TARGET IDENTIFICATION IN OFFLINE IMITATION LEARNING: THE EFFECT OF STATE CONDITIONING

Anonymous authors
Paper under double-blind review

ABSTRACT

We study reward-free offline imitation learning with logs from a preferred and a comparison behavior source. Their differences can be used to score candidate behavior. The resulting *target* is the state–action visitation distribution that maximizes this score while remaining feasible under the environment's dynamics and close to the logged behavior. We ask whether source visitation frequencies alone determine that target when the audit does not use linked transitions. A joint score uses both where the sources go and what actions they choose there; conditioning on state removes the first kind of information. We show that this removal can make the target ambiguous: the same two source occupancies uniquely determine every joint target in a two-state example, yet permit different conditional targets under different compatible dynamics. Under positivity and when the sources do not span all feasible flow directions, we prove that a Kullback–Leibler (KL) regularized target is uniquely determined exactly when its unconstrained reweighting is a nonnegative occupancy in the affine span of the sources. For finite samples, State-Conditional Occupancy Reweighting (SCOR) returns a target estimate and a confidence ball covering every target compatible with the true source occupancies, without estimating transitions. Sampling uncertainty shrinks with more data for identified targets with bounded affine coefficients, while uncertainty caused by indistinguishable dynamics can persist. Experiments across 30 source geometries examine these claims and the value of signed affine combinations. The guarantees concern the target specified by the learning objective, not task return.

1 INTRODUCTION

Offline imitation learning often uses demonstrations from contrasting behavior sources, such as a preferred and a comparison policy, without a numerical task reward. A common approach scores state–action pairs by how much more often they occur in the preferred source and trains a policy to favor high-scoring behavior (Jang et al., 2023; Hoang et al., 2025; 2026). Such a score may be defined jointly over states and actions or conditioned on the current state. Conditioning removes state-visitation differences and isolates action choice, but raises a basic question: can removing state information also remove information needed to determine the behavior induced by the score?

Consider two navigation policies that visit different rooms and choose different actions within each room. A joint contrast reflects both which rooms a source visits and what it does there, whereas a conditional contrast compares actions only within the same room. Conditioning can isolate action choice, but visitation frequencies also constrain which behaviors are compatible with the underlying dynamics. Removing them may therefore leave several dynamics models consistent with the same sources but producing different optimized behavior.

This occurs because a score and its target are different objects. A score ranks state–action pairs, whereas an occupancy describes how often a policy visits them along a discounted trajectory (Puterman, 1994). We define the target as the achievable occupancy maximizing the source-defined score while remaining KL-regularized toward a reference source mixture. Because reweightings need not satisfy the unknown flow constraints, even a known score need not determine a unique target.

We study this question under a *source-only* interface: the auditor observes source state–action occupancies but not linked successor observations. Different transition kernels can reproduce the same sources while inducing different targets. A target is *identified* when every compatible kernel

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

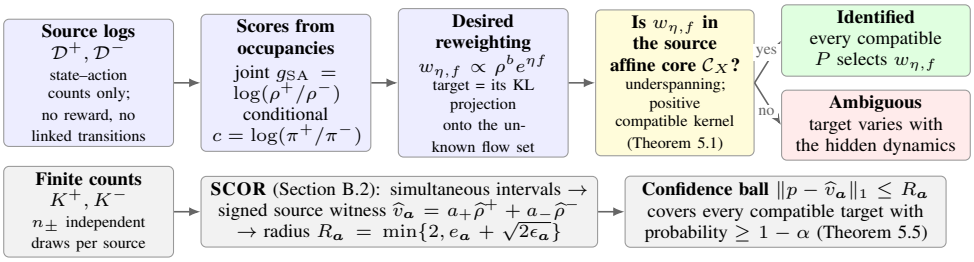


Figure 1: Overview. Source occupancies define joint and conditional scores and desired re-weightings. Under Theorem 5.1, the target is identified exactly when the reweighting lies in the source affine core. With finite counts, SCOR returns a confidence ball covering compatible targets.

selects the same optimizer. Linked transition data provide information, so our impossibility results concern this source-only setting. Figure 1 summarizes the resulting audit.

Our main result characterizes identification through the *source affine core*, the nonnegative part of the affine span of the observed source occupancies and hence a set feasible under every compatible dynamics model. Under positivity and underspanning conditions, a fixed-η target is identified exactly when its unconstrained exponential reweighting lies in this core. A positive two-state construction gives a sharp reversal: the same two sources identify the entire joint target path, while every conditional target at $\eta > 0$ depends on the hidden dynamics. Thus conditioning can remove precisely the information that made the joint target identifiable. For finite data, State-Conditional Occupancy Reweighting (SCOR) constructs a signed source-affine center and an L_1 confidence radius covering every compatible target without estimating transitions. For identified targets with source-affine representations within the allowed coefficient budget, the minimum radius contracts at rate $n^{-1/2}$; when indistinguishable dynamics imply different targets, a nonvanishing structural ambiguity remains. The same procedure audits either the joint or conditional score. Our contributions are:

- **Target identification.** We give a necessary-and-sufficient source-only criterion and a same-data construction in which conditioning destroys identification.
- **Finite-data confidence.** We construct source-affine balls covering all compatible targets and prove when sampling uncertainty shrinks or structural ambiguity persists.
- **Controlled evaluation.** Across 30 source geometries, we test the identification reversal and evaluate finite-source confidence across dimension, source overlap, and restrictions on affine source combinations.

2 Related Work

Prior work uses contrasting desirable and undesirable demonstrations for occupancy correction or separation (Jang et al., 2023; Hoang et al., 2025; 2026; 2024), and conditional or policy-ratio signals for imitation (Han et al., 2020; Watson et al., 2023; Gong et al., 2025; Jaegle et al., 2021). We ask what target these source frequencies determine when the transition kernel is unobserved. This is a target-specific data-informativity question (van Waarde et al., 2020). Reward-equivalence work addresses which scores induce the same behavior within a fixed process (Cao et al., 2021; Gleave et al., 2021; Skalse et al., 2024); our results concern target uncertainty across compatible processes. Schlaginhaufen & Kamgarpour (2023, Theorem 4.12) study when rewards recovered by IRL agree on optimal behavior under new dynamics. Here a fixed score must select the same target across dynamics that preserve every source occupancy. The source-preserving perturbations yield our exact criterion. Proto successor measures learn an affine behavior basis from transition triples (Agarwal et al., 2025); our affine core uses the supplied occupancies. Causal partial-identification methods address hidden confounding and expert performance (Ruan et al., 2024); we identify a declared KL target from source frequencies.

3 PRELIMINARIES

We describe policies by state–action visitation frequencies and then define the two source-derived scores. The positive and negative labels identify behavior sources, not numerical rewards.

MDP representation. We model the environment at the level of the decisions made by the observed sources: the index k advances from one decision to the next, and any intervening actions by a shared controller are absorbed into the transition between decisions. The environment is a finite Markov decision process (MDP) $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, \mu_0, \gamma \rangle$, where $\mathcal{S}$ is the decision-state space with $S = |\mathcal{S}|$, $\mathcal{A}$ is the action space, $P(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})$ is the next-decision transition kernel, μ_0 is the initial-state distribution, and $\gamma \in (0, 1)$ is the decision-level discount factor (Puterman, 1994). Here $\mathbf{s}$ and $\mathbf{a}$ denote a decision state and action. Let $\mathcal{E} = \mathcal{S} \times \mathcal{A}$ be the state–action space, $E = |\mathcal{E}|$, and write $e = (\mathbf{s}, \mathbf{a}) \in \mathcal{E}$ for a state–action pair. A policy $\pi(\mathbf{a} \mid \mathbf{s})$ induces the normalized discounted occupancy

$$\rho^\pi(\mathbf{s}, \mathbf{a}) = (1 - \gamma) \sum\nolimits_{k \geq 0} \gamma^k \Pr_\pi(\mathbf{s}_k = \mathbf{s}, \mathbf{a}_k = \mathbf{a}), \qquad d^\pi(\mathbf{s}) = \sum\nolimits_{\mathbf{a} \in \mathcal{A}} \rho^\pi(\mathbf{s}, \mathbf{a}). \tag{1}$$

A valid occupancy must satisfy *Bellman-flow equations*:

$$\sum\nolimits_{\mathbf{a} \in \mathcal{A}} \rho(\mathbf{s}, \mathbf{a}) = (1 - \gamma)\mu_0(\mathbf{s}) + \gamma \sum\nolimits_{\bar{\mathbf{s}}, \bar{\mathbf{a}}} P(\mathbf{s} \mid \bar{\mathbf{s}}, \bar{\mathbf{a}})\rho(\bar{\mathbf{s}}, \bar{\mathbf{a}}), \tag{2}$$

These equations distinguish realizable occupancies from arbitrary probability distributions. For probability vectors p, q on a common finite support, $D_{\mathrm{KL}}(p\|q) = \sum_e p(e) \log[p(e)/q(e)]$ and $\|p - q\|_1 = \sum_e |p(e) - q(e)|$.

Demonstration data. We consider two unpaired trajectory datasets, $\mathcal{D}^+$ from the source labeled positive or preferred and $\mathcal{D}^-$ from the comparison source. The labels indicate which behavior generated a trajectory; they do not encode observed numerical rewards. We retain genuine source decision events and represent them by a sufficient decision state, yielding the shared decision-level kernel P. The sources share the dynamics, initial law, recording mechanism, and decision-state representation. The primary audit uses only the resulting state–action occupancies, not linked successor observations. The identification result treats the population occupancies as known; the finite-data procedure receives independent state–action draws from each occupancy.

Source occupancies and conditional evidence. Let $\rho^y(\mathbf{s}, \mathbf{a}) = d^y(\mathbf{s})\pi^y(\mathbf{a} \mid \mathbf{s}),\ y \in \{+, -\}$, denote the source occupancies. Here $d^y(\mathbf{s}) = \sum_{\mathbf{a} \in \mathcal{A}} \rho^y(\mathbf{s}, \mathbf{a})$ is the decision-state marginal of source y, and $\pi^y(\mathbf{a} \mid \mathbf{s}) = \rho^y(\mathbf{s}, \mathbf{a})/d^y(\mathbf{s})$ is its conditional action distribution. Let $\rho^b = \lambda_b \rho^+ + (1 - \lambda_b)\rho^-,\ \lambda_b \in [0, 1]$, be a reference occupancy, with mixture weight λ_b. We assume both source occupancies are strictly positive on $\mathcal{E}$ and define

$$g_{\mathrm{SA}}(\mathbf{s}, \mathbf{a}) = \log \frac{\rho^+(\mathbf{s}, \mathbf{a})}{\rho^-(\mathbf{s}, \mathbf{a})}, \qquad g_{\mathrm{S}}(\mathbf{s}) = \log \frac{d^+(\mathbf{s})}{d^-(\mathbf{s})}, \qquad c(\mathbf{s}, \mathbf{a}) = g_{\mathrm{SA}}(\mathbf{s}, \mathbf{a}) - g_{\mathrm{S}}(\mathbf{s}). \tag{3}$$

Since $\rho^y = d^y \pi^y$, $c = \log(\pi^+/\pi^-)$ on $\mathcal{E}$. Thus c isolates action-choice differences at a fixed state, whereas g_{SA} also contains state-visitation differences. These are the two source-derived scores whose targets we audit; the reference ρ^b anchors the regularization in Equation (4).

4 FROM SOURCE LOGS TO AN OCCUPANCY TARGET

The audit has three steps: define the score-preferred behavior, characterize dynamics consistent with the sources, and ask whether they imply the same target. Comparing joint and conditional scores then reveals whether removing state evidence changes what the sources identify.

Let $B \in \mathbb{R}^{S \times E}$ map a state–action occupancy to its state marginal: $B_{s,e} = 1$ when coordinate e belongs to state s, and zero otherwise. We use $P \in \mathbb{R}^{E \times S}$ for the row-stochastic matrix whose entry $P_{e,s'}$ is the probability of next state s' from coordinate e. For a candidate transition kernel P, define

$$A_P = B - \gamma P^\top, \qquad b = (1 - \gamma)\mu_0, \qquad \mathcal{C}_P = \{\rho \in \mathbb{R}^E_+ : A_P \rho = b\}.$$

Thus $\mathcal{C}_P$ is the set of state–action occupancies that satisfy the Bellman-flow equations under P. For a strictly positive reference occupancy ρ^b, a real-valued score vector $f \in \mathbb{R}^E$, and $\eta > 0$, we define

$$\mathcal{T}^P_\eta(f) = \arg\max\nolimits_{\rho \in \mathcal{C}_P} \left\{ \mathbb{E}_\rho[f] - \eta^{-1} D_{\mathrm{KL}}(\rho\|\rho^b) \right\}, \qquad \mathbb{E}_\rho[f] = \sum\nolimits_{e \in \mathcal{E}} \rho(e) f(e). \tag{4}$$

The first term favors high-scoring behavior; the second penalizes departures from the reference occupancy ρ^b. The constraint $\rho \in \mathcal{C}_P$ requires that a policy can realize this behavior under P. Their optimizer $\mathcal{T}_\eta^P(f)$ is the *target occupancy*. Larger η gives the score more influence relative to the penalty. The score f is a surrogate signal from the sources, so this optimizer is not necessarily optimal for an unobserved task reward. When P is fixed and clear from context, we write $\mathcal{T}_\eta(f)$ and $\mathcal{C}$.

Source-only information and compatible dynamics. Suppose the source occupancies are known exactly. Let $X = [\rho_0, \dots, \rho_m]$, $Z = [\rho_1 - \rho_0, \dots, \rho_m - \rho_0]$ collect the supplied positive occupancies and their differences; in the two-source setting, $X = [\rho^-, \rho^+]$. With known μ_0 and γ, define

$$\begin{aligned} \mathcal{K}_X &= \left\{ P \geq 0 : P\mathbf{1} = \mathbf{1},\ (B - \gamma P^\top)X = b\mathbf{1}^\top \right\}, \\ \mathcal{C}_X &= (\rho_0 + \operatorname{range} Z) \cap \mathbb{R}_+^E. \end{aligned} \tag{5}$$

Thus $\mathcal{K}_X$ contains all kernels compatible with the supplied occupancies, while the *source affine core* $\mathcal{C}_X$ is their nonnegative affine span and is feasible under every compatible kernel. For two sources,

$$\mathcal{C}_X = \{u\rho^+ + (1-u)\rho^- : u \in \mathbb{R},\ u\rho^+ + (1-u)\rho^- \geq 0\},$$

allowing signed extrapolation whenever the resulting occupancy is nonnegative. For fixed f and η, the target is *point identified* if $\{\mathcal{T}_\eta^P(f) : P \in \mathcal{K}_X\}$ is a singleton, and partially identified otherwise. This is weaker than identifying the dynamics.

Comparing regularization strengths. Finally, we also compare scores at equal excess-KL radius to remove arbitrary positive scaling; the formal definition is in Section A.

5 Theoretical Analysis

The analysis answers three questions about the target occupancy that optimizes the regularized score in Equation (4): *when do exact source frequencies determine it, can conditioning change that answer, and how much uncertainty remains when the frequencies must be estimated?*

5.1 When Do the Source Frequencies Determine the Target?

Why target-specific identification is possible. A useful special case explains the main result. First ignore the flow constraints and compute the probability distribution preferred by the score. If this distribution lies in the source affine core $\mathcal{C}_X$, it is feasible under every compatible kernel. Imposing the unknown dynamics then cannot change the optimizer. The theorem shows that, under its assumptions, this sufficient condition is also necessary.

The unconstrained tilt. For a finite score f, strictly positive occupancy ρ^b, and $\eta > 0$, define

$$w_{\eta,f}(e) = \frac{\rho^b(e)\exp\{\eta f(e)\}}{\sum_{e' \in \mathcal{E}} \rho^b(e')\exp\{\eta f(e')\}}. \tag{6}$$

This *unconstrained exponential tilt*, or desired reweighting, depends only on the score and reference. The actual target is its closest feasible occupancy in KL divergence:

$$\mathcal{T}_\eta^P(f) = \arg\min_{\rho \in \mathcal{C}_P} D_{\mathrm{KL}}(\rho \| w_{\eta,f}).$$

Thus all dependence on the unknown dynamics enters through the feasible set $\mathcal{C}_P$.

For $0 < \gamma < 1$, the Bellman-flow affine space has dimension $E - S$. We call $\operatorname{rank} Z < E - S$ *underspanning*: the supplied sources do not span all flow-feasible directions. The theorem also assumes a compatible kernel with positive transition probabilities everywhere, allowing small source-preserving changes to the dynamics.

