WHEN DELAYED FEEDBACK CHANGES THE OBJECTIVE: BENCHMARK VALIDITY IN SEQUENTIAL DECISION MAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern recommender, advertising, and online experimentation systems often act several times before a consequential outcome matures. Delayed-learning theory usually treats this lag as an information constraint while taking the evaluation loss as given. We study the prior question: when does the route from delayed observations to action-level losses define a valid benchmark? We separate structural decision time from observation time and show that validity is governed by preservation of action gaps, not by delay length. For any realized action sequence, route-level performance transfers to the structural benchmark up to the cumulative distortion of action gaps. When structural margins are present, this gap-control certificate also recovers rankings and optimal actions; when persistent conflict crosses those margins, a learner may optimize the routed objective while accumulating structural regret. We further prove that fully unlabelled delayed observations can leave route validity non-identified even under adaptive data collection, motivating evidence-qualified audits rather than assumption-free corrections. Controlled experiments stress-test the predicted transfer and failure regimes; logged-data studies diagnose attribution sensitivity and illustrate the distinction between predictive fit and decision recovery on observed support. Reliable delayed-feedback learning therefore requires two certificates: optimization of the operational route and validation of the benchmark it induces.

1 INTRODUCTION

Sequential decision systems often act before outcomes mature. A recommender serves items before retention is measured; an advertising platform records a conversion after several exposures; and an online experiment adapts assignments before a delayed endpoint. Standard analyses treat this lag as a speed problem: the correct reward arrives late. But an interface must attribute, bind, or reconstruct the observation before scoring a decision. A fully observed outcome may then score the wrong action, letting the learner optimize its operational score while worsening the intended intervention. The prior problem is benchmark validity, not delay alone.

Regret is the standard language for sequential decisions (Auer et al., 2002; Bubeck & Cesa-Bianchi, 2012). Classical delayed-learning analyses fix a loss process and study learning under delayed, anonymous, aggregated, or partial observations (Joulani et al., 2013; Pike-Burke et al., 2018; Li et al., 2019; Howson et al., 2023; Wan et al., 2024). They apply once the loss is accepted as the target, but typically take as given the preceding link from an observed arrival to the decision it scores. We ask:

When does a loss map constructed from delayed observations preserve the decision-time comparisons of an independently specified structural target?

The question is sharpest when several interventions precede one outcome. Suppose a user sees advertisements at $t, t+1, t+2$ and converts at $t+3$. Arrival-time assignment credits the last exposure; a learned rule may distribute credit; source binding returns the outcome to its generating intervention. Despite sharing delay and predictive fit, these routes can induce different action rankings. Source labels establish factual correspondence for observed outcomes, not counterfactual losses of uncho-

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.

sen actions; a history proxy represents latent decision-time state. Delay determines when evidence arrives; the route determines which objective it creates.

We therefore separate structural decision time from observation time. At round t, state S_t and intervention A_t generate an outcome. An independently specified structural loss L_t^c defines contemporaneous intervention comparisons, while a route r maps delayed information to a routed loss $\widetilde{L}_t^r$. Structural causal regret and route regret evaluate the same realized action sequence under these two loss maps. This separates optimization of the routed objective from validity with respect to the structural benchmark.

Because decisions depend on relative action comparisons rather than absolute loss levels, route validity is governed by distortion of action gaps. We quantify this discrepancy by $\delta_t^r := \max_{a,b\in\mathcal{A}}|G_t^r(a,b) - G_t^c(a,b)|$, where G_t^c and G_t^r denote the structural and routed action gaps. Theorem 1 shows that, for every realized action sequence, cumulative action-gap distortion sharply controls the discrepancy between route and structural regret. Structural margins further convert gap control into ranking and choice recovery. Conversely, Theorem 2 shows that persistent, margin-separated conflict can make route and structural consistency incompatible, so vanishing route regret may coexist with linear structural regret.

Can delayed data tell us whether the route is valid? Theorem 3 constructs two structural worlds that induce the same fully unlabelled law under every common adaptive policy but have different route validity. Under the common route-greedy policy, they can also induce different structural regret. Thus validity can fail to be identified even with unlimited data from the observed interface. The response is an evidence contract specifying the source correspondences or structural comparisons actually observed, their supported comparison graph, and audit uncertainty. If the available evidence identifies the complete cardinal gap map, the audit bound and transfer theorem yield a structural-regret certificate.

Our contributions follow this sequence.

- **A benchmark-validity problem setting.** We separate structural target, operational route, and unavailable-state cost, then define cardinal, ordinal, and choice validity through action-gap geometry (Sections 3–4).
- **A transfer-to-conflict theory.** Writing $\mathfrak{A}_T^r := \sum_{t=1}^T \delta_t^r$, we establish the sharp pathwise bound $|R_T^c - R_T^r| \le \mathfrak{A}_T^r$. We further derive margin-based ranking and choice recovery, and characterize persistent conflict in which route optimization forces linear structural regret (Section 4).
- **An identification-to-audit boundary.** We establish adaptive-policy non-identification, derive audit guarantees under explicit evidence contracts, and probe their predictions in controlled and logged studies (Sections 5–6).

The paper follows the same progression. Section 2 positions the problem, and Section 3 separates target from route. Section 4 asks when performance transfers and when persistent conflict prevents it. Section 5 asks whether validity is knowable and turns sufficient evidence into an audit certificate. Section 6 probes each link; Sections 7 and 8 delimit the scope and return to the two-certificate principle. Proofs follow in the appendix.

2 Literature Review

Delayed learning versus benchmark validity. Delayed-feedback bandits quantify the statistical or adversarial price of observing a specified reward process late (Joulani et al., 2013; Vernade et al., 2017; Howson et al., 2023). Aggregated anonymous variants hide source identity while retaining the arm-generated reward process as the regret target (Pike-Burke et al., 2018; Cesa-Bianchi et al., 2022). We instead hold an independently specified structural target fixed and ask whether the route from arrivals to action-level losses preserves it. The distinction is orthogonal: once a route is valid, delayed-learning results may control its optimization error; without validity, faster optimization can reinforce the wrong benchmark.

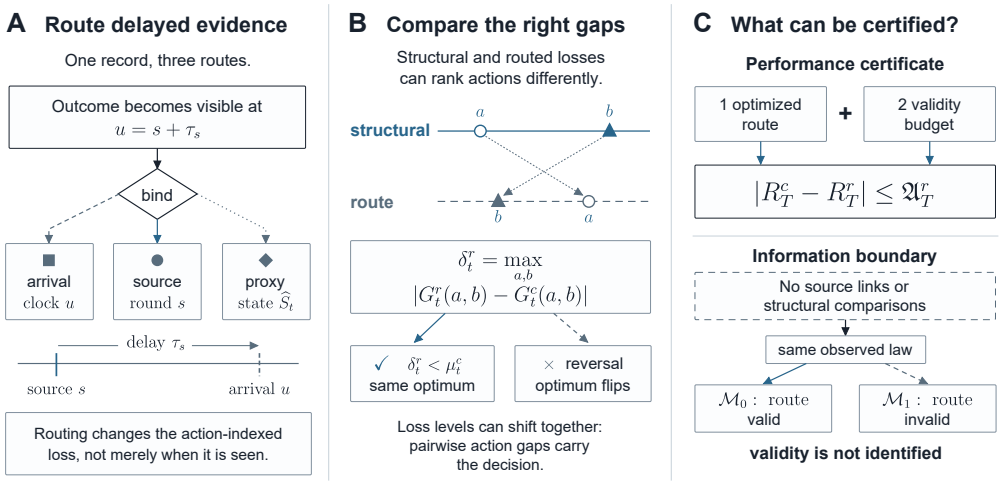

Figure 1: **The two-certificate view of delayed feedback.** Routing determines the operational objective induced by delayed evidence, while action-gap alignment determines whether optimizing that objective transfers to the structural benchmark. Without additional source or comparison evidence, validity may remain non-identified.

Observation mismatch and decision equivalence. Partial monitoring, corrupted rewards, and evolving observations separate incurred losses from feedback signals (Bartók et al., 2014; Everitt et al., 2017; Bar-On & Mansour, 2024). Reward shaping and STARC characterize transformations or pseudometrics that preserve policies or control downstream regret (Ng et al., 1999; Skalse et al., 2024). Our quotient geometry shares their decision-centered perspective but is local to the action map at each structural round and explicitly connects route distortion, regret transfer, and delayed source ambiguity.

Attribution, prediction, and identification. Credit-assignment and advertising-attribution methods redistribute delayed outcomes under model-specific assumptions (Arjona-Medina et al., 2019; Diemert et al., 2017; Yao et al., 2022; Applebaum et al., 2026). Decision-focused prediction shows that predictive fit and downstream loss are different objectives (Elmachtoub & Grigas, 2022). The closest studies respectively control regret by feedback accuracy (Bar-On & Mansour, 2024), learn from candidate attribution sets and priors (Applebaum et al., 2026), or identify delayed potential outcomes under a specified causal model (Zheng et al., 2026). These studies address complementary parts of the problem.

Research gap. We instead make preservation between a route-generated loss map and an external, state-conditioned action benchmark the primary object, and jointly study regret transfer, failure under persistent conflict, and the information needed to audit validity. Appendix B gives the full comparison.

3 Structural Target and Feedback Routes

Delay becomes a benchmark problem only after feedback is assigned to a decision. We therefore define the structural benchmark before the delayed route, separating benchmark validity from route optimization and unavailable-state cost from misrouting.

Fix a horizon T and $2 \le |\mathcal{A}| < \infty$.At round t, a possibly latent state S_t and disturbances define the potential outcomes $\{Y_t(a) : a \in \mathcal{A}\}$; the intervention A_t then realizes $Y_t = Y_t(A_t)$. The kernel $P_t^Y(dy \mid s, a)$ defines the decision-time loss

$$L_t^c(a) := \int \ell(a, y) P_t^Y(dy \mid S_t, a), \qquad r_t^c := L_t^c(A_t) - \min_a L_t^c(a), \quad R_T^c := \sum_{t=1}^{T} r_t^c.$$

This intended on-trajectory benchmark compares actions under the realized contemporaneous state. It is not policy regret or a long-horizon counterfactual value.

Outcome Y_t becomes observable at clock $t + \tau_t$. The interface may label all, some, or no arrivals, or expose only an aggregate. A prespecified route r maps the interface history available at its evaluation time to a complete evaluation vector $\widetilde{L}_t^r$, including attribution, proxy, empty-arrival, and fallback rules. Route regret is

$$r_t^r := \widetilde{L}_t^r(A_t) - \min_a \widetilde{L}_t^r(a), \qquad R_T^r := \sum_{t=1}^{T} r_t^r.$$

Any route quantity used to choose A_t must be measurable with respect to pre-action information $\mathcal{F}_t$. Admissible randomization satisfies

$$\mathbb{P}(A_t = a | \mathcal{F}_t \vee \sigma\{L_t^c(b) : b \in \mathcal{A}\}) = \pi_t(a \mid \mathcal{F}_t) \quad \text{a.s. for every } a \in \mathcal{A},$$

which rules out hidden coupling between policy randomization and the current structural loss vector beyond $\mathcal{F}_t$.

What the benchmark does and does not demand. Let $L_t^{\mathcal{F}}(a) = \mathbb{E}[L_t^c(a) \mid \mathcal{F}_t]$ and

$$\iota_t := \min_a L_t^{\mathcal{F}}(a) - \mathbb{E}[\min_a L_t^c(a) \mid \mathcal{F}_t] \geq 0.$$

For every admissible policy,

$$\mathbb{E}[r_t^c \mid \mathcal{F}_t] = \underbrace{\sum_a \pi_t(a \mid \mathcal{F}_t) L_t^{\mathcal{F}}(a) - \min_a L_t^{\mathcal{F}}(a)}_{\text{information-feasible decision regret}} + \underbrace{\iota_t}_{\text{unavailable-state cost}}.$$

Thus the state oracle need not be attainable: in the iid Bernoulli-state example of Appendix C.2, every admissible policy incurs expected regret $T/2$ even with immediate, correctly labelled feedback. This unavailable-state cost is distinct from route misalignment.

We assume a finite horizon and action set, a specified observation interface, and non-anticipating policies with fresh randomization. Delay is observation-only: it may depend on state, action, outcome, or history but does not alter outcomes, transitions, or feasible actions. Source binding, proxy sufficiency, and comparison coverage enter the validity and identification analysis below. Formal definitions and the decomposition proof appear in Appendices C and H. We next ask when route performance transfers to the structural benchmark.

4 Route Validity and Structural Regret

Having separated target and route, we organize validity around action-gap geometry. We first define cardinal, ordinal, and choice distortions and the cumulative alignment budget. This geometry yields a sharp pathwise regret-transfer certificate (Theorem 1). Structural margins then convert gap control into ranking and choice recovery. Persistent margin-separated disagreement gives the complementary failure regime (Theorem 2). We finally specialize the same geometry to arrival-time assignment and proxy reconstruction.

4.1 Decision geometry

We first encode route validity through the structural and route gaps $G_t^c(a, b) = L_t^c(a) - L_t^c(b)$ and $G_t^r(a, b) = \widetilde{L}_t^r(a) - \widetilde{L}_t^r(b)$. The cardinal defect and alignment budget are

$$\delta_t^r := \max_{a,b} |G_t^r(a, b) - G_t^c(a, b)|, \qquad \mathfrak{A}_T^r := \sum_{t=1}^{T} \delta_t^r.$$

Write $\mathcal{A}_t^{c,\star} := \arg\min_{a \in \mathcal{A}} L_t^c(a)$ and $\mathcal{A}_t^{r,\star} := \arg\min_{a \in \mathcal{A}} \widetilde{L}_t^r(a)$ for the structural and route-optimal action sets. If $e_t^r(a) = \widetilde{L}_t^r(a) - L_t^c(a)$, then $\delta_t^r = \max_a e_t^r(a) - \min_a e_t^r(a)$. Hence loss

maps live naturally in $\mathbb{R}^{|\mathcal{A}|}/\operatorname{span}\{\mathbf{1}\}$: arbitrary action-invariant shifts disappear, but a small action-dependent perturbation can reverse a near tie. We also record ordinal disagreement ρ_t^r (the fraction of pairwise signs that differ) and choice distortion χ_t^r (the fraction of route-optimal actions that are not structurally optimal). Exact cardinal validity implies ordinal validity, which implies choice validity; neither converse holds. Their horizon averages are $\overline{\rho}_T^r := T^{-1}\sum_{t=1}^T \rho_t^r$ and $\overline{\chi}_T^r := T^{-1}\sum_{t=1}^T \chi_t^r$.

Figure 2 summarizes this roadmap and its audit extension.

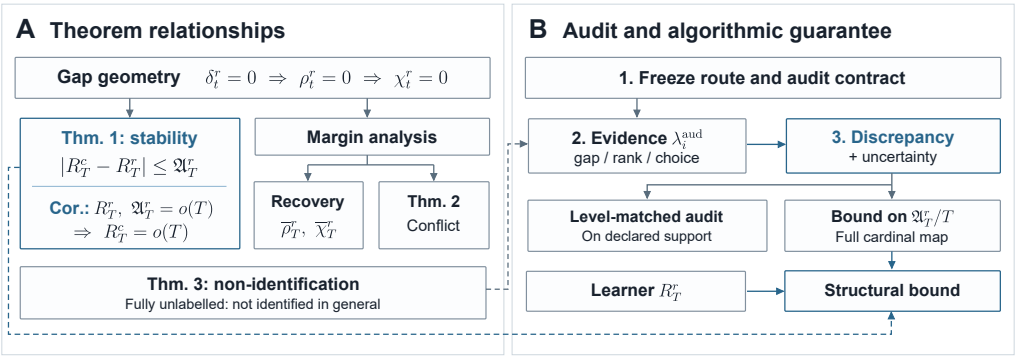

Figure 2: **Theoretical roadmap and audit pipeline.** Panel A links action-gap validity to regret stability, margin recovery, persistent conflict, and non-identification. Panel B shows how declared evidence and audit uncertainty yield level-matched validity certificates and, under full cardinal support, a structural-regret bound.

4.2 Sharp transfer, recovery, and failure

We begin with the pathwise route-to-structural transfer certificate.

Theorem 1 (Sharp route-to-structural regret stability)**.** *For every realized action sequence and every route-induced loss sequence,*

$$|r_t^c - r_t^r| \le \delta_t^r \quad (t = 1, \dots, T), \qquad \boxed{|R_T^c - R_T^r| \le \mathfrak{A}_T^r.} \tag{1}$$

The constant one is sharp.

The bound bridges route optimization and structural performance: $R_T^r = o(T)$ and $\mathfrak{A}_T^r = o(T)$ imply $R_T^c = o(T)$. Its proof bounds the regret difference by the oscillation of route error; two actions attain equality. This pathwise comparison scores one action sequence and does not equate trajectories trained under different routes. It is sufficient, not necessary: if $\widetilde{L}_t^r = \alpha_t L_t^c + \beta_t \mathbf{1}$ with $\alpha_t > 0$, optima agree and within-round regret rescales despite nonzero cardinal defect. Appendix Theorems D.4 and D.13 give the formal statement and witnesses.

Decision recovery additionally depends on structural margins. Let $\mu_t^c = \min_{a \notin \mathcal{A}_t^{c,\star}} [L_t^c(a) - \min_b L_t^c(b)]$, with $\mu_t^c = +\infty$ when every action is optimal. If $\delta_t^r < \mu_t^c$, every route optimum is structurally optimal. More generally, for $\varepsilon > 0$,

$$\overline{\rho}_T^r \le H_{T,\mathrm{ord}}(\varepsilon) + \frac{\mathfrak{A}_T^r}{\varepsilon T}, \qquad \overline{\chi}_T^r \le H_{T,\mathrm{ch}}(\varepsilon) + \frac{\mathfrak{A}_T^r}{\varepsilon T},$$

where, letting $\mathcal{P}$ contain each unordered action pair once,

$$H_{T,\mathrm{ord}}(\varepsilon) := \frac{1}{T|\mathcal{P}|}\sum_{t=1}^T \sum_{(a,b)\in\mathcal{P}} \mathbf{1}\{|G_t^c(a,b)| \le \varepsilon\}, \qquad H_{T,\mathrm{ch}}(\varepsilon) := \frac{1}{T}\sum_{t=1}^T \mathbf{1}\{\mu_t^c \le \varepsilon\}$$

are respectively the fractions of small-gap action pairs and small-margin rounds. If $H_{T,\ell}(\varepsilon) \le C_\ell \varepsilon^{\kappa_\ell}$ uniformly in T, optimizing ε yields the rate $O[(\mathfrak{A}_T^r/T)^{\kappa_\ell/(\kappa_\ell+1)}]$, and a horizon-indexed construction attains this exponent. Thus average gap error alone gives no recovery rate when structural margins collapse.

