{-1}(b)}(\theta';\pi) \neq 0.$$

Thus each $\mathcal{O}^{\pi}_{ab}$ is dense. Since the edge set of G is finite and

$$\mathcal{V}_{G,s,\pi} = \bigcap_{a \to b \text{ in } G} \mathcal{O}^{\pi}_{ab},$$

$\mathcal{V}_{G,s,\pi}$ is open and dense. Similarly, for each ordered pair (j,i), the map

$$\theta \longmapsto D_{ji}\big(W^{\mathrm{can}}_{\theta,\pi}\big)$$

is continuous. Define, for $v \in [0,1]^n$,

$$p^0_\theta(v) = \prod_{\ell=1}^{n} p_\ell(v_\ell \mid v_{\mathrm{pa}_G(\ell)}), \qquad p^{\pi(j)}_\theta(v) = q_{\pi(j)}(v_{\pi(j)}) \prod_{\ell \neq \pi(j)} p_\ell(v_\ell \mid v_{\mathrm{pa}_G(\ell)}),$$

and set

$$\Gamma^0_{\theta,\pi,i}(v) = p^0_\theta(v)\,\phi\!\left(\frac{q_{\pi(i)}(v_{\pi(i)})}{p_{\pi(i)}(v_{\pi(i)} \mid v_{\mathrm{pa}_G(\pi(i))})}\right),$$

$$\Gamma^j_{\theta,\pi,i}(v) = p^{\pi(j)}_\theta(v)\,\phi\!\left(\frac{q_{\pi(i)}(v_{\pi(i)})}{p_{\pi(i)}(v_{\pi(i)} \mid v_{\mathrm{pa}_G(\pi(i))})}\right).$$

The observational and interventional mean embeddings in Definition 7 have the fixed-cube representations

$$\int \phi(r)\,d\mathsf{P}^0_{\theta,\pi,i}(r) = \int_{[0,1]^n} \Gamma^0_{\theta,\pi,i}(v)\,dv, \qquad \int \phi(r)\,d\mathsf{P}^j_{\theta,\pi,i}(r) = \int_{[0,1]^n} \Gamma^j_{\theta,\pi,i}(v)\,dv.$$

Consequently

$$D_{ji}(W^{\mathrm{can}}_{\theta,\pi}) = \left\| \int_{[0,1]^n} \Gamma^0_{\theta,\pi,i}(v)\,dv - \int_{[0,1]^n} \Gamma^j_{\theta,\pi,i}(v)\,dv \right\|_H.$$

It remains to justify the displayed continuity. Fix $\theta_\star \in \Theta_{G,s}$. By strict positivity on the compact cube, there is a relative neighborhood $\mathcal{N}^{\mathrm{loc}}_{\theta_\star}$ and a number $m_{\theta_\star} > 0$ such that, for every $\eta \in \mathcal{N}^{\mathrm{loc}}_{\theta_\star}$, every $\ell \in [n]$, and all admissible arguments,

$$p^\eta_\ell(v_\ell \mid v_{\mathrm{pa}_G(\ell)}) \ge m_{\theta_\star}.$$

The coordinate maps $\eta \mapsto p^\eta_\ell$ and $\eta \mapsto q^\eta_\ell$ are continuous in the relative product C^2 topology of Definition 3, hence uniformly continuous into C^0 on this local neighborhood after shrinking it. Therefore, as $\eta \to \theta_\star$,

$$\sup_{v\in[0,1]^n} \left| \frac{q^\eta_{\pi(i)}(v_{\pi(i)})}{p^\eta_{\pi(i)}(v_{\pi(i)} \mid v_{\mathrm{pa}_G(\pi(i))})} - \frac{q^{\theta_\star}_{\pi(i)}(v_{\pi(i)})}{p^{\theta_\star}_{\pi(i)}(v_{\pi(i)} \mid v_{\mathrm{pa}_G(\pi(i))})} \right| \longrightarrow 0,$$

and likewise

$$\sup_{v\in[0,1]^n} |p^0_\eta(v) - p^0_{\theta_\star}(v)| \longrightarrow 0, \qquad \sup_{v\in[0,1]^n} |p^{\pi(j)}_\eta(v) - p^{\pi(j)}_{\theta_\star}(v)| \longrightarrow 0.$$

The ratios above stay in a common compact interval. For the Gaussian feature map, $\|\phi(x)\|_H = 1$ and

$$\|\phi(x) - \phi(y)\|_H^2 = 2 - 2\exp\{-(x-y)^2\},$$

so ϕ is uniformly continuous on that interval. Combining the preceding uniform convergences gives

$$\sup_{v\in[0,1]^n} \|\Gamma^0_{\eta,\pi,i}(v) - \Gamma^0_{\theta_\star,\pi,i}(v)\|_H \longrightarrow 0, \qquad \sup_{v\in[0,1]^n} \|\Gamma^j_{\eta,\pi,i}(v) - \Gamma^j_{\theta_\star,\pi,i}(v)\|_H \longrightarrow 0.$$