Theorem 5.1 (Source-only target identification)**.** *Suppose $S \geq 2$, $0 < \gamma < 1$, $\operatorname{rank} Z < E - S$, and $\mathcal{K}_X$ contains a kernel with $P(\boldsymbol{s}' \mid \boldsymbol{s}, \boldsymbol{a}) > 0$ for all $(\boldsymbol{s}, \boldsymbol{a}) \in \mathcal{E}$ and $\boldsymbol{s}' \in \mathcal{S}$. Then, for any finite score f, strictly positive reference occupancy ρ^b, and $\eta > 0$,*

$$\mathcal{T}_\eta^P(f) \text{ is point identified over } P \in \mathcal{K}_X \quad \Longleftrightarrow \quad w_{\eta,f} \in \mathcal{C}_X.$$

When identified, $\mathcal{T}_\eta^P(f) = w_{\eta,f}$. If also $\rho^b \in \mathcal{C}_X$ and $z^\top f \neq 0$ for some $z \in \operatorname{range} Z$, the same criterion holds at any common attainable radius $\delta > 0$, using the unique tilt with $D_{\mathrm{KL}}(w_{\eta,f} \| \rho^b) = \delta$. If no finite $\eta > 0$ gives an unconstrained tilt at δ, the matched-radius target is not point identified.

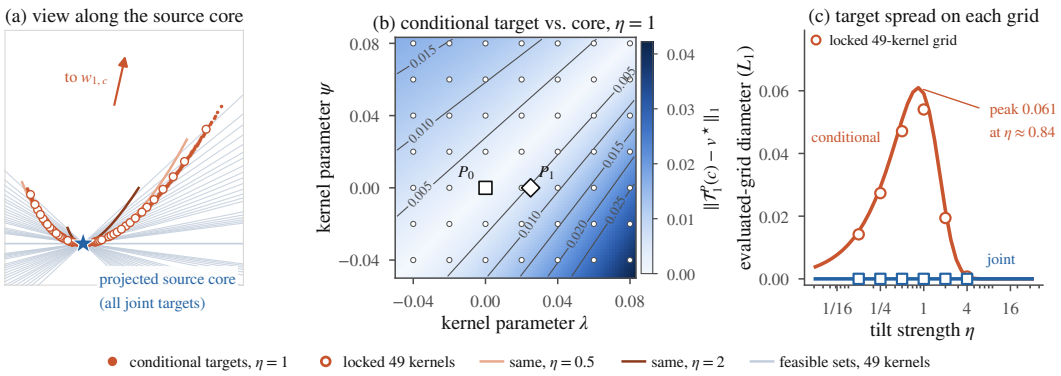


Figure 2: Conditioning breaks identification in the two-state example (exact population computations, not confidence regions). (a) Coordinates orthogonal to the source core: the core ($\rho^{\pm}$, ρ^b, and every joint target) collapses to the star, and each compatible feasible set becomes a line through it; conditional targets over a 41×41 kernel grid trace the orange curves (circles: the 49 locked kernels at $\eta = 1$), while the unconstrained conditional tilt lies far off the core along the arrow. (b) L_1 distance from the conditional target at $\eta = 1$ to the source-only center $v^\star = \arg\min_{v \in \mathcal{C}_X} D_{\mathrm{KL}}(v \| w_{1,c})$ on an 81×81 grid of compatible kernels $P_{\lambda,\psi}$ (Section C.1): largest observed distance 0.042, versus the uniform bound $\sqrt{2\epsilon_{v^\star}} = 0.246$; the joint map is identically zero. (c) Evaluated-grid target diameter versus η (curve: 21×21 kernels; markers: the locked 49-kernel grid).

The criterion asks whether the desired reweighting lies in the sources' affine span, allowing signed coefficients but requiring nonnegative probabilities. If so, every compatible dynamics model selects the same target; otherwise, some compatible models select different targets despite agreeing on all supplied source occupancies. The necessity proof perturbs the transition kernel while preserving every source occupancy, showing that a target outside the criterion cannot remain optimal under all such perturbations. This source-preserving construction differs from unrestricted kernel changes (Schlaginhaufen & Kamgarpour, 2023, Theorem 4.12). See Section B.1 for the detailed proof.

Exact identification is a yes-or-no property, but nearly identified targets can still be useful. Choose a distribution v in the source core that is close to the desired reweighting. Since v is feasible under every compatible kernel, it bounds how far any compatible target can lie from it. For any $v \in \mathcal{C}_X$, let $\epsilon_v = D_{\mathrm{KL}}(v \| w_{\eta,f})$. Then every compatible target satisfies

$$\|\mathcal{T}_\eta^P(f) - v\|_1 \le \sqrt{2\epsilon_v}, \qquad F_{\eta,f}(\mathcal{T}_\eta^P(f)) - F_{\eta,f}(v) \le \epsilon_v/\eta, \tag{7}$$

where $F_{\eta,f}$ is the objective in Equation (4). Minimizing ϵ_v over $\mathcal{C}_X$ gives the tightest bound of this form.

5.2 How Conditioning Can Make the Target Ambiguous

We now apply the criterion to the two scores from the same sources. In the example below, every joint reweighting lies in the source core, while every conditional reweighting at positive strength leaves it. Thus removing the state term changes whether the target is determined without the dynamics. Consider two states and two actions, $\gamma = 1/2$, uniform μ_0, and state–action order $(1, L), (1, R), (2, L), (2, R)$:

$$\rho^- = (5, 9, 3, 15)/32, \qquad \rho^+ = (15, 3, 9, 5)/32, \qquad \rho^b = (5, 3, 3, 5)/16. \tag{8}$$

Writing each kernel as the four probabilities of transitioning to state 1 in the displayed state–action order, the strictly positive kernels $P_0 = (3/4, 1/4, 3/4, 1/4)$ and $P_1 = (33/40, 1/4, 5/8, 1/4)$ preserve the same sources but induce different feasible sets. Their source core is

$$\mathcal{C}_X = \{(5h, 3(1-h), 3h, 5(1-h))/8 : 0 \le h \le 1\}. \tag{9}$$

The joint score $g_{\mathrm{SA}} = \log(3)(1, -1, 1, -1)$ has a tilt in the core for all $\eta > 0$, whereas the conditional tilt satisfies $w_{\eta,c}(1)/w_{\eta,c}(3) = (5/3)(7/9)^{2\eta} \neq 5/3$ and therefore leaves the core.

For two distinct strictly positive sources with midpoint reference, the entire joint-tilt path lies in their affine core exactly when the source likelihood ratio takes exactly two values. Under the hypotheses of Theorem 5.1, this also characterizes identification of the joint-target path; see Section B.1.

Corollary 5.2 (Same sources, different target identification)**.** *For Equation* (8)*, every joint target at $\eta > 0$ is point identified, whereas no conditional target at $\eta > 0$ is. The same distinction holds at every matched radius $0 < \delta < \log 2$. At $\eta = 1$, the identified joint target is $(45, 3, 27, 5)/80 = (13/10)\rho^+ - (3/10)\rho^-$, outside the source convex hull.*

An auditor sees the same sources under P_0 and P_1 and therefore computes the same joint and conditional scores. Yet the conditional target can vary with the hidden dynamics, while the joint target remains fixed because its desired reweighting is feasible under every compatible kernel. Even infinitely many additional samples from these occupancies cannot resolve the ambiguity. Figure 2 illustrates the geometry: every compatible feasible set passes through the source core, where the joint target is fixed, while conditional targets vary across compatible kernels. Using the comparison source as reference gives the same reversal at unit strength whenever the sources have different state marginals and satisfy the identification assumptions; it does not require the two-level likelihood ratio of Equation (8).

Corollary 5.3 (Reversal under the comparison reference)**.** *Let $\rho^b = \rho^-$ and $\eta = 1$, and let $\rho^+ \neq \rho^-$ be strictly positive. Then $w_{1,g_{\mathrm{SA}}} = \rho^+ \in \mathcal{C}_X$, so the joint target is point identified and equals ρ^+ under every compatible kernel. The conditional tilt is $w_{1,c}(\boldsymbol{s}, \boldsymbol{a}) = d^-(\boldsymbol{s})\pi^+(\boldsymbol{a} \mid \boldsymbol{s})$. If $d^+ \neq d^-$, it lies outside $\mathcal{C}_X$, so under the hypotheses of Theorem 5.1 the conditional target is not point identified.*

The joint audit asks for the preferred source's own behavior, which every compatible model realizes. The conditional audit asks for the preferred action choices at the comparison source's state frequencies. Under any compatible kernel, the preferred policy has state marginal d^+, so it cannot realize this combination when $d^+ \neq d^-$. The proof is in Section B.1. The regularizer also affects target identification. Occupancy KL separates into state-marginal and conditional-policy terms:

$$D_{\mathrm{KL}}(\rho \| \rho^-) = D_{\mathrm{KL}}(d^\rho \| d^-) + \sum\nolimits_{\mathbf{s}} d^\rho(\mathbf{s}) D_{\mathrm{KL}}(\pi^\rho(\cdot \mid \mathbf{s}) \| \pi^-(\cdot \mid \mathbf{s})).$$

Keeping only the second term gives the log-policy-ratio construction used in coherent imitation (Watson et al., 2023), with the following target.

Proposition 5.4 (Target under conditional policy regularization)**.** *Let $\rho^b = \rho^-$ and $\eta = 1$, with strictly positive sources, and write d^ρ and π^ρ for the marginal and conditional of ρ, using any conditional at zero-mass states. Replace the divergence in Equation* (4) *by $\sum_{\boldsymbol{s}} d^\rho(\boldsymbol{s}) D_{\mathrm{KL}}(\pi^\rho(\cdot \mid \boldsymbol{s}) \| \pi^-(\cdot \mid \boldsymbol{s}))$. For the conditional score c, the objective becomes $-\sum_{\boldsymbol{s}} d^\rho(\boldsymbol{s}) D_{\mathrm{KL}}(\pi^\rho(\cdot \mid \boldsymbol{s}) \| \pi^+(\cdot \mid \boldsymbol{s}))$, whose unique maximizer over $\mathcal{C}_P$ is ρ^+ for every compatible P.*

At this reference and strength, conditional policy regularization selects the preferred policy and its occupancy under every compatible kernel. Occupancy KL penalizes changes in state visitation relative to d^-. For that objective, Corollary 5.3 shows the conditional target is ambiguous when state marginals differ and the identification assumptions hold. Our audit concerns occupancy-KL objectives (Jang et al., 2023; Hoang et al., 2026); Theorem 5.1 characterizes their targets. The comparison shows identification depends jointly on the score, reference, and regularizer. Section B.1 provides the proof.

5.3 What Can We Report from Finite Source Counts?

With finite logs, exact affine membership is not a stable test. Instead, SCOR reports a center $\widehat{v}$ and radius R covering every target compatible with the true sources; radius 2 gives no localization. A *source-affine witness* is a source combination that remains a probability distribution, and the radius accounts for its estimation error and mismatch from the desired reweighting.

For the finite-sample result, fix the midpoint reference $\lambda_b = 1/2$. Observe n_y independent draws from each strictly positive source occupancy ρ^y, where each draw returns a state–action pair e with probability $\rho^y(e)$, and write $\widehat{\rho}^y$ for the empirical frequency vector. Equivalently, one may sample one decision index k per independent rollout with probability $(1-\gamma)\gamma^k$; pooling consecutive records from the same trajectory does not satisfy the stated independence assumption. Fix $\eta > 0$, failure probability $\alpha \in (0, 1)$, the true midpoint reference $\rho^b = (\rho^+ + \rho^-)/2$, and either true-source score $f \in \{g_{\mathrm{SA}}, c\}$. The target $\mathcal{T}_\eta^P(f)$ is defined by these true source-derived quantities, not by treating the empirical score as exact. We construct simultaneous probability intervals for source occupancies and conditional action frequencies, then propagate them to the desired reweighting. Let $[L_e^y, U_e^y]$

enclose source y and $[L_e^w, U_e^w]$ enclose the true-source tilt. The $4E$ binomial intervals share error level $\alpha/(4E)$; their construction is in Section B.2.

Computing the confidence ball. For $\boldsymbol{a} = (u, 1-u)$, $u \in [-1, 2]$, let $v_{\boldsymbol{a}} = \sum_y a_y \rho^y$ and $\widehat{v}_{\boldsymbol{a}} = \sum_y a_y \widehat{\rho}^y$. With $a_y^+ = \max(a_y, 0)$ and $a_y^- = \max(-a_y, 0)$, define

$$L^v_{\boldsymbol{a},e} = \sum\nolimits_y (a_y^+ L_e^y - a_y^- U_e^y), \qquad U^v_{\boldsymbol{a},e} = \sum\nolimits_y (a_y^+ U_e^y - a_y^- L_e^y),$$
$$e_{\boldsymbol{a}} = \sum\nolimits_y |a_y| \sum\nolimits_e \max\{\widehat{\rho}_e^y - L_e^y, U_e^y - \widehat{\rho}_e^y\},$$
$$r_{\boldsymbol{a},e} = \max\{U^v_{\boldsymbol{a},e} - L_e^w, U_e^w - L^v_{\boldsymbol{a},e}, 0\}, \qquad \epsilon_{\boldsymbol{a}} = \sum\nolimits_e r^2_{\boldsymbol{a},e}/L_e^w.$$

Select the checked candidate with $L^v_{\boldsymbol{a}} \geq 0$ and smallest evaluated $e_{\boldsymbol{a}} + \sqrt{2\epsilon_{\boldsymbol{a}}}$; if support or numerical evaluation is unavailable, return the empirical midpoint with radius 2.

Theorem 5.5 (Source-only target confidence)**.** *Under the assumptions above, with probability at least $1-\alpha$, simultaneously for every data-selected coefficient vector satisfying $\sum_y a_y = 1$ and $L^v_{\boldsymbol{a}} \geq 0$ componentwise, and every kernel $P \in \mathcal{K}_X$ compatible with the true source occupancies,*

$$\|\mathcal{T}_\eta^P(f) - \widehat{v}_{\boldsymbol{a}}\|_1 \leq R_{\boldsymbol{a}} := \min\{2, e_{\boldsymbol{a}} + \sqrt{2\epsilon_{\boldsymbol{a}}}\}. \tag{10}$$

Here $L^w > 0$ is required componentwise for the displayed budget. When support or numerical enclosures are unavailable, the procedure returns the empirical midpoint with the trivial radius 2.

The count-only procedure chooses a certified source combination to minimize $e_{\boldsymbol{a}} + \sqrt{2\epsilon_{\boldsymbol{a}}}$; its steps and numerical controls are in Section B.2. The theorem covers every dynamics model consistent with the sources. The radius adds the error in estimating a feasible source combination to its mismatch from the desired reweighting. Coverage holds for every admissible coefficient vector at once, so coefficients may be chosen from the same counts. Signed coefficients matter: an identified target can lie outside the convex hull of the sources. The reported center need not obey unknown flow equations. Appendix B.3 also bounds the occupancy induced by a policy normalized from that center. The next result separates finite-sampling uncertainty from missing dynamical information.

Proposition 5.6 (Statistical consequence of conditioning)**.** *Fix $0 < \alpha < 1/2$, strictly positive source occupancies, finite dimension, and $\eta > 0$. If $w_{\eta,f} = \sum_y a_y^* \rho^y$ with coefficients in the allowed budget, the minimum certified radius is locally $O_{\Pr}(n_{\min}^{-1/2})$ at fixed confidence, where $n_{\min} = \min_y n_y$, up to optimization error. This rate cannot be uniformly improved over a full-support two-source MDP subfamily. For the sources in Equation* (8)*, two compatible conditional targets are separated by $D_\eta > 0$. Hence any confidence set with the stated coverage under both models has L_1 diameter at least D_η with probability at least $1 - 2\alpha$, for every sample size.*

Sampling uncertainty versus structural ambiguity. For identified targets with coefficients within the allowed budget, sampling uncertainty shrinks at the inverse-square-root rate. In the conditional reversal, observationally identical dynamics imply different targets, producing a nonvanishing *structural ambiguity floor*. Linked transitions can restrict the compatible kernel class, while additional sources can enlarge the affine core. The upper-rate constants depend on support, η, and coefficient size. Signed extrapolation can be necessary—at $\eta = 1$, the identified joint target $1.3\rho^+ - 0.3\rho^-$ lies outside the source convex hull—but larger coefficients increase sampling error. A radius of 2 means that the procedure gives no nontrivial localization at that sample size, not that the target is unidentified.

6 Experiments

We test the identification reversal by varying hidden dynamics with source occupancies fixed, then evaluate confidence radii and coverage from finite counts. Dynamics generate data and oracle targets only; SCOR receives source counts and state–action labels. Detailed results are in Appendix C.