We next turn to persistent margin crossings, where route and structural optima are separated. For each round on which these sets are disjoint, define $\gamma_t^r := \min_{b \in \mathcal{A}_t^{r,\star}} [L_t^c(b) - \min_a L_t^c(a)]$ and $\eta_t^r := \min_{a \notin \mathcal{A}_t^{r,\star}} [\widetilde{L}_t^r(a) - \min_b \widetilde{L}_t^r(b)]$.

Theorem 2 (Persistent conflict). *Let $\mathcal{T}_0 \subseteq \{t : \mathcal{A}_t^{r,\star} \cap \mathcal{A}_t^{c,\star} = \varnothing\}$. If $\gamma_t^r \geq \gamma > 0$ and $\eta_t^r \geq \eta > 0$ for every $t \in \mathcal{T}_0$, then every realized action sequence obeys*

$$\frac{R_T^c}{\gamma} + \frac{R_T^r}{\eta} \geq |\mathcal{T}_0|. \tag{2}$$

Consequently, if $|\mathcal{T}_0| = \Omega(T)$ and $R_T^r = o(T)$, then $R_T^c = \Omega(T)$.

This is the negative counterpart to transfer. On each conflict round, an action is either route-optimal and pays at least γ structurally, or is not route-optimal and pays at least η operationally. Hence linear conflict and sublinear route regret entail linear structural regret: optimizing the operational benchmark can systematically fail on the intended target. The full result is Theorem D.11.

4.3 Validity levels and evidence requirements

Validity strength and evidence strength are distinct. Cardinal validity supports the regret-transfer certificate, while ordinal and choice validity preserve progressively coarser decision objects; the cardinal condition is sufficient, not necessary, for structural-regret transfer. Evidence strength instead determines what can be certified: source correspondence need not reveal unchosen-action comparisons, and calibration does not strengthen identification. Section 5 formalizes these requirements.

4.4 Delayed-feedback specializations

For arrival-time and proxy routes, let $L_t^c(a) = L^c(a; S_t)$, define the decision-relevant pseudometric

$$d_{\mathrm{dec}}(s, s') := \max_{a,b \in \mathcal{A}} \left| \left[L^c(a; s) - L^c(b; s)\right] - \left[L^c(a; s') - L^c(b; s')\right] \right|.$$

For a single-source arrival route that credits source round $\kappa(t) \leq t$ at clock t, the route map is $L^c(\cdot; S_{\kappa(t)})$, so $\delta_t^{\mathrm{arr}} = d_{\mathrm{dec}}(S_t, S_{\kappa(t)})$. Exact cardinal validity therefore holds exactly when the source-to-arrival change is action invariant, equivalently when $d_{\mathrm{dec}}(S_t, S_{\kappa(t)}) = 0$. If $d_{\mathrm{dec}}(s, s') \leq L_\Psi \|s - s'\|$, then $\delta_t^{\mathrm{arr}} \leq L_\Psi \|S_t - S_{\kappa(t)}\|$; elapsed delay alone is thus not the relevant quantity. For an implementable proxy route, let $\widehat{S}_t = \psi_t(\mathcal{F}_t)$ be a pre-action proxy and let $\widetilde{L}_t^r(a) = \widehat{L}(a; \widehat{S}_t)$ be its fitted action-level map. With $\mathrm{osc}_{a \in \mathcal{A}}\, z(a) := \max_{a \in \mathcal{A}} z(a) - \min_{a \in \mathcal{A}} z(a)$, the route satisfies

$$\delta_t^r \leq d_{\mathrm{dec}}(S_t, \widehat{S}_t) + \mathrm{osc}_{a \in \mathcal{A}} \{\widehat{L}(a; \widehat{S}_t) - L^c(a; \widehat{S}_t)\}.$$

The right-hand side separates proxy-state and fitted-kernel gap errors; their underlying actionwise gap errors can offset, so it is an upper bound rather than an exact decomposition. Source binding and state reconstruction solve different problems. Complete results and proofs appear in Appendices D and H. One question remains: can the observed interface identify route validity?

5 What the Interface Can Identify and Audit

Section 4 established how route alignment controls structural regret; here we ask whether the observation interface can identify or certify that alignment. We first show that fully unlabelled feedback can leave route validity non-identified even under adaptive data collection (Theorem 3). This information boundary makes evidence strength part of the audit target: cardinal, ordinal, and choice evidence support different claims. We then bound finite-sample audit error under a fixed evidence contract (Proposition E.10) and show that only a complete cardinal reference over the structural horizon yields a structural-regret certificate (Corollary E.11).

5.1 An adaptive-policy information boundary

We first establish this information boundary.

Theorem 3 (Adaptive-policy non-identification under fully unlabelled feedback)**.** *For every even horizon $T = 2m \geq 2$, under fully unlabelled feedback there exist two structural worlds $\mathcal{M}_0, \mathcal{M}_1$ and a fixed clock-only route r such that, for every policy admissible in both worlds,*

$$P^{\pi, I_{\mathrm{unlab}}}_{\mathcal{M}_0} = P^{\pi, I_{\mathrm{unlab}}}_{\mathcal{M}_1}, \qquad \mathsf{Val}^r_T(\mathcal{M}_0) \neq \mathsf{Val}^r_T(\mathcal{M}_1).$$

Hence no estimator based only on this interface can consistently determine route validity. Here Val^r_T is the horizon-level validity profile defined in Section 4, consisting of cardinal gap distortion, ordinal disagreement, and choice distortion. In particular, a two-action alternating-state construction gives

$$\mathsf{Val}^r_T(\mathcal{M}_0) = (0, 0, 0), \qquad \mathsf{Val}^r_T(\mathcal{M}_1) = (2, 1, 1).$$

The theorem separates identification from estimation. In the witness construction, alternating latent states and action-independent outcomes produce identical observable laws under adaptive histories. A route-optimal learner has $R^r_T = 0$ in both worlds, while R^c_T is 0 in one and T in the other. Operational success therefore does not identify structural success. For any scalar functional separated by Δ_Θ, the two-world minimax absolute-error risk is $\Delta_\Theta/2$ when source labels are fully absent; retaining each of nT source labels independently with probability p_{src} changes it, in this construction, to $\frac{\Delta_\Theta}{2}(1 - p_{\mathrm{src}})^{nT}$. This interpolation is construction-specific. The full witness and coarsening result appear in Appendix H.6.

5.2 Evidence-adaptive audit and structural certification

We therefore make the evidence level part of the estimand. Let $\mathcal{I}_M = \{1, \ldots, M\}$ be a finite population of audit units and $\mathcal{P}^{\mathrm{aud}}_i \subseteq \mathcal{P}$ the supported, canonically oriented action pairs for unit i. Let $\lambda^{\mathrm{aud}}_i \in \{\mathrm{cardinal}, \mathrm{ordinal}, \mathrm{choice}, \mathrm{unsupported}\}$ be the strongest level supported by the validation design. For an audited pair, let $\Delta^r_i(a, b)$ denote the route-induced action gap, the unit-level analogue of $G^r_t(a, b)$ in Section 4, and let $\phi^{a,b}_i$ denote the corresponding source-grounded validation gap. The corresponding audit-side discrepancies are

$$d^{r,\mathrm{obs}}_{i,\delta} := \max_{(a,b)\in\mathcal{P}^{\mathrm{aud}}_i} \left|\Delta^r_i(a, b) - \phi^{a,b}_i\right|, \qquad \lambda^{\mathrm{aud}}_i = \mathrm{cardinal},$$

$$d^{r,\mathrm{obs}}_{i,\rho} := \frac{1}{|\mathcal{P}^{\mathrm{aud}}_i|} \sum_{(a,b)\in\mathcal{P}^{\mathrm{aud}}_i} \mathbf{1}\Big\{\operatorname{sgn}\Delta^r_i(a, b) \neq \operatorname{sgn}\phi^{a,b}_i\Big\}, \quad \lambda^{\mathrm{aud}}_i \in \{\mathrm{cardinal}, \mathrm{ordinal}\},$$

$$d^{r,\mathrm{obs}}_{i,\chi} := \frac{|\mathcal{A}^{r,\star}_i \setminus \mathcal{A}^{\mathrm{val},\star}_i|}{|\mathcal{A}^{r,\star}_i|}, \qquad \lambda^{\mathrm{aud}}_i \in \{\mathrm{cardinal}, \mathrm{ordinal}, \mathrm{choice}\}.$$

The choice contribution requires both optimal-action sets to be identified. These discrepancies match the cardinal, ordinal, and choice levels of Section 4 only on their declared support; ordinal evidence supplies no numerical gaps, and choice evidence alone yields no structural-regret certificate.

The evidence contract fixes the route, supported population, comparison coverage, reference construction, and sampling design before sampling. Exact compatible cardinal gaps identify the complete gap map on a connected comparison graph; ordinal and choice identification require uniqueness of the corresponding decision object. Sampling, reference-quality, and unsupported-mass uncertainties remain separate.

For a fixed supported audit level $\ell \in \{\delta, \rho, \chi\}$, relabel the supported units as $1, \ldots, M$ and define $D^{r,\mathrm{obs}}_{M,\ell} := M^{-1}\sum_i d^{r,\mathrm{obs}}_{i,\ell}$ and $\widehat{D}^{r,\mathrm{obs}}_{M,\ell} := n^{-1}\sum_{i\in\mathcal{S}_{\mathrm{aud}}} d^{r,\mathrm{obs}}_{i,\ell}$. If a simple random sample without replacement of size n is drawn and $0 \leq d^{r,\mathrm{obs}}_{i,\ell} \leq B_\ell$, then with probability at least $1 - \alpha$,

$$D^{r,\mathrm{obs}}_{M,\ell} \leq \widehat{D}^{r,\mathrm{obs}}_{M,\ell} + B_\ell\sqrt{\frac{\log(2/\alpha)}{2n}}.$$

Here $B_\rho = B_\chi = 1$; cardinal boundedness is an audit condition, not a global assumption of the structural theory. Unequal-probability designs, reference-quality guarantees, and calibration details are given in Appendix E.

Only a complete cardinal reference over the structural horizon can turn this audit into a structural-regret certificate. Under a frozen reference, simple random sampling, and a full-map reference

error bound $\overline{V}^{\mathrm{full}}_{\delta,\beta}$ holding with probability at least $1-\beta$, let $\widehat{D}^{r,\mathrm{full,obs}}_{T,\delta}$ be the sampled mean of the complete-map cardinal discrepancy. Then, for $\alpha, \beta \in (0,1)$ with $\alpha+\beta<1$, with probability at least $1-\alpha-\beta$,

$$R^c_T \le R^r_T + T\left[\widehat{D}^{r,\mathrm{full,obs}}_{T,\delta} + B^{\mathrm{full}}_{\delta}\sqrt{\frac{\log(2/\alpha)}{2n}} + \overline{V}^{\mathrm{full}}_{\delta,\beta}\right].$$

The bracket also bounds $\mathfrak{A}^r_T/T$; covered-pair or ordinal-only evidence cannot replace this complete cardinal object. Appendix E gives formal definitions, reference-quality bridges, and proofs.

6 Theory-Indexed Evidence

Controlled complete-map experiments probe regret transfer, margin recovery, and persistent conflict; a Criteo delayed-conversion log (Diemert et al., 2017) examines ambiguity among source-time routes when multiple candidate sources exist; and a controlled audit experiment separates information that changes the route from evidence used to audit a fixed route. Complete experimental designs, data processing, evaluation metrics, and additional diagnostics are reported in Appendix F and Appendix I.

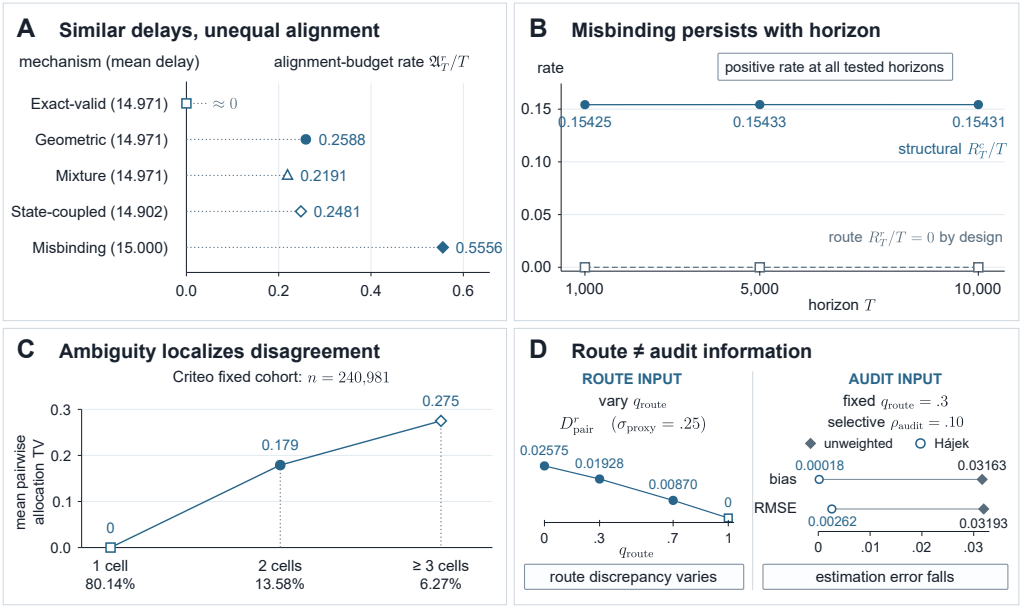

Figure 3: **Theory-indexed empirical evidence.** (a) Controlled complete maps show that similar delays can induce different action-gap distortions, including an exactly valid delayed shift. (b) Under systematic misbinding, route regret is zero while the structural-regret rate remains positive across horizons, exhibiting the persistent-conflict regime. (c) In the Criteo delayed-conversion log, disagreement among source-time routes concentrates on journeys with multiple candidate source cells. (d) Route information changes the underlying discrepancy, whereas audit weighting changes estimation accuracy for a fixed route.

Controlled validity and structural failure. Figure 3a–b compares six mechanisms under fixed learners or route constructions, varying only how delayed outcomes map to action evaluations. The exact-valid shift has mean delay 14.9709 but zero alignment-budget rate and route-greedy structural regret; three stochastic-delay mechanisms with nearly identical mean delay yield rates 0.2588, 0.2191, and 0.2481. A matched perturbation sweep confirms action-invariant cancellation to machine precision, with choice error appearing at the predicted boundary $\delta^T_t/\mu^c_t = 1$. Under systematic misbinding, route regret is zero, complete conflict occurs on roughly half the rounds, and structural-regret rate stays near 0.1543 for $T \in \{10^3, 5\times10^3, 10^4\}$. These paths realize the transfer and conflict regimes in Equations (1)–(2).

Natural delayed-conversion log. Figure 3c evaluates four source-time routes on the same retained Criteo conversion journeys: first-click-or-touch, last-click-or-touch, linear, and time-decay. Relative to arrival-time accounting, allocation total variation ranges from 0.103 to 0.138 and Kendall's τ_b from 0.482 to 0.589. The routes agree on singleton-candidate journeys (80.14% of the cohort), while mean pairwise allocation variation rises to 0.179 with two candidate cells and 0.275 with three or more. Thus, the source-assignment ambiguity in Section 3 is visible in logged data: the same delayed conversion log supports different route-induced allocations and rankings.

Route information versus audit information. Figure 3d tests the distinction in Section 5 between information that changes the route and evidence used to audit a fixed route. At $\sigma_{\text{proxy}} = .25$, increasing route-label retention q_{route} changes the route and lowers mean pairwise discrepancy from 0.02575 to zero. Holding the route fixed at $q_{\text{route}} = .3$ and auditing only 10% of the evidence, ambiguity-selective unweighted estimation has bias/RMSE $0.03163/0.03193$, whereas Hájek weighting with known inclusion probabilities reduces them to $0.00018/0.00262$. Thus, better audit estimation does not itself establish structural validity; obtaining the additive structural-regret certificate through the alignment budget requires control of the worst-pair cardinal gap distortion characterized by the theory.

A semi-synthetic KuaiRand-1K study (Gao et al., 2022) in Appendix F.4 shows that lower predictive MAE need not yield better decision recovery on common logged support: Ridge has lower MAE but worse gap, sign, and top-action recovery, with the same direction appearing in most user resamples.

7 Discussion

Delay, predictive fit, and route regret characterize different properties of the feedback pipeline; none alone establishes benchmark validity. Routes with similar delays can induce different action-gap distortions, while a well-optimized route objective can remain persistently misaligned with the structural target. What matters for benchmark validity is whether routing preserves the relevant action comparisons. Learning performance and benchmark validity are therefore distinct properties of the same feedback pipeline.

Accordingly, evaluating a delayed-feedback system requires two kinds of evidence. Optimization evidence measures how well the learner optimizes a given route objective, whereas validity evidence measures how well that objective preserves the comparisons required by the structural benchmark. Validity claims of different strength require evidence at corresponding levels: supported pairwise evidence may justify local gap, ranking, or choice claims, whereas the additive structural-regret certificate through the alignment budget requires complete cardinal comparison evidence. More accurate auditing can improve estimation of supported objects, but cannot upgrade weaker evidence into stronger structural identification.

Our analysis considers a fixed operational route under observation-only delay and evaluates its agreement with an on-trajectory, decision-time benchmark. When the route or audit process becomes adaptive, validity analysis becomes a joint problem of route construction, information acquisition, and sequential decision making; when delay also changes system transitions or feasible actions, the structural decision problem itself changes. These cases motivate validity-aware route design and adaptive auditing.

8 Conclusion

Delayed feedback does not merely postpone evidence: once observations are routed, they may change the action comparisons being optimized. We separate benchmark validity from unavailable state information, represent it through action-gap geometry, and obtain a sharp transfer certificate controlled by cumulative gap distortion. Persistent conflict gives the failure regime, where vanishing route regret can coexist with linear structural regret. Observational equivalence then shows that structural validity itself may be unknowable from unlabelled feedback alone. A complete-cardinal, comparison-identifying audit closes the chain by turning observed discrepancy and its sampling and reference uncertainty into a structural-regret certificate. The resulting principle is simple but demanding: optimize the route, and independently validate the objective it creates.

AI USE STATEMENT

During the preparation of this submission, the authors used AI-based tools to aid language editing, manuscript organization, coding, and formal-proof scaffolding. Every retained claim, proof, and code contribution remains the authors' responsibility. No AI system is listed as an author.