The Bochner integral over the unit cube is continuous under this uniform convergence because

$$\left\| \int_{[0,1]^n} \left(\Gamma^0_{\eta,\pi,i}(v) - \Gamma^0_{\theta_\star,\pi,i}(v)\right) dv \right\|_H \le \sup_{v\in[0,1]^n} \|\Gamma^0_{\eta,\pi,i}(v) - \Gamma^0_{\theta_\star,\pi,i}(v)\|_H,$$

and the same estimate holds for Γ^j. Taking the difference of the two mean embeddings and then the Hilbert norm proves continuity of $\theta \mapsto D_{ji}(W^{\mathrm{can}}_{\theta,\pi})$ at $\theta_\star$. Since $\theta_\star$ was arbitrary, the map is continuous on $\Theta_{G,s}$. A finite intersection of the open sets $\{D_{ji} > 0\}$, indexed by the ancestral covers of Definition 7, gives openness of $\mathcal{U}_{G,s,\pi}$.

3. Since $\mathcal{V}_{G,s,\pi}$ is dense and $\mathcal{V}_{G,s,\pi} \subseteq \mathcal{U}_{G,s,\pi}$, every nonempty relatively open subset of $\Theta_{G,s}$ meets $\mathcal{U}_{G,s,\pi}$. Hence $\mathcal{U}_{G,s,\pi}$ is dense in $\Theta_{G,s}$.

4. The complement of $\mathcal{U}_{G,s,\pi}$ has the asserted topology. Since $\mathcal{U}_{G,s,\pi}$ is open, its complement is closed. Since $\mathcal{U}_{G,s,\pi}$ is dense, the complement has empty interior; being closed with empty interior, it is nowhere dense. Every nowhere dense set is meagre, so

$$\Theta_{G,s} \setminus \mathcal{U}_{G,s,\pi}$$

is meagre, closed, and nowhere dense.

5. Finally assume G has no directed edges. The defining condition for $\mathcal{V}_{G,s,\pi}$ in Definition 5 is then vacuous. Also, if $a, b \in [n]$ and $a \lessdot_G b$, then the cover-to-edge implication gives the directed edge $a \to b$, which the no-edge hypothesis rules out. Thus

$$\text{there are no } a, b \in [n] \text{ with } a \lessdot_G b.$$

The defining condition for $\mathcal{U}_{G,s,\pi}$ in Definition 7 is vacuous. Thus

$$\mathcal{V}_{G,s,\pi} = \Theta_{G,s}, \qquad \mathcal{U}_{G,s,\pi} = \Theta_{G,s}.$$

Combining the preceding conclusions gives all conclusions. □

Proof of Theorem 3. Put $W = \mathsf{world}(\theta)$, let $\mathcal{P} = (P^0, \dots, P^n)$, and write

$$H_D = \{j \to i : j \neq i \text{ and } 0 < D_{ji}\}.$$

Throughout the proof the superscript 0 is reserved for the observational environment, while for $i \in [n]$ the notation P^i denotes the single-target intervention environment whose target is $\pi(i)$; hence $R_i = dP^i/dP^0$ always compares intervention environment i with the observational law. The assumptions in Assumptions 1 to 3 make each P^e a probability law. They also give the coherence of $\mathcal{P}$ with the observed world W, in the following concrete sense. If $\mathcal{R}_i^W$ denotes the likelihood-ratio function specified by the world W for environment i, and if R_i denotes the observed-law likelihood ratio selected from (P^0, P^i), then

$$\mathcal{R}_i^W(x) = R_i(x) \qquad \text{for } P^0\text{-almost every } x.$$

The same coherence statement records that selected rank coordinates are functions of the common law family whenever the two worlds have the same observed laws. More precisely, if W' is compatible with W, so that both worlds have the same environment-law family $\mathcal{P}$, then for every order τ, every i, and every point x in the observed support of W,

$$C_i^{W'}\left(L_i^{W'}(x) \mid L_{B_i^\tau}^{W'}(x)\right) = C_i^W\left(L_i^W(x) \mid L_{B_i^\tau}^W(x)\right),$$

where each side is the selected rank coordinate computed from the same observed law family. Separately, the pointwise ratio formula in Assumption 3, together with the continuous observed-law representative of the likelihood ratio, gives, for every i and every $v \in [0,1]^n$,

$$L_i(f(v)) = \log R_i(f(v)) = \log \frac{q_{\pi(i)}(v_{\pi(i)})}{p_{\pi(i)}(v_{\pi(i)} \mid v_{\mathrm{pa}_G(\pi(i))})}. \tag{1}$$

If $n = 0$, all indexed conclusions below are over empty finite sets and the same displayed identities are vacuous; the argument is otherwise unchanged.

The cover-separation premise is stated for the canonical identity-mixing world. For an ancestral cover $a \lessdot_G b$, Definition 7 gives

$$0 < D_{\pi^{-1}(a),\pi^{-1}(b)}(W^{\mathrm{can}}_{\theta,\pi}).$$

The transfer from the canonical world to the supplied world is needed only on ancestral covers. Applied to the displayed cover, it rewrites the positive canonical population discrepancy as the corresponding supplied-world observed-law discrepancy, using Assumption 3 for the environment pushforwards, Assumption 2 for the inverse on observed support, and the almost-sure ratio coherence established above. Hence

$$0 < D_{\pi^{-1}(a),\pi^{-1}(b)}(W) \qquad (a \lessdot_G b). \tag{2}$$

Thus every transported ancestral cover appears as an edge of H_D.