Exact Identification Reversal We build 49 strictly positive kernels sharing the source occupancies in Eq. 8. For each, we compute joint and conditional targets at six strengths and six matched radii, then measure the maximum cross-kernel L_1 distance, or target-set *diameter*. The joint diameter stays

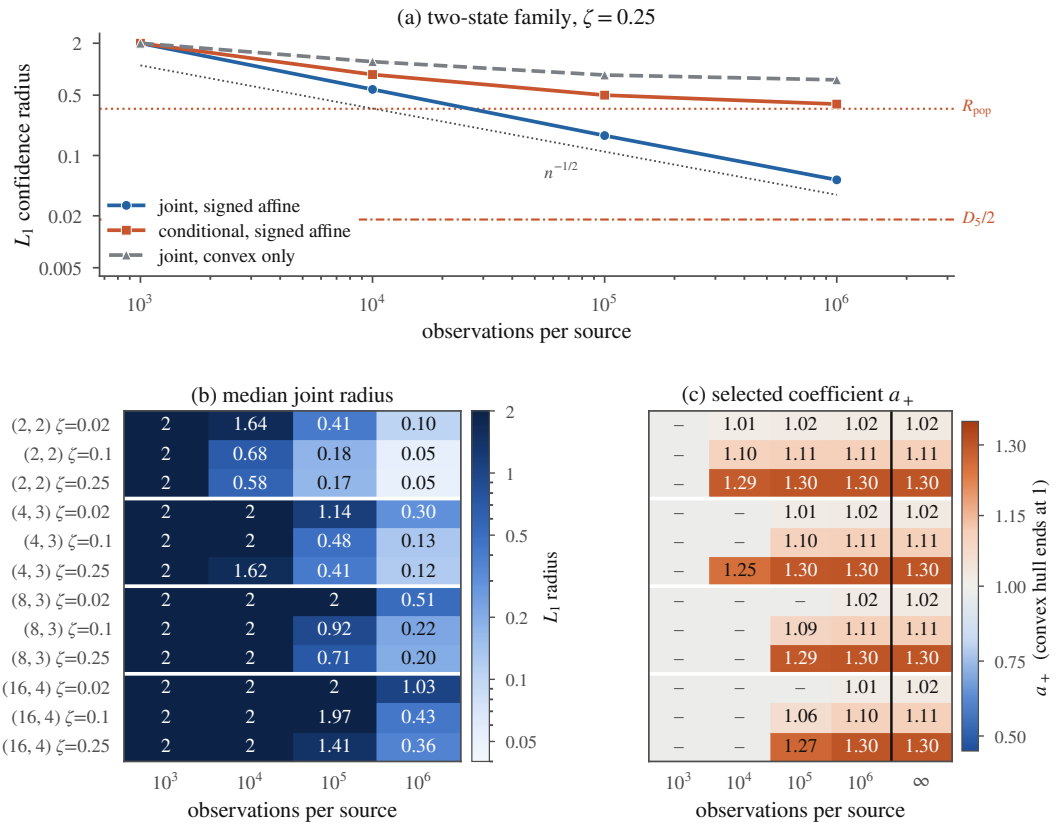


Figure 3: Count-only confidence at $\eta = 1$. (a) Two-state family, $\zeta = 0.25$: medians over 100 datasets with 10–90% bands, thinner than the median lines at this scale; R_{pop} is the conditional population certificate and $D_5/2$ the five-kernel half-diameter (Section C.4); the plateau is a conservative certificate, not the sharp ambiguity floor. (b) Median joint signed-affine radius; "2" is the trivial ball. (c) Median selected coefficient a_+ on $\widehat{\rho}^+$ (gray: trivial balls; ∞: population value). The population targets require $a_+ > 1$, so the convex-only certificate retains a positive limiting radius.

below 2.04×10^{-14}, consistent with identification up to precision. The conditional diameter ranges from 0.00496 to 0.05446 across matched radii; on the locked 49-kernel, six-strength grid its largest fixed-strength value is 0.0540 at $\eta = 1$. The denser curve peaks near 0.061 at $\eta = 0.84$ (Figure 2c) and vanishes as $\eta \to 0$ or $\eta \to \infty$, since compatible conditional targets approach the always-L core point as $\eta \to \infty$. At $\eta = 1$, the largest distance from source-only center $v^\star$ on the 81×81 grid is 0.042, versus the uniform bound 0.246 from Equation (7) (Figure 2b). Thus identical source frequencies permit different conditional targets before sampling error. An independent constrained-KL solver agrees with every fixed-strength solution within 2.42×10^{-13} in L_1.

Finite-Source Confidence Across Dimension and Overlap The count-only study uses 12 identified-family and 18 random full-support geometries (4–64 state–action pairs, three overlap levels), with 100 datasets per geometry at each $n \in \{10^3, 10^4, 10^5, 10^6\}$ per source. Signed-affine, convex-only and fixed-midpoint methods share counts and intervals; we compare their radii and oracle coverage. See Section C.2 for generation and audit details. For moderate overlap in the identified family, the joint radius contracts across all four dimensions. At $n = 10^6$, signed-affine median radii are 0.052, 0.121, 0.200 and 0.359, versus 0.753, 0.813, 0.879 and 1.019 for convex-only. The gap reflects the theory: an identified target can lie in the affine core but outside the convex hull, so forbidding signed extrapolation adds uncertainty. When informative, the selected coefficient on $\widehat{\rho}^+$ concentrates at its population value of 1.02, 1.11 and 1.30 for $\zeta = 0.02$, 0.1 and 0.25 (Figure 3c). Figure 3b shows all identified-family settings. At 64 state–action pairs and lowest overlap, the joint median remains 1.029; random-pair medians are 2 at $n = 10^3$ and 0.291–2 at $n = 10^6$.

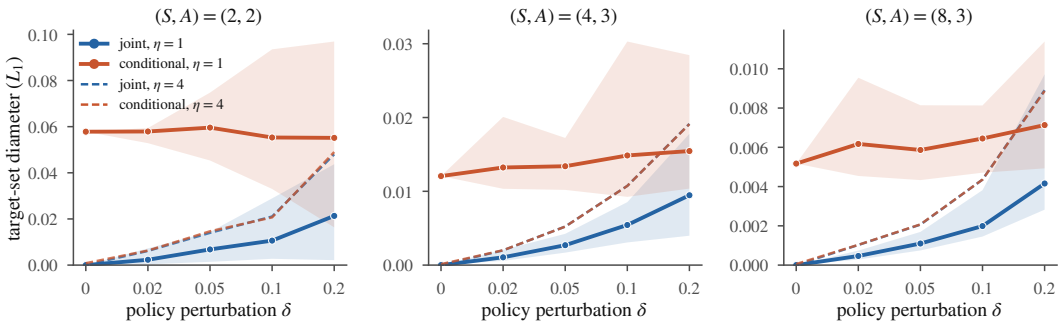


Figure 4: Evaluated target-set L_1 diameter at each policy-perturbation level δ (median over 20 draws; bands are the full range at $\eta = 1$; $\delta = 0$ uses the unperturbed pair). Two states use a compatible-kernel grid; $(4, 3)$ and $(8, 3)$ use 201 compatible kernels. Every evaluated diameter is a lower bound on its full-class value. The median joint diameter grows from zero, while the conditional diameter is positive already at $\delta = 0$; at $\eta = 4$ the two are comparable.

Across the grid, 899 datasets have unavailable support intervals and return the trivial radius (Figure 6 maps informative balls); every setting is retained. No evaluated target or extracted-policy occupancy falls outside its bound; the largest per-setting one-sided 95% Monte Carlo upper miss rate rounds to 0.0295. The five kernels per geometry are an evaluation set; uniformity over the compatible class follows from Theorem 5.5. At $n = 10^6$, affine policy bounds are nontrivial in all 1,200 identified-family datasets for each score, and in 1,491/1,800 joint and 1,400/1,800 conditional random-family evaluations.

Beyond the exactly identified family. A wide ball alone does not establish ambiguity because the certificate can be conservative; Section C.4 separates sampling error from certificate slack. For the two-state, $\zeta = 0.25$ case at $n = 10^6$, the joint median radius is 0.0522 with median tested error 0.00144. For the conditional score, the finite radius is 0.3941, its population certificate is 0.3478, and the five-kernel half-diameter is 0.0181. Thus the spread lower-bounds the required radius, while the larger values reflect finite-data inflation and certificate slack.

Robustness of the Identification Gap We perturb both policies of the identified family at $\zeta = 0.25$ by multiplying action probabilities by $e^{\delta\xi}$, with independent standard normal ξ, and renormalizing within each state. Recomputing occupancies under the generating kernel preserves compatibility. We use 20 draws at each $\delta \in \{0.02, 0.05, 0.1, 0.2\}$. Two states use a compatible-kernel grid; $(S, A) = (4, 3), (8, 3)$ use the generating kernel and 200 source-preserving perturbations. Every evaluated diameter lower-bounds its full-class value. Section C.6 includes the full protocol and $\delta = 0.4$.

Figure 4 shows the outcome. Both tilts leave the core for every perturbed pair at the tested strengths. At $\eta = 1$ and $\delta \in \{0.02, 0.05\}$, the median ratio of conditional to joint evaluated diameter ranges from 8.8 to 25.2 for two states, 4.7 to 12.3 for $(4, 3)$, and 5.7 to 12.1 for $(8, 3)$, using the same kernel set for each pair. The conditional diameter is larger in 233 of 240 perturbed pairs at $\eta = 1$ and $\delta \leq 0.2$. As perturbation grows, the median joint diameter increases from zero while the conditional diameter changes less. Thus small perturbations preserve a large gap after exact joint identification is lost. At $\eta = 4$, the two diameters are comparable.

7 Conclusion

Source occupancies can identify a KL-regularized imitation target despite unknown dynamics: under our conditions, identification holds exactly when the desired reweighting lies in the sources' affine core, and our two-state example shows conditioning can turn an identified joint target into an ambiguous conditional one. For finite logs, SCOR reports a count-only confidence ball covering compatible targets. Signed source combinations can localize identified targets outside the source convex hull while preserving flow feasibility.

STATEMENT OF USE OF AI

The research questions, the main ideas, the methodology, and the experimental protocols of this paper were chosen by the authors, who divided the work among themselves, supervised every stage of the project, and made all decisions on claims and conclusions. Under the authors' direction, we used generative AI tools as assistants for the following tasks with required disclosure: refining the conceptual framing, discussing candidate mathematical claims, and implementing the experiments and figure scripts. For the theoretical component, generative AI was used in part as a "theory attacker": it assisted in implementing computational searches and large-scale simulations over extensive collections of generated examples in order to identify potential counterexamples, edge cases, or weaknesses in candidate theorem statements and proof arguments. These searches were used to stress-test and refine the theoretical claims and proofs. The failure of such searches to identify a counterexample does not constitute a proof, and no theorem in this paper is considered established on the basis of AI-assisted simulation, search, or testing.

The authors independently verified all theorem statements and proofs, citations, experimental specifications, and AI-assisted text against their research artifacts. No experimental data or results were produced by generative AI tools themselves: every reported number comes from experiments that were actually executed, and translation and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools for literature discovery and for drafting and editing the manuscript, including checks of notation and of the consistency of reported numbers. The authors take full responsibility for every claim, proof, citation, and experimental result in this paper; no AI system is an author.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 specify the assumptions and observation interface; Section B gives the proofs. The appendix records the exact-source identification grid, the 12,000-dataset count-only confidence study, independent solver checks, population source-core diagnostics, and the 303-pair source-perturbation study. The inference procedure uses source counts and state–action labels; compatible kernels are used only to evaluate coverage and target spread. The supplementary material provides the source code for all experiments and figures, together with the recorded outputs behind every reported result.

REFERENCES

Siddhant Agarwal, Harshit Sikchi, Peter Stone, and Amy Zhang. Proto successor measure: Representing the behavior space of an RL agent. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 566–586, 2025. URL https://proceedings.mlr.press/v267/agarwal25e.html.

Haoyang Cao, Samuel N. Cohen, and Łukasz Szpruch. Identifiability in inverse reinforcement learning. In *Advances in Neural Information Processing Systems 34*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/671f0311e2754fcdd37f70a8550379bc-Abstract.html.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Adam Gleave, Michael Dennis, Shane Legg, Stuart Russell, and Jan Leike. Quantifying differences in reward functions. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=LwEQnp6CYev.

Ze Gong, Akshat Kumar, and Pradeep Varakantham. Offline safe reinforcement learning using trajectory classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16):16880–16887, 2025. doi: 10.1609/aaai.v39i16.33855. URL https://ojs.aaai.org/index.php/AAAI/article/view/33855.

Dong-Sig Han, Hyunseo Kim, Hyundo Lee, Je-Hwan Ryu, and Byoung-Tak Zhang. Unbiased learning with state-conditioned rewards in adversarial imitation learning, 2020. URL https://openreview.net/forum?id=bIwkmDnSeu. ICLR 2021 conference submission.

Huy Hoang, Tien Mai, and Pradeep Varakantham. Imitate the good and avoid the bad: An incremental approach to safe reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12439–12447, 2024. doi: 10.1609/aaai.v38i11.29136. URL https://ojs.aaai.org/index.php/AAAI/article/view/29136.

Huy Hoang, Tien Mai, and Pradeep Varakantham. No experts, no problem: Avoidance learning from bad demonstrations. In *Advances in Neural Information Processing Systems 38*, 2025. doi: 10.52202/085713-0956. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/292451739ffc63510bc8aac64c9ad23f-Abstract-Conference.html. Earlier arXiv title: "UNIQ: Offline Inverse Q-Learning for Avoiding Undesirable Demonstrations".

Huy Hoang, Tien Mai, Pradeep Varakantham, and Tanvi Verma. DualCOIL: Offline imitation learning from contrasting demonstrations. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2505.21182. Accepted paper; released as arXiv:2505.21182 under the earlier title "Learning What to Do and What Not To Do: Offline Imitation from Expert and Undesirable Demonstrations".

Andrew Jaegle, Yury Sulsky, Arun Ahuja, Jake Bruce, Rob Fergus, and Greg Wayne. Imitation by predicting observations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4665–4676, 2021. URL https://proceedings.mlr.press/v139/jaegle21b.html.

Youngsoo Jang, Geon-Hyeong Kim, Jongmin Lee, Sungryull Sohn, Byoungjip Kim, Honglak Lee, and Moontae Lee. SafeDICE: Offline safe imitation learning with non-preferred demonstrations. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/ed2fb79f2664c3d9ba878be7e575b2af-Abstract-Conference.html.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994.

Kangrui Ruan, Junzhe Zhang, Xuan Di, and Elias Bareinboim. Causal imitation for markov decision processes: a partial identification approach. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2781.

URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9f7f2f57d8eaf44b2f09020f64ff6d96-Abstract-Conference.html.

Andreas Schlaginhaufen and Maryam Kamgarpour. Identifiability and generalizability in constrained inverse reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 30224–30251. PMLR, 2023. URL https://proceedings.mlr.press/v202/schlaginhaufen23a.html.

Joar Max Viktor Skalse, Lucy Farnik, Sumeet Ramesh Motwani, Erik Jenner, Adam Gleave, and Alessandro Abate. STARC: A general framework for quantifying differences between reward functions. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/20bd42d82998bc61732c00452228e814-Abstract-Conference.html.

Henk J. van Waarde, Jaap Eising, Harry L. Trentelman, and M. Kanat Camlibel. Data informativity: A new perspective on data-driven analysis and control. *IEEE Transactions on Automatic Control*, 65(11):4753–4768, 2020. doi: 10.1109/TAC.2020.2966717. URL https://arxiv.org/abs/1908.00468.

Joe Watson, Sandy Huang, and Nicolas Heess. Coherent soft imitation learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 14540–14583. Curran Associates, Inc., 2023. doi: 10.52202/075280-0640. URL https://papers.neurips.cc/paper_files/paper/2023/hash/2f0435cffef91068ced08d7c7d8e643e-Abstract-Conference.html.

APPENDIX

Appendix roadmap. The appendix first defines response-matched targets used in the two-state example. It then proves the identification, approximation, and finite-source confidence results. The final section gives protocols and full results for the identification, confidence, and source-perturbation experiments.

A MATCHED-RADIUS TARGETS

The main result fixes a score and its strength η. The two-state audit also retunes the strength to compare targets at the same excess KL divergence from the reference. This removes arbitrary positive score scaling. For fixed P, let $\rho_P^{\mathrm{base}} = \arg\min_{\rho \in \mathcal{C}_P} D_{\mathrm{KL}}(\rho \| \rho^b)$ be the feasible occupancy closest to the reference. Define its excess-KL response radius by

$$\delta_P(\rho) = D_{\mathrm{KL}}(\rho \| \rho^b) - D_{\mathrm{KL}}(\rho_P^{\mathrm{base}} \| \rho^b).$$

A radius $\delta > 0$ is *attainable* for score f if some $\eta > 0$ produces a target at that radius; it is *common attainable* over a kernel class if this holds for every kernel in the class. When attainable, we choose $\eta_{f,P}(\delta)$ so that

$$\delta_P\big(\mathcal{T}^P_{\eta_{f,P}(\delta)}(f)\big) = \delta, \qquad \mathcal{P}^P_f(\delta) = \mathcal{T}^P_{\eta_{f,P}(\delta)}(f).$$

We call $\mathcal{P}^P_f$ the *matched-radius target path*. When f is nonconstant over $\mathcal{C}_P$, the optimizer changes with η: otherwise subtracting its two interior KKT equations would make f a flow-null score. For $0 < \eta_1 < \eta_2$, the two optimality inequalities then give $\langle p_2, f\rangle > \langle p_1, f\rangle$ and $D_{\mathrm{KL}}(p_2\|\rho^b) > D_{\mathrm{KL}}(p_1\|\rho^b)$, where $p_i = \mathcal{T}^P_{\eta_i}(f)$. Continuity and strict monotonicity make $\eta_{f,P}(\delta)$ unique at every attainable positive radius.

B PROOFS OF TARGET IDENTIFICATION AND CONFIDENCE

The proofs follow the main-paper claims: exact target identification, the conditioning example, and confidence from finite source counts.