REPRODUCIBILITY STATEMENT

The manuscript and appendix provide formal definitions, assumptions, proofs, and experiment-specific protocols. An anonymized code archive is supplied as supplementary material, containing experiment implementations, configurations, random seeds, environment specifications, frozen derived results, and analysis and per-experiment plotting scripts. Execution instructions and external data requirements are documented in the archive. Experiments 1 and 4 use controlled simulations. Experiment 2 uses the Criteo delayed-conversion log (Diemert et al., 2017), and Experiment 3 uses KuaiRand-1K (Gao et al., 2022); reproducing these logged-data studies requires obtaining the external datasets under their original release conditions.

ETHICS STATEMENT

Experiments 2 and 3 use previously released datasets, and no new human-subject data were collected for this work. The analyses report aggregate statistics and do not seek to identify individual users. Dataset access and use followed the conditions associated with the original dataset releases.

REFERENCES

Nan An, Weian Li, Qi Qi, Changyuan Yu, and Liang Zhang. Beyond last-click: An optimal mechanism for ad attribution. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 167544–167562. Curran Associates, Inc., 2025. doi: 10.52202/085713-5583.

Lorne Applebaum, Robert Busa-Fekete, August Chen, Claudio Gentile, Tomer Koren, and Aryan Mokhtari. Statistical learning from attribution sets. In *Proceedings of Thirty Ninth Conference on Learning Theory*, June 2026.

Jose A. Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. RUDDER: Return decomposition for delayed rewards. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, May 2002. ISSN 1573-0565. doi: 10.1023/A:1013689704352.

Yogev Bar-On and Yishay Mansour. Non-stochastic bandits with evolving observations, 2024. URL https://arxiv.org/abs/2405.16843.

Gábor Bartók, Dean P. Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, November 2014. ISSN 0364-765X. doi: 10.1287/moor.2014.0663.

Ron Berman. Beyond the last touch: Attribution in online advertising. *Marketing Science*, 37(5):771–792, September 2018. ISSN 0732-2399. doi: 10.1287/mksc.2018.1104.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, December 2012. ISSN 1935-8237. doi: 10.1561/2200000024.

N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Regret minimization under partial monitoring. In *2006 IEEE Information Theory Workshop - ITW '06 Punta del Este*, pp. 72–76, 2006. doi: 10.1109/ITW.2006.1633784.

Nicolò Cesa-Bianchi, Tommaso Cesari, Roberto Colomboni, Claudio Gentile, and Yishay Mansour. Nonstochastic bandits with composite anonymous feedback. *Journal of Machine Learning Research*, 23(277):1–24, 2022.

Stephen R. Cole and Miguel A. Hernán. Constructing inverse probability weights for marginal structural models., 2008. URL https://api.semanticscholar.org/CorpusID:8121173.

Brian Dalessandro, Claudia Perlich, Ori Stitelman, and Foster Provost. Causally motivated attribution for online advertising. In *ACM Conferences*, pp. 1–9, Beijing China, August 2012. ACM. ISBN 978-1-4503-1545-6. doi: 10.1145/2351356.2351363.

Eustache Diemert, Julien Meynet, Pierre Galland, and Damien Lefortier. Attribution modeling increases efficiency of bidding in display advertising. In *Proceedings of the ADKDD'17*, ADKDD'17, pp. 1–6, New York, NY, USA, August 2017. Association for Computing Machinery. ISBN 978-1-4503-5194-2. doi: 10.1145/3124749.3124752.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68 (1):9–26, January 2022. ISSN 0025-1909. doi: 10.1287/mnsc.2020.3922.

Tom Everitt, Victoria Krakovna, Laurent Orseau, and Shane Legg. Reinforcement learning with a corrupted reward channel. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pp. 4705–4713, Melbourne, Australia, August 2017. AAAI Press. ISBN 978-0-9992411-0-3.

Chongming Gao, Shijun Li, Yuan Zhang, Jiawei Chen, Biao Li, Wenqiang Lei, Peng Jiang, and Xiangnan He. KuaiRand: An unbiased sequential recommendation dataset with randomly exposed videos. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, pp. 3953–3957, New York, NY, USA, October 2022. Association for Computing Machinery. ISBN 978-1-4503-9236-5. doi: 10.1145/3511808.3557624.

Jaroslav Hajek. Asymptotic Theory of Rejective Sampling with Varying Probabilities from a Finite Population. *The Annals of Mathematical Statistics*, 35(4):1491 – 1523, 1964. doi: 10.1214/aoms/1177700375. URL https://doi.org/10.1214/aoms/1177700375.

Anna Harutyunyan, Will Dabney, Thomas Mesnard, Mohammad Gheshlaghi Azar, Bilal Piot, Nicolas Heess, Hado van Hasselt, Gregory Wayne, Satinder Singh, Doina Precup, and Remi Munos. Hindsight credit assignment. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963. ISSN 0162-1459. doi: 10.1080/01621459.1963.10500830.

Benjamin Howson, Ciara Pike-Burke, and Sarah Filippi. Delayed feedback in generalised linear bandits revisited. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, April 2023.

Sheng Ji. Inventory-constrained online learning for revenue management with delayed feedback. *Transportation Research Part E: Logistics and Transportation Review*, 208:104649, April 2026. ISSN 1366-5545. doi: 10.1016/j.tre.2025.104649.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvari. Online learning under delayed feedback. In *Proceedings of the 30th International Conference on Machine Learning*, May 2013.

Jacek Karwowski, Oliver Hayman, Xingjian Bai, Klaus Kiendlhofer, Charlie Griffin, and Joar Skalse. Goodharts law in reinforcement learning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 24546–24576, 2024.

Sang Su Lee, Vineeth Loganathan, Shishir Dash, and Vijay Raghavan. When offline evaluation misleads: A diagnostic protocol for reward and policy selection in delayed-feedback contextual bandits, 2026. URL https://arxiv.org/abs/2608.11560.

Bingcong Li, Tianyi Chen, and Georgios B. Giannakis. Bandit online learning with unknown delays. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, April 2019.

Lingling Li, Changyu Shen, Xiaochun Li, and James M Robins. On weighting approaches for missing data. *Statistical Methods in Medical Research*, 22(1):14–30, February 2013. ISSN 0962-2802. doi: 10.1177/0962280211403597.

Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pp. 278–287, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1558606122.

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed, aggregated anonymous feedback. In *Proceedings of the 35th International Conference on Machine Learning*, July 2018.

Joar Skalse and Alessandro Abate. Partial identifiability and misspecification in inverse reinforcement learning. *Artificial Intelligence*, 356:104525, July 2026. ISSN 0004-3702. doi: 10.1016/j.artint.2026.104525.

Joar Skalse, Lucy Farnik, Sumeet Ramesh Motwani, Erik Jenner, Adam Gleave, and Alessandro Abate. Starc: A general framework for quantifying differences between reward functions, 2024. URL https://arxiv.org/abs/2309.15257.

Joar Max Viktor Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *Proceedings of the 40th International Conference on Machine Learning*, July 2023.

Dirk Van Der Hoeven and Nicolò Cesa-Bianchi. Nonstochastic bandits and experts with arm-dependent delays. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 2022–2044. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/van-der-hoeven22a.html.

Claire Vernade, Olivier Cappé, and Vianney Perchet. Stochastic bandit models for delayed conversions, June 2017.

Yuanyu Wan, Chang Yao, Mingli Song, and Lijun Zhang. Improved regret for bandit convex optimization with delayed feedback. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 169–196. Curran Associates, Inc., 2024. doi: 10.52202/079017-0005.

Zongqi Wan, Xiaoming Sun, and Jialin Zhang. Bounded memory adversarial bandits with composite anonymous delayed feedback. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 3501–3507, Vienna, Austria, July 2022. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-1-956792-00-3. doi: 10.24963/ijcai.2022/486.

Di Yao, Chang Gong, Lei Zhang, Sheng Chen, and Jingping Bi. Causalmta: Eliminating the user confounding bias for causal multi-touch attribution, 2022. URL https://arxiv.org/abs/2201.00689.

Chunyuan Zheng, Anpeng Wu, Chuan Zhou, Taojun Hu, Qingying Chen, Hongyi Liu, Chenxi Li, Huiyou Jiang, Haoxuan Li, and Zhouchen Lin. Uplift modeling with delayed feedback: Identifiability and algorithms. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, volume 40 of *AAAI'26/IAAI'26/EAAI'26*, pp. 16468–16476. AAAI Press, June 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i19.38686.

APPENDIX CONTENTS

	Appendix Title	Page
A	Overview of Theoretical Results	13
B	Related Work	13
C	Problem Setup and Information Structure	15
D	Route Validity and Structural-Regret Stability	17
E	Identification and Evidence-Qualified Route Audit	22
F	Experiments	31
G	Timing and Notation Support	38
H	Proofs and Technical Derivations	38
I	Experimental Protocols and Supplementary Results	49

A OVERVIEW OF THEORETICAL RESULTS

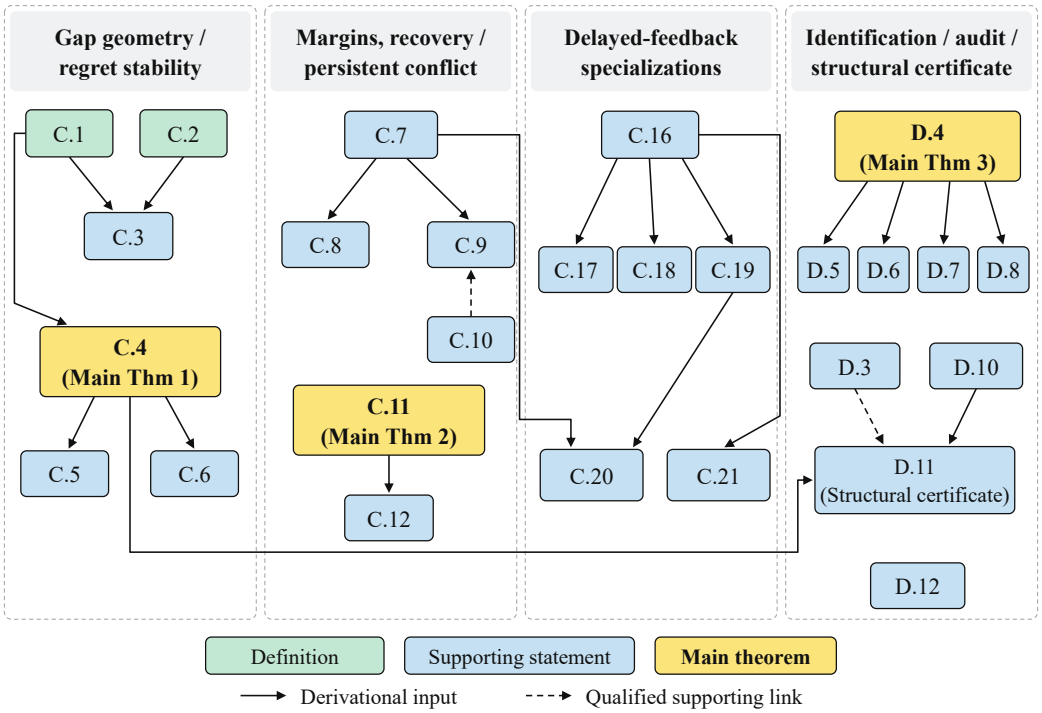

Figure 4: We include this chart as a guide to selected theoretical results in the appendix. Appendix C specifies the problem setup and information structure. Appendix D develops route validity, regret stability, and delayed-feedback specializations, while Appendix E addresses identification and evidence-qualified route auditing. Solid arrows indicate selected derivational inputs under the target statements' assumptions. Dashed links denote sharpness (D.10 to D.9) or a sufficient exact-reference construction (E.3 to E.11). The highlighted results correspond to the three main theorems. Detailed proofs are collected in Appendix H. The supplementary Lean formalization covers selected deterministic finite-action results only, not identification or statistical guarantees.

B RELATED WORK

We compare related work by the decision targets and observation interfaces taken as given in each setting.

B.1 Delayed and Anonymous Feedback with a Fixed Target

Classical delayed-feedback models study how late observations affect learning once the reward or loss target has been defined at the action level. General reductions quantify the regret cost of delay in stochastic and adversarial online learning (Joulani et al., 2013); subsequent work considers delayed conversions, unknown or arm-dependent delays, generalized linear bandits, and related structured settings (Vernade et al., 2017; Li et al., 2019; Van Der Hoeven & Cesa-Bianchi, 2022; Howson et al., 2023). Anonymous and aggregate models remove direct source identity while retaining regret with respect to the underlying generated reward process, including settings with rewards aggregated at arrival times and composite-anonymous adversaries (Pike-Burke et al., 2018; Cesa-Bianchi et al., 2022; Wan et al., 2022). Recent revenue-management work similarly studies delayed and anonymous demand or revenue observations while keeping the pricing objective fixed (Ji, 2026).

B.2 Partial, Corrupted, and Evolving Feedback Channels

More general feedback models separate losses from observations. In partial monitoring, the loss function and feedback signal are specified separately, and minimax regret depends on observability of decision-relevant loss differences (Cesa-Bianchi et al., 2006; Bartók et al., 2014). Corrupted-reward models distinguish a true reward from a potentially corrupted reward channel and show that reliable optimization of the true objective requires additional structure or cross-checking information (Everitt et al., 2017). Evolving-observation models likewise allow observations to differ from the true incurred loss and parameterize regret by feedback accuracy relative to that loss (Bar-On & Mansour, 2024).

B.3 Reward Equivalence, Invariance, and Objective Misspecification

A neighboring line of work asks when distinct reward functions induce equivalent decisions. Potential-based reward shaping gives classical policy-invariance results (Ng et al., 1999). STARC develops pseudometrics on reward functions with upper and lower control of worst-case regret and characterizes corresponding soundness and completeness properties (Skalse et al., 2024). Reward-learning work studies invariances and partial identifiability from behavioral data, including ambiguities that persist with unlimited observations (Skalse et al., 2023; Skalse & Abate, 2026). Objective-misspecification and Goodhart-type analyses further formalize how optimizing an imperfect proxy can degrade performance on an intended objective (Karwowski et al., 2024).

B.4 Delayed Credit Assignment and Source Attribution

Delayed consequences often motivate reassignment to earlier decisions. RUDDER constructs return-equivalent reward redistributions that preserve optimal policies in delayed-reward reinforcement learning, while Hindsight Credit Assignment derives value identities from hindsight-conditioned credit (Arjona-Medina et al., 2019; Harutyunyan et al., 2019). Advertising attribution studies related source-assignment problems through causal and counterfactual multi-touch methods, analyses of last-touch incentives under delayed conversion, and peer-validated attribution mechanisms (Dalessandro et al., 2012; Diemert et al., 2017; Yao et al., 2022; Berman, 2018; An et al., 2025).

B.5 Decision-Focused Prediction and Identification under Coarse Evidence

Prediction-to-decision methods distinguish statistical fit from downstream decision quality. Smart Predict, then Optimize defines prediction-induced decision loss and derives the SPO+ surrogate, with statistical consistency for the SPO loss under stated conditions (Elmachtoub & Grigas, 2022). A recent delayed-feedback contextual-bandit preprint similarly separates reward alignment with a downstream "north-star" outcome from learnability of the reward-optimal policy (Lee et al., 2026). Identification, however, depends on the observation interface and maintained assumptions. Reward-learning results characterize invariance classes; attribution-set learning uses explicit candidate sets and a candidate prior; and delayed-uplift results use a delayed potential-outcome model, with response-time identification restricted to the stated identifiable stratum (Skalse et al., 2023; Skalse & Abate, 2026; Applebaum et al., 2026; Zheng et al., 2026).

The closest comparisons clarify the distinction. Evolving-observation bandits take the true incurred loss as the target and control regret through average feedback accuracy (Bar-On & Mansour, 2024); delayed aggregated anonymous bandits retain the arm-generated reward process as the benchmark (Pike-Burke et al., 2018); and attribution-set learning estimates population prediction loss from candidate sets and priors (Applebaum et al., 2026). Our object is different: the action-level map is itself generated by a delayed-feedback route and is compared with an independently specified structural benchmark. Source correspondence and predictive fit must therefore be distinguished from preservation of structural action comparisons; neither alone establishes route validity.

C Problem Setup and Information Structure

We distinguish *structural decision time*, when an intervention generates an outcome, from *clock time*, when it becomes observable, without presuming valid arrival-time binding.

C.1 Sequential Setting

Capital and lower-case letters denote random variables and realizations. Fix a finite horizon $T \in \mathbb{N}$, an action space satisfying $2 \le |\mathcal{A}| < \infty$, and probability simplex $\Delta(\mathcal{A})$. The state and outcome spaces are standard Borel, and all structural maps, kernels, and losses below are measurable. Structural rounds are $t = 1, \dots, T$; observation clocks are $u \in \mathbb{N}$. State $S_t \in \mathcal{S}$ need not be observed. Optional pre-action context $X_t \in \mathcal{X}$ may provide only partial information about S_t; $X_t \equiv \varnothing$ denotes no current context.

C.2 Structural Decision Process

The learner selects $A_t \in \mathcal{A}$. For each $a \in \mathcal{A}$, intervention $A_t = a$ generates $Y_t(a)$, with

$$Y_t(a) = g_t(S_t, a, \varepsilon_t), \qquad Y_t = Y_t(A_t), \qquad S_{t+1} = f_t(S_t, A_t, \xi_t),$$

where ε_t and ξ_t are exogenous disturbances. The interventional outcome kernel is $P_t^Y(dy \mid s, a) := \mathbb{P}(Y_t \in dy \mid do(A_t = a), S_t = s)$. Actions may affect later states.

C.3 Delayed Observation Interface

Outcome Y_t arrives at $t + \tau_t$, where $\tau_t \in \mathbb{N}_0$ may depend on state, action, realized outcome, or prior history. Define the arriving source set $\mathcal{B}_u := \{s \in \{1, \dots, T\} : s + \tau_s = u\}$, batch $\mathcal{Y}_u^{\mathrm{arr}} := \{\{Y_s : s \in \mathcal{B}_u\}\}$, and subset $\mathcal{L}_u \subseteq \mathcal{B}_u$ of sources with observed labels. The labelled, fully unlabelled, and partially labelled interfaces are

$$\mathcal{O}_u^{\mathrm{lab}} = \{(s, Y_s) : s \in \mathcal{B}_u\}, \qquad \mathcal{O}_u^{\mathrm{unlab}} = \{\{Y_s : s \in \mathcal{B}_u\}\},$$
$$\mathcal{O}_u^{\mathrm{part}} = \left(\{(s, Y_s) : s \in \mathcal{L}_u\}, \{\{Y_s : s \in \mathcal{B}_u \setminus \mathcal{L}_u\}\}\right).$$

Write $\mathcal{O}_u$ for the chosen interface output, including aggregation $\mathcal{O}_u^{\mathrm{agg}} = \Gamma_u(\mathcal{Y}_u^{\mathrm{arr}})$ with specified map Γ_u. Unlabelled feedback retains individual values and multiplicities within clock u, not source rounds; aggregation may discard individual-level information. Source labels are interface information; binding is a route's assignment of observations to generating interventions.