Conversely, Lemma 2 applies under Assumptions 1 to 3. Hence, whenever $j \neq i$ and $\pi(j)$ is not an ancestor of $\pi(i)$ in G, the Gaussian-kernel discrepancy of ratio i between the observational environment and intervention environment j satisfies

$$D_{ji} = 0.$$

The lemma gives this conclusion by marginalizing over the complement of the ancestral set $\mathrm{an}_G(\pi(i)) \cup \{\pi(i)\}$: the observational and intervention-j marginal laws agree on the coordinates on which R_i depends, so the two ratio pushforward laws, and therefore their Gaussian mean embeddings, coincide. We now pass from these discrepancy statements to the graph identity. Let G^π denote the environment-label graph

$$j \to i \text{ in } G^\pi \quad \Longleftrightarrow \quad \pi(j) \to \pi(i) \text{ in } G.$$

If $j \to i$ is an edge of H_D, then $D_{ji} > 0$. The zero-discrepancy conclusion above shows that $\pi(j)$ must be an ancestor of $\pi(i)$ in G; hence every edge of H_D is contained in the transitive closure of G^π. Therefore

$$\mathrm{TC}(H_D) \subseteq \mathrm{TC}(G^\pi). \tag{3}$$

For the reverse inclusion, take any ancestral relation $\pi(j) \rightsquigarrow_G \pi(i)$. In the finite ancestral order, choose a saturated chain

$$\pi(j) = a_0 \lessdot_G a_1 \lessdot_G \cdots \lessdot_G a_m = \pi(i),$$

where each step is an ancestral cover. By Equation (2), each transported pair

$$\pi^{-1}(a_{r-1}) \to \pi^{-1}(a_r), \qquad r = 1, \ldots, m,$$

is an edge of H_D. Thus j reaches i in H_D, and

$$\mathrm{TC}(G^\pi) \subseteq \mathrm{TC}(H_D). \tag{4}$$

Combining Equations (3) and (4) gives

$$\mathrm{TC}(H_D) = \mathrm{TC}(G^\pi). \tag{5}$$

The graph reconstruction statement also supplies the decoder's selected order as a topological ordering of H_D. If τ is any topological ordering of H_D and $a \in \mathrm{pa}_G(\pi(i))$, then $\pi^{-1}(a)$ reaches i in H_D by Equation (5); applying the topological-order inequalities along the path gives

$$\pi^{-1}\big(\mathrm{pa}_G(\pi(i))\big) \subseteq B_i^\tau := \{j : \tau(j) < \tau(i)\}. \tag{6}$$

Fix such an order τ. By Equation (5), τ also respects every transported latent edge. For $i \in [n]$, set

$$A_i^\tau = \{\pi(i)\} \cup \pi(B_i^\tau).$$

Let

$$\mathcal{Z}_i^\tau = \prod_{b \in B_i^\tau} [0,1],$$

with the usual one-point product when $B_i^\tau = \varnothing$. Embed $\chi \in \mathcal{Z}_i^\tau$ into the full latent cube by the zero-filled map

$$E_i^\tau(\chi)_a = \begin{cases} \chi_b, & a = \pi(b) \text{ for some } b \in B_i^\tau, \\ 0, & a \notin \pi(B_i^\tau). \end{cases}$$

Define the predecessor score map and its range by

$$\Lambda_i^\tau : \mathcal{Z}_i^\tau \to \mathbb{R}^{B_i^\tau}, \qquad \Lambda_i^\tau(\chi)_b = \log \frac{q_{\pi(b)}\big(E_i^\tau(\chi)_{\pi(b)}\big)}{p_{\pi(b)}\big(E_i^\tau(\chi)_{\pi(b)} \mid E_i^\tau(\chi)_{\mathrm{pa}_G(\pi(b))}\big)}, \qquad \mathcal{R}_i^\tau = \Lambda_i^\tau(\mathcal{Z}_i^\tau).$$

If $v \in [0,1]^n$, then $\chi_b = v_{\pi(b)}$ gives

$$\Lambda_i^\tau(\chi)_b = L_b(f(v)), \qquad b \in B_i^\tau.$$

Indeed, every parent of $\pi(b)$ has environment label preceding b, hence belongs to B_i^τ; on the own coordinate and on those parent coordinates the zero-filled point $E_i^\tau(\chi)$ agrees with v, and the local dependence of $p_{\pi(b)}$ removes all other coordinates. By Lemma 3, applied to the present order τ,

$$\Lambda_i^\tau : \mathcal{Z}_i^\tau \longrightarrow \mathcal{R}_i^\tau \quad \text{is a homeomorphism onto its image.} \tag{7}$$

In particular, $\mathcal{R}_i^\tau$ is compact and the inverse $(\Lambda_i^\tau)^{-1} : \mathcal{R}_i^\tau \to \mathcal{Z}_i^\tau$ is continuous.