B.1 TARGET IDENTIFICATION FROM THE SOURCE AFFINE CORE

Proof strategy. The main theorem rests on two geometric facts. First, the occupancies feasible under every source-compatible dynamics are exactly the nonnegative part of the affine span of the observed sources. Second, under underspanning, the only score component that is flow-null for every compatible dynamics is a constant. Together these facts force any optimizer shared by all compatible models to be the unconstrained KL tilt, and then point identification reduces to observable affine membership.

We prove Theorem 5.1 for the complete compatible class in Equation (5). It also holds if that class is restricted to its strictly positive kernels. Write $A_P = B - \gamma P^\top$, $\mathcal{C}_P = \{p \geq 0 : A_P p = b\}$, and $d = \operatorname{rank} Z$. Every state has at least one allowed action.

The common feasible set. For $S \geq 2$ and a strictly positive compatible P_0,

$$\bigcap_{P \in \mathcal{K}_X} \mathcal{C}_P = \mathcal{C}_X. \tag{11}$$

The right-to-left inclusion follows by affine combination of the common source flow equations. For the converse put $U = \ker X^\top$. For every $u \in U$ and $h \perp \mathbf{1}_S$, sufficiently small nonzero ε makes $P_\varepsilon = P_0 + \varepsilon u h^\top$ strictly positive, row stochastic, and source compatible. If p is feasible under all such kernels, subtraction of its flow equations gives $h(u^\top p) = 0$. Since $S \geq 2$, choose $h \neq 0$ and obtain $p \perp U$, hence $p \in \operatorname{range} X$. Every feasible occupancy is normalized: $\mathbf{1}_S^\top A_P = (1-\gamma)\mathbf{1}_E^\top$ implies $\mathbf{1}_E^\top p = 1$. Thus a representation $p = X\alpha$ has $\mathbf{1}^\top \alpha = 1$, so $p \in \mathcal{C}_X$. This argument does not require underspanning.

The common normal space. When $d < E - S$,

$$\bigcap_{P \in \mathcal{K}_X} \operatorname{range}(A_P^\top) = \operatorname{span}\{\mathbf{1}_E\}. \tag{12}$$

Each A_P has row rank S: $A_P^\top V = 0$ implies $|V(s)| \le \gamma \|V\|_\infty$ at every state, hence $V = 0$. Normalization gives $\operatorname{rank} X = d + 1$, so $\dim U = E - d - 1$. Put $R_0 = \operatorname{range}(A_{P_0}^\top)$. Since $X^\top A_{P_0}^\top V = \mathbf{1}(b^\top V)$ and $b \neq 0$, injectivity of $A_{P_0}^\top$ yields $\dim(U \cap R_0) = S - 1$. The strict inequality $d < E - S$ therefore provides $u \in U \setminus R_0$.

Let $\xi = A_{P_0}^\top V$ belong to every compatible normal space. For each $h \perp \mathbf{1}$ use the source-preserving perturbation above. Membership of ξ in its normal space supplies W satisfying

$$A_{P_0}^\top (W - V) = \gamma \varepsilon u (h^\top W). \tag{13}$$

The left side lies in R_0 and $u \notin R_0$, forcing $h^\top W = 0$. Injectivity then gives $W = V$. This holds for every $h \perp \mathbf{1}$, so V is constant. Conversely, $A_P^\top \mathbf{1} = (1 - \gamma)\mathbf{1}$ for every stochastic kernel, proving Equation (12). Intersections of shaping spaces under kernel changes also underlie Schlaginhaufen & Kamgarpour (2023, Theorem 4.12 and Appendix C.6); here the perturbation directions are constrained by $X^\top u = 0$.

Proof of Theorem 5.1. Compactness and strict convexity give a unique target in each $\mathcal{C}_P$. It is positive because mixing any proposed boundary optimum with a common strictly positive source occupancy yields an entropy gain of order $-\varepsilon \log \varepsilon$ that dominates the linear score change. At a common target p_*,

$$\log(p_*/\rho^b) + \mathbf{1} - \eta f \in \operatorname{range}(A_P^\top) \qquad \text{for every } P. \tag{14}$$

Equation (12) makes the left side constant, which forces $p_* = w_{\eta,f}$. Equation (11) then forces $w_{\eta,f} \in \mathcal{C}_X$. Conversely the unconstrained tilt is the unique optimizer over the probability simplex. If it lies in $\mathcal{C}_X$, it is feasible and optimal under every compatible kernel.

For radius matching, feasibility of ρ^b gives the same projection base ρ^b in every model. Pair the KKT equations at a proposed common target p_* with the stipulated observed displacement z:

$$z^\top \log(p_*/\rho^b) = \eta_{f,P}(\delta) z^\top f. \tag{15}$$

Because $z^\top f \neq 0$, the values $\eta_{f,P}(\delta)$ coincide across compatible kernels. The fixed-η result applies, and its tilt must have radius δ. Conversely a feasible tilt of that radius supplies the common target. The simplex-tilt radius increases strictly from zero to $\delta_\Delta(f) = -\log \sum_{e: f(e) = \max f} \rho^b(e)$. Thus its finite positive inverse exists exactly for $0 < \delta < \delta_\Delta(f)$; a common constrained radius at or beyond that endpoint cannot have a point-identified finite-η target. □

The hypotheses delimit the converse. At full rank $d = E - S$, all compatible feasible sets equal $\mathcal{C}_X$ and every target is identified by constrained optimization, even when the unconstrained tilt is outside that set. When $S = 1$ or $\gamma = 0$, the occupancy geometry is already fixed independently of these perturbations. Known transition zeros or absence of a strictly positive compatible kernel can likewise prevent the converse. With an arbitrary nonfeasible reference, the fixed-η result still holds, but unknown projection bases preclude the stated radius reduction. Affine membership concerns exact source occupancies; numerical rank of sampled tables does not certify equality.

Proof of Equation (7). For $w = w_{\eta,f}$, completing the KL expression gives $F_{\eta,f}(p) = \eta^{-1} \log \sum_e \rho^b(e) e^{\eta f(e)} - \eta^{-1} D_{\mathrm{KL}}(p \| w)$. Thus $p_P = \mathcal{T}_\eta^P(f)$ is the KL projection of w onto $\mathcal{C}_P$. Since $v \in \mathcal{C}_X \subseteq \mathcal{C}_P$ and $p_P > 0$, its KKT equation and common flow yield the exact Pythagorean identity

$$D_{\mathrm{KL}}(v \| w) = D_{\mathrm{KL}}(v \| p_P) + D_{\mathrm{KL}}(p_P \| w). \tag{16}$$

Consequently $F_{\eta,f}(p_P) - F_{\eta,f}(v) = \eta^{-1} D_{\mathrm{KL}}(v \| p_P) \le \epsilon_v / \eta$. Pinsker's inequality gives the distance bound. The entire compatible target set therefore has L_1 diameter at most $2\sqrt{2\epsilon_v}$. These upper bounds need only a common strictly positive source occupancy and a feasible v, not underspanning or a strictly positive transition kernel. Optimizing v over $\mathcal{C}_X$ tightens the certificate, while any verified

feasible approximate solution remains valid. The bounds concern the declared regularized objective at fixed η; they imply neither task-return performance nor automatic matched-radius bounds. □

Proof of Corollary 5.2. The source displacement has rank one, below $E - S = 2$. Direct flow checks give the stated two strictly positive compatible kernels. Uniform occupancy is feasible for P_0 but violates the first flow equation of P_1 by $1/160$, so the feasible sets differ. The common source core is Equation (9), and the joint tilt has $h_\eta = 3^{2\eta}/(1 + 3^{2\eta})$. It lies in that core for all finite $\eta > 0$. At $\eta = 1$, $h_\eta = 9/10$, giving the claimed affine extrapolation outside the observed convex hull.

Put $k = \log(9/7) > 0$. The state score is $(k, k, -k, -k)$ and $c = (\log(7/3), \log(7/27), \log(27/7), \log(3/7))$. For the source displacement $z = \rho^+ - \rho^-$, $z^\top c = \log(63)/4 > 0$ and $z^\top g_{\mathrm{SA}} = \log 3 > 0$; thus the matched-radius condition in Theorem 5.1 holds for both scores. The conditional tilt violates $p_1/p_3 = 5/3$ for every $\eta > 0$. Theorem 5.1 proves the claimed point-identification distinction for fixed η and for common positive radii.

Both scores have unique optimal always-L occupancy $p_H = (5/8, 0, 3/8, 0)$ in every compatible model. For the conditional score, the value V_L of always choosing L obeys $\operatorname{span} V_L \le 2k/(1-\gamma) = 4k$. Its L-over-R Bellman advantage at either state is at least $2\log 3 - \gamma \operatorname{span} V_L \ge 2\log 3 - 2k > 0$. Thus always-$L$ is strictly optimal for c, as it plainly is for g_{SA}. Since $D_{\mathrm{KL}}(p_H \| \rho^b) = \log 2$, both positive η-paths attain exactly $(0, \log 2)$ in radius.

The displayed kernels themselves witness the conditional ambiguity. Their normal spaces are $\operatorname{span}\{\mathbf{1}, a_0\}$ and $\operatorname{span}\{\mathbf{1}, a_1\}$, where $a_0 = (5, 7, -3, -1)/8$ and $a_1 = (47, 70, -25, -10)/80$. These spaces intersect only in the constants, while their common feasible set is the source core. A common conditional optimizer at fixed η would therefore equal its unconstrained tilt, which lies outside that core. At a matched radius, pairing the two KKT equations with z first forces the two strengths to agree because $z^\top c > 0$; the same contradiction applies. The joint path is common to both kernels, while their conditional targets differ throughout $0 < \delta < \log 2$. □

Proof of Corollary 5.3. With $\rho^b = \rho^-$ and $\eta = 1$, Equation (6) gives $w_{1,g_{\mathrm{SA}}} \propto \rho^- \cdot \rho^+/\rho^- = \rho^+$, which is normalized, and $w_{1,c}(\mathbf{s}, \mathbf{a}) \propto \rho^-(\mathbf{s}, \mathbf{a})\pi^+(\mathbf{a} \mid \mathbf{s})/\pi^-(\mathbf{a} \mid \mathbf{s}) = d^-(\mathbf{s})\pi^+(\mathbf{a} \mid \mathbf{s})$, which is also normalized. The first tilt is the core point with $u = 1$. It is feasible under every compatible kernel and already maximizes the objective over all probability vectors, so every compatible target equals ρ^+. The second tilt has state marginal d^-. A core point $\rho^- + u(\rho^+ - \rho^-)$ has state marginal $d^- + u(d^+ - d^-)$, so equality of marginals forces $u = 0$ when $d^+ \ne d^-$, that is $w_{1,c} = \rho^-$ and $\pi^+ = \pi^-$ at every state. Two policies that agree everywhere have the same occupancy under the shared kernel and initial law, contradicting $\rho^+ \ne \rho^-$. Hence $w_{1,c} \notin \mathcal{C}_X$, and the converse direction of Theorem 5.1 applies. □

Proof of Proposition 5.4. Write $\rho = d^\rho \pi^\rho$, choosing π^ρ arbitrarily at zero-mass states and interpreting zero-probability terms by continuity. Since $c = \log(\pi^+/\pi^-)$,

$$\begin{aligned}\mathbb{E}_\rho[c] - \sum_{\mathbf{s}} d^\rho(\mathbf{s}) D_{\mathrm{KL}}(\pi^\rho \| \pi^-) &= \sum_{\mathbf{s}} d^\rho(\mathbf{s}) \sum_{\mathbf{a}} \pi^\rho(\mathbf{a} \mid \mathbf{s}) \Big[\log \frac{\pi^+(\mathbf{a} \mid \mathbf{s})}{\pi^-(\mathbf{a} \mid \mathbf{s})} - \log \frac{\pi^\rho(\mathbf{a} \mid \mathbf{s})}{\pi^-(\mathbf{a} \mid \mathbf{s})} \Big] \\ &= -\sum_{\mathbf{s}} d^\rho(\mathbf{s}) D_{\mathrm{KL}}(\pi^\rho \| \pi^+) \le 0,\end{aligned}$$

The feasible occupancy ρ^+ attains zero. Equality holds exactly when $\pi^\rho = \pi^+$ at every state with $d^\rho(\mathbf{s}) > 0$. Fix a compatible kernel P and write $P_\pi(\mathbf{s}' \mid \mathbf{s}) = \sum_{\mathbf{a}} \pi(\mathbf{a} \mid \mathbf{s}) P(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})$. The state marginal of any $\rho \in \mathcal{C}_P$ solves $d = (1-\gamma)\mu_0 + \gamma P_{\pi^\rho}^\top d$, and states with $d^\rho(\mathbf{s}) = 0$ contribute nothing to the right-hand side, so an optimal ρ has the marginal that π^+ induces under P. Because ρ^+ satisfies the flow equations of P with conditional π^+, that marginal is d^+, which is strictly positive. Every state is therefore visited, $\pi^\rho = \pi^+$ everywhere, and $\rho = \rho^+$. □

When two sources identify the entire unconstrained joint-tilt path. For two distinct strictly positive sources and midpoint reference, put $r_e = \rho_e^+/\rho_e^-$. The ratio $w_{\eta, g_{\mathrm{SA}}}(e)/\rho_e^-$ is proportional to $(1 + r_e) r_e^\eta$. At $\eta = 1$, affine membership requires a nonzero quadratic in r_e to equal an affine function, allowing at most two distinct ratio values. Normalization and distinctness exclude a constant ratio. Conversely, with two ratio groups, both sources and every unconstrained tilt share the conditional

proportions within each group. Probability vectors with these fixed within-group proportions form a line spanned by the two sources, so every unconstrained tilt belongs to their source affine core. Under the underspanning and compatibility hypotheses of Theorem 5.1, this also characterizes the entire identified joint-target path. It is not a converse for full-span constrained targets. For the higher-ratio group H, the identified finite-η radii range over $0 < \delta < -\log \rho^b(H)$; the endpoint is attained only as $\eta \to \infty$.

B.2 Count-only confidence construction

The following details specify the intervals, radius budget, and procedure used in Theorem 5.5.

Allocate error $\alpha/(4E)$ to each of $2E$ source-probability and $2E$ conditional-action Clopper–Pearson intervals, i.e., exact binomial confidence intervals (Clopper & Pearson, 1934). Their simultaneous coverage event has probability at least $1 - \alpha$. Let $[L_e^y, U_e^y]$ denote the source intervals and $[L_e^w, U_e^w]$ the induced intervals for the true tilt Equation (6). Section B.4 specifies their log-ratio and normalized exponential formulas, including random state-count denominators.

For signed coefficients $\boldsymbol{a} = (a_+, a_-)$ with $a_+ + a_- = 1$, form $\widehat{v}_{\boldsymbol{a}} = \sum_y a_y \widehat{\rho}^y$ and sign-aware intervals $[L^v_{\boldsymbol{a},e}, U^v_{\boldsymbol{a},e}]$ for the unobserved population combination $v_{\boldsymbol{a}} = \sum_y a_y \rho^y$. For $a_y^+ = \max(a_y, 0)$ and $a_y^- = \max(-a_y, 0)$, these are

$$L^v_{\boldsymbol{a},e} = \sum\nolimits_y (a_y^+ L_e^y - a_y^- U_e^y), \qquad U^v_{\boldsymbol{a},e} = \sum\nolimits_y (a_y^+ U_e^y - a_y^- L_e^y). \tag{17}$$

Allowing negative coefficients lets the witness extend beyond a simple mixture of the sources, as the identified target in the two-state example requires. This also amplifies estimation error, so the construction must account for coefficient signs and magnitudes. Negative coefficients reverse interval endpoints. Requiring $L^v_{\boldsymbol{a},e} \geq 0$ for every coordinate certifies that the population affine witness remains a probability vector. The following quantities turn that intuition into a radius: $e_{\boldsymbol{a}}$ bounds the error in estimating the witness, $r_{\boldsymbol{a},e}$ bounds its coordinatewise mismatch from the desired reweighting, and $\epsilon_{\boldsymbol{a}}$ combines those mismatches into an upper bound on their KL divergence. Define

$$\begin{aligned} e_{\boldsymbol{a}} &= \sum\nolimits_y |a_y| \sum\nolimits_e \max\{\widehat{\rho}_e^y - L_e^y, U_e^y - \widehat{\rho}_e^y\}, \\ r_{\boldsymbol{a},e} &= \max\{U^v_{\boldsymbol{a},e} - L_e^w, U_e^w - L^v_{\boldsymbol{a},e}, 0\}, \qquad \epsilon_{\boldsymbol{a}} = \sum\nolimits_e \frac{r^2_{\boldsymbol{a},e}}{L_e^w}. \end{aligned} \tag{18}$$

B.3 Policy extraction from the confidence center

Optional policy extraction from the reported center. The confidence ball concerns an occupancy target, but its empirical center can also be normalized into a count-computable policy. This is a secondary consequence rather than a new identification claim. Set $x = \widehat{v}_{\boldsymbol{a}}$, $d_x(s) = \sum_a x(s,a)$ and $\widehat{\pi}(a \mid s) = x(s,a)/d_x(s)$ when $d_x(s) > 0$, choosing any action distribution at zero-mass states. Write $\rho_P^{\widehat{\pi}}$ for its actual discounted occupancy under a compatible kernel. On the event of Theorem 5.5,

$$\|\rho_P^{\widehat{\pi}} - \mathcal{T}_\eta^P(f)\|_1 \leq \min\left\{2, \frac{2e_{\boldsymbol{a}}}{1-\gamma} + \sqrt{2\epsilon_{\boldsymbol{a}}}\right\} \quad \text{for every } P \in \mathcal{K}_X. \tag{19}$$

Normalizing the center gives a policy, but its actual visitation frequencies need not equal that center: the empirical center can violate the unknown flow equations. The factor involving $(1-\gamma)^{-1}$ accounts for this error accumulating along the discounted trajectory. Thus the displayed bound connects the estimated target to behavior produced by a policy, while still making no claim about task return. Section B.4 provides the derivation.