C.4 Information Structure and Admissible Policies

Pre-action information is $\mathcal{F}_t = \sigma(\{X_s : s \le t\}, \{A_s : s < t\}, \{\mathcal{O}_u : u < t\})$. Feedback arriving at clock t, including $\tau_t = 0$, is post-action and may inform only later decisions. The admissible class Π_{adm} comprises policies $\pi = \{\pi_t\}_{t=1}^T$ with $\mathcal{F}_t$-measurable kernels $\pi_t(\cdot \mid \mathcal{F}_t) \in \Delta(\mathcal{A})$. Formally, admissibility requires

$$\mathbb{P}(A_t = a | \mathcal{F}_t \vee \sigma\{L_t^c(b) : b \in \mathcal{A}\}) = \pi_t(a \mid \mathcal{F}_t) \quad \text{a.s. for every } a \in \mathcal{A}.$$

A sufficient implementation is to generate actions by an inverse-CDF sampler from π_t using a seed $V_t \sim \mathrm{Unif}(0, 1)$ independent of the joint sigma-field generated by $\mathcal{F}_t$, the current potential outcomes and state, and current and future structural disturbances. The policy randomization therefore cannot

be coupled with latent information beyond $\mathcal{F}_t$. This clause is stronger than merely requiring the correct conditional marginal law for A_t, which would not exclude hidden coupling with S_t. History-dependent representations for action selection must also be constructed from $\mathcal{F}_t$, which need not identify the benchmark state S_t.

C.5 Decision-Time Structural Benchmark

Let $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ be the realized loss, with $\ell(a, Y_t(a))$ integrable for every t, a. We fix the measurable kernel version of the interventional loss

$$L_t^c(a) := \int \ell(a, y) P_t^Y(dy \mid S_t, a) = \mathbb{E}[\ell(a, Y_t) \mid do(A_t = a), S_t]$$

and define the structural optimal-action set $\mathcal{A}_t^{c,\star} := \arg\min_{a \in \mathcal{A}} L_t^c(a)$. For realized actions under $\pi \in \Pi_{\mathrm{adm}}$, define

$$r_t^c := L_t^c(A_t) - \min_{a \in \mathcal{A}} L_t^c(a), \qquad R_T^c := \sum_{t=1}^{T} r_t^c, \qquad \mathcal{R}_T^c(\pi) := \mathbb{E}_\pi[R_T^c].$$

Proposition C.1 (Exact information–decision decomposition)**.** *Define the information-feasible loss* $L_t^{\mathcal{F}}(a) := \mathbb{E}[L_t^c(a) \mid \mathcal{F}_t]$*, the state-information gap*

$$\iota_t := \min_{a \in \mathcal{A}} L_t^{\mathcal{F}}(a) - \mathbb{E}\left[\min_{a \in \mathcal{A}} L_t^c(a) \middle| \mathcal{F}_t\right] \geq 0,$$

and, for $\pi \in \Pi_{\mathrm{adm}}$*, its information-feasible decision regret*

$$d_t^{\mathcal{F}}(\pi) := \sum_{a \in \mathcal{A}} \pi_t(a \mid \mathcal{F}_t) L_t^{\mathcal{F}}(a) - \min_{a \in \mathcal{A}} L_t^{\mathcal{F}}(a).$$

Then

$$\mathbb{E}[r_t^c \mid \mathcal{F}_t] = d_t^{\mathcal{F}}(\pi) + \iota_t, \qquad \mathcal{R}_T^c(\pi) = \sum_{t=1}^{T} \mathbb{E}[d_t^{\mathcal{F}}(\pi) + \iota_t].$$

Thus routing error and unavailable state information are logically distinct: no feedback route can eliminate ι_t*, whereas a decision-sufficient context makes* $\iota_t = 0$*.*

Remark C.2 (Oracle attainability). The state-conditioned oracle does not presume pre-action observability of S_t. It is attainable if its optimal-action correspondence admits an $\mathcal{F}_t$-measurable selector; equivalently, the decomposition permits both $d_t^{\mathcal{F}} = 0$ and $\iota_t = 0$. Context can therefore be decision-sufficient without reconstructing the full state.

The distinction is consequential. If $S_t \sim \mathrm{Bernoulli}(1/2)$ iid and independent of $\mathcal{F}_t$, $\mathcal{A} = \{0, 1\}$, $Y_t(a) = S_t$, and $\ell(a, y) = \mathbf{1}\{a \neq y\}$, then the state oracle has zero loss but $L_t^{\mathcal{F}}(0) = L_t^{\mathcal{F}}(1) = \iota_t = 1/2$. Every admissible policy has $d_t^{\mathcal{F}} = 0$ and nevertheless incurs expected structural regret $T/2$, even with immediate, correctly source-labelled feedback.

This *on-trajectory, decision-time* benchmark evaluates contemporaneous intervention comparisons under the realized state S_t. Its structural-time index remains t regardless of when the associated feedback becomes observable.

C.6 Scope Conditions

Assumption C.3 (Finite decision interface)**.** The horizon T is finite and $2 \leq |\mathcal{A}| < \infty$.

Assumption C.4 (Observation-only delay)**.** Delay changes when Y_t becomes observable, but does not alter the structural outcome kernel $P_t^Y(dy \mid s, a)$, the transition law of S_{t+1}, or the set of feasible actions.

Assumption C.5 (Specified observation interface)**.** The interface is source-labelled, partially labelled, fully unlabelled, or aggregated; in the last case, Γ_u is specified.

Assumption C.6 (Non-anticipation and admissibility)**.** For every t, π_t is measurable with respect to $\mathcal{F}_t$, and the induced action A_t satisfies the admissibility condition in Section C.4.

These conditions define the decision and observation environment used by the subsequent analysis. Finiteness guarantees attained minima and finite comparison graphs; compact action spaces with measurable minimizers and uniform suprema provide the standard extension. Observation-only delay separates delayed observation from changes in the underlying structural intervention. The specified interface fixes the information available to each route, while non-anticipation ensures that action selection is based on decision-time information $\mathcal{F}_t$. Source–intervention binding, proxy sufficiency, comparison-identifying coverage, and structural identification enter later as additional conditions for the corresponding validity and identification results.

D Route Validity and Structural-Regret Stability

Section C defines the decision-time structural benchmark. We now treat the feedback route itself as the object of analysis: first by identifying the decision-relevant geometry of route-induced loss maps, then by quantifying regret stability and the consequences of persistent conflict, and finally by deriving decision-preserving transformation classes and delayed-feedback specializations. The central quantity is therefore route-to-benchmark action-gap distortion, not delay magnitude by itself.

Throughout the main analysis, the feasible action set is the common finite set $\mathcal{A}$, as in the experiments. A round-specific feasible set can be accommodated by replacing $\mathcal{A}$ below with the common feasible set of the structural and route maps being compared at that round.

D.1 Generic Operational Routes and Validity Geometry

Operational routes and interface admissibility. Fix a feedback interface I. Let $H_t^{r,I}$ denote the route-side information exposed by that interface by the clock at which the round-t evaluation map is constructed. An operational route r is a fixed measurable rule $\widetilde{L}_t^r = \Phi_t^r(H_t^{r,I}, U^r) \in \mathbb{R}^{|\mathcal{A}|}$, where U^r denotes any route-side algorithmic randomization. The measurable rule Φ_t^r and the law of U^r are held fixed across structural models being compared, and U^r is independent of latent structural variables not contained in $H_t^{r,I}$. We call the route *interface-admissible* when $H_t^{r,I}$ contains only information exposed by interface I under the declared route timing convention.

The route specification includes source binding, inferred attribution, proxy construction, batch aggregation, initialization, carry-forward, decay, and prespecified no-arrival or fallback rules. We require $\widetilde{L}_t^r$ to be defined for every $t = 1, \dots, T$, so empty-arrival clocks and batch arrivals are handled by the route itself.

The theory depends only on the induced map $\widetilde{L}_t^r$. A retrospective evaluation map may use feedback available by its declared evaluation clock; a route-derived quantity used to select A_t must also be measurable with respect to the pre-action information set $\mathcal{F}_t$.

For $a, b \in \mathcal{A}$, define the structural and route-induced gaps by $G_t^c(a,b) := L_t^c(a) - L_t^c(b)$ and $G_t^r(a,b) := \widetilde{L}_t^r(a) - \widetilde{L}_t^r(b)$.

Definition D.1 (Action-gap defect and alignment budget)**.** The action-gap defect and cumulative alignment budget are

$$\delta_t^r := \max_{a,b \in \mathcal{A}} |G_t^r(a,b) - G_t^c(a,b)|, \qquad \mathfrak{A}_T^r := \sum_{t=1}^{T} \delta_t^r.$$

Let $e_t^r(a) := \widetilde{L}_t^r(a) - L_t^c(a)$. Then $\delta_t^r = \max_a e_t^r(a) - \min_a e_t^r(a)$.

Decision equivalence and quotient geometry. Let $K = |\mathcal{A}| \geq 2$, identify a loss map with a vector in $\mathbb{R}^K$, and write $[x] := \{x + c\mathbf{1} : c \in \mathbb{R}\}$. Loss maps in the same class induce identical action gaps and identical regret values. The corresponding decision-relevant cardinal space is $\mathcal{Q}_K := \mathbb{R}^K / \operatorname{span}\{\mathbf{1}\}$. For $z \in \mathbb{R}^K$, define $\operatorname{osc}(z) := \max_a z(a) - \min_a z(a)$. The induced gap distance on $\mathcal{Q}_K$ is $d_{\mathrm{gap}}([x],[y]) := \operatorname{osc}(x - y)$, so $\delta_t^r = d_{\mathrm{gap}}([\widetilde{L}_t^r], [L_t^c])$.

Hence action-independent shifts are invariant in the decision-relevant geometry. The oscillation distance is used because it equals the largest perturbation of a pairwise action gap. A two-action

contrast makes the distinction explicit. If $L^c = (0, 1)$, then $\widetilde{L}^r = (5, 6)$ has zero gap distance despite a large level shift, whereas $\widetilde{L}^r = (1, 0)$ reverses the unique action comparison and has gap distance two. Thus loss-level displacement and decision displacement are different objects.

Cardinal, ordinal, and choice validity. Fix an arbitrary deterministic ordering of $\mathcal{A}$ and let $\mathcal{P} := \{(a, b) \in \mathcal{A}^2 : a < b\}$ contain each unordered action pair once.

Recall the structural optimal-action set $\mathcal{A}_t^{c,\star} = \arg\min_{a\in\mathcal{A}} L_t^c(a)$ from Section C.5, and define $\mathcal{A}_t^{r,\star} := \arg\min_{a\in\mathcal{A}} \widetilde{L}_t^r(a)$.

Using the standard sign map with $\operatorname{sgn}(0) = 0$, define

$$\rho_t^r := \frac{1}{|\mathcal{P}|} \sum_{(a,b)\in\mathcal{P}} \mathbf{1}\{\operatorname{sgn} G_t^r(a, b) \neq \operatorname{sgn} G_t^c(a, b)\}, \qquad \chi_t^r := \frac{|\mathcal{A}_t^{r,\star} \setminus \mathcal{A}_t^{c,\star}|}{|\mathcal{A}_t^{r,\star}|}.$$

Definition D.2 (Route-validity profile)**.** The per-round profile is $\mathsf{Val}_t^r := (\delta_t^r, \rho_t^r, \chi_t^r)$, corresponding to cardinal, ordinal, and choice-level distortion. Its horizon-level counterpart is

$$\mathsf{Val}_T^r := \left(\frac{\mathfrak{A}_T^r}{T}, \overline{\rho}_T^r, \overline{\chi}_T^r\right), \qquad \overline{\rho}_T^r := \frac{1}{T}\sum_{t=1}^T \rho_t^r, \quad \overline{\chi}_T^r := \frac{1}{T}\sum_{t=1}^T \chi_t^r.$$

The three exact validity notions are $\delta_t^r = 0$, $\rho_t^r = 0$, and $\chi_t^r = 0$. At the choice level, $\chi_t^r = 0$ if and only if

$$\mathcal{A}_t^{r,\star} \subseteq \mathcal{A}_t^{c,\star},$$

so choice validity means that every route-optimal action is structurally optimal.

The corresponding information reduction is $\mathcal{Q}_K \xrightarrow{\pi_{\text{rank}}} \mathfrak{O}(\mathcal{A}) \xrightarrow{\pi_{\text{opt}}} 2^{\mathcal{A}} \setminus \{\varnothing\}$, where $\mathfrak{O}(\mathcal{A})$ is the space of weak action orderings. The first map discards cardinal gap magnitude, and the second discards all ordering information except the optimal-action set.

Proposition D.3 (Hierarchy of decision preservation)**.** *For every route r and round t, $\delta_t^r = 0 \Rightarrow \rho_t^r = 0 \Rightarrow \chi_t^r = 0$. Neither converse holds in general.*

At the horizon level, $\mathfrak{A}_T^r/T = 0 \Rightarrow \overline{\rho}_T^r = 0 \Rightarrow \overline{\chi}_T^r = 0$.

The reverse implications fail even in small action spaces. For example, $L^c = (0, 1)$ and $\widetilde{L}^r = (0, 2)$ preserve the complete ordering and optimal action but have nonzero cardinal defect. Likewise, $L^c = (0, 1, 2)$ and $\widetilde{L}^r = (0, 2, 1)$ have the same optimal action while reversing the ordering of two suboptimal actions.

Approximate cardinal validity does not imply approximate ordinal or choice validity without structural separation. For instance, with two actions, $L_t^c = (0, 1/t)$ and $\widetilde{L}_t^r = (1/t, 0)$ give $\delta_t^r = 2/t$ and $\mathfrak{A}_T^r/T = O(\log T/T) \to 0$, while $\rho_t^r = \chi_t^r = 1$ for every t. Accordingly, the exact hierarchy is unconditional, whereas approximate decision preservation requires margin separation.

D.2 Sharp Regret Stability under Route Misalignment

Route-side optimization is informative about structural performance through the alignment of the two evaluation maps, and the next result quantifies the additive same-sequence certificate supplied by that alignment. Define the route-level per-round regret $r_t^r := \widetilde{L}_t^r(A_t) - \min_{a\in\mathcal{A}} \widetilde{L}_t^r(a)$, the realized route regret $R_T^r := \sum_{t=1}^T r_t^r$, and its expectation under a common admissible policy π, $\mathcal{R}_T^r(\pi) := \mathbb{E}_\pi[R_T^r]$.

Theorem D.4 (Sharp pathwise regret stability)**.** *For any realized action sequence $A_{1:T}$ and any route-induced loss sequence,*

$$|r_t^c - r_t^r| \le \delta_t^r \quad (t = 1, \dots, T), \qquad \boxed{|R_T^c - R_T^r| \le \mathfrak{A}_T^r}.$$

The constant 1 is sharp. For any common admissible policy π such that R_T^c, R_T^r, and $\mathfrak{A}_T^r$ are integrable, $|\mathcal{R}_T^c(\pi) - \mathcal{R}_T^r(\pi)| \le \mathbb{E}_\pi[\mathfrak{A}_T^r]$.

The theorem gives the additive same-sequence certificate

$$R_T^c \le R_T^r + \mathfrak{A}_T^r,$$

where the structural and route-induced evaluation maps are applied to the same realized action sequence. Its oscillation argument controls both the perturbation at the realized action and the route-induced change in the within-round minimum. Appendix H.2 gives a two-action construction attaining equality. Theorem D.13 provides a complementary consistency-preserving regime under positive-affine transformations, where cardinal defect may accumulate linearly while optima and consistency remain preserved.

Corollary D.5 (Structural consistency transfer)**.** *Pathwise, along any sequence of realized instances for which $R_T^r = o(T)$ and $\mathfrak{A}_T^r = o(T)$, one has $R_T^c = o(T)$. Likewise, if $\mathcal{R}_T^r(\pi) = o(T)$ and $\mathbb{E}_\pi[\mathfrak{A}_T^r] = o(T)$, then $\mathcal{R}_T^c(\pi) = o(T)$.*

A realized policy may avoid highly distorted actions; the next result gives the corresponding uniform characterization over actions.

For $a \in \mathcal{A}$, let $q_t^c(a) := L_t^c(a) - \min_b L_t^c(b)$, $q_t^r(a) := \widetilde{L}_t^r(a) - \min_b \widetilde{L}_t^r(b)$, and $d_{t,\mathrm{reg}}^r := \max_a |q_t^r(a) - q_t^c(a)|$.

Proposition D.6 (Uniform regret-map characterization)**.** *For every route and round, $\frac{1}{2}\delta_t^r \le d_{t,\mathrm{reg}}^r \le \delta_t^r$. Hence $\mathfrak{A}_T^r = o(T)$ if and only if $\sum_{t=1}^T d_{t,\mathrm{reg}}^r = o(T)$. Thus the gap defect is equivalent, up to a factor of two, to the worst actionwise perturbation of the regret map.*

D.3 Margins, Persistent Conflict, and Structural Failure

The validity profile records disagreement, but its decision consequence depends on structural separation. Structural margins convert cardinal perturbation into ordinal and choice guarantees; under complete conflict they also quantify how much route distortion is required to reverse the decision. Define the structural optimality margin $\mu_t^c := \min_{a \notin \mathcal{A}_t^{c,\star}} [L_t^c(a) - \min_b L_t^c(b)]$, with $\mu_t^c = +\infty$ when every action is structurally optimal.

Proposition D.7 (Margin bridge between gap, ranking, and choice validity)**.** *For any route r and round t:*

1. *If $|G_t^c(a,b)| > \delta_t^r$, then $\operatorname{sgn} G_t^r(a,b) = \operatorname{sgn} G_t^c(a,b)$. Thus a pairwise ranking error can occur only on a structural pair whose absolute gap is at most δ_t^r.*
2. *If $\delta_t^r < \mu_t^c$, then $\mathcal{A}_t^{r,\star} \subseteq \mathcal{A}_t^{c,\star}$, hence $\chi_t^r = 0$.*
3. *If $\mathcal{A}_t^{r,\star} \cap \mathcal{A}_t^{c,\star} = \varnothing$, define $\gamma_t^r := \min_{b \in \mathcal{A}_t^{r,\star}} q_t^c(b)$ and $\eta_t^r := \min_{a \notin \mathcal{A}_t^{r,\star}} q_t^r(a)$. Then $\delta_t^r \ge \gamma_t^r + \eta_t^r$. The bound is sharp.*

Part (2) requires a strict inequality: at equality, a structurally suboptimal action may tie for route optimality. The following corollary gives the horizon-level consequence under a uniform margin.