Now work under the target-$\pi(i)$ environment. For $\zeta \in [0,1]^{[n]\setminus A_i^\tau}$, let $u^{v,\tau}(\zeta) \in [0,1]^n$ be the filled-in latent point defined by

$$u^{v,\tau}(\zeta)_a = \begin{cases} v_a, & a \in A_i^\tau, \\ \zeta_a, & a \notin A_i^\tau. \end{cases}$$

The marginalization identity in Lemma 4 gives, for $v \in [0,1]^n$,

$$\int_{[0,1]^{[n]\setminus A_i^\tau}} q_{\pi(i)}\big(u^{v,\tau}(\zeta)_{\pi(i)}\big) \prod_{\ell \neq \pi(i)} p_\ell\big(u^{v,\tau}(\zeta)_\ell \mid u^{v,\tau}(\zeta)_{\mathrm{pa}_G(\ell)}\big)\, d\zeta = q_{\pi(i)}(v_{\pi(i)}) \prod_{\ell \in A_i^\tau \setminus \{\pi(i)\}} p_\ell(v_\ell \mid v_{\mathrm{pa}_G(\ell)}). \tag{8}$$

The fiber version in Lemma 5 gives the same factorization after replacing $v_{\pi(i)}$ by an arbitrary scalar w. These identities are the density input for Lemma 7. Applying that result to the order τ, the law-selected conditional CDF $C_i(t \mid L_{B_i^\tau})$ used by $\mathscr{D}$ is a continuous conditional-CDF version on the observed conditional-ratio support, and there is a continuous version C such that, for all $t \in \mathbb{R}$ and $v \in [0,1]^n$,

$$C\big(t \mid L_{B_i^\tau}(f(v))\big) = C^{\mathrm{str}}_{\theta,\pi(i)}(t \mid v), \tag{9}$$

where

$$C^{\mathrm{str}}_{\theta,\pi(i)}(t \mid v) = \int_0^1 \mathbf{1}\left\{\log \frac{q_{\pi(i)}(w)}{p_{\pi(i)}(w \mid v_{\mathrm{pa}_G(\pi(i))})} \le t\right\} q_{\pi(i)}(w)\, dw. \tag{10}$$

Here $p_{\pi(i)}(w \mid v_{\mathrm{pa}_G(\pi(i))})$ is the value $p_{\pi(i)}(v^{(w)})$ of the mechanism at the point $v^{(w)}$ obtained from v by replacing coordinate $\pi(i)$ with w; the two agree because a mechanism depends only on its own coordinate and its parents. The same cited result states that every continuous conditional-CDF version agrees with the selected version on the observed conditional-ratio support and that the selected version itself satisfies the structural identity at every mixed latent point:

$$C_i\big(t \mid L_{B_i^\tau}(f(v))\big) = C^{\mathrm{str}}_{\theta,\pi(i)}(t \mid v), \qquad v \in [0,1]^n. \tag{11}$$

Evaluating Equation (11) at $t = L_i(f(v))$ and using the strict monotonicity from Assumption 4 gives

$$U_i(f(v)) = \begin{cases} Q_{\pi(i)}(v_{\pi(i)}), & s_{\pi(i)} = 1, \\ 1 - Q_{\pi(i)}(v_{\pi(i)}), & s_{\pi(i)} = -1. \end{cases} \tag{12}$$

The right-hand side is a componentwise continuous, strictly monotone transform of the latent coordinate.

We next prove the pruning characterization used by the decoder. Let

$$P_i^\pi = \pi^{-1}\big(\mathrm{pa}_G(\pi(i))\big), \qquad S_i^\tau = B_i^\tau.$$

For every A with $A \subseteq S_i^\tau$, the target assertion is

$$U_i \perp\!\!\!\perp U_{S_i^\tau \setminus A} \mid U_A \iff P_i^\pi \subseteq A. \tag{13}$$

The pruning characterization is supplied by the rank conditional-independence characterization applied at the present order. Its hypotheses are exactly the positivity, causal minimality, shared mixing, one-intervention, sign, and rank-identity clauses already established, together with the subset bookkeeping $A \subseteq S_i^\tau = B_i^\tau$. It yields

$$U_i \perp\!\!\!\perp U_{B_i^\tau \setminus A} \mid U_A \iff P_i^\pi \subseteq A. \tag{14}$$

Equation (14) says exactly that the admissible sets are the supersets of P_i^π. Indeed, by Algorithm 1, a set $A \subseteq B_i^\tau$ is admissible precisely when the left side of Equation (14) holds. Thus P_i^π is admissible by Equation (6), and if A is any admissible set then $P_i^\pi \subseteq A$. Conversely every superset of P_i^π contained in B_i^τ is admissible. Hence P_i^π is the unique inclusion-minimal admissible parent set for i, for every topological ordering τ of H_D. Applying this extraction to the selected order in Algorithm 1, the parent-pruned graph has edge relation

$$a \to i \iff a \in P_i^\pi \iff \pi(a) \in \mathrm{pa}_G(\pi(i)),$$

which is exactly

$$G^\pi_{\mathscr{D}} = G^\pi. \tag{15}$$

It remains to prove representation uniqueness. Let G', θ', and W' satisfy the same regularity, causal-minimality, sign, mixing, and one-intervention assumptions, and suppose W' has the same observed law family as W. Write $\widetilde{f}$ and $\widetilde{\pi}$ for the mixing map and target permutation of W'; these are in general different from f and π. Let υ' be the canonical topological ordering of G', and define the environment-label order

$$\tau'(e) = \upsilon'(\widetilde{\pi}(e)), \qquad e \in [n].$$

The common-order lemma for equal observed law families gives that τ' is a topological ordering of the common observable ratio graph. Since Equation (5) holds for W, the same τ' also respects the transported graph of W; by construction through υ', it respects the transported graph of W'.