B.4 Finite-sample confidence from source occupancies

Proof strategy. The finite-sample argument mirrors the population geometry without testing an exact affine equality. We construct one simultaneous event for the source and conditional-action probabilities, propagate that event to an enclosure of the true source-defined tilt, certify a source-affine

Procedure 1: Count-only source-target confidence

Input: source count vectors K^+, K^- and state–action labels; score choice $f \in \{g_{\mathrm{SA}}, c\}$, fixed $\eta > 0$, and error level α. *Output:* a probability-vector center $\widehat{v}_{\boldsymbol{a}}$, a confidence radius $R_{\boldsymbol{a}}$, and the selected source coefficients $\boldsymbol{a}$.

1. Compute empirical source occupancies and state totals. Construct the $4E$ Clopper–Pearson intervals with error $\alpha/(4E)$ each, using state totals as denominators for conditional-action intervals.
2. Enclose the true-source score, midpoint reference and normalized tilt using Equation (22). If required log bounds are not finite or any tilt lower bound is nonpositive, return the empirical midpoint and radius 2.
3. Write $\boldsymbol{a} = (u, 1-u)$ and intersect each sign region $[-1, 0]$, $[0, 1]$, $[1, 2]$ with $L^v_{\boldsymbol{a}} \geq 0$. On each feasible interval, minimize $e_{\boldsymbol{a}} + \sqrt{2\epsilon_{\boldsymbol{a}}}$ by bounded scalar optimization; also evaluate numerically inset interval endpoints and the midpoint candidate $u = 1/2$ (use the interval midpoint if its width is below the numerical tolerance).
4. Recheck candidate feasibility and finite arithmetic. Among checked candidates, select the smallest uncapped radius and return $(\widehat{v}_{\boldsymbol{a}}, R_{\boldsymbol{a}}, \boldsymbol{a})$. If numerical evaluation is unavailable, return the empirical midpoint and radius 2 with its fallback status.

Matched controls: replace $[-1, 2]$ by $[0, 1]$ or fix $u = 1/2$; retain the same observations, intervals and confidence allocation.

witness with signed coefficients that is nonnegative and feasible under every compatible dynamics, and then use KL projection geometry to convert tilt mismatch into target distance. A final centering term moves the guarantee from the latent affine witness to the reported empirical center.

We keep $\eta > 0$ fixed throughout. The true source occupancies ρ^+ and ρ^- are strictly positive; their reference is $\rho^b = (\rho^+ + \rho^-)/2$, and $f \in \{g_{\mathrm{SA}}, c\}$ is computed from these true occupancies. Write $w = w_{\eta,f}$ and $p_P = \mathcal{T}^P_\eta(f)$ for the resulting population tilt and target. The compatible class $\mathcal{K}_X$ uses the true source occupancies, with common μ_0, γ as in Equation (5).

Simultaneous binomial intervals. For each $y \in \{+, -\}$, let K^y_e count state–action pair $e = (s, a)$ among n_y independent categorical draws from ρ^y, and put $\widehat{\rho}^y_e = K^y_e/n_y$ and $N^y_s = \sum_{a:(s,a)\in\mathcal{E}} K^y_{(s,a)}$. An independent geometric-time observation from each independent rollout has this sampling law for normalized discounted occupancy. Consecutive observations within one trajectory require a different concentration argument. Let $E = |\mathcal{E}|$ and $\alpha_0 = \alpha/(4E)$. Writing $\mathrm{Beta}^{-1}(u; r, t)$ for the beta quantile, define

$$\begin{aligned} \ell(k, N) &= \begin{cases} 0, & k = 0, \\ \mathrm{Beta}^{-1}(\alpha_0/2; k, N-k+1), & k > 0, \end{cases} \\ u(k, N) &= \begin{cases} 1, & k = N, \\ \mathrm{Beta}^{-1}(1-\alpha_0/2; k+1, N-k), & k < N. \end{cases} \end{aligned} \tag{20}$$

For $N = 0$ these definitions give $[0, 1]$. Use $[L^y_e, U^y_e] = [\ell(K^y_e, n_y), u(K^y_e, n_y)]$ for the $2E$ source probabilities. For the $2E$ conditional action probabilities, use $[L^{\pi,y}_{(s,a)}, U^{\pi,y}_{(s,a)}] = [\ell(K^y_{(s,a)}, N^y_s), u(K^y_{(s,a)}, N^y_s)]$. Conditionally on $N^y_s = N$, the state–action count is $\mathrm{Bin}(N, \pi^y(a \mid s))$. Its conditional Clopper–Pearson failure probability is at most α_0; averaging over the random total preserves this bound. The $2E$ unconditional intervals have the same coverage guarantee. Thus the event $\mathcal{H}$ that all $4E$ intervals cover satisfies

$$\Pr(\mathcal{H}) \geq 1 - 4E\alpha_0 = 1 - \alpha. \tag{21}$$

The union bound needs no independence among the intervals or between the source datasets. It does require the stated sampling law within each source.

Bounding the true-source score and tilt. For $f = g_{\mathrm{SA}}$ put $L^f_e = \log L^+_e - \log U^-_e$ and $U^f_e = \log U^+_e - \log L^-_e$. For $f = c$, replace L^y_e, U^y_e in these two expressions by $L^{\pi,y}_e, U^{\pi,y}_e$. When their

required lower endpoints are positive, set

$$\begin{aligned} L_e^z &= \log \frac{L_e^+ + L_e^-}{2} + \eta L_e^f, \qquad U_e^z = \log \frac{U_e^+ + U_e^-}{2} + \eta U_e^f, \\ L_e^w &= \frac{\exp L_e^z}{\exp L_e^z + \sum_{j \neq e} \exp U_j^z}, \qquad U_e^w = \frac{\exp U_e^z}{\exp U_e^z + \sum_{j \neq e} \exp L_j^z}. \end{aligned} \tag{22}$$

A softmax coordinate increases with its own logit and decreases with every other logit. Consequently $L^w \leq w \leq U^w$ on $\mathcal{H}$. The same event controls the sampled reference and the sampled score; their dependence incurs no additional confidence allocation. It also supports reporting both score choices from the same data. If finite log bounds or positive L_e^w are unavailable, the procedure returns radius 2 about the empirical midpoint, which contains the whole probability simplex. Replacing an unavailable lower bound by an arbitrary positive floor would not preserve this argument.

Proof of Theorem 5.5. For $\sum_y a_y = 1$ write $a_y^+ = \max(a_y, 0)$, $a_y^- = \max(-a_y, 0)$, and define

$$\begin{aligned} L_{\boldsymbol{a},e}^v &= \sum_y (a_y^+ L_e^y - a_y^- U_e^y), \quad U_{\boldsymbol{a},e}^v = \sum_y (a_y^+ U_e^y - a_y^- L_e^y), \\ v_{\boldsymbol{a}} &= \sum_y a_y \rho^y, \qquad\qquad \widehat{v}_{\boldsymbol{a}} = \sum_y a_y \widehat{\rho}^y. \end{aligned} \tag{23}$$

Fix any observed coefficient vector with $L_{\boldsymbol{a}}^v \geq 0$ componentwise. On $\mathcal{H}$, the latent combination lies between its displayed endpoints, is nonnegative, and has total mass one. Moreover, for every $P \in \mathcal{K}_X$,

$$A_P v_{\boldsymbol{a}} = \sum_y a_y A_P \rho^y = \Big(\sum_y a_y \Big) b = b. \tag{24}$$

Clopper–Pearson intervals contain the empirical proportions, so the same endpoints enclose $\widehat{v}_{\boldsymbol{a}}$, which is also a probability vector. With

$$E_y = \sum_e \max\{\widehat{\rho}_e^y - L_e^y, U_e^y - \widehat{\rho}_e^y\}, \qquad e_{\boldsymbol{a}} = \sum_y |a_y| E_y, \tag{25}$$

we have $\|v_{\boldsymbol{a}} - \widehat{v}_{\boldsymbol{a}}\|_1 \leq e_{\boldsymbol{a}}$ on $\mathcal{H}$. The coordinate mismatch and budget are

$$\begin{aligned} r_{\boldsymbol{a},e} &= \max\{U_{\boldsymbol{a},e}^v - L_e^w, U_e^w - L_{\boldsymbol{a},e}^v, 0\}, \\ \epsilon_{\boldsymbol{a}} &= \sum_e \frac{r_{\boldsymbol{a},e}^2}{L_e^w}. \end{aligned} \tag{26}$$

Since $|v_{\boldsymbol{a},e} - w_e| \leq r_{\boldsymbol{a},e}$, the inequality $\log t \leq t - 1$ gives

$$D_{\mathrm{KL}}(v_{\boldsymbol{a}} \| w) \leq \sum_e \frac{(v_{\boldsymbol{a},e} - w_e)^2}{w_e} \leq \epsilon_{\boldsymbol{a}}. \tag{27}$$

Zero coordinates of $v_{\boldsymbol{a}}$ are allowed by continuity. A common strictly positive source occupancy is feasible, so the entropy projection p_P is strictly positive. Its affine KKT equation, paired with $v_{\boldsymbol{a}} - p_P$, yields

$$D_{\mathrm{KL}}(v_{\boldsymbol{a}} \| w) = D_{\mathrm{KL}}(v_{\boldsymbol{a}} \| p_P) + D_{\mathrm{KL}}(p_P \| w). \tag{28}$$

Pinsker's inequality and the triangle inequality therefore give

$$\|p_P - \widehat{v}_{\boldsymbol{a}}\|_1 \leq R_{\boldsymbol{a}} := \min\{2, e_{\boldsymbol{a}} + \sqrt{2\epsilon_{\boldsymbol{a}}}\}. \tag{29}$$

The cap uses normalization of both vectors. Completing the KL expression in $F_{\eta,f}$ as in Equation (7) also gives

$$\begin{aligned} F_{\eta,f}(p_P) - F_{\eta,f}(v_{\boldsymbol{a}}) &= \eta^{-1} D_{\mathrm{KL}}(v_{\boldsymbol{a}} \| p_P) \leq \epsilon_{\boldsymbol{a}} / \eta, \\ \mathrm{diam}_1\{p_P : P \in \mathcal{K}_X\} &\leq \min\{2, 2\sqrt{2\epsilon_{\boldsymbol{a}}}\}. \end{aligned} \tag{30}$$

The diameter bound uses the same latent feasible combination for every kernel. Every step holds for all admissible $\boldsymbol{a}$ and all $P \in \mathcal{K}_X$ on the single event $\mathcal{H}$, proving the claimed coverage even when the coefficients are selected on these data. No source rank test, underspanning condition, or positive transition kernel is needed. The objective gap concerns $v_{\boldsymbol{a}}$; $\widehat{v}_{\boldsymbol{a}}$ need not satisfy an unknown flow equation. □

Coefficient selection. For two sources the implemented selector writes $\boldsymbol{a} = (u, 1-u)$ and searches $u \in [-1, 2]$, using the feasible sign regions in Procedure 1. The convex and midpoint controls use $[0, 1]$ and $u = 1/2$. Every verified feasible candidate has the stated coverage, regardless of optimization accuracy.

The policy obtained from the empirical center. Put $x = \widehat{v}_{\boldsymbol{a}}$, $d_x = Bx$, and $\widehat{\pi}(a \mid s) = x(s,a)/d_x(s)$ where $d_x(s) > 0$, choosing any action distribution at other states. This policy is computable from the counts. If $\rho_P^{\widehat{\pi}}$ is its true discounted occupancy, then on $\mathcal{H}$, uniformly over compatible P,

$$\|\rho_P^{\widehat{\pi}} - p_P\|_1 \le \min\left\{2, \frac{2e_{\boldsymbol{a}}}{1-\gamma} + \sqrt{2\epsilon_{\boldsymbol{a}}}\right\}. \tag{31}$$

Indeed $x = d_x \widehat{\pi}$ even at zero-mass states, and, writing $P_{\widehat{\pi}}$ for its state transition matrix,

$$(I - \gamma P_{\widehat{\pi}}^\top) d_x - b = A_P(x - v_{\boldsymbol{a}}). \tag{32}$$

The stochastic resolvent has induced 1-norm at most $(1-\gamma)^{-1}$, while $\|A_P\|_1 \le 1 + \gamma$. The state-marginal difference and occupancy difference between $\rho_P^{\widehat{\pi}}$ and x have equal 1-norm because they share the same conditional policy. Hence $\|\rho_P^{\widehat{\pi}} - x\|_1 \le (1+\gamma)e_{\boldsymbol{a}}/(1-\gamma)$. Adding the bounds for $x - v_{\boldsymbol{a}}$ and $v_{\boldsymbol{a}} - p_P$ proves Equation (31).

Clopper–Pearson widths with random state totals. For $N > 0$, both endpoints in Equation (20) are within

$$h(N) = \sqrt{\frac{\log(8E/\alpha)}{2N}} \tag{33}$$

of k/N. For example, a nontrivial lower endpoint satisfies $\Pr_\ell\{\mathrm{Bin}(N, \ell) \ge k\} = \alpha_0/2$. The binomial Hoeffding bound makes this probability at most $\exp\{-2N(k/N - \ell)^2\}$, proving the lower-endpoint bound. The upper endpoint follows by the opposite tail; the trivial endpoint at a boundary count equals k/N. This bounds the exact binomial intervals used above, rather than changing their construction.

Let $d_{\min} = \min_{y,s} d^y(s) > 0$ and $n_{\min} = \min_y n_y$. Write $\mathcal{G}$ for the event that $N_s^y \ge n_y d^y(s)/2$ for every source and state. Binomial tails give

$$\Pr(\mathcal{G}^c) \le 2S \exp(-n_{\min} d_{\min}/8). \tag{34}$$

On $\mathcal{G}$, all source and conditional endpoints are within $h_n = \sqrt{\log(8E/\alpha)/(n_{\min} d_{\min})}$ of their empirical proportions. On $\mathcal{H} \cap \mathcal{G}$, they are therefore within $2h_n$ of the corresponding true probabilities.

Local contraction at an identified unconstrained tilt. Suppose $w = v_{\boldsymbol{a}_*}$ for fixed coefficients with $\sum_y a_{*,y} = 1$, $\|\boldsymbol{a}_*\|_1 \le A_* \le A$, and $\omega = \min_e w_e > 0$. Fix the strictly positive source occupancies, E, η, α, A. Let $\kappa = \min_{y,e} \rho_e^y > 0$; the true conditional probabilities and reference coordinates are also at least κ. If $h_n \le \kappa/4$, every relevant log endpoint is within $4h_n/\kappa$ of its truth on $\mathcal{H} \cap \mathcal{G}$. For either score, the logit endpoint error is at most $(4 + 8\eta)h_n/\kappa$. A softmax coordinate has gradient 1-norm at most $1/2$, so the tilt endpoint errors are at most Kh_n, where $K = (2 + 4\eta)/\kappa$. At the oracle coefficients,

$$L_{\boldsymbol{a}_*}^v \ge w - 2A_* h_n \mathbf{1}, \quad U_{\boldsymbol{a}_*}^v \le w + 2A_* h_n \mathbf{1}, \quad e_{\boldsymbol{a}_*} \le A_* E h_n. \tag{35}$$

Thus $\boldsymbol{a}_*$ is certified feasible eventually, and $L^w \ge (\omega/2)\mathbf{1}$. Consequently

$$\begin{aligned} \epsilon_{\boldsymbol{a}_*} &\le \frac{2E}{\omega}(2A_* + K)^2 h_n^2, \\ \inf_{\boldsymbol{a}}\{e_{\boldsymbol{a}} + \sqrt{2\epsilon_{\boldsymbol{a}}}\} &\le \left[A_* E + 2\sqrt{E/\omega}(2A_* + K)\right] h_n, \end{aligned} \tag{36}$$

where the infimum is over certified coefficients within budget A. For sufficiently large sample sizes this bound holds with probability at least $1 - \alpha - 2S\exp(-n_{\min}d_{\min}/8)$. An optimizer within τ_n of the infimum adds at most τ_n to its radius. In particular, the ideal minimum has local $O_{\Pr}(n_{\min}^{-1/2})$ radius at fixed confidence level. For this probability-order statement, empirical source and conditional proportions have root-$n_{\min}$ fluctuations, the state totals converge to positive fractions, and Equation (33) bounds the endpoint deviations from those empirical proportions. These facts give the same oracle argument in probability without conditioning on the fixed-probability event $\mathcal{H}$.