Corollary D.8 (Uniform-margin choice-disagreement bound)**.** *If $\mu_t^c \ge \underline{\mu} > 0$ for all t, then $\overline{\chi}_T^r \le \mathfrak{A}_T^r/(\underline{\mu}T)$.*

The uniform-margin condition is sufficient but stronger than necessary. To quantify recovery when small structural margins occur on only part of the horizon, define for $\varepsilon > 0$

$$H_{T,\mathrm{ord}}(\varepsilon) := \frac{1}{T|\mathcal{P}|} \sum_{t=1}^T \sum_{(a,b)\in\mathcal{P}} \mathbf{1}\{|G_t^c(a,b)| \le \varepsilon\}, \qquad H_{T,\mathrm{ch}}(\varepsilon) := \frac{1}{T} \sum_{t=1}^T \mathbf{1}\{\mu_t^c \le \varepsilon\}.$$

Proposition D.9 (Soft-margin decision recovery)**.** *For every route r, horizon T, and $\varepsilon > 0$,*

$$\overline{\rho}_T^r \le H_{T,\mathrm{ord}}(\varepsilon) + \frac{\mathfrak{A}_T^r}{\varepsilon T}, \qquad \overline{\chi}_T^r \le H_{T,\mathrm{ch}}(\varepsilon) + \frac{\mathfrak{A}_T^r}{\varepsilon T}.$$

Consequently, suppose that for each $\ell \in \{\mathrm{ord}, \mathrm{ch}\}$ there exist constants $C_\ell > 0$ and $\kappa_\ell > 0$, independent of the horizon T, such that

$$H_{T,\ell}(\varepsilon) \le C_\ell \varepsilon^{\kappa_\ell} \qquad \text{for every horizon } T \text{ under consideration and every } \varepsilon > 0.$$

Then

$$\overline{d}^r_{T,\ell} = O\left[\left(\frac{\mathfrak{A}^r_T}{T}\right)^{\kappa_\ell/(\kappa_\ell+1)}\right],$$

where $\overline{d}^r_{T,\mathrm{ord}} = \overline{\rho}^r_T$ *and* $\overline{d}^r_{T,\mathrm{ch}} = \overline{\chi}^r_T$.

The implied constants in this $O(\cdot)$ bound are uniform in T.

Proposition D.10 (Sharpness of the soft-margin exponent)**.** *For every $\kappa > 0$ and every horizon $T \geq 2$, there exists a two-action horizon-T structural–route instance $\mathcal{I}_T$ satisfying $H_{T,\mathrm{ord}}(\varepsilon) \leq \varepsilon^\kappa$ and $H_{T,\mathrm{ch}}(\varepsilon) \leq \varepsilon^\kappa$ for which*

$$\overline{\rho}^r_T = \overline{\chi}^r_T = \Theta\left[\left(\frac{\mathfrak{A}^r_T}{T}\right)^{\kappa/(\kappa+1)}\right].$$

Thus the exponent in Proposition D.9 cannot be improved uniformly under only the stated margin-mass condition. The instances $\{\mathcal{I}_T : T \geq 2\}$ form a horizon-indexed family; the statement does not assert that they are restrictions of a single nested infinite trajectory.

These results sharpen the exact hierarchy in Section D.1: vanishing average cardinal defect can coexist with persistent decision error when structural margins collapse too quickly, while controlled small-margin mass converts cardinal alignment into quantitative ranking and choice recovery.

A round exhibits complete route conflict when $\mathcal{A}^{r,\star}_t \cap \mathcal{A}^{c,\star}_t = \varnothing$. Equivalently, $\chi^r_t = 1$. Partial optimal-set disagreement, $0 < \chi^r_t < 1$, does not by itself force a regret trade-off because the two optimal sets still share at least one action.

Theorem D.11 (Persistent route conflict forces a regret trade-off)**.** *Let $\mathcal{T}_0 \subseteq \{t : \mathcal{A}^{r,\star}_t \cap \mathcal{A}^{c,\star}_t = \varnothing\}$. Suppose there exist constants $\gamma > 0$ and $\eta > 0$ such that $\gamma^r_t \geq \gamma$ and $\eta^r_t \geq \eta$ for every $t \in \mathcal{T}_0$. Then every realized action sequence satisfies*

$$\frac{R^c_T}{\gamma} + \frac{R^r_T}{\eta} \geq |\mathcal{T}_0|.$$

Equivalently, $R^c_T \geq \gamma|\mathcal{T}_0| - (\gamma/\eta)R^r_T$. Consequently, if $|\mathcal{T}_0| = \Omega(T)$ and $R^r_T = o(T)$, then $R^c_T = \Omega(T)$.

At each complete-conflict round, following a route optimum incurs at least the structural margin, whereas departing from the route-optimal set incurs at least the route margin. The theorem therefore applies to arbitrary realized action sequences, not only route-greedy learners. Under persistent margin-separated complete conflict, route-level and structural consistency are incompatible.

Corollary D.12 (Route-greedy systematic misbinding)**.** *Under the conditions of Theorem D.11, if $A_t \in \mathcal{A}^{r,\star}_t$ for every $t \in \mathcal{T}_0$, then $R^c_T \geq \gamma|\mathcal{T}_0|$.*

D.4 Decision-Preserving Transformation Classes

The generic validity theory does not require a functional relationship between $\widetilde{L}^r_t$ and L^c_t. The results below are supporting sufficient conditions for common transformations that preserve regret, ordering, or choice; they are not additional validity criteria.

Consider $\widetilde{L}^r_t(a) = h_t(L^c_t(a)) + \zeta^r_t(a)$, where h_t is common to all actions within round t, while $\zeta^r_t(a)$ contains action-specific residual distortion. This residual is distinct from the total route error $e^r_t = \widetilde{L}^r_t - L^c_t$ defined in Section D.1. The transformation may vary across rounds.

Theorem D.13 (Positive-affine regret transfer)**.** *Suppose $\widetilde{L}^r_t(a) = \alpha_t L^c_t(a) + \beta_t + \zeta^r_t(a)$ with $\alpha_t > 0$, and define $\varepsilon^r_t := \mathrm{osc}(\zeta^r_t)$. Then $|r^r_t - \alpha_t r^c_t| \leq \varepsilon^r_t$. Consequently, if $0 < \underline{\alpha} \leq \alpha_t \leq \overline{\alpha} < \infty$ for all t, then $\left|R^r_T - \sum_{t=1}^T \alpha_t r^c_t\right| \leq \sum_{t=1}^T \varepsilon^r_t$ and $\underline{\alpha} R^c_T - \sum_t \varepsilon^r_t \leq R^r_T \leq \overline{\alpha} R^c_T + \sum_t \varepsilon^r_t$. In particular, $R^c_T \leq (R^r_T + \sum_t \varepsilon^r_t)/\underline{\alpha}$.*

The positive-affine result extends to common nonlinear transformations when the transformation is quantitatively nondegenerate on the structural-loss range. The following statement records the resulting bound; its derivation is given in Appendix H.4.

Proposition D.14 (Controlled nonlinear regret transfer)**.** *Let* $I_t := [\min_{a\in\mathcal{A}} L_t^c(a), \max_{a\in\mathcal{A}} L_t^c(a)]$, *and suppose* $\widetilde{L}_t^r(a) = h_t(L_t^c(a)) + \zeta_t^r(a)$, *where* h_t *is common to all actions and, for all distinct* $x, y \in I_t$, $0 < m_t \le \frac{h_t(x)-h_t(y)}{x-y} \le M_t < \infty$. *With* $\varepsilon_t^r := \mathrm{osc}(\zeta_t^r)$, *one has* $m_t r_t^c - \varepsilon_t^r \le r_t^r \le M_t r_t^c + \varepsilon_t^r$; *if* $m_t \ge \underline{m} > 0$ *for all* t, *then* $R_T^c \le (R_T^r + \sum_t \varepsilon_t^r)/\underline{m}$.

Corollary D.15 (Strict-monotone decision preservation)**.** *Suppose* $\widetilde{L}_t^r(a) = h_t(L_t^c(a)) + c_t$, *where* h_t *is strictly increasing and* c_t *does not depend on* a. *Then* $\rho_t^r = \chi_t^r = 0$, *while* δ_t^r *may be nonzero.*

Positive-affine maps support quantitative regret transfer, whereas strictly increasing common maps preserve ordering and choice. The additive case $\alpha_t = 1$ with action-invariant residual distortion recovers exact cardinal preservation. Thus a route can be cardinally misaligned while remaining exactly ordinal- and choice-valid.

D.5 Arrival and Proxy Routes as Delayed-Feedback Specializations

The preceding theory applies to any route-induced action-level map. This subsection gives decision-relevant specializations for the two route mechanisms emphasized in the paper: temporal source misassignment and state-valued proxy reconstruction. The state-indexed representation introduced here is used only for these specializations and is not a maintained assumption of the structural benchmark. Whenever a norm $\|s - s'\|$ is used below, we additionally assume that the relevant state representation lies in a normed vector space and that the displayed Lipschitz condition holds on its support.

Suppose there exists a common state-indexed loss map $L^c(a; s)$ such that $L_t^c(a) = L^c(a; S_t)$. Fix $a_0 \in \mathcal{A}$ and define

$$\Psi_a(s) := \begin{cases} 0, & a = a_0, \\ L^c(a; s) - L^c(a_0; s), & a \ne a_0, \end{cases} \qquad d_{\mathrm{dec}}(s, s') := \mathrm{osc}(\Psi(s) - \Psi(s')).$$

This is a pseudometric on the original state space: two different structural states may have zero decision distance when they induce identical action gaps. Full-state recovery is therefore sufficient but not necessary for route validity; only the decision-relevant comparison representation must be preserved.

Proposition D.16 (Decision-relevant state replacement)**.** *Suppose a route replaces* S_t *by a state value* S_t' *and uses* $\widetilde{L}_t^r(a) = L^c(a; S_t')$. *Then* $\delta_t^r = d_{\mathrm{dec}}(S_t, S_t')$.

Arrival-time assignment. In the single-source case, let $\kappa(t) \le t$ denote the source round of the feedback assigned to arrival clock t, and set $L_t^{\mathrm{arr}}(a) := L_{\kappa(t)}^c(a) = L^c(a; S_{\kappa(t)})$. Then $\delta_t^{\mathrm{arr}} = d_{\mathrm{dec}}(S_t, S_{\kappa(t)})$.

Theorem D.17 (Exact cardinal validity of arrival-time evaluation)**.** *For a fixed round* t, *the following are equivalent: (i)* $\delta_t^{\mathrm{arr}} = 0$; *(ii)* $G_t^{\mathrm{arr}}(a, b) = G_t^c(a, b)$ *for all* $a, b \in \mathcal{A}$; *and (iii) there exists* $c_t \in \mathbb{R}$ *such that* $L_{\kappa(t)}^c(a) = L_t^c(a) + c_t$ *for all* $a \in \mathcal{A}$. *Thus positive delay is compatible with exact cardinal validity whenever the source-to-arrival change is action invariant.*

Observed source binding resolves factual assignment ambiguity but does not by itself identify counterfactual structural action gaps.

Proposition D.18 (Arrival-route alignment bound)**.** *If* $d_{\mathrm{dec}}(s, s') \le L_\Psi \|s - s'\|$ *for the relevant states, then* $\delta_t^{\mathrm{arr}} \le L_\Psi \|S_t - S_{\kappa(t)}\|$ *and* $\mathfrak{A}_T^{\mathrm{arr}} \le L_\Psi \sum_{t=1}^T \|S_t - S_{\kappa(t)}\|$.

Under the stronger condition $|L^c(a; s) - L^c(a; s')| \le L_S \|s - s'\|$ *for all* a, *the same bounds hold with* L_Ψ *replaced by* $2L_S$.

Hence delay magnitude alone need not predict arrival-route defect; the relevant quantity is displacement in decision space, not elapsed clock time.

Batch arrival routes. Suppose $\widetilde{L}_t^{\mathrm{batch}}(a) = \sum_{s\in\mathcal{B}_t} w_{t,s} L^c(a; S_s)$, with $w_{t,s} \ge 0$ and $\sum_s w_{t,s} = 1$. Then $\delta_t^{\mathrm{batch}} \le \sum_{s\in\mathcal{B}_t} w_{t,s} d_{\mathrm{dec}}(S_t, S_s)$. Under the decision-space Lipschitz condition above,

$\delta_t^{\mathrm{batch}} \le L_\Psi \sum_{s\in\mathcal{B}_t} w_{t,s}\|S_t - S_s\|$. Thus the single-source arrival route is the special case in which all weight is placed on one source round.

More generally, partially attributed routes, learned batch rules, carry-forward procedures, and nonlinear aggregate interfaces are handled directly by the generic theory once their route specification produces a complete action-level map $\widetilde{L}_t^r$.

Proxy reconstruction. When S_t is latent, let an executable state-valued proxy be $\widehat{S}_t = \psi_t(\mathcal{F}_t)$. The oracle proxy route is defined by $L_t^{\mathrm{proxy}}(a) := L^c(a; \widehat{S}_t)$, and therefore assumes that the structural loss kernel $L^c(a; s)$ is evaluable at the proxy state $\widehat{S}_t$. It follows that $\delta_t^{\mathrm{proxy}} = d_{\mathrm{dec}}(S_t, \widehat{S}_t)$.

Proposition D.19 (Oracle proxy-route alignment bound)**.** *If $d_{\mathrm{dec}}(s, s') \le L_\Psi\|s - s'\|$, then $\delta_t^{\mathrm{proxy}} \le L_\Psi\|S_t - \widehat{S}_t\|$ and $\mathfrak{A}_T^{\mathrm{proxy}} \le L_\Psi \sum_{t=1}^T \|S_t - \widehat{S}_t\|$.*

Under the stronger uniform condition $|L^c(a; s) - L^c(a; s')| \le L_S\|s - s'\|$ for all a, the same bounds hold with L_Ψ replaced by $2L_S$.

Corollary D.20 (Oracle proxy optimal-action preservation)**.** *If $d_{\mathrm{dec}}(S_t, \widehat{S}_t) < \mu_t^c$, then $\arg\min_a L_t^{\mathrm{proxy}}(a) \subseteq \arg\min_a L_t^c(a)$. A stronger sufficient condition is $2L_S\|S_t - \widehat{S}_t\| < \mu_t^c$.*

For an implemented route, the structural kernel may instead be replaced by a fitted action-level map. The next proposition records the corresponding bound and separates proxy-state error from fitted-kernel error.

Proposition D.21 (Fitted proxy-route alignment bound)**.** *Suppose that $\widehat{S}_t = \psi_t(\mathcal{F}_t)$ and that the implemented route uses*

$$\widetilde{L}_t^r(a) = \widehat{L}(a; \widehat{S}_t).$$

Define the fitted-kernel gap error

$$\varepsilon_{t,\mathrm{ker}}^r := \mathrm{osc}_{a\in\mathcal{A}} \left\{ \widehat{L}(a; \widehat{S}_t) - L^c(a; \widehat{S}_t) \right\}.$$

Then

$$\delta_t^r \le d_{\mathrm{dec}}(S_t, \widehat{S}_t) + \varepsilon_{t,\mathrm{ker}}^r.$$

Consequently,

$$\mathfrak{A}_T^r \le \sum_{t=1}^T d_{\mathrm{dec}}(S_t, \widehat{S}_t) + \sum_{t=1}^T \varepsilon_{t,\mathrm{ker}}^r.$$

The two terms are not an exact error decomposition: their actionwise gap contributions may cancel. The displayed inequalities are therefore upper bounds. When $\widehat{L}(a; \widehat{S}_t) = L^c(a; \widehat{S}_t)$, the oracle proxy-route identity is recovered.

Proxy accuracy matters through the action comparisons it preserves, not generic state-space fidelity alone. A proxy can therefore be inaccurate as a reconstruction of S_t but exact for the decision, while an accurate predictor can still be decision-wrong if it distorts the relevant action gaps.

A generic operational representation need not reconstruct a structural state. If $M_t = \psi_t(\mathcal{F}_t)$ directly induces an action-level map through the operational route, that map is evaluated using the generic quantities G_t^r, δ_t^r, ρ_t^r, and χ_t^r. No artificial distance between M_t and S_t is required.

E Identification and Evidence-Qualified Route Audit

Section D asks whether a route is structurally aligned. This section asks the distinct epistemic question: whether the available observation interface can identify or support that claim, and with what finite-sample accuracy. The evidential chain runs from observable information to an identified or supported decision object, then to a finite-sample audit and, when cardinal comparison information is sufficient, to a structural-regret certificate. The final implication requires comparison evidence that identifies the cardinal structural object entering the alignment budget.

E.1 Identification, Audit Support, and Estimability

Identification and finite-sample estimation answer different questions. Identification asks whether the observable law determines the structural object within the maintained model class; audit support and estimability concern what a validation design exposes and how precisely that supported object can be estimated.

Let $\mathfrak{M}$ denote the maintained class of structural models and let I denote a specified feedback interface. Fix an interface-admissible operational route r. For a specified admissible data-collection policy π, write $P_{\mathcal{M}}^{\pi,I}$ for the observable law induced by model $\mathcal{M} \in \mathfrak{M}$ under interface I. Here $P_{\mathcal{M}}^{\pi,I}$ is the law of the observable record generated under the specified interface and evaluation window, including observed contexts, actions, clock information, and interface outputs.

For $\mathcal{M} \in \mathfrak{M}$, let $\mathsf{Val}_T^r(\mathcal{M})$ denote the horizon-level route-validity profile of Definition D.2, evaluated under structural model $\mathcal{M}$. More generally, an identification claim below concerns a prespecified deterministic functional $\Theta^r : \mathfrak{M} \to \mathcal{V}$, where $\mathcal{V}$ is its specified target space. If a validity quantity is random under a structural model, Θ^r must specify the object being identified, such as its induced law or a prespecified functional of that law.

Definition E.1 (Interface-relative identification)**.** Fix an interface I, route r, policy π, and structural model class $\mathfrak{M}$. A structural route-validity object Θ^r is *identified from interface I* if

$$P_{\mathcal{M}_0}^{\pi,I} = P_{\mathcal{M}_1}^{\pi,I} \quad \Longrightarrow \quad \Theta^r(\mathcal{M}_0) = \Theta^r(\mathcal{M}_1)$$

for every $\mathcal{M}_0, \mathcal{M}_1 \in \mathfrak{M}$.