For any coordinate label a, define the signed intervention distribution functions

$$S_a(u) = \begin{cases} Q_a(u), & s_a = 1, \\ 1 - Q_a(u), & s_a = -1, \end{cases} \qquad S'_a(u) = \begin{cases} Q'_a(u), & s_a = 1, \\ 1 - Q'_a(u), & s_a = -1. \end{cases}$$

Here Q'_a is the intervention distribution function for coordinate a of θ', and both signed functions are indexed by the same ambient sign vector s. Using the continuous-rank construction at this common order gives, for $v, z \in [0,1]^n$,

$$U_e(f(v)) = S_{\pi(e)}(v_{\pi(e)}), \qquad U_e(\widetilde{f}(z)) = S'_{\widetilde{\pi}(e)}(z_{\widetilde{\pi}(e)}). \tag{16}$$

Applying Equation (14) to W and to W' at the same order shows that both representations have the same unique minimal admissible parent sets. Therefore, with

$$\sigma = \widetilde{\pi} \circ \pi^{-1}, \qquad \sigma(\pi(e)) = \widetilde{\pi}(e),$$

where π and $\widetilde{\pi}$ are the target permutations of W and W', respectively, the map σ applies the inverse target permutation of W first and then the target permutation of W', and carries G-labels to G'-labels, one has, for every j, i,

$$\sigma(j) \to \sigma(i) \text{ in } G' \quad \Longleftrightarrow \quad j \to i \text{ in } G. \tag{17}$$

Equality of the observational laws gives equality of observed supports,

$$f([0,1]^n) = \widetilde{f}([0,1]^n). \tag{18}$$

It remains to assemble the common ranks, aligned edges, and common support into componentwise C^2 equivalence. We use the following implication, proved in the formal development: if two worlds have the common rank identities in Equation (16), equal observed supports, and the edge alignment in Equation (17), then there are interval maps $h_i, \bar{h}_i : [0,1] \to [0,1]$, one pair for each latent coordinate i, such that

$$h_i \in C^2([0,1]), \qquad \bar{h}_i \in C^2([0,1]), \qquad \bar{h}_i \circ h_i = \mathrm{id}_{[0,1]}, \qquad h_i \circ \bar{h}_i = \mathrm{id}_{[0,1]},$$

and, for every observed point $x \in f([0,1]^n) = \widetilde{f}([0,1]^n)$,

$$\widetilde{f}^{-1}(x)_{\sigma(i)} = h_i\big(f^{-1}(x)_i\big). \tag{19}$$

The smoothness of these charts follows by composing the one-coordinate face embedding $z \mapsto 0_n^{i\leftarrow z}$, the C^2 mixing map, the inverse mixing map on the common observed support, and the coordinate projection; the inverse charts are obtained by interchanging the two worlds. Together with Equation (17), this is exactly the componentwise C^2 equivalence of W and W' up to the simultaneous relabeling $\sigma = \widetilde{\pi} \circ \pi^{-1}$.

When $n = 2$, every edge is an ancestral cover. Therefore Equation (2) gives, for each edge $j \to i$ of G,

$$0 < D_{\pi^{-1}(j),\pi^{-1}(i)}. \tag{20}$$

The established graph equality Equation (15), the order and predecessor containment Equations (5) and (6), the rank identity Equation (12), the uniqueness statement above, and the bivariate positivity Equation (20) supply the packaged comparison clauses: within the positive compact-cube C^3 mechanism class with shared C^2 mixing, the decoder identifies the latent representation and structure from one perfect intervention per node, supplies the continuous law-separation witness used by von Kügelgen et al. [2023, Theorem 3.2] when $n = 2$, and recovers the graph, topological order, and intervention alignment relative to Wendong et al. [2023, Theorem 4.2] and Yao et al. [2025, Corollary D.1]. Collecting Equations (5), (6), (11), (12), (14) and (15), the coherence of $\mathcal{P}$ with W, the componentwise equivalence statement, Equation (20), and these comparison clauses gives every assertion of Theorem 3. □

Proof of Theorem 4. Let $\mathcal{H}_k$ be the Hilbert space of the Gaussian unit-norm feature map Φ. Write

$$\underline{N} = \min_{0\le e\le n} N_e, \qquad \delta_N(\alpha) = \frac{1 + \sqrt{2\log\{n(n+1)/\alpha\}}}{\sqrt{\underline{N}}},$$

so that

$$\varepsilon_N(\alpha,\eta) = 2\sqrt{2}\, a_N(\eta) + 2\delta_N(\alpha).$$

For $e \in \{0,\dots,n\}$, $i \in [n]$, and an outcome ω, set

$$\widehat{\mu}_{ei}(\omega) = \frac{1}{N_e}\sum_{r=1}^{N_e} \Phi\left(\widehat{R}_i(X_r^{(e)}(\omega))\right),$$

$$\widetilde{\mu}_{ei}(\omega) = \int \Phi\left(\widehat{R}_i(x)\right)\, dP^e(x), \qquad \mu_{ei} = \int \Phi(R_i(x))\, dP^e(x).$$