The quadratic budget in Equation (27) is what preserves this rate after Pinsker. Uniform constants over a varying model family require bounds on small source and tilt masses, η, and the norm of an identifying coefficient vector. A coefficient budget that excludes every such representation can prevent contraction. The fixed numerical tolerances of the scalar implementation are not a proof that τ_n vanishes; coverage only uses its verified candidates. None of these arguments controls a dynamics-dependent matched-radius η.

Proof of Proposition 5.6. We construct a full-support identified subfamily and then treat estimation risk, confidence diameter, and conditional ambiguity separately.

Expected estimation risk on an identified subfamily. Use two states, two actions, $\gamma = 1/2$, uniform μ_0, and $\theta \in [1/16, 3/16]$. In state–action order $(1, L), (1, R), (2, L), (2, R)$, set

$$p_h(\theta) = \big((1/2+\theta)h, (1/2-\theta)(1-h), (1/2-\theta)h, (1/2+\theta)(1-h)\big). \tag{37}$$

Take $\rho^- = p_{1/4}$, $\rho^+ = p_{3/4}$ and $\rho^b = p_{1/2}$. The strictly positive kernel with next-state-1 probabilities

$$(1/2+2\theta,\ 1/2-2\theta,\ 1/2+2\theta,\ 1/2-2\theta) \tag{38}$$

realizes every p_h: its state-1 flow equation is $\sum_a p_h(1,a) = 1/2+\theta(2h-1)$. The source displacement has rank $1 < E - S = 2$. The joint score is $\log 3(1,-1,1,-1)$ and its unconstrained tilt is

$$T(\theta) = p_{h_\eta}(\theta), \qquad h_\eta = \frac{3^{2\eta}}{1+3^{2\eta}}, \qquad \boldsymbol{a}_* = (2h_\eta - 1/2,\ 3/2 - 2h_\eta). \tag{39}$$

Its coefficient norm is at most 2, so it is in the source affine core and identified by Theorem 5.1. All required source, kernel and target probabilities stay positive on this compact family at fixed finite η.

For two parameters,

$$\|T(\theta) - T(\theta')\|_1 = 2|\theta - \theta'|, \qquad D_{\mathrm{KL}}(p_h(\theta)\|p_h(\theta')) = D_{\mathrm{Ber}}(1/2+\theta\|1/2+\theta'). \tag{40}$$

The second equality follows by grouping the four coordinates into their two state-within-action Bernoulli factors; the factors for the other action are complementary and have the same divergence. Let $\mathbb{P}_\theta^{(n)}$ be the data law for n independent draws from each source. Product additivity and the Bernoulli chi-square bound give

$$D_{\mathrm{KL}}(\mathbb{P}_\theta^{(n)}\|\mathbb{P}_{\theta'}^{(n)}) = 2nD_{\mathrm{Ber}}(1/2+\theta\|1/2+\theta') \le 10n(\theta-\theta')^2. \tag{41}$$

Choose $\theta_0 = 1/8$ and $\theta_1 = \theta_0 + c_0/\sqrt{n}$, where $0 < c_0 \le 1/16$. Their target separation is $\Delta_n = 2c_0/\sqrt{n}$ and their data total variation V_n is at most $\sqrt{5}\,c_0$ by Pinsker. For any point estimator $\widehat{T}$, choosing the closer of the two targets defines a test. A test error implies estimation loss at least $\Delta_n/2$, and the sum of its two error probabilities is at least $1 - V_n$. Hence

$$\max_{j\in\{0,1\}} \mathbb{E}_{\theta_j}\|\widehat{T} - T(\theta_j)\|_1 \ge \frac{\Delta_n}{4}(1 - V_n). \tag{42}$$

This proves the stated $\Omega(n^{-1/2})$ minimax expected-risk lower bound. The estimator may know this parametric family but observes only source samples; observing the numerical transition kernel would reveal θ and change the experiment.

Uniform confidence diameter on the identified subfamily. This requires a separate coverage argument. Let $\widehat{C}_n$ have coverage at least $1-\alpha$ at each parameter above. Transferring the second coverage event to the first data law gives

$$\begin{aligned}\mathbb{P}_{\theta_0}^{(n)}\{\mathrm{diam}_1(\widehat{C}_n) \geq \Delta_n\} &\geq \mathbb{P}_{\theta_0}^{(n)}\{T(\theta_0), T(\theta_1) \in \widehat{C}_n\} \\ &\geq 1 - 2\alpha - V_n.\end{aligned} \tag{43}$$

For $\alpha < 1/2$, choosing $c_0 \leq \min\{1/16, (1-2\alpha)/(2\sqrt{5})\}$ makes the right side at least $(1-2\alpha)/2 > 0$. This is a confidence-diameter lower bound in probability, distinct from Equation (42).

The conditional ambiguity floor. At $\theta = 1/8$, Equation (37) gives the sources in Equation (8). Their two displayed compatible kernels P_0, P_1 have normal spaces $\mathrm{span}\{\mathbf{1}, a_0\}$ and $\mathrm{span}\{\mathbf{1}, a_1\}$, where

$$a_0 = (5, 7, -3, -1)/8, \qquad a_1 = (47, 70, -25, -10)/80. \tag{44}$$

The vectors $\mathbf{1}, a_0, a_1$ are independent, so the two normal spaces intersect only in constants. Their common affine feasible set has dimension one and contains both sources, hence its nonnegative part is the source core. A common fixed-η conditional optimizer would, by its two KKT equations, have to equal the unconstrained conditional tilt. That tilt lies outside the core for every $\eta > 0$, as shown in Section B.1. Thus $D_\eta = \|\mathcal{T}_\eta^{P_0}(c) - \mathcal{T}_\eta^{P_1}(c)\|_1 > 0$.

The two kernels induce exactly the same source-data distribution at every sample size. Under this common law, the triangle inequality yields

$$\max_{j\in\{0,1\}} \mathbb{E}\|\widehat{T} - \mathcal{T}_\eta^{P_j}(c)\|_1 \geq D_\eta/2. \tag{45}$$

For a confidence set with coverage at least $1-\alpha$ separately under both kernels, the two coverage events intersect with probability at least $1-2\alpha$. Its diameter is therefore at least D_η with that probability. If the set promises simultaneous coverage of all compatible targets, as in Theorem 5.5, the probability is at least $1-\alpha$. In either case a containing ball has radius at least $D_\eta/2$ on the respective event. This floor persists with arbitrarily many source samples. □

Numerical scope. The coverage theorem uses the exact interval formulas and checked coefficient constraints. The experiments use floating-point evaluation with the feasibility and evaluation tolerances specified below. Underflow or nonfinite bounds trigger the radius-2 fallback. Uniform coverage follows from the common coverage event and flow argument.

C Experimental Details

The experiments below check the exact identification reversal, confidence coverage from finite counts, and the tightness of source-core bounds. Simulator kernels are evaluation tools, not inputs to the count-only method.

C.1 Exact target-identification reversal

The exact-source grid varies only the hidden dynamics while holding the audited source occupancies fixed. It directly checks the population identification claim in Corollary 5.2; no sampled-data confidence statement is involved.

The population experiment fixes the rational sources, discount and initial law in Equation (8). Compatible kernels are $P_{\lambda,\psi}$, whose probabilities of moving to state 1 from $(1,L),(1,R),(2,L),(2,R)$ are $(3/4+3\lambda,\ 1/4+5\psi,\ 3/4-5\lambda,\ 1/4-3\psi)$. The two source flow equations leave exactly this two-parameter affine family, which is strictly positive precisely on $(\lambda,\psi) \in (-1/20, 1/12)^2$; the kernels P_0 and P_1 of Section 5 are $(0,0)$ and $(1/40,0)$. The locked grid takes $\lambda, \psi \in \{-2,-1,0,1,2,3,4\}/50$, and exact rational arithmetic verifies every source flow and strict kernel positivity. The inverse-regularization grid is $\eta \in \{1/8, 1/4, 1/2, 1, 2, 4\}$, and the radius grid is $\delta \in \{0.001, 0.005, 0.02, 0.05, 0.1, 0.2\}$. For both scores, each of the 49 kernels is evaluated at all 12 operating points. No kernel, radius, or failed evaluation is dropped.

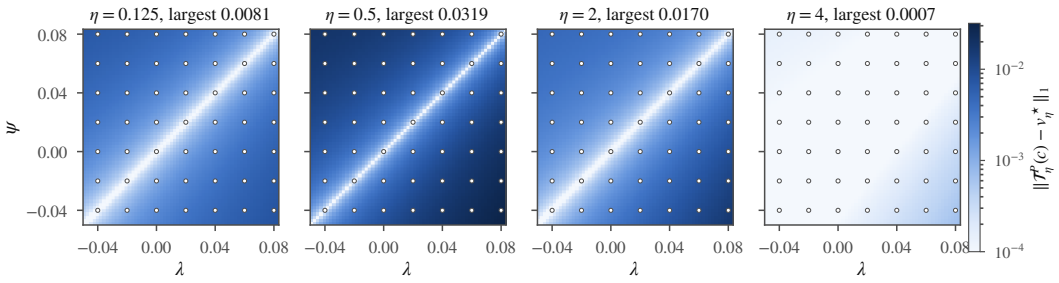


Figure 5: Conditional-target displacement $\|\mathcal{T}_\eta^P(c) - v_\eta^\star\|_1$ over the strictly positive compatible class $P_{\lambda,\psi}$ at four strengths (61×61 interior kernels each, shared logarithmic color scale; white dots are the locked 7×7 grid). Titles report the largest observed grid displacement. The pale band marks small distances from the optimizer to the common feasible center; every compatible feasible set contains that center. Values below 10^{-4} share the lowest color.

In two states, normalization reduces flow to one scalar constraint $\ell^\top \rho = b_1$. The primary solver finds the multiplier v satisfying $\ell^\top \operatorname{softmax}(\log \rho^b + \eta f + v\ell) = b_1$. A second scalar solve matches KL radius when required. Original flow, normalization, stationarity and radius residuals are checked at 10^{-9}. All 588 fixed-η targets are also solved by the independent general flow-constrained Newton routine, with L_1 agreement tolerance 10^{-8}. The maximum primary flow, stationarity and radius errors are 6.67×10^{-15}, 2.67×10^{-15} and 5.54×10^{-15}; the maximum cross-solver target error is 2.42×10^{-13}. The complete record retains all target coordinates, multipliers, η values, residuals and affine-membership calculations. An independent audit replays every saved evaluation and all 588 Newton solves. There are no sampled datasets in this experiment, and its finite grid does not replace the universal proof.

Figure 5 evaluates denser interior grids of the compatible class at four strengths. Each coordinate grid runs from $-1/20 + 0.001$ to $1/12 - 0.001$. Figure 2 uses 41×41 kernels in panel (a), 81×81 in panel (b), and 21×21 at 41 log-spaced strengths from $1/32$ to 32 in panel (c). For each η, the source-only center $v_\eta^\star = \arg\min_{v \in \mathcal{C}_X} D_{\mathrm{KL}}(v \| w_{\eta,c})$ is computed from the sources alone, and the map shows how far each compatible conditional target moves from it. The largest observed grid displacements are 0.008, 0.032, 0.042, 0.017 and 0.001 at $\eta = 1/8, 1/2, 1, 2, 4$, always inside the uniform budget $\sqrt{2\epsilon_{v_\eta^\star}}$ of Equation (7). The joint map is identically zero because the joint tilt is itself a core point. Every recomputed target at a locked grid point agrees with the locked record within 10^{-9} in L_1; the fine grids are exact-population illustrations and add no sampled data.

C.2 Finite-source target confidence

This study evaluates the finite-sample consequence of the primary identification theory. Its selector receives only source counts and state–action labels; compatible kernels are used only for independent oracle evaluation.

Inputs, confidence rule, and comparisons. The study evaluates target confidence from two source count tables at fixed $\eta = 1$, for both g_{SA} and c, with true reference $\rho^b = (\rho^+ + \rho^-)/2$. The selector receives only counts, their state–action indexing, the score choice, and η. Generating source occupancies, kernels, and oracle targets enter simulation and evaluation only. Equal-tailed Clopper–Pearson intervals (Clopper & Pearson, 1934) cover the $2E$ source-coordinate and $2E$ conditional-action probabilities, each with error probability $0.05/(4E)$. Conditional intervals use the observed source/state count as denominator; a zero denominator gives $[0, 1]$. Both endogenous scores and the unknown reference are enclosed using these same intervals.

Write $\boldsymbol{a} = (a, 1-a)$. All methods select $v_{\boldsymbol{a}} = a\rho^+ + (1-a)\rho^-$ while checking its interval lower bound for nonnegativity. The signed method allows $a \in [-1, 2]$, the convex baseline restricts $a \in [0, 1]$, and the midpoint fixes $a = 1/2$. Bounded scalar minimization on each coefficient-sign region uses tolerance 10^{-9} and at most 100 iterations. Endpoints are inset by at most 10^{-10}, and every proposed coefficient is checked before reporting. The selected candidate has the smallest checked target radius $e_{\boldsymbol{a}} + \sqrt{2\epsilon_{\boldsymbol{a}}}$, where $e_{\boldsymbol{a}}$ bounds the empirical-centering error and $\epsilon_{\boldsymbol{a}}$ is the

interval-derived KL budget in Equation (26). All methods share counts and confidence endpoints. Nonpositive required support bounds or invalid numerical arithmetic return the empirical midpoint and trivial radius two; no state–action pairs are removed or assigned pseudocounts. All reported L_1 radii are capped at two.

Identified joint-target family: 12 geometries. For $(S, A) \in \{(2,2), (4,3), (8,3), (16,4)\}$, let $J = S/2$ be the number of equally weighted blocks, each with internal states $i = 0, 1$. Fix $\theta = 1/8$, $\gamma = 1/2$, and uniform initial state law. There is one action L and $A-1$ actions R_k. For each block j, $h \in (0,1)$, and $k = 1, \ldots, A-1$, define

$$\begin{aligned} \rho_h(j,0,L) &= \frac{(1/2+\theta)h}{J}, & \rho_h(j,1,L) &= \frac{(1/2-\theta)h}{J}, \\ \rho_h(j,0,R_k) &= \frac{(1/2-\theta)(1-h)}{J(A-1)}, & \rho_h(j,1,R_k) &= \frac{(1/2+\theta)(1-h)}{J(A-1)}. \end{aligned} \tag{46}$$

The next block is uniform. Its internal state is zero with probability $1/2 + 2\theta$ after L and $1/2 - 2\theta$ after any R_k. This positive kernel preserves the displayed occupancy family. For each $\zeta \in \{0.02, 0.10, 0.25\}$, take $\rho^+ = \rho_{1-\zeta}$ and $\rho^- = \rho_\zeta$. The joint tilt remains in their affine core at every finite positive η. All coordinates, including the rare actions, are retained.

Random source pairs: 18 geometries. Use $(S, A) \in \{(4,3), (8,3), (16,4)\}$, uniform-policy weights $u \in \{0.05, 0.25, 0.60\}$, and replicates $r \in \{0, 1\}$. For each shape, mixing level and replicate, draw independently, in order, kernel rows from $\mathrm{Dirichlet}(\mathbf{1}_S)$, an initial law $\mu_0 = \frac{1}{2}\,\mathrm{Unif}(\mathcal{S}) + \frac{1}{2}d$ with $d \sim \mathrm{Dirichlet}(\mathbf{1}_S)$, and two arrays of independent $\mathcal{N}(0, 2^2)$ policy logits. Each policy is $u\,\mathrm{Unif}(\mathcal{A}) + (1-u)\,\mathrm{softmax}(\mathrm{logits})$ within each state. At $\gamma = 0.9$, solving their discounted state-flow equations gives the two source occupancies. No source pair is selected by its target gap or confidence radius.

Five compatible kernels per geometry. Let $X = [\rho^+, \rho^-]$ and let P_0 be the generating kernel. For each additional kernel $k = 1, \ldots, 4$, draw a standard Gaussian coefficient vector independently in a numerical orthonormal basis of $\ker X^\top$, giving z, and an independent standard Gaussian state vector centered to give $\nu \perp \mathbf{1}_S$. Set

$$\Delta = z\nu^\top, \qquad t = 0.8 \min_{\Delta_{es} < 0} \frac{(P_0)_{es}}{|\Delta_{es}|}, \qquad P_k = P_0 + t\Delta. \tag{47}$$

The perturbation preserves both source flows and row sums, while each decreasing entry stays positive. All five kernels are checked for positivity, normalization, and source-flow residual at tolerance 10^{-9}. Geometry indices 0–11 enumerate the identified shapes and overlap levels in the displayed order; indices 12–29 enumerate random shapes, mixing levels, and replicates in that order, with the last index varying fastest. These kernels form an evaluation set for each source pair. Uniformity over the entire compatible class follows from Theorem 5.5.

Sampling and measured errors. Each source contributes $n \in \{10^3, 10^4, 10^5, 10^6\}$ independent observations, with 100 paired count datasets per geometry and size. We draw the multinomial sufficient statistics of independently geometrically stopped observations from Equation (1), without simulating full trajectories. Use independent sampling streams for each geometry, sample size and source. Both scores and all three methods share each dataset: there are 12,000 independent paired datasets and 72,000 correlated certificate evaluations. The sampling cost per dataset is $2n$ terminal observations.