An empirical comparison may therefore be computable on logged support without identifying the corresponding structural object over the full model class.

We therefore use the following terminology throughout this section.

- *Validity* describes the actual structural alignment of the implemented route.
- *Identification* asks whether the observable law uniquely determines the structural decision object relevant to that validity claim.
- *Audit support* describes which decision comparisons are exposed by an external, labelled, randomized, controlled, or otherwise validated comparison design.
- *Estimability* describes the finite-sample accuracy with which the resulting audit target can be estimated.

A route may therefore be structurally valid but weakly auditable, or structurally invalid but precisely audited.

E.2 Decision Objects and Comparison-Identifying Evidence

The cardinal, ordinal, and choice levels correspond to progressively coarser structural decision objects and therefore impose different comparison-identification requirements. To formalize this distinction, consider an evaluation unit i with action set $\mathcal{A}$ and validation evidence $\mathcal{E}_i$. Let $\mathfrak{C}_i(\mathcal{E}_i)$ denote the set of loss maps in $\mathbb{R}^{|\mathcal{A}|}$ compatible with all validated comparisons in $\mathcal{E}_i$. Because an action-independent shift does not change decision-relevant gaps, cardinal identification concerns the quotient class $[L] \in \mathbb{R}^{|\mathcal{A}|}/\operatorname{span}\{\mathbf{1}\}$.

Definition E.2 (Level-specific comparison identification)**.** Validation evidence $\mathcal{E}_i$ is called compatible only if $\mathfrak{C}_i(\mathcal{E}_i) \neq \varnothing$. Incompatible evidence is reported as a failed validation contract and is assigned none of the identification levels below. Compatible evidence is:

1. *cardinal-identifying* if all $L, L' \in \mathfrak{C}_i(\mathcal{E}_i)$ satisfy $[L] = [L']$;
2. *ordinal-identifying* if all $L, L' \in \mathfrak{C}_i(\mathcal{E}_i)$ induce the same complete weak ordering over $\mathcal{A}$, including ties;
3. *choice-identifying* if all $L, L' \in \mathfrak{C}_i(\mathcal{E}_i)$ satisfy $\arg\min_{a \in \mathcal{A}} L(a) = \arg\min_{a \in \mathcal{A}} L'(a)$.

The choice-identifying notion above is route-independent: it identifies the full structural optimal-action set. It is sufficient, but need not be minimal, for identifying the directed choice discrepancy χ_t^r of a fixed route, since $\chi_t^r = 0$ requires only that every route-optimal action be structurally optimal.

The three information levels obey cardinal-identifying ⇒ ordinal-identifying ⇒ choice-identifying, and neither converse holds in general. This information hierarchy is distinct from the validity hierarchy in Section D: identification concerns ambiguity across compatible structural models, whereas δ_t^r, ρ_t^r, and χ_t^r measure realized route discrepancy.

A useful exact characterization is available when validation supplies cardinal pairwise gaps. Let $\mathcal{G}_i^{\mathrm{aud}} = (\mathcal{A}, E_i)$ be the action-comparison graph, where $\{a, b\} \in E_i$ whenever the exact structural gap between a and b is available.

Proposition E.3 (Connected cardinal-comparison coverage)**.** *Suppose the cardinal validation information for unit i consists of exact and mutually compatible pairwise structural gaps on the edges of $\mathcal{G}_i^{\mathrm{aud}}$, with no additional constraints linking disconnected components. Then the complete structural action-gap map is identified if and only if $\mathcal{G}_i^{\mathrm{aud}}$ is connected.*

Thus observing every action pair is unnecessary: a spanning tree is sufficient, whereas disconnected coverage leaves cross-component gaps unidentified.

The connected-graph criterion is specific to exact cardinal gaps. Connectedness of a graph containing only pairwise signs does not in general identify the complete weak ordering. For ordinal and choice evidence, the relevant condition remains uniqueness of the corresponding decision object across $\mathfrak{C}_i(\mathcal{E}_i)$.

E.3 Non-Identification under Fully Unlabelled Feedback

The fully unlabelled interface of Section C.3 retains the clock at which feedback arrives, the number of outcomes arriving at that clock, their individual values, and their multiplicities. What it removes is the source-to-outcome correspondence within an arriving batch.

This loss of source correspondence can be sufficient to destroy identification of route validity.

Theorem E.4 (Non-identification under fully unlabelled feedback)**.** *Under the fully unlabelled feedback interface, structural route validity is not identified in general. Specifically, for every even horizon $T = 2m \geq 2$, there exist a maintained structural model class $\mathfrak{M}$, a prespecified clock-measurable, interface-admissible route r, and models $\mathcal{M}_0, \mathcal{M}_1 \in \mathfrak{M}$. For every common admissible policy π,*

$$P_{\mathcal{M}_0}^{\pi, I_{\mathrm{unlab}}} = P_{\mathcal{M}_1}^{\pi, I_{\mathrm{unlab}}}, \qquad \mathsf{Val}_T^r(\mathcal{M}_0) = (0,0,0), \quad \mathsf{Val}_T^r(\mathcal{M}_1) = (2,1,1).$$

The obstruction is informational rather than statistical: if competing structural worlds induce the same observable law under every common admissible policy, additional fully unlabelled trajectories cannot separate their validity conclusions.

A witness uses two actions and an alternating latent state. Let $T = 2m$, $\mathcal{A} = \{0,1\}$, $X_t \equiv \varnothing$, $S_{t+1} = 1 - S_t$, $Y_t(a) = S_t$, and $\ell(a, y) = \mathbf{1}\{a \neq y\}$. The two worlds differ only in their initial state: $\mathcal{M}_0 : S_1 = 0$ and $\mathcal{M}_1 : S_1 = 1$. With delays $\tau_{2k-1} = 1$ and $\tau_{2k} = 0$, both generate the same fully unlabelled record, $\mathcal{O}_{2k-1}^{\mathrm{unlab}} = \varnothing$ and $\mathcal{O}_{2k}^{\mathrm{unlab}} = \{\{0,1\}\}$.

For the clock-measurable route $\widetilde{L}_t^r = (0,1)$ on odd rounds and $\widetilde{L}_t^r = (1,0)$ on even rounds, the route agrees with the structural comparison in $\mathcal{M}_0$ and reverses it in $\mathcal{M}_1$. Because the state transition and observed outcome process are action-independent, the observable laws coincide under every common admissible policy. The clock-only route is deliberately chosen to remove route-estimation error from the witness: the ambiguity comes from the missing structural source correspondence itself. Appendix H.6 gives the full construction and proof.

Corollary E.5 (Operational success does not identify structural success)**.** *Under the witness construction above, let the learner choose a route-optimal action at every round. Then $R_T^r(\mathcal{M}_0) = R_T^r(\mathcal{M}_1) = 0$, while $R_T^c(\mathcal{M}_0) = 0$ and $R_T^c(\mathcal{M}_1) = T$. Thus identical fully unlabelled observations and identical route-level optimization success can coexist with fundamentally different structural performance.*

The same witness yields a quantitative information floor. Let $\theta_j := \Theta^r(\mathcal{M}_j) \in \mathbb{R}$ for the two observationally equivalent worlds and $\Delta_\Theta := |\theta_1 - \theta_0| > 0$. For n independent trajectories, define minimax absolute-error risk over these two worlds by $\mathcal{R}^\star_{n,T}(p_{\mathrm{src}}) := \inf_{\widehat{\Theta}} \max_{j\in\{0,1\}} \mathbb{E}_{\mathcal{M}_j}|\widehat{\Theta} - \theta_j|$, where p_{src} is the independent probability that a source label is retained for each source outcome in the stated construction.

Corollary E.6 (Fully unlabelled minimax floor)**.** *For the two-world construction in Theorem E.4, under the fully unlabelled interface,*

$$\mathcal{R}^\star_{n,T}(0) = \frac{\Delta_\Theta}{2} \qquad \textit{for every } n \geq 1.$$

Thus additional trajectories do not reduce the ambiguity when the observable laws remain identical.

Corollary E.7 (Exact partial-label interpolation in the two-world construction)**.** *In the same construction, suppose every source label is retained independently with world-independent probability $p_{\mathrm{src}} \in [0,1]$. Every observed source–outcome pair separates the two worlds, and each length-T trajectory contains T such opportunities. Then*

$$\mathcal{R}^\star_{n,T}(p_{\mathrm{src}}) = \frac{\Delta_\Theta}{2}(1 - p_{\mathrm{src}})^{nT}.$$

The equality is specific to this two-world construction and retention model; it is not a universal interpolation formula for arbitrary structural classes.

The impossibility also propagates through deterministic aggregation.

Corollary E.8 (Coarsening cannot restore identification)**.** *Suppose interface J is generated from interface I by a fixed Markov kernel K that does not depend on the structural model. Fix a target functional that is well-defined under both interfaces and a J-admissible route. Any J-admissible policy is lifted to I by ignoring the extra information, and its observable law under J is the pushforward $K_\# P^{\pi,I}_{\mathcal{M}}$. If $P^{\pi,I}_{\mathcal{M}_0} = P^{\pi,I}_{\mathcal{M}_1}$, then $P^{\pi,J}_{\mathcal{M}_0} = P^{\pi,J}_{\mathcal{M}_1}$. Consequently, the fixed J-route validity object remains non-identified under such coarsening whenever the two structural worlds assign different values to it. Deterministic aggregation $\mathcal{O}^{\mathrm{agg}} = \Gamma(\mathcal{O}^{\mathrm{unlab}})$ is the special case in which K is induced by Γ.*

E.4 Recovery under Additional Information

Identification can be recovered when additional information or restrictions eliminate ambiguity in the structural decision object. Three mechanisms are relevant.

Source-binding information. Source labels, known return maps, or externally matched records identify the generating intervention, resolving factual source correspondence without identifying unobserved counterfactual action gaps.

Comparison-identifying information. Controlled structural maps, randomized action comparisons, valid causal identification designs, or external source-grounded comparison samples may identify the cardinal, ordinal, or choice decision object directly. The required information depends on the validity level being claimed. For example, exact cardinal pair references with connected comparison coverage identify the full gap geometry by Proposition E.3.

Structural restrictions. A sufficiently restrictive structural or proxy model can also eliminate the two-world ambiguity if it identifies the decision-relevant action comparisons rather than only predicting observed delayed outcomes.

These mechanisms need not coincide: source binding may be identified while counterfactual action comparisons remain unidentified, and a structural comparison design may identify a choice-level object without reconstructing complete source correspondence.

The following boundary case reconnects the framework to standard delayed learning.

Remark E.9 (Conditional reduction to source-indexed delayed learning). Suppose that:

1. arriving feedback is correctly bound to its generating source intervention;

2. the resulting source-indexed observation is valid feedback for the structural loss sequence; and

3. the remaining assumptions of the selected delayed-learning result, including its delay and observation model, are satisfied.

As a concrete comparator-matched instance, consider a finite contextual model in which $X_t \in \mathcal{F}_t$ is observed before action selection and there exists a policy $\pi^\star : \mathcal{X} \to \mathcal{A}$ such that

$$\pi^\star(X_t) \in \arg\min_{a \in \mathcal{A}} L_t^c(a) \qquad \text{almost surely for every } t.$$

Thus, the context is decision-sufficient: it need not reconstruct the full state S_t, but it identifies a structural optimal action.

Suppose that a delayed contextual-learning guarantee for a finite policy class $\Pi \ni \pi^\star$, using the correctly source-indexed feedback, takes the form

$$\mathbb{E}_\pi\left[\sum_{t=1}^{T}\{L_t^c(A_t) - L_t^c(\pi^\star(X_t))\}\right] \le G(T), \qquad \frac{G(T)}{T} \to 0.$$

Because

$$L_t^c(\pi^\star(X_t)) = \min_{a \in \mathcal{A}} L_t^c(a) \quad \text{almost surely},$$

this guarantee is exactly a guarantee for the structural regret benchmark defined in Section C.5:

$$\mathcal{R}_T^c(\pi) \le G(T).$$

The comparator here is a context-dependent structural oracle. It is therefore different from a best-fixed-action comparator of the form

$$\min_{a \in \mathcal{A}} \sum_{t=1}^{T} L_t^c(a),$$

which does not imply a guarantee against the state-oracle benchmark. This is a comparator-matching reduction: it imports a delayed contextual-learning guarantee under the stated source-indexed feedback assumptions. It does not assert a new delayed-learning algorithm, and a guarantee against the best fixed action does not imply a guarantee against the state-oracle comparator.

Source labels restore the factual binding required by this reduction; the imported delayed-learning assumptions must still hold.

E.5 Evidence-Adaptive Audit under Partial Validation

We now fix an implemented route r and consider partial source-grounded validation.

Let $\mathcal{I}_M = \{1, \dots, M\}$ be a fixed finite target population of audit units. An audit unit is the primary population unit. Action pairs form comparison support *within* that unit.

For unit i, let $\mathcal{P}_i^{\mathrm{aud}} \subseteq \mathcal{P}$ denote the canonically oriented action pairs for which the validation design provides usable comparison evidence, where $\mathcal{P}$ is the one-pair-per- unordered-pair set defined in Section D.1. The orientation is used only to give signed pairwise gaps an unambiguous representation; comparison-graph connectivity below remains an undirected property.

A pair-pooled average and a unit-level average are different estimands. The validity audit developed here treats the audit unit as primary; units with more observed pairs do not automatically receive larger population weight.

Route-side and audit-side information. Let Z_i^{route} denote source information available when the operational route is constructed, and let Z_i^{aud} denote information introduced at the subsequent validation stage. Information in Z_i^{route} enters Φ^r and can therefore change the induced route-validity profile. Information in Z_i^{aud} is used to evaluate the resulting fixed route against the validation reference. If additional source information is incorporated into Φ^r, the resulting mapping is treated as a distinct route under the same validity framework.

Evidence level. The audit follows the strongest comparison evidence available for each unit. Cardinal references on the route's loss scale support both numerical gap discrepancy and sign discrepancy. When validation supplies ordering but no common numerical scale, the audit retains sign comparisons on supported pairs without assigning unobserved gap magnitudes. Ordinal audit is therefore a weaker evaluation level that remains available without cardinal evidence, not a numerical substitute for cardinal discrepancy.

For each audit unit, the validation design determines the strongest supported comparison level: $\lambda_i^{\mathrm{aud}} \in \{\mathrm{cardinal}, \mathrm{ordinal}, \mathrm{choice}, \mathrm{unsupported}\}$. The evidence contract fixes this level before candidate routes are compared.

Implemented route comparisons. For an audited pair (a, b), let $\Delta_i^r(a, b)$ denote the pairwise comparison actually induced by the fixed implemented route on audit unit i. Whenever audit unit i has a well-defined structural comparison target, let $\Delta_i^c(a, b)$ denote that target structural pairwise gap. When $i = t$ is a structural decision round and the audit uses the same action-level route map and structural benchmark as Section D, $\Delta_t^r(a, b) = G_t^r(a, b)$ and $\Delta_t^c(a, b) = G_t^c(a, b)$. The generic Δ_i^c notation separates audit units from structural rounds; Section E.6.2 later specializes $i = t$. Any source inference, fitted proxy, carry-forward rule, or model fitting is already part of the fixed route Φ^r.

Let $\phi_i^{a,b}$ denote a source-grounded validation comparison when cardinal or ordinal pairwise evidence is available. For cardinal-supported units with $|\mathcal{P}_i^{\mathrm{aud}}| > 0$, define

$$d_{i,\delta}^{r,\mathrm{obs}} := \max_{(a,b)\in\mathcal{P}_i^{\mathrm{aud}}} |\Delta_i^r(a, b) - \phi_i^{a,b}|.$$

For cardinal- or ordinal-supported units with $|\mathcal{P}_i^{\mathrm{aud}}| > 0$, define

$$d_{i,\rho}^{r,\mathrm{obs}} := \frac{1}{|\mathcal{P}_i^{\mathrm{aud}}|} \sum_{(a,b)\in\mathcal{P}_i^{\mathrm{aud}}} \mathbf{1}\Big\{\operatorname{sgn}\Delta_i^r(a, b) \neq \operatorname{sgn}\phi_i^{a,b}\Big\}.$$

For a choice-level audit, the fixed route specification must also determine a route-supported optimal-action set $\mathcal{A}_i^{r,\star}$ on unit i. If no such action-level route object is defined, choice audit is unsupported for that unit. If the validation design identifies a reference optimal-action set $\mathcal{A}_i^{\mathrm{val},\star}$, define the directed choice discrepancy as $d_{i,\chi}^{r,\mathrm{obs}} := |\mathcal{A}_i^{r,\star} \setminus \mathcal{A}_i^{\mathrm{val},\star}|/|\mathcal{A}_i^{r,\star}|$.

These quantities measure route disagreement with the validation evidence at its supported level; structural interpretation additionally requires the corresponding reference-quality and comparison-identification conditions. Ordinal-only evidence can support an audit of sign agreement on covered pairs, but it does not supply the cardinal comparison information required by the structural-regret certificate below.

E.6 Finite-Evidence Certificates

We first separate finite-sample audit uncertainty from uncertainty in the validation reference itself.

For each fixed audit level $\ell \in \{\delta, \rho, \chi\}$, the finite target population below is restricted to units for which the prespecified contribution $d_{i,\ell}^{r,\mathrm{obs}}$ is defined. In particular, the two pairwise audit domains are

$$\mathcal{I}_\delta := \{i \in \mathcal{I}_M : \lambda_i^{\mathrm{aud}} = \mathrm{cardinal},\ |\mathcal{P}_i^{\mathrm{aud}}| > 0\},$$
$$\mathcal{I}_\rho := \{i \in \mathcal{I}_M : \lambda_i^{\mathrm{aud}} \in \{\mathrm{cardinal}, \mathrm{ordinal}\},\ |\mathcal{P}_i^{\mathrm{aud}}| > 0\},$$
$$\mathcal{I}_\chi := \{i \in \mathcal{I}_M : \lambda_i^{\mathrm{aud}} \in \{\mathrm{cardinal}, \mathrm{ordinal}, \mathrm{choice}\},\ \mathcal{A}_i^{r,\star} \text{ and } \mathcal{A}_i^{\mathrm{val},\star} \text{ are defined}\}.$$

Thus $\mathcal{I}_\delta \subseteq \mathcal{I}_\rho$. For $\ell \in \{\delta, \rho, \chi\}$ with $|\mathcal{I}_\ell| > 0$, the supported-population mean is

$$\frac{1}{|\mathcal{I}_\ell|} \sum_{i\in\mathcal{I}_\ell} d_{i,\ell}^{r,\mathrm{obs}}.$$

No population mean is reported when its support is empty. The three means may describe different populations and are not interchangeable summaries of a common set of units. In the results below, we fix one audit level, relabel its supported units as $1, \ldots, M$, and use M for that population's size. Let $D_{M,\ell}^{r,\mathrm{obs}} := M^{-1} \sum_{i=1}^{M} d_{i,\ell}^{r,\mathrm{obs}}$ denote the corresponding finite-population target.