Define the evaluation concentration event

$$\mathcal{E}_{\mathrm{emb}} = \left\{\omega : \forall i \in [n],\ \forall e \in \{0,\dots,n\},\ \|\widehat{\mu}_{ei}(\omega) - \widetilde{\mu}_{ei}(\omega)\|_{\mathcal{H}_k} \le \delta_N(\alpha)\right\}.$$

The independent-environment sampling assumption supplies the conditional evaluation-sampling hypotheses in Lemma 8. Hence, conditionally on the training fold, $\mathcal{E}_{\mathrm{emb}}$ has probability at least $1-\alpha$: for every training-fold measurable event B,

$$(1-\alpha)\Pr(B) \le \Pr(\mathcal{E}_{\mathrm{emb}} \cap B).$$

For the bridge from fitted to population embeddings, first note that the Gaussian normalization gives, for all $a, b \in \mathbb{R}$,

$$\|\Phi(a) - \Phi(b)\|^2_{\mathcal{H}_k} = 2 - 2\exp\{-(a-b)^2\} \le 2(a-b)^2,$$

and therefore

$$\|\Phi(a) - \Phi(b)\|_{\mathcal{H}_k} \leq \sqrt{2}\,|a - b|.$$

On $\mathcal{E}_{\text{fs}}$, the defining $L^1(P^e)$ bound gives, for every i and e,

$$\begin{aligned}
\|\widetilde{\mu}_{ei}(\omega) - \mu_{ei}\|_{\mathcal{H}_k} &= \left\| \int \left\{ \Phi(\widehat{R}_i(x)) - \Phi(R_i(x)) \right\} dP^e(x) \right\|_{\mathcal{H}_k} \\
&\leq \int \|\Phi(\widehat{R}_i(x)) - \Phi(R_i(x))\|_{\mathcal{H}_k}\, dP^e(x) \\
&\leq \sqrt{2} \int |\widehat{R}_i(x) - R_i(x)|\, dP^e(x) \\
&\leq \sqrt{2}\, a_N(\eta).
\end{aligned}$$

The measurability and integrability requirements here are exactly the training-fold and finite-integral clauses in Assumption 7 and in the definition of $\mathcal{E}_{\text{fs}}$.

Now fix $\omega \in \mathcal{E}_{\text{emb}} \cap \mathcal{E}_{\text{fs}}$ and $j \neq i$. The empirical and population discrepancies are

$$\widehat{D}_{ji}(\omega) = \|\widehat{\mu}_{ji}(\omega) - \widehat{\mu}_{0i}(\omega)\|_{\mathcal{H}_k}, \qquad D_{ji} = \|\mu_{0i} - \mu_{ji}\|_{\mathcal{H}_k}.$$

By the reverse triangle inequality,

$$\begin{aligned}
|\widehat{D}_{ji}(\omega) - D_{ji}| &\leq \|(\widehat{\mu}_{ji} - \widehat{\mu}_{0i}) - (\mu_{ji} - \mu_{0i})\|_{\mathcal{H}_k} \\
&\leq \|\widehat{\mu}_{ji} - \widetilde{\mu}_{ji}\|_{\mathcal{H}_k} + \|\widetilde{\mu}_{ji} - \mu_{ji}\|_{\mathcal{H}_k} \\
&\quad + \|\widehat{\mu}_{0i} - \widetilde{\mu}_{0i}\|_{\mathcal{H}_k} + \|\widetilde{\mu}_{0i} - \mu_{0i}\|_{\mathcal{H}_k} \\
&\leq 2\delta_N(\alpha) + 2\sqrt{2}\, a_N(\eta) = \varepsilon_N(\alpha, \eta).
\end{aligned}$$

Thus

$$\mathcal{E}_{\text{emb}} \cap \mathcal{E}_{\text{fs}} \subseteq \mathcal{E}_{\text{MMD}}.$$

Applying the conditional bound for $\mathcal{E}_{\text{emb}}$ to $\mathcal{E}_{\text{fs}} \cap B$, which is training-fold measurable by Assumption 7, yields

$$(1 - \alpha) \Pr(\mathcal{E}_{\text{fs}} \cap B) \leq \Pr(\mathcal{E}_{\text{MMD}} \cap \mathcal{E}_{\text{fs}} \cap B)$$

for every training-fold measurable B. This is the stated conditional probability bound.