The oracle evaluator uses each known simulation kernel to solve all $30 \times 5 \times 2 = 300$ true targets with the constrained-KL Newton solver. An independent check fits each target's KKT multiplier and verifies its flow; maximum residuals are 1.73×10^{-14} for stationarity and 6.84×10^{-14} for flow. For each empirical center $\widehat{v} = \widehat{v}_{\boldsymbol{a}}$, target error is the largest $\|\widehat{v} - \mathcal{T}_1^{P_k}(f)\|_1$ over the five kernels. The policy check normalizes $\widehat{v}$ within each state, using a uniform policy at a zero state marginal, and computes its actual occupancy under each kernel. Its largest target error is compared with the separate bound $2e_{\boldsymbol{a}}/(1-\gamma) + \sqrt{2\epsilon_{\boldsymbol{a}}}$, capped at two. Numerical miss checks allow tolerance 10^{-9}; these floating-point evaluations are not outward-rounded machine certificates.

Table 1: All 12 identified-family geometries; level denotes ζ. The four sample-size columns give median signed-affine joint-target radii for $n = 10^3, 10^4, 10^5, 10^6$ observations per source. The convex and conditional columns give, respectively, median convex joint-target and signed-affine conditional-target radii at $n = 10^6$. Support is the fraction of datasets with positive required support bounds at $n = 10^3$. Each entry uses 100 replications; radius two is trivial.

(S, A) / level	10^3	10^4	10^5	10^6	Convex 10^6	Cond. 10^6	Support 10^3
$(2, 2)$ / 0.02	2.000	1.642	0.408	0.103	1.458	0.770	1.00
$(2, 2)$ / 0.1	2.000	0.680	0.178	0.054	1.185	0.618	1.00
$(2, 2)$ / 0.25	2.000	0.583	0.170	0.052	0.753	0.394	1.00
$(4, 3)$ / 0.02	2.000	2.000	1.135	0.297	1.577	0.901	0.36
$(4, 3)$ / 0.1	2.000	2.000	0.481	0.129	1.251	0.679	1.00
$(4, 3)$ / 0.25	2.000	1.618	0.409	0.121	0.813	0.454	1.00
$(8, 3)$ / 0.02	2.000	2.000	2.000	0.507	1.713	1.083	0.00
$(8, 3)$ / 0.1	2.000	2.000	0.919	0.217	1.323	0.762	0.89
$(8, 3)$ / 0.25	2.000	2.000	0.714	0.200	0.879	0.526	1.00
$(16, 4)$ / 0.02	2.000	2.000	2.000	1.029	2.000	1.518	0.00
$(16, 4)$ / 0.1	2.000	2.000	1.974	0.432	1.490	0.913	0.00
$(16, 4)$ / 0.25	2.000	2.000	1.412	0.359	1.019	0.662	0.64

Table 2: All 18 random source-pair geometries, including both fixed replicates at every shape and policy-mixing weight u. Columns and sampling units follow Table 1; all unavailable and trivial outcomes are included in the medians.

(S, A) / level	10^3	10^4	10^5	10^6	Convex 10^6	Cond. 10^6	Support 10^3
$(4, 3)$ / 0.05 / r0	2.000	2.000	2.000	1.680	1.815	2.000	0.95
$(4, 3)$ / 0.05 / r1	2.000	2.000	1.530	1.181	1.230	1.404	1.00
$(4, 3)$ / 0.25 / r0	2.000	2.000	0.805	0.440	0.594	0.355	1.00
$(4, 3)$ / 0.25 / r1	2.000	2.000	1.782	1.421	1.865	0.845	1.00
$(4, 3)$ / 0.6 / r0	2.000	1.838	0.648	0.356	0.632	0.389	1.00
$(4, 3)$ / 0.6 / r1	2.000	1.904	0.576	0.291	0.428	0.326	1.00
$(8, 3)$ / 0.05 / r0	2.000	2.000	2.000	1.508	1.637	1.405	0.40
$(8, 3)$ / 0.05 / r1	2.000	2.000	2.000	2.000	2.000	2.000	0.42
$(8, 3)$ / 0.25 / r0	2.000	2.000	1.747	1.125	1.470	1.180	1.00
$(8, 3)$ / 0.25 / r1	2.000	2.000	1.933	1.339	1.570	1.177	1.00
$(8, 3)$ / 0.6 / r0	2.000	2.000	0.930	0.407	0.592	0.406	1.00
$(8, 3)$ / 0.6 / r1	2.000	2.000	0.986	0.459	0.647	0.451	1.00
$(16, 4)$ / 0.05 / r0	2.000	2.000	2.000	2.000	2.000	2.000	0.00
$(16, 4)$ / 0.05 / r1	2.000	2.000	2.000	2.000	2.000	2.000	0.00
$(16, 4)$ / 0.25 / r0	2.000	2.000	2.000	1.380	1.524	1.366	0.60
$(16, 4)$ / 0.25 / r1	2.000	2.000	2.000	1.382	1.548	1.308	0.47
$(16, 4)$ / 0.6 / r0	2.000	2.000	1.800	0.587	0.736	0.554	0.97
$(16, 4)$ / 0.6 / r1	2.000	2.000	1.761	0.560	0.668	0.531	1.00

Complete results and uncertainty. Tables 1 and 2 retain all 30 geometries. At $\zeta = 0.25$ and $n = 10^6$, all four identified-family cases meet the prespecified criterion: median signed target radius below 0.5 and at least 25% smaller than the matched convex radius. Across the full grid, 899 datasets have unavailable support and return trivial balls; none has a numerical-unavailable outcome. The simultaneous source-interval event fails in 405 of 12,000 datasets; those datasets remain in every applicable summary. There are zero tested target-bound or policy-bound misses in every setting. With 100 replications per setting, the exact two-sided 95% binomial interval for each observed zero miss rate is $[0, 0.0362]$, and its one-sided 95% upper bound is 0.0295. These are per-setting Monte Carlo intervals, not a simultaneous statement across all 720 settings.

All saved counts, 72,000 coefficient selections, target and policy errors, and 720 summaries are replayed independently. The complete record retains unavailable and uninformative outputs along with all oracle targets. The study measures fixed-η target confidence in simulated finite MDPs; uniform coverage over compatible dynamics follows from Theorem 5.5.

To distinguish statistical inflation from approximation slack, we additionally compare the finite radius with the population version of the same certificate and with the evaluated target spread at the same η (Section C.4). For the two-state, $\zeta = 0.25$ case at $n = 10^6$, the joint median radius is 0.0522 and median tested error is 0.00144; its population budget is zero. The conditional median radius is 0.3941, versus a population certificate of 0.3478 and a five-kernel half-diameter of 0.0181. Thus its numerical plateau must not be identified with the sharp ambiguity floor. All 60 score/geometry diagnostics include the midpoint control and paired radius-difference quantiles.

Table 3: Random-source confidence at $n = 10^6$ per source and $\eta = 1$. Entries are minimum–maximum across both fixed replicates at each (S, A, u), covering all 18 geometries. R_n includes every trivial outcome; R_{pop} removes source-estimation error from the same certificate. Ranges span process replicates, not confidence intervals.

		Joint score		Conditional score	
(S, A)	u	R_n	R_{pop}	R_n	R_{pop}
$(4, 3)$	0.05	1.181–1.680	1.036–1.489	1.404–2.000	1.268–1.886
$(4, 3)$	0.25	0.440–1.421	0.293–1.266	0.355–0.845	0.228–0.712
$(4, 3)$	0.60	0.291–0.356	0.175–0.234	0.326–0.389	0.223–0.278
$(8, 3)$	0.05	1.508–2.000	1.242–2.000	1.405–2.000	1.166–2.000
$(8, 3)$	0.25	1.125–1.339	0.885–1.104	1.177–1.180	0.957–0.962
$(8, 3)$	0.60	0.407–0.459	0.205–0.247	0.406–0.451	0.230–0.264
$(16, 4)$	0.05	2.000–2.000	1.631–2.000	2.000–2.000	1.712–2.000
$(16, 4)$	0.25	1.380–1.382	0.955–0.957	1.308–1.366	0.922–0.986
$(16, 4)$	0.60	0.560–0.587	0.173–0.194	0.531–0.554	0.179–0.193

C.3 Random-source confidence summary

Beyond the exactly identified family. In Table 3, R_n is each geometry's median radius over 100 datasets; R_{pop} evaluates the same signed-affine certificate with true sources and no centering error (Equation (48)). The latter is an oracle diagnostic, not an inference input. Both scores give informative balls outside the exact family. Bounds are smaller at larger mixing weights u in this grid, but each level redraws the kernel, initial law and policy logits: this is not a fixed-process mixing ablation. R_{pop} is a conservative upper bound, whereas the tested half-diameter $D_5/2$ is a lower bound on the radius needed to contain all compatible targets; neither specifies the exact ambiguity set.

C.4 Confidence-Radius Diagnostics

This descriptive analysis uses the same locked observations and targets, without new sampling, at $n = 10^6$ per source and $\eta = 1$. For each source geometry and score, define the population version of the same signed-coefficient certificate by

$$R_{\text{pop}} = \min\left\{2, \sqrt{2 \min_{a \in [-1,2]:\, v(a) \geq 0} \sum_e \frac{(v_e(a) - w_e)^2}{w_e}}\right\}, \qquad v(a) = \rho^- + a(\rho^+ - \rho^-). \quad (48)$$

This oracle diagnostic uses population source occupancies; it is not supplied to the inference routine. The minimization is a scalar convex quadratic. With $z = \rho^+ - \rho^-$, its unconstrained minimizer is $-\sum_e z_e(\rho_e^- - w_e)/w_e \,/\, \sum_e z_e^2/w_e$, clipped to the feasible interval. R_{pop} is a conservative certificate, not the sharp model-ambiguity radius. We also compute $D_5 = \max_{j,k} \|\mathcal{T}_1^{P_j}(f) - \mathcal{T}_1^{P_k}(f)\|_1$. Then $D_5/2$ is a lower bound on the radius of a ball containing all compatible targets, because it must contain these five. The evaluation set need not attain the full compatible diameter.

Tables 4 and 5 give every geometry/score combination. R_n is the median signed-affine reported radius; Error is the median largest target-center error across the five kernels. Midpoint reports its own same-data median radius. The final column reports the median and empirical 10th/90th percentiles of the within-dataset difference $\Delta R = R_{\text{affine}} - R_{\text{convex}}$; negative values favor signed coefficients. These are descriptive paired quantiles over 100 datasets, not confidence intervals. All trivial outcomes are included.

The population comparison separates the conservative certificate remaining without source-estimation error from its finite-data inflation. It does not decompose that population budget into sharp ambiguity and avoidable relaxation error. Neither small observed errors on five kernels nor the population budget proves that the finite confidence construction is optimal. Figure 6 plots all 60 finite radii against their population certificates and maps, for every geometry and sample size, the fraction of datasets whose joint ball is informative.

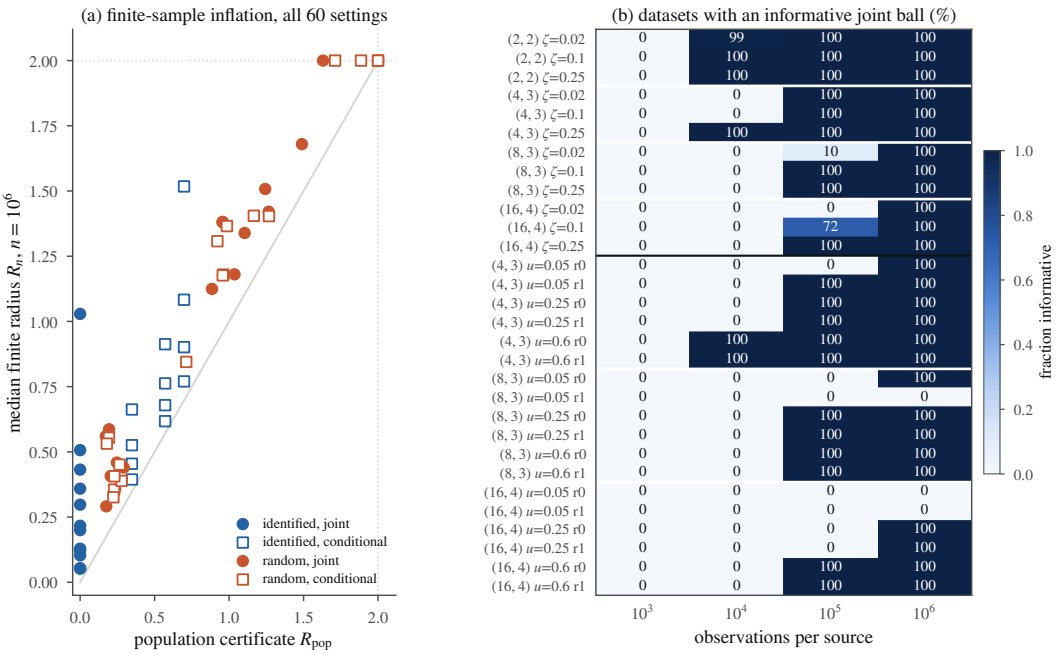


Figure 6: (a) Median signed-affine radius at $n = 10^6$ against the population certificate R_{pop} for all 30 geometries and both scores. Points lie above the diagonal by the finite-sample inflation; identified joint targets have $R_{\text{pop}} = 0$, and the dotted lines mark the trivial radius. (b) Percentage of the 100 datasets per cell whose joint signed-affine ball is nontrivial. Positive support bounds are one prerequisite for a useful ball; interval width and population mismatch also determine its radius. Every cell, including uninformative ones, is retained.

Table 4: Fixed-η tightness diagnostics for all identified-family settings. J/C denote joint/conditional. Radii and errors are rounded; half-diameters below 10^{-12} are displayed as zero. Paired brackets are empirical quantiles, not confidence intervals.

Sources	Score	R_n	Error	R_{pop}	$D_5/2$	Midpoint	Paired ΔR [10,90%]
(2,2)/0.02	J	0.103	0.001	0.000	0	2.000	-1.353 [-1.367,-1.341]
(2,2)/0.02	C	0.770	0.001	0.697	1.4e-04	2.000	-0.862 [-0.872,-0.853]
(2,2)/0.1	J	0.054	0.001	0.000	0	2.000	-1.131 [-1.134,-1.126]
(2,2)/0.1	C	0.618	0.004	0.571	0.003	2.000	-0.663 [-0.666,-0.660]
(2,2)/0.25	J	0.052	0.001	0.000	0	1.949	-0.700 [-0.702,-0.698]
(2,2)/0.25	C	0.394	0.024	0.348	0.018	1.888	-0.377 [-0.379,-0.374]
(4,3)/0.02	J	0.297	0.004	0.000	0	2.000	-1.278 [-1.297,-1.256]
(4,3)/0.02	C	0.901	0.004	0.697	3.6e-05	2.000	-0.881 [-0.892,-0.870]
(4,3)/0.1	J	0.129	0.002	0.000	0	2.000	-1.122 [-1.126,-1.118]
(4,3)/0.1	C	0.679	0.003	0.571	5.0e-04	2.000	-0.675 [-0.679,-0.672]
(4,3)/0.25	J	0.121	0.003	0.000	0	2.000	-0.692 [-0.693,-0.689]
(4,3)/0.25	C	0.454	0.007	0.348	0.003	1.970	-0.374 [-0.376,-0.371]
(8,3)/0.02	J	0.507	0.006	0.000	0	2.000	-1.206 [-1.225,-1.171]
(8,3)/0.02	C	1.083	0.006	0.697	1.4e-05	2.000	-0.869 [-0.882,-0.851]
(8,3)/0.1	J	0.217	0.003	0.000	0	2.000	-1.107 [-1.109,-1.101]
(8,3)/0.1	C	0.762	0.004	0.571	4.2e-04	2.000	-0.677 [-0.680,-0.673]
(8,3)/0.25	J	0.200	0.005	0.000	0	2.000	-0.679 [-0.681,-0.677]
(8,3)/0.25	C	0.526	0.009	0.348	0.001	2.000	-0.364 [-0.366,-0.362]
(16,4)/0.02	J	1.029	0.012	0.000	0	2.000	-0.971 [-1.009,-0.917]
(16,4)/0.02	C	1.518	0.012	0.697	5.8e-06	2.000	-0.482 [-0.513,-0.442]
(16,4)/0.1	J	0.432	0.012	0.000	0	2.000	-1.059 [-1.073,-1.040]
(16,4)/0.1	C	0.913	0.008	0.571	1.3e-04	2.000	-0.705 [-0.709,-0.700]
(16,4)/0.25	J	0.359	0.007	0.000	0	2.000	-0.661 [-0.663,-0.659]
(16,4)/0.25	C	0.662	0.020	0.348	7.1e-04	2.000	-0.357 [-0.359,-0.355]

C.5 SOURCE-CORE APPROXIMATION BUDGETS

For the canonical reversal pair and five retained random pairs, the source-only routine parametrizes the affine core as $v(a) = \rho^- + a(\rho^+ - \rho^-)$ on its full nonnegative interval. It minimizes $D_{\text{KL}}(v(a)\|w_{\eta,f})$

Table 5: The same diagnostics for all random source pairs. Source labels give (S, A), policy-mixing weight, and replicate. Large population budgets, trivial finite bounds, and all matched comparisons are retained.