Audit timing and fitted references. The audit draw determines which unit-level contributions are observed, not their values. The route, supported population, and validation references must determine a fixed collection $\{d_{i,\ell}^{r,\mathrm{obs}}\}_{i=1}^{M}$ before the draw. Fixed does not mean already observed or computed: a contribution may be observed only when its unit is audited. For learned references, one sufficient protocol is to fit the reference using independent training data, freeze all fitted or tuned components used in the contribution, and then draw the audit sample independently. Directly specified references require no training step.

Cross-fitting removes a unit from the fit used to evaluate that unit, but this does not by itself freeze the contributions of all units before the audit draw. If the fitted reference, support set, or calibration changes when the set of audited units changes, then the resulting contributions are not a fixed finite population for Proposition E.10.

The following result uses representative sampling and a conservative Hoeffding bound for sampling without replacement (Hoeffding, 1963); selective sampling is treated in Section E.7.

Proposition E.10 (Finite-population audit concentration)**.** *Condition on a fixed implemented route and a fixed finite audit population. Suppose* $0 \le d_{i,\ell}^{r,\mathrm{obs}} \le B_\ell$ *for every* i*. Let* $\mathcal{S}_{\mathrm{aud}} \subseteq \mathcal{I}_M$ *be a simple random sample without replacement of size* n*, and define* $\widehat{D}_{M,\ell}^{r,\mathrm{obs}} = \frac{1}{n}\sum_{i\in\mathcal{S}_{\mathrm{aud}}} d_{i,\ell}^{r,\mathrm{obs}}$*. Then, for every* $\alpha \in (0,1)$*, with probability at least* $1-\alpha$*,*

$$D_{M,\ell}^{r,\mathrm{obs}} \le \widehat{D}_{M,\ell}^{r,\mathrm{obs}} + B_\ell\sqrt{\frac{\log(2/\alpha)}{2n}}.$$

For ordinal and choice audit contributions one may take $B_\rho = B_\chi = 1$. For cardinal comparisons, boundedness is required only for the audit contribution used in this finite-sample proposition; it is not a global bounded-loss assumption for the structural theory.

The route and unit-level audit variable are fixed before audit sampling, so the result applies to either covered contributions or the complete reconstructed cardinal contribution defined below. References or calibration rules learned on the evaluation units do not automatically define a fixed finite population for the audit concentration result. They require an additional sampling-design argument establishing that the resulting contributions are fixed with respect to the audit draw. Cross-fitting alone only controls in-sample optimism and does not by itself justify the finite-population concentration bound.

E.6.1 Reference-quality bridge

Finite audit sampling controls the difference between a sample audit statistic and its finite-population target. Structural interpretation additionally requires the validation reference to approximate the corresponding structural comparison. These are distinct uncertainty sources: sampling uncertainty concerns which audit units are observed, whereas reference uncertainty concerns whether the validation comparison itself represents the structural target.

We first consider the comparisons directly covered by the validation design. For cardinal-supported units, define $d_{i,\delta}^{r,\mathrm{cov}} := \max_{(a,b)\in\mathcal{P}_i^{\mathrm{aud}}} |\Delta_i^r(a,b) - \Delta_i^c(a,b)|$ and $v_{i,\delta}^{\mathrm{cov}} := \max_{(a,b)\in\mathcal{P}_i^{\mathrm{aud}}} |\phi_i^{a,b} - \Delta_i^c(a,b)|$. Then $|d_{i,\delta}^{r,\mathrm{obs}} - d_{i,\delta}^{r,\mathrm{cov}}| \le v_{i,\delta}^{\mathrm{cov}}$.

Define $D_{M,\delta}^{r,\mathrm{cov}} = \frac{1}{M}\sum_{i=1}^{M} d_{i,\delta}^{r,\mathrm{cov}}$, $\quad V_{M,\delta}^{\mathrm{cov}} = \frac{1}{M}\sum_{i=1}^{M} v_{i,\delta}^{\mathrm{cov}}$. Therefore $D_{M,\delta}^{r,\mathrm{cov}} \le D_{M,\delta}^{r,\mathrm{obs}} + V_{M,\delta}^{\mathrm{cov}}$.

If the validation design supplies a bound $V_{M,\delta}^{\mathrm{cov}} \le \overline{V}_{\delta,\beta}^{\mathrm{cov}}$ with probability at least $1-\beta$, then Proposition E.10 and a union bound imply, with probability at least $1-\alpha-\beta$, that $D_{M,\delta}^{r,\mathrm{cov}} \le \widehat{D}_{M,\delta}^{r,\mathrm{obs}} + B_\delta\sqrt{\log(2/\alpha)/(2n)} + \overline{V}_{\delta,\beta}^{\mathrm{cov}}$.

The covered certificate controls only pairs in $\mathcal{P}_i^{\mathrm{aud}}$. A structural alignment certificate requires a comparison-identifying reconstruction of the complete cardinal gap map, whose true discrepancy is $d_{i,\delta}^{r,\mathrm{full}} := \max_{a,b\in\mathcal{A}} |\Delta_i^r(a,b) - \Delta_i^c(a,b)|$.

Complete cardinal reference. Suppose the validation evidence is cardinal comparison-identifying and the validation protocol specifies a reconstruction rule that produces a complete refer-

ence gap map $\phi^{a,b}_{i,\mathrm{full}}$ for all $a, b \in \mathcal{A}$. For exact mutually compatible pair references on a connected comparison graph, Proposition E.3 implies that this complete gap map is uniquely determined. Estimated references require a full-map reference-quality guarantee for the reconstruction.

Define $d^{r,\mathrm{full,obs}}_{i,\delta} := \max_{a,b\in\mathcal{A}} |\Delta^r_i(a,b) - \phi^{a,b}_{i,\mathrm{full}}|$ and $v^{\mathrm{full}}_{i,\delta} := \max_{a,b\in\mathcal{A}} |\phi^{a,b}_{i,\mathrm{full}} - \Delta^c_i(a,b)|$. Then $|d^{r,\mathrm{full,obs}}_{i,\delta} - d^{r,\mathrm{full}}_{i,\delta}| \le v^{\mathrm{full}}_{i,\delta}$.

Let $D^{r,\mathrm{full}}_{M,\delta} := M^{-1}\sum_i d^{r,\mathrm{full}}_{i,\delta}$, $D^{r,\mathrm{full,obs}}_{M,\delta} := M^{-1}\sum_i d^{r,\mathrm{full,obs}}_{i,\delta}$, and $V^{\mathrm{full}}_{M,\delta} := M^{-1}\sum_i v^{\mathrm{full}}_{i,\delta}$. Then $D^{r,\mathrm{full}}_{M,\delta} \le D^{r,\mathrm{full,obs}}_{M,\delta} + V^{\mathrm{full}}_{M,\delta}$.

If $0 \le d^{r,\mathrm{full,obs}}_{i,\delta} \le B^{\mathrm{full}}_\delta$, Proposition E.10 applies to the full-map contribution. If, in addition, the validation design guarantees $V^{\mathrm{full}}_{M,\delta} \le \overline{V}^{\mathrm{full}}_{\delta,\beta}$ with probability at least $1-\beta$, then a union bound gives, with probability at least $1-\alpha-\beta$, $D^{r,\mathrm{full}}_{M,\delta} \le \widehat{D}^{r,\mathrm{full,obs}}_{M,\delta} + B^{\mathrm{full}}_\delta\sqrt{\log(2/\alpha)/(2n)} + \overline{V}^{\mathrm{full}}_{\delta,\beta}$. Exact comparison-identifying references give $V^{\mathrm{full}}_{M,\delta} = 0$; estimated references require a quality guarantee for the reconstructed complete object.

Ordinal and choice reference quality. Reference quality remains level-specific. For an ordinal-supported unit with $|\mathcal{P}^{\mathrm{aud}}_i| > 0$, let $d^{r,\mathrm{cov}}_{i,\rho}$ be the fraction of audited pairs on which the route and structural gap signs disagree, and let $v_{i,\rho}$ be the corresponding fraction on which the validation-reference and structural signs disagree. The disagreement triangle inequality gives $d^{r,\mathrm{cov}}_{i,\rho} \le d^{r,\mathrm{obs}}_{i,\rho} + v_{i,\rho}$, so reference quality remains validity-level specific.

At the choice level, structural interpretation requires the validation-supported optimal-action set to identify, or otherwise control its discrepancy from, the structural optimal-action set.

E.6.2 From audit validity to structural regret

A horizon-level alignment certificate requires a complete cardinal comparison object.

Suppose that:

1. the audit units correspond to structural decision rounds, so that $M = T$ and $i = t$;
2. the target audit population is the complete structural horizon;
3. the route-side and structural audit comparisons coincide with the Section D objects, so that $\Delta^r_t(a,b) = G^r_t(a,b)$ and $\Delta^c_t(a,b) = G^c_t(a,b)$ for all $a, b \in \mathcal{A}$;
4. the validation evidence is cardinal comparison-identifying and the validation protocol produces the complete reconstructed reference $\phi^{a,b}_{t,\mathrm{full}}$ for all $a, b \in \mathcal{A}$;
5. the full-map audit contribution is bounded by B^{full}_δ, and the reconstructed reference satisfies the stated full-map reference-quality guarantee.
6. conditional on the implemented route, supported horizon, and frozen reference construction, the audit is a simple random sample without replacement of size $1 \le n \le T$, as required by Proposition E.10; the full-map contributions and all fitted reference components are fixed before that draw.

Under these conditions, $d^{r,\mathrm{full}}_{t,\delta} = \delta^r_t$ and $D^{r,\mathrm{full}}_{T,\delta} = T^{-1}\sum_t \delta^r_t = \mathfrak{A}^r_T/T$. Applying the full-map finite-evidence bound above therefore yields the following structural certificate.

Corollary E.11 (Evidence-qualified structural-regret certificate)**.** *Under the conditions above, for $\alpha, \beta \in (0,1)$ with $\alpha + \beta < 1$, with probability at least $1-\alpha-\beta$,*

$$\frac{\mathfrak{A}^r_T}{T} \le \widehat{D}^{r,\mathrm{full,obs}}_{T,\delta} + B^{\mathrm{full}}_\delta\sqrt{\frac{\log(2/\alpha)}{2n}} + \overline{V}^{\mathrm{full}}_{\delta,\beta}.$$

Consequently, the sharp regret-stability result of Section D gives

$$R^c_T \le R^r_T + T\left[\widehat{D}^{r,\mathrm{full,obs}}_{T,\delta} + B^{\mathrm{full}}_\delta\sqrt{\frac{\log(2/\alpha)}{2n}} + \overline{V}^{\mathrm{full}}_{\delta,\beta}\right].$$

Without cardinal comparison-identifying evidence, Corollary E.11 is unavailable; an estimated complete reference additionally requires the full-map quality guarantee above.

E.7 Extensions: Selective Validation and Calibration Recoverability

Selective validation is a selection problem over validation evidence, not an intervention on the route being audited. The route and its unit-level discrepancy are fixed before audit sampling; the inclusion indicator therefore controls whether the discrepancy is observed, rather than changing the underlying route-validity object. This separation places selective audit in the same formal family as inverse-probability-of-observation and inverse-probability-of-sampling problems, while the estimand here remains a route-discrepancy functional rather than a treatment effect (Cole & Hernán, 2008; Li et al., 2013).

Selection assumptions and support. Let $R_i \in \{0,1\}$ denote audit inclusion and let W_i^{aud} collect the pre-audit design information used to govern inclusion. Write $\pi_i = \mathbb{P}(R_i = 1 \mid W_i^{\mathrm{aud}})$. When the audit is a controlled probability design and π_i is known by construction, justification is design-based unequal-probability sampling rather than causal treatment-effect identification (Hajek, 1964). When inclusion probabilities are instead estimated from an observational validation process, population interpretation additionally requires that the observed design information be sufficient for the selection mechanism–equivalently, after conditioning on W_i^{aud}, inclusion must not retain residual dependence on the unobserved audit contribution–together with positivity and adequate specification of the inclusion model (Cole & Hernán, 2008; Li et al., 2013).

The positivity requirement $\pi_i > 0$ is therefore a support condition, not merely a numerical safeguard against extreme weights. If $\pi_i = 0$ on part of the target population, no reweighting scheme can recover that unsupported region without additional structural assumptions; the target population must be restricted or the validation design must be changed. Likewise, if audit inclusion depends on unobserved discrepancy even after conditioning on the recorded design information, inverse-probability weighting alone does not restore population identification.

Design-aware weighting. Fix a supported unit-level audit contribution d_i for the target population, such as $d_{i,\ell}^{r,\mathrm{obs}}$ at a prespecified evidence level. For represented units, inverse-probability weights are $w_i = \pi_i^{-1}$. The Horvitz–Thompson finite-population mean is $\widehat{D}_{\mathrm{HT}} = M^{-1}\sum_i R_i w_i d_i$. The implemented audit instead uses the self-normalized Hájek IPW mean $\widehat{D}_{\mathrm{H}} = \{\sum_i R_i w_i d_i\}/\{\sum_i R_i w_i\}$; Appendix I.5 gives its experiment-specific form. The normalization changes finite-sample behavior but does not create identification: the validity of the weighting argument still comes from the sampling or selection assumptions above.

Theorem E.12 (Design-matched inference for independent audit inclusion)**.** *Fix $\alpha \in (0,1)$ and condition on fixed $d_1, \dots, d_M \in [0, B]$. Suppose the inclusion indicators are independent with $R_i \sim \mathrm{Bernoulli}(\pi_i)$ and $\pi_i \geq \pi_{\min} > 0$. Define*

$$D_M = \frac{1}{M}\sum_{i=1}^{M} d_i, \quad \widehat{D}_{\mathrm{HT}} = \frac{1}{M}\sum_{i=1}^{M}\frac{R_i d_i}{\pi_i}, \quad \widehat{W} = \frac{1}{M}\sum_{i=1}^{M}\frac{R_i}{\pi_i},$$

and $\widehat{D}_{\mathrm{H}} = \widehat{D}_{\mathrm{HT}}/\widehat{W}$ on $\{\widehat{W} > 0\}$. Then

$$\mathbb{E}[\widehat{D}_{\mathrm{HT}}] = D_M, \qquad V_D := \mathrm{Var}(\widehat{D}_{\mathrm{HT}}) = \frac{1}{M^2}\sum_{i=1}^{M}\frac{1-\pi_i}{\pi_i} d_i^2.$$

Let $V_W = M^{-2}\sum_i (1-\pi_i)/\pi_i$, $x = \log(4/\alpha)$,

$$\epsilon_D = \sqrt{2V_D x} + \frac{Bx}{3M\pi_{\min}}, \qquad \epsilon_W = \sqrt{2V_W x} + \frac{x}{3M\pi_{\min}}.$$

Because $V_D \leq B^2 V_W$, the fully design-computable replacement $\bar{\epsilon}_D := B\sqrt{2V_W x} + Bx/(3M\pi_{\min})$ satisfies $\bar{\epsilon}_D \geq \epsilon_D$. If $\epsilon_W < 1$, then with probability at least $1-\alpha$,

$$\widehat{W} > 0, \qquad |\widehat{D}_{\mathrm{H}} - D_M| \leq \frac{\epsilon_D + B\epsilon_W}{1-\epsilon_W} \leq \frac{\bar{\epsilon}_D + B\epsilon_W}{1-\epsilon_W}.$$

The empty-audit probability satisfies $\mathbb{P}(\widehat{W} = 0) = \prod_i (1 - \pi_i) \le \exp(-\sum_i \pi_i)$. *Finally, along a triangular array with* $M \to \infty$, $V_D > 0$, *and*

$$\max_{i \le M} \frac{d_i}{M \pi_i \sqrt{V_D}} \longrightarrow 0,$$

one has $(\widehat{D}_{\mathrm{HT}} - D_M)/\sqrt{V_D} \Rightarrow N(0,1)$.

The theorem distinguishes exact design unbiasedness of Horvitz–Thompson from the self-normalization used by Hájek. It also makes the empty-audit event explicit, so implementation includes a prespecified fallback for $\widehat{W} = 0$. When inclusion probabilities are estimated from observational audit data, population-level inference additionally depends on nuisance estimation and the maintained selection model.

For nonnegative weights, we also report the conventional effective-support diagnostic n_{eff}, defined in Appendix I.5. It summarizes weight concentration and the effective amount of labelled support. The inferential quantities remain tied to their corresponding objects: Proposition E.10 uses the audit sample size, Theorem E.12 uses the design variance, and population interpretation additionally depends on positivity, the sampling or selection model, and reference quality.

Calibration recoverability. We use pair-specific affine calibration only as a pairwise recoverability diagnostic, not as an action-level loss map. For a fixed route comparison $\Delta_i^r(a,b)$, consider

$$\widehat{G}_{i,\mathrm{cal}}^r(a,b) = \widehat{\alpha}_{r,a,b} + \widehat{\beta}_{r,a,b} \Delta_i^r(a,b).$$

Independent fits across pairs need not satisfy antisymmetry or cycle consistency and therefore need not be differences of a common action-level loss vector.

The calibration parameters are estimated without using the held-out unit on which the calibrated comparison is evaluated.

Let $\widehat{D}_{\mathrm{raw}}^r$ and $\widehat{D}_{\mathrm{cal}}^r$ denote the corresponding out-of-fold pre- and post-calibration comparison discrepancies on the same supported audit target. When $\widehat{D}_{\mathrm{raw}}^r > 0$, define $\widehat{\mathrm{Rec}}^r := 1 - \widehat{D}_{\mathrm{cal}}^r / \widehat{D}_{\mathrm{raw}}^r$.

A positive value means that the prespecified calibration family reduces out-of-fold comparison discrepancy relative to the specified validation reference on the declared support. It does not establish route validity, structural identification, or improved decision performance. Calibratability and structural correctness are therefore distinct properties. Deployment would require a jointly consistent action-level representation, obtained by imposing antisymmetry and cycle consistency or by a separately specified reconstruction. Any resulting route must satisfy the declared information interface and undergo separate validity evaluation. Constructing and evaluating such a route is outside the scope of this calibration audit. Cross-fitting addresses in-sample optimism, not identification or support failures.