Taking $B = \Omega$ gives

$$(1 - \alpha) \Pr(\mathcal{E}_{\text{fs}}) \leq \Pr(\mathcal{E}_{\text{MMD}} \cap \mathcal{E}_{\text{fs}}) \leq \Pr(\mathcal{E}_{\text{MMD}}).$$

Since $\Pr(\mathcal{E}_{\text{fs}}) \geq 1 - \eta$ and $\alpha, \eta \in (0, 1)$,

$$\Pr(\mathcal{E}_{\text{MMD}}) \geq (1 - \alpha)(1 - \eta) = 1 - \alpha - \eta + \alpha\eta \geq 1 - \alpha - \eta.$$

It remains to identify the graph-theoretic consequences on $\mathcal{E}_{\text{MMD}}$. Fix $\omega \in \mathcal{E}_{\text{MMD}}$, and suppose $j \to i$ is selected by $\widehat{H}(\omega)$. Then $j \neq i$ and

$$0 < \widehat{D}_{ji}(\omega) - \varepsilon_N(\alpha, \eta).$$

If $\pi(j)$ were not an ancestor of $\pi(i)$ in G, the structural Assumptions 1 to 3 imply that the ratio coordinate R_i has the same law under P^0 and P^j:

$$(R_i)_\# P^0 = (R_i)_\# P^j.$$

Indeed, in the latent factorization, $R_i(f(v))$ depends only on $v_{\pi(i)}$ and the parents of $\pi(i)$; an intervention at a non-ancestor of $\pi(i)$ leaves the joint law of these ancestral coordinates unchanged. Therefore

$$D_{ji} = \left\| \int \Phi \, d((R_i)_{\#} P^0) - \int \Phi \, d((R_i)_{\#} P^j) \right\|_{\mathcal{H}_k} = 0.$$

The event $\mathcal{E}_{\mathrm{MMD}}$ then gives

$$\widehat{D}_{ji}(\omega) = |\widehat{D}_{ji}(\omega) - 0| \le \varepsilon_N(\alpha, \eta),$$

contradicting the selection inequality. Hence every selected arrow $j \to i$ satisfies that $\pi(j)$ is an ancestor of $\pi(i)$ in G.

Assume now the cover margin

$$2\varepsilon_N(\alpha, \eta) < D_{ji} \qquad \text{whenever } \pi(j) \lessdot_G \pi(i).$$

The soundness just proved gives the forward inclusion

$$\mathrm{TC}(\widehat{H}(\omega)) \subseteq \mathrm{TC}(G^\pi):$$

a selected edge $j \to i$ yields an ancestral chain from $\pi(j)$ to $\pi(i)$ in G, and applying π^{-1} to each edge of that chain gives a directed path from j to i in G^π.

For the reverse inclusion, let $a \lessdot_G b$ be an ancestral cover and put

$$j = \pi^{-1}(a), \qquad i = \pi^{-1}(b).$$

The cover relation implies $a \neq b$, hence $j \neq i$. By the margin and the event $\mathcal{E}_{\mathrm{MMD}}$,

$$\widehat{D}_{ji}(\omega) \ge D_{ji} - \varepsilon_N(\alpha, \eta) > \varepsilon_N(\alpha, \eta),$$

so $j \to i$ is selected by $\widehat{H}(\omega)$. Thus every transported ancestral cover of G is an edge of $\widehat{H}(\omega)$.

Finally, every strict ancestral relation in the finite ancestor order of G factors into covers:

$$a <_G b \quad \Longrightarrow \quad \exists a = a_0 \lessdot_G a_1 \lessdot_G \cdots \lessdot_G a_m = b.$$

This follows by induction on the finite interval $\{c : a \le_G c \le_G b\}$: if $a \lessdot_G b$ there is nothing to prove, while otherwise one chooses an intermediate c with $a <_G c <_G b$ and concatenates the two shorter saturated chains. Hence each edge of G^π, and therefore each path in G^π, maps to a path in $\widehat{H}(\omega)$. We obtain

$$\mathrm{TC}(G^\pi) \subseteq \mathrm{TC}(\widehat{H}(\omega)).$$

Together with the forward inclusion,

$$\mathrm{TC}(\widehat{H}(\omega)) = \mathrm{TC}(G^\pi).$$

□

References

Emre Acartürk, Burak Varıcı, Karthikeyan Shanmugam, and Ali Tajer. Sample complexity of interventional causal representation learning. In *Advances in Neural Information Processing Systems 37*. Curran Associates, Inc., 2024. doi: 10.5555/3737916.3739159.

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*, pages 372–407. PMLR, 2023.

Kartik Ahuja, Amin Mansouri, and Yixin Wang. Multi-domain causal representation learning via weak distributional invariances. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*. PMLR, 2024.

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*, pages 843–867. PMLR, 2024.

Philip Boeken, Patrick Forré, and Joris M. Mooij. Are bayesian networks typically faithful?, 2024. arXiv version 3, 2026.

Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems 35*. Curran Associates, Inc., 2022.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. In *Advances in Neural Information Processing Systems 36*, pages 45419–45462. Curran Associates, Inc., 2023.

Guangyi Chen, Yunlong Deng, Peiyuan Zhu, Yan Li, Yifan Sheng, Zijian Li, and Kun Zhang. CausalVerse: Benchmarking causal representation learning with configurable high-fidelity simulations, 2025.

Haoyue Dai, Ignavier Ng, Jianle Sun, Zeyu Tang, Gongxu Luo, Xinshuai Dong, Peter Spirtes, and Kun Zhang. When selection meets intervention: Additional complexities in causal discovery, 2025.