Sources	Score	R_n	Error	R_{pop}	$D_5/2$	Midpoint	Paired ΔR [10,90%]
(4,3)/0.05/r0	J	1.680	0.610	1.489	0.005	2.000	-0.135 [-0.136,-0.134]
(4,3)/0.05/r0	C	2.000	0.670	1.886	0.005	2.000	0.000 [0.000,0.000]
(4,3)/0.05/r1	J	1.181	0.360	1.036	0.026	2.000	-0.049 [-0.050,-0.049]
(4,3)/0.05/r1	C	1.404	0.388	1.268	0.031	2.000	-0.028 [-0.029,-0.028]
(4,3)/0.25/r0	J	0.440	0.067	0.293	0.002	0.868	-0.155 [-0.156,-0.153]
(4,3)/0.25/r0	C	0.355	0.056	0.228	0.001	0.830	-0.177 [-0.180,-0.175]
(4,3)/0.25/r1	J	1.421	0.372	1.266	0.006	2.000	-0.444 [-0.447,-0.442]
(4,3)/0.25/r1	C	0.845	0.275	0.712	0.004	2.000	-0.477 [-0.480,-0.474]
(4,3)/0.6/r0	J	0.356	0.124	0.234	0.004	1.238	-0.276 [-0.278,-0.275]
(4,3)/0.6/r0	C	0.389	0.113	0.278	0.002	1.153	-0.196 [-0.198,-0.194]
(4,3)/0.6/r1	J	0.291	0.073	0.175	0.007	0.903	-0.137 [-0.138,-0.136]
(4,3)/0.6/r1	C	0.326	0.074	0.223	0.011	0.889	-0.116 [-0.117,-0.115]
(8,3)/0.05/r0	J	1.508	0.283	1.242	2.8e-04	2.000	-0.129 [-0.131,-0.127]
(8,3)/0.05/r0	C	1.405	0.265	1.166	2.9e-04	2.000	-0.132 [-0.135,-0.131]
(8,3)/0.05/r1	J	2.000	0.240	2.000	0.003	2.000	0.000 [0.000,0.000]
(8,3)/0.05/r1	C	2.000	0.245	2.000	0.003	2.000	0.000 [0.000,0.000]
(8,3)/0.25/r0	J	1.125	0.260	0.885	5.4e-04	2.000	-0.346 [-0.350,-0.343]
(8,3)/0.25/r0	C	1.180	0.242	0.962	7.0e-04	2.000	-0.295 [-0.298,-0.292]
(8,3)/0.25/r1	J	1.339	0.257	1.104	0.006	2.000	-0.231 [-0.233,-0.229]
(8,3)/0.25/r1	C	1.177	0.251	0.957	0.004	2.000	-0.231 [-0.234,-0.229]
(8,3)/0.6/r0	J	0.407	0.089	0.205	1.7e-04	1.106	-0.184 [-0.185,-0.183]
(8,3)/0.6/r0	C	0.406	0.075	0.230	1.3e-04	1.069	-0.160 [-0.161,-0.158]
(8,3)/0.6/r1	J	0.459	0.088	0.247	1.1e-04	1.101	-0.188 [-0.189,-0.186]
(8,3)/0.6/r1	C	0.451	0.084	0.264	7.2e-05	1.052	-0.157 [-0.159,-0.155]
(16,4)/0.05/r0	J	2.000	0.393	1.631	4.4e-05	2.000	0.000 [0.000,0.000]
(16,4)/0.05/r0	C	2.000	0.393	1.712	4.6e-05	2.000	0.000 [0.000,0.000]
(16,4)/0.05/r1	J	2.000	0.469	2.000	2.9e-04	2.000	0.000 [0.000,0.000]
(16,4)/0.05/r1	C	2.000	0.466	2.000	2.4e-04	2.000	0.000 [0.000,0.000]
(16,4)/0.25/r0	J	1.380	0.272	0.955	1.2e-04	2.000	-0.144 [-0.146,-0.142]
(16,4)/0.25/r0	C	1.366	0.266	0.986	1.4e-04	2.000	-0.136 [-0.138,-0.134]
(16,4)/0.25/r1	J	1.382	0.299	0.957	9.9e-05	2.000	-0.167 [-0.169,-0.164]
(16,4)/0.25/r1	C	1.308	0.299	0.922	8.3e-05	2.000	-0.169 [-0.172,-0.166]
(16,4)/0.6/r0	J	0.587	0.090	0.194	2.6e-05	1.200	-0.149 [-0.150,-0.148]
(16,4)/0.6/r0	C	0.554	0.089	0.193	5.2e-05	1.163	-0.148 [-0.149,-0.146]
(16,4)/0.6/r1	J	0.560	0.078	0.173	7.7e-05	1.054	-0.108 [-0.108,-0.107]
(16,4)/0.6/r1	C	0.531	0.078	0.179	7.8e-05	1.023	-0.108 [-0.109,-0.107]

by a scalar derivative root, using only the two sources, reference and score. The candidate is not restricted to a convex source mixture. The five random pairs are prespecified four-state, three-action examples with $\gamma = 0.9$, separate from the 18 random count-study geometries. For each state–action row, the next state around a ring is included among three distinct successors; the others are sampled uniformly without replacement. Independent $\mathrm{Unif}(0.25, 1.25)$ weights on those successors are normalized. The initial law normalizes independent $\mathrm{Unif}(0.1, 1.1)$ weights. Independent standard-normal arrays L, Z give two policies $\mathrm{softmax}(L \pm 0.6Z)$ within each state, each then mixed with weight 0.4 on the uniform action policy. Their discounted occupancies form the source pairs. All five prescribed draws are retained without selection by target error or budget size.

For each pair, the oracle check uses its generating kernel and four additional compatible kernels. Each additional kernel averages the original kernel with a compatible stochastic kernel obtained by minimizing a random linear objective over Equation (5). Each geometry and additional kernel uses an independent objective draw. Every kernel is checked for stochasticity and both source flow equations at tolerance 10^{-9}. These are selected compatible completions, not independent tasks or the whole compatible set. Strict kernel positivity is not required for the approximation bound.

Both scores and all $\eta \in \{0.25, 1, 2\}$ are evaluated: 180 true target solves and 36 geometry/score/η summaries. Every candidate flow, target flow, Pythagorean identity, target-distance bound and objective bound passes tolerance 10^{-8}. The largest identity residual is 8.46×10^{-15}. Table 6 reports every setting; the bound is uniform by the proof, not by the finite kernel grid. The data here are exact population laws and the comparison is at fixed η. No estimated-core or matched-radius guarantee is inferred.

Table 6: Source-core approximation: maximum observed error over five compatible kernels / source-only uniform bound. Objective columns refer to the regularized objective, not a task reward. All 36 settings are shown; the canonical joint path uses the exact zero errors and budgets proved in Corollary 5.2; raw floating-point residuals are retained.

Sources	η	Joint L_1	Joint objective	Conditional L_1	Conditional objective
Canonical	0.25	0 / 0	0 / 0	0.0080 / 0.0609	1.39e-04 / 7.42e-03
Canonical	1	0 / 0	0 / 0	0.0163 / 0.2462	2.81e-04 / 3.03e-02
Canonical	2	0 / 0	0 / 0	0.0055 / 0.4967	9.47e-05 / 6.17e-02
Design 1	0.25	0.0017 / 0.0040	7.17e-06 / 3.26e-05	0.0041 / 0.0139	5.45e-05 / 3.89e-04
Design 1	1	0.0258 / 0.0706	5.28e-04 / 2.49e-03	0.0304 / 0.0935	7.18e-04 / 4.37e-03
Design 1	2	0.0954 / 0.2400	3.77e-03 / 1.44e-02	0.1046 / 0.2739	4.35e-03 / 1.88e-02
Design 2	0.25	0.0048 / 0.0075	6.72e-05 / 1.13e-04	0.0065 / 0.0153	1.34e-04 / 4.68e-04
Design 2	1	0.0678 / 0.1201	4.34e-03 / 7.21e-03	0.0665 / 0.0931	3.91e-03 / 4.33e-03
Design 2	2	0.2150 / 0.4454	3.01e-02 / 4.96e-02	0.2110 / 0.3724	2.87e-02 / 3.47e-02
Design 3	0.25	0.0027 / 0.0040	2.60e-05 / 3.17e-05	0.0046 / 0.0134	5.85e-05 / 3.60e-04
Design 3	1	0.0513 / 0.0732	2.34e-03 / 2.68e-03	0.0488 / 0.0767	2.18e-03 / 2.94e-03
Design 3	2	0.1889 / 0.2750	1.63e-02 / 1.89e-02	0.1813 / 0.2656	1.52e-02 / 1.76e-02
Design 4	0.25	0.0044 / 0.0068	5.51e-05 / 9.36e-05	0.0047 / 0.0076	6.45e-05 / 1.16e-04
Design 4	1	0.0615 / 0.0945	2.70e-03 / 4.47e-03	0.0624 / 0.0962	2.78e-03 / 4.62e-03
Design 4	2	0.2178 / 0.3361	1.68e-02 / 2.82e-02	0.2186 / 0.3389	1.70e-02 / 2.87e-02
Design 5	0.25	0.0045 / 0.0082	7.78e-05 / 1.33e-04	0.0049 / 0.0090	9.09e-05 / 1.63e-04
Design 5	1	0.0553 / 0.1059	2.74e-03 / 5.60e-03	0.0487 / 0.0997	2.47e-03 / 4.97e-03
Design 5	2	0.1707 / 0.3766	1.48e-02 / 3.55e-02	0.1566 / 0.3588	1.38e-02 / 3.22e-02

C.6 Perturbation robustness of the identification gap

This exact-population study supports Section 6. It starts from the identified block family of Equation (46) at $\zeta = 0.25$ for $(S, A) \in \{(2,2), (4,3), (8,3)\}$; the two-state member is the pair of Equation (8). For each perturbation level δ and draw, both source policies are perturbed by $\pi_\delta(\mathbf{a} \mid \mathbf{s}) \propto \pi(\mathbf{a} \mid \mathbf{s}) e^{\delta \xi(\mathbf{s},\mathbf{a})}$ with independent standard normal ξ, and the perturbed occupancies are their discounted occupancies under the generating kernel and initial law, so the generating kernel remains compatible with both. All perturbed occupancies are strictly positive.

The full study uses $\delta \in \{0, 0.02, 0.05, 0.1, 0.2, 0.4\}$, with one unperturbed pair and 20 independent draws per positive level for each shape. For two states, the compatible class is the two-parameter family of kernels $Q_0 + uh^\top$ with $u \perp \{\rho^+, \rho^-\}$ and $h = (1, -1)/\sqrt{2}$. We evaluate a 41×41 coordinate grid over the polygon's bounding box, computed by linear programming, and keep kernels whose entries lie in $(10^{-6}, 1 - 10^{-6})$. This gives 1,031–1,513 kernels per pair across the full study. For $(4, 3)$ and $(8, 3)$, the generating kernel is joined by 200 kernels $Q_0 + t\, uh^\top$ with u a standard Gaussian vector in the orthogonal complement of the two sources, h a centered standard Gaussian state vector, and t a uniform fraction in $[0.2, 0.95]$ of the largest step that keeps every entry nonnegative. Every kernel is checked for strict positivity, row sums and both source flow equations at tolerance 10^{-9}; the largest flow violation is 2.5×10^{-16}.

Targets use the midpoint reference and both true-source scores at $\eta \in \{1/4, 1/2, 1, 2, 4\}$. All targets for one source pair, score and strength are solved together by batched Newton descent on the flow dual. With $A_P = B - \gamma P^\top$ and $b = (1 - \gamma)\mu_0$, the dual minimizes

$$D_P(\nu) = \sum_e \rho_e^b \exp\{\eta f_e + (A_P^\top \nu)_e - 1\} - b^\top \nu.$$

Its gradient is $A_P p_\nu - b$, where p_ν is the vector inside the sum, so the stopping residual directly checks the Bellman-flow equations. The solver uses double precision on a GPU and warm-started continuation in η; every solve reaches flow residual below 10^{-9}. Before the study, the same solver re-solves the locked 49-kernel, six-strength grid of Section C.1 and agrees with every stored occupancy within 1.41×10^{-12} in L_1. The evaluated diameter is the largest pairwise L_1 distance among targets on the same kernel set; it is a lower bound on the full-class diameter. The study has 303 source pairs and 3,030 target sets, of which 243 pairs and 2,430 sets belong to the $\delta \le 0.2$ neighborhood in the main text. Table 7 includes every perturbation level, and Figure 7 shows all strengths. At $\delta = 0.4$ and $\eta = 1$, the conditional diameter is larger in 42 of 60 pairs. Under the comparison reference $\rho^b = \rho^-$ at $\eta = 1$, every one of these pairs has joint tilt equal to ρ^+ to machine precision and a conditional tilt outside the core, as Corollary 5.3 states; for the pair of Equation (8) the conditional tilt is at Euclidean distance 0.1299 from the core.

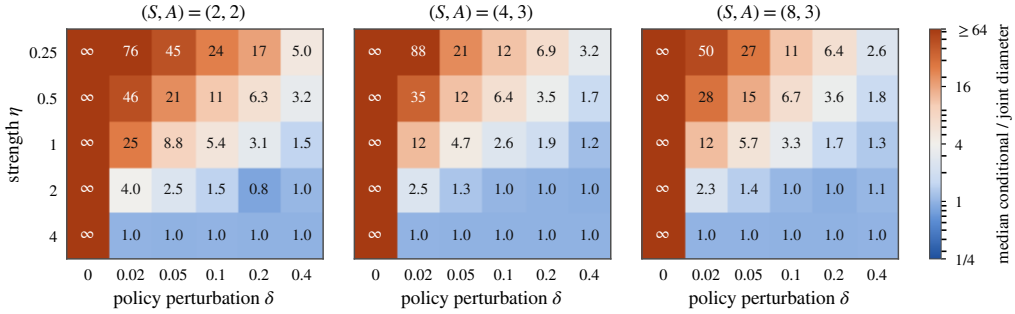


Figure 7: Median paired ratio of conditional to joint evaluated diameter over the perturbation draws, including every strength and perturbation level (∞: the unperturbed pair, whose joint diameter is zero). Each ratio uses the same compatible-kernel set for both scores. Labels give the measured ratios; color saturates at 64. Ratios are largest at small strength and small perturbation, and approach one at larger strengths.

Table 7: Median evaluated target-set L_1 diameter over 20 perturbation draws ($\delta = 0$ is the unperturbed pair), for both scores at three strengths, and the number of draws in which the conditional diameter exceeds the joint one at $\eta = 1$. Two-state rows use a compatible-kernel grid; the other rows use 201 compatible kernels. Every diameter is a lower bound on its full-class value.

		$\eta = 1/4$		$\eta = 1$		$\eta = 4$		
(S, A)	δ	Joint	Cond.	Joint	Cond.	Joint	Cond.	Cond. > Joint
$(2, 2)$	0	0.0000	0.0289	0.0000	0.0578	0.0000	0.0008	1/1
$(2, 2)$	0.02	0.0004	0.0288	0.0023	0.0579	0.0061	0.0063	20/20
$(2, 2)$	0.05	0.0006	0.0289	0.0068	0.0596	0.0140	0.0147	20/20
$(2, 2)$	0.1	0.0012	0.0281	0.0106	0.0553	0.0211	0.0207	20/20
$(2, 2)$	0.2	0.0018	0.0282	0.0213	0.0551	0.0480	0.0490	17/20
$(2, 2)$	0.4	0.0067	0.0273	0.0334	0.0641	0.0703	0.0683	13/20
$(4, 3)$	0	0.0000	0.0068	0.0000	0.0121	0.0000	0.0001	1/1
$(4, 3)$	0.02	0.0001	0.0076	0.0011	0.0132	0.0020	0.0020	20/20
$(4, 3)$	0.05	0.0004	0.0074	0.0027	0.0134	0.0052	0.0052	20/20
$(4, 3)$	0.1	0.0007	0.0082	0.0054	0.0149	0.0107	0.0107	20/20
$(4, 3)$	0.2	0.0013	0.0080	0.0095	0.0155	0.0191	0.0191	17/20
$(4, 3)$	0.4	0.0023	0.0081	0.0210	0.0236	0.0490	0.0458	12/20
$(8, 3)$	0	0.0000	0.0034	0.0000	0.0052	0.0000	0.0001	1/1
$(8, 3)$	0.02	0.0001	0.0035	0.0005	0.0062	0.0010	0.0010	20/20
$(8, 3)$	0.05	0.0001	0.0036	0.0011	0.0059	0.0021	0.0021	20/20
$(8, 3)$	0.1	0.0003	0.0036	0.0020	0.0064	0.0044	0.0044	20/20
$(8, 3)$	0.2	0.0006	0.0035	0.0042	0.0071	0.0089	0.0089	19/20
$(8, 3)$	0.4	0.0014	0.0043	0.0087	0.0108	0.0166	0.0168	17/20