F Experiments

F.1 Empirical Design and Evidence Map

The experiments ask four questions about the same organizing object: whether an operational route preserves the decision-time comparisons that define the structural benchmark. Experiment 1 uses controlled complete maps to test whether delay magnitude orders benchmark damage and to confront the action-gap theory directly. Experiment 2 asks when a fixed delayed-conversion log admits multiple source-time routes. Experiment 3 asks whether lower predictive MAE also yields better recovery of the action gaps and choices supported by the log. Experiment 4 separates information that changes a route from evidence used to audit or calibrate a fixed route. The evidence regimes are deliberately different: Experiments 1 and 4 expose structural comparison information by construction, whereas Experiments 2 and 3 remain restricted to logged support.

Table 1 summarizes the claim tested, the evidence available for testing it, and the primary diagnostic. The experiments are complementary rather than repeated estimates of a common quantity.

Table 1: Empirical claims and the evidence used to assess them.

Exp.	Claim tested	Evidence regime	Primary diagnostic
1	Claim A: Delay magnitude does not determine structural validity.	Controlled complete maps plus a fixed delayed-feedback learner	$\mathfrak{A}_T^r/T$, R_T^c/T, choice/conflict diagnostics, matched cancellation, realized stability utilization, arrival–source learner contrast
2	Claim B: A delayed log does not uniquely determine a source-time evaluation route.	Fixed delayed-conversion log with prespecified candidate-source construction	Allocation TV by route and candidate-source multiplicity
3	Claim C: Lower score-prediction MAE need not imply more accurate action-gap recovery on common logged support.	Logged held-out targets on a common supported action set	MAE $\rightarrow$ gap/sign recovery $\rightarrow$ top action and paired value
4	Claim D: Good audit/calibration statistics do not themselves certify structural validity.	Controlled route, audit, and calibration interventions	D_{pair}^r / audit RMSE / calibration recoverability

Table 2: **Mapping theoretical results to empirical evidence and evidence scope.**

Theoretical object	Formal guarantee	Direct evidence	Boundary of evidence
Quotient geometry and margin bridge	Common shifts cancel; choice is preserved below the structural margin	Matched cancellation and threshold sweep	Controlled complete loss maps
Sharp transfer and persistent conflict	Theorems 1–2	Same-sequence route-greedy stress test across three horizons	Does not estimate the sharp constant or learner regret
Unlabelled non-identification	Theorem 3	Closed-form probability calculation is numerically replayed in the appendix	Identity check, not an independent test of observational equivalence
Finite-evidence audit	SRS and independent-inclusion guarantees under frozen references	Controlled repeated audit sampling with known propensities	Pair-average target; not the worst-pair regret certificate
Logged route implications	No assumption-free structural identification claim	Criteo source-route non-uniqueness; semi-synthetic KuaiRand action-gap/choice diagnostic	Conditional on converters or common logged support

F.2 Experiment 1: Delay, Route Conflict, and Structural Regret

Experiment 1 asks whether benchmark damage is governed by delay magnitude or by the action comparisons induced by the route. **Claim tested (Claim A).** *Delay magnitude does not determine structural validity.* An action-invariant route error can move loss levels without changing a decision comparison, whereas an action-dependent component enters the action-gap defect and can cross the margin boundary at which the preferred action changes. The experiment also asks whether persistent route conflict can make route-side optimization structurally misleading.

The complete-map component uses six controlled mechanisms: zero delay, an exact-valid delayed shift, three stochastic delays with generated mean near 15, and systematic misbinding. A route-greedy diagnostic selects a route-optimal action at each round, so route regret is zero by construction and the same action sequence can be evaluated against the structural benchmark. A separate delayed-feedback learner keeps the algorithm fixed and changes only whether factual feedback is bound to the arrival clock or returned to its generating source round. The primary mechanisms are unchanged; the matched cancellation, horizon, and utilization diagnostics are targeted route-map checks reported in Appendix I.2.

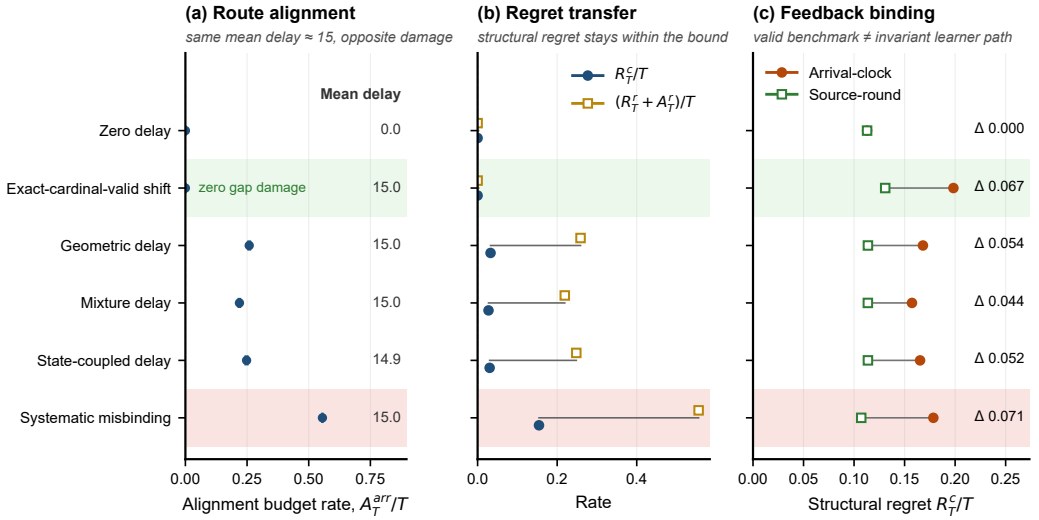

Figure 5: **Experiment 1: comparable delay can produce negligible or substantial benchmark damage, and persistent conflict can make route success structurally misleading.** Panel (a) compares alignment budgets at similar mean delay and separates the exact-valid shift from action-dependent distortions. Panel (b) compares route-greedy structural regret with the sharp transfer bound; systematic misbinding realizes the persistent-conflict regime while route regret is zero by construction. Panel (c) contrasts arrival-clock and source-round feedback bindings for a fixed learner; the exact-cardinal-valid shift shows that benchmark invariance need not imply learner-path invariance. Whiskers in Panels (a) and (c) are 95% seed-bootstrap intervals over 30 shared evaluation seeds.

The exact-cardinal-valid control first separates delay magnitude from benchmark validity. Its generated mean delay is 14.9709, while its alignment-budget rate is numerically zero and its route-greedy structural-regret rate is zero. The geometric, mixture, and state-coupled mechanisms have nearly the same mean delay but alignment-budget rates of 0.2588, 0.2191, and 0.2481. Comparable delay therefore need not imply comparable action-gap distortion.

The targeted matched sweep then isolates the component that the theory predicts should cancel. At each supported round, it adds a shared component identically to every action and an action-dependent residual scaled to the structural margin. The shared component is tested both as a constant and as a state-varying function of S_t. Holding the residual fixed, changing the shared amplitude or profile changes loss levels but leaves the action-gap defect, sign disagreement, choice disagreement, optimal-action masks, and all-action regret map unchanged up to 2.22×10^{-16}. By contrast, the action-dependent residual is constructed so that $\delta_t^r / \mu_t^c = \alpha_{\mathrm{dep}}$: choice disagreement is zero below one, equals one half at one, and equals one above one. The experiment therefore perturbs the margin boundary directly. The supported conclusion is about within-round action invariance, not state variation in general.

Systematic misbinding realizes the stronger persistent-conflict regime. Its alignment-budget rate is 0.5556 and its route-greedy structural-regret rate is 0.1543, while route regret remains zero. Complete conflict occurs on approximately one half of the rounds. The same structural-regret rate remains near 0.1543 at $T \in \{10^3, 5\times10^3, 10^4\}$, so the cumulative structural regret grows linearly over this horizon range while the route objective remains perfectly optimized. The full conflict-margin arithmetic and horizon table are reported in Appendix I.2.

The controlled paths also quantify how much of the same-sequence stability budget is realized. For routes with a numerically positive alignment budget, the mean realized utilization is approximately 0.122–0.126 for the three stochastic-delay mechanisms and 0.278 for systematic misbinding. It is undefined, rather than set to zero, when the alignment budget is numerically zero. This is a descriptive path-level diagnostic; sharpness is established analytically in Theorem D.4.

The fixed learner answers a different question. The exact-cardinal-valid shift has zero complete-map action-gap defect but still produces an arrival-minus-source learner contrast of 0.0675, because the arriving scalar loss changes a selected action–context update rather than an entire action-loss vector. Benchmark invariance therefore does not imply learner-path invariance under partial-feedback updates. The learner contrast and route-greedy structural regret are different estimands and are not added or interpreted as components of a common total. Appendix I.2 reports their matched comparison.

Experiment 1 therefore distinguishes three layers that delay magnitude alone cannot separate: route-map validity, structural consequence under route-side optimization, and learner-path sensitivity. The first two are governed by preservation of action comparisons; the third is an algorithmic consequence of when scalar feedback enters the update path.

F.3 Experiment 2: Source Ambiguity and Route Non-Uniqueness

Experiment 2 asks whether a fixed delayed-conversion log determines a unique source-time evaluation route. **Claim tested (Claim B).** *A delayed log does not uniquely determine a source-time evaluation route.* The analysis uses the Criteo delayed-conversion log (Diemert et al., 2017), treats impressions as candidate source events, and defines source-time decision cells by campaign and source calendar day. First-click-or-touch, Last-click-or-touch, Linear credit, and Time-decay credit are evaluated on the same retained journeys and aggregate cell universe. Allocation total variation measures displacement in aggregate credit, while Kendall's τ_b measures changes in active-cell ordering. Construction and robustness details are in Appendix I.3.

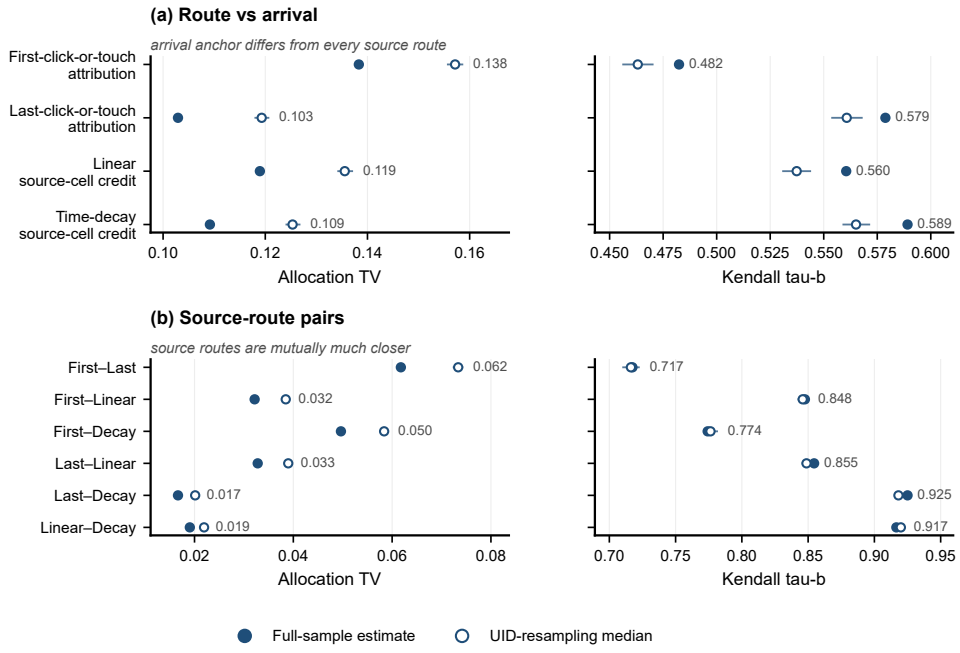

Figure 6: **Experiment 2: source ambiguity localizes route non-uniqueness.** Row group (a) compares each source-time route with the common arrival-time accounting anchor, with allocation total variation on the left and Kendall's τ_b on the right. Row group (b) reports the same two statistics for every pair of source-time routes, in the same left–right order. Filled markers are full-sample estimates, open markers are UID-cluster resampling medians, and horizontal ranges summarize resampling sensitivity. Disagreement against the arrival anchor is systematically larger than disagreement between source-time routes. The resampling ranges are empirical pipeline-sensitivity ranges, not confidence intervals.

Relative to arrival-time accounting, source-time allocation-TV distances range from 0.103 to 0.138, while Kendall's τ_b ranges from 0.482 to 0.589. The disagreement is concentrated where source

correspondence is ambiguous. Singleton-candidate journeys account for 80.14% of the retained cohort and force the four source-time routes to agree. Mean pairwise allocation TV rises to 0.179 for two candidate cells and 0.275 for three or more.

The mechanism is therefore localized rather than global: when the log admits multiple plausible source cells, the same delayed conversion evidence can induce different route-level allocations and rankings. Additional pairwise, entropy, candidate-window, support-threshold, and ranking-depth diagnostics preserve this pattern (Appendix I.3 and Appendix I.7). The logged cohort therefore establishes source-route non-uniqueness on the retained support.

F.4 Experiment 3: Semi-Synthetic Logged-Data Decision Recovery

Experiment 3 asks whether lower held-out score MAE translates into more accurate recovery of the pairwise comparisons and choices that define the supported decision problem. **Claim tested (Claim C).** *Lower score-prediction MAE need not imply more accurate action-gap recovery on common logged support.* Using KuaiRand-1K (Gao et al., 2022), the history period defines the top-20 tag action space and fits the history-based routes; a later evaluation period supplies a constructed six-hour weighted target, and a deterministic pseudo-random rule assigns synthetic arrival lags of six to ten hours. The experiment is therefore semi-synthetic, with evaluation restricted to common supported action cells. Arrival carrier is a deliberate source-misbinding control, Historical mean uses matured history-side averages, and Ridge proxy adds a history-fitted linear proxy using the same pre-decision information. Details are given in Appendix I.4.

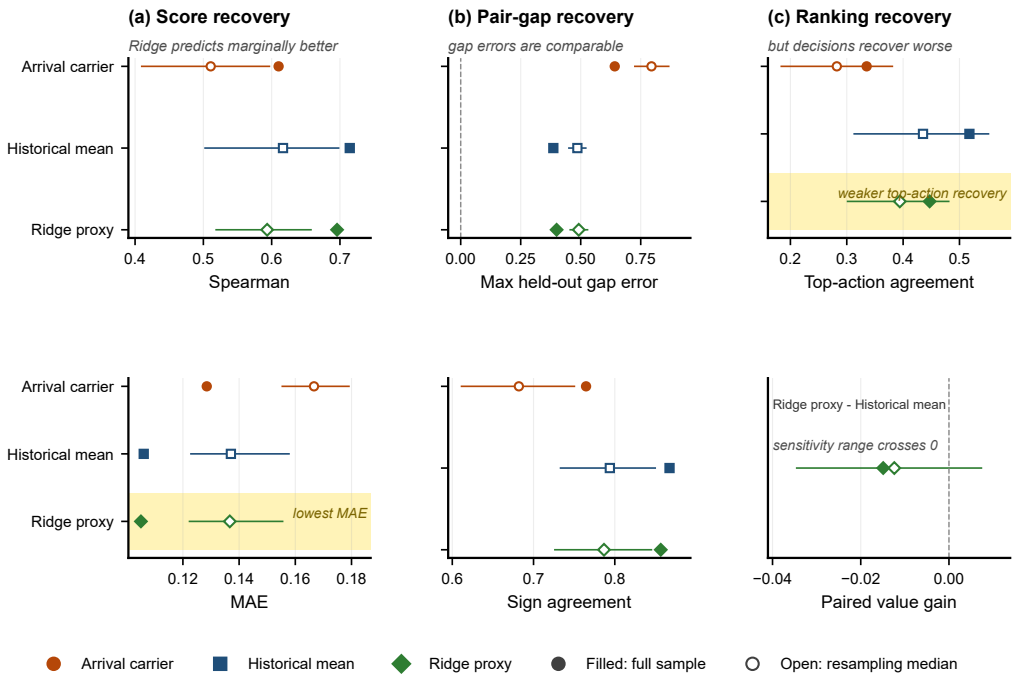

Figure 7: **Experiment 3: full-sample score and decision recovery on logged support.** Panel (a) reports slightly lower full-sample MAE for Ridge than Historical mean (0.105 vs. 0.106), together with lower score-rank correlation. Panel (b) reports higher full-sample maximum pair-gap error and lower pairwise sign agreement for Ridge. Panel (c) reports lower full-sample top-action agreement and a Ridge-minus-Historical paired value contrast of -0.0149; the horizontal range summarizes empirical user-resampling sensitivity for this value contrast only. Table 12 reports the paired resampling summaries for MAE, maximum gap error, sign agreement, and top-action agreement. All four empirical 2.5%–97.5% ranges cross zero; these are pipeline-sensitivity ranges, not formal confidence intervals.

In the full sample, Ridge has the lowest MAE, 0.105 versus 0.106 for Historical mean, but the decision-relevant ordering reverses. Its maximum held-out pair-gap error is 0.400 versus 0.386, gap-sign agreement is 0.857 versus 0.867, and top-action agreement is 0.447 versus 0.518. The Ridge-minus-Historical paired value contrast is -0.0149, with empirical user-resampling sensitivity range $[-0.0347, 0.0075]$.

In the paired resampling analysis, all four metric-difference ranges cross zero (Table 12); Ridge has lower MAE in only 54.6% of the paired resamples. The full-sample decision-side disadvantage is directionally prevalent across the paired user-resampling pipeline: Ridge has higher maximum gap error in 74.2%, lower sign agreement in 91.0%, and lower top-action agreement in 84.9% of resamples. However, all empirical metric-difference ranges cross zero, so the experiment does not establish a stable reversal under user resampling. The resampling ranges are sensitivity summaries rather than formal confidence intervals.

F.5 Experiment 4: Route Information, Audit Information, and Calibration

Experiment 4 asks whether information that improves route construction has the same empirical role as evidence used to audit or calibrate a fixed route. **Claim tested (Claim D).** *Good audit/calibration statistics do not themselves certify structural validity.* The controlled simulator exposes the structural action-gap reference and varies three components separately: route-label retention q_{route}, audit-evidence sampling ρ_{audit}, and an out-of-fold affine calibration step. Here q_{route} controls source information available to route construction, whereas p_{src} in Section E.3 controls source-label observability for a fixed route. The population alignment statistic is the mean pairwise gap discrepancy D^r_{pair}, a controlled audit diagnostic rather than the worst-pair alignment budget in Section D. Complete settings are reported in Appendix I.5.

Route information changes the population object being evaluated. At $\sigma_{\mathrm{proxy}} = 0.25$, D^r_{pair} decreases from