Jianqing Fan and Irene Gijbels. *Local Polynomial Modelling and Its Applications*. Chapman and Hall/CRC, London, 1996. ISBN 9780412983214.

Jianqing Fan, Qiwei Yao, and Howell Tong. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83(1):189–206, 1996. doi: 10.1093/biomet/83.1.189.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, 2007.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

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

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., 2016.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. Wiley, New York, 2001. ISBN 9780471405405. doi: 10.1002/0471221317.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 859–868. PMLR, 2019.

Aapo Hyvärinen, Ilyes Khemakhem, and Hiroshi Morioka. Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning. *Patterns*, 4(10):100844, 2023. doi: 10.1016/j.patter.2023.100844.

Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. In *Advances in Neural Information Processing Systems 33*, pages 9551–9561. Curran Associates, Inc., 2020.

Yibo Jiang and Bryon Aragam. Learning nonparametric latent causal graphs with unknown interventions. In *Advances in Neural Information Processing Systems 36*, pages 60468–60513. Curran Associates, Inc., 2023.

Jikai Jin and Vasilis Syrgkanis. Learning causal representations from general environments: Identifiability and intrinsic ambiguity. In *Advances in Neural Information Processing Systems 37*. Curran Associates, Inc., 2024.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2207–2217. PMLR, 2020.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Learning latent causal graphs via mixture oracles. In *Advances in Neural Information Processing Systems 34*. Curran Associates, Inc., 2021.

Pranamya Kulkarni, Puranjay Datta, Burak Varıcı, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. ROPES: Robotic pose estimation via score-based causal representation learning, 2025.

Inbeom Lee, Tongtong Jin, and Bryon Aragam. Beyond identifiability: Learning causal representations with few environments and finite samples, 2026.

Ying-Ying Lee. Partial mean processes with generated regressors: Continuous treatment effects and nonseparable models, 2018.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 13557–13603. PMLR, 2022.

Enno Mammen, Christoph Rothe, and Melanie Schienle. Nonparametric regression with nonparametrically generated covariates. *The Annals of Statistics*, 40(2):1132–1170, 2012. doi: 10.1214/12-AOS995.

Christopher Meek. Strong completeness and faithfulness in bayesian networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 411–418. Morgan Kaufmann, 1995.

Ignavier Ng, Shaoan Xie, Xinshuai Dong, Peter Spirtes, and Kun Zhang. Causal representation learning from general environments under nonparametric mixing. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 3700–3708. PMLR, 2025.

Adrian Pagan. Econometric issues in the analysis of regressions with generated regressors. *International Economic Review*, 25(1):221–247, 1984. doi: 10.2307/2648877.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2 edition, 2009. ISBN 9780521895606.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, 2017. ISBN 9780262037310.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

Antonin Schrab, Ilmun Kim, Mélisande Albert, Béatrice Laurent, Benjamin Guedj, and Arthur Gretton. MMD aggregated two-sample test. *Journal of Machine Learning Research*, 24(194): 1–81, 2023.

Alexander J. Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *Algorithmic Learning Theory*, volume 4754 of *Lecture Notes in Computer Science*, pages 13–31. Springer, 2007. doi: 10.1007/978-3-540-75225-7_5.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2 edition, 2000. ISBN 9780262284158. doi: 10.7551/mitpress/1754.001.0001.

Chandler Squires, Anna Seigal, Salil Bhate, and Caroline Uhler. Linear causal disentanglement via interventions, 2022.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(70):2389–2410, 2011.

Yuewen Sun, Lingjing Kong, Guangyi Chen, Loka Li, Gongxu Luo, Zijian Li, Yixuan Zhang, Yujia Zheng, Mengyue Yang, Petar Stojanov, Eran Segal, Eric P. Xing, and Kun Zhang. Causal representation learning from multimodal biomedical observations, 2024.

Burak Varıcı, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General identifiability and achievability for causal representation learning. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research. PMLR, 2024.

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.

Julius von Kügelgen, Michel Besserve, Liang Wendong, 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 36*, pages 48603–48638. Curran Associates, Inc., 2023. doi: 10.52202/075280-2110.

Kun Wang, Sumanth Varambally, Duncan Watson-Parris, Yi-An Ma, and Rose Yu. Discovering latent causal graphs from spatiotemporal data, 2024.

Liang Wendong, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal component analysis. In *Advances in Neural Information Processing Systems 36*. Curran Associates, Inc., 2023.

Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. *Journal of Machine Learning Research*, 2024.

Haitian Xie. Uniform convergence results for the local linear regression estimation of the conditional distribution. *Statistics and Probability Letters*, 201:109886, 2023. doi: 10.1016/j.spl.2023.109886.

Karren D. Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal DAGs under interventions. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*. PMLR, 2018.

Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. In *International Conference on Learning Representations*, 2025.

Jiaqi Zhang, Chandler Squires, Kristjan Greenewald, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. In *Advances in Neural Information Processing Systems 36*. Curran Associates, Inc., 2023.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024.

Mingxuan Zhang, Khushi Desai, Sopho Kevlishvili, and Elham Azizi. Scalable contrastive causal discovery under unknown soft interventions, 2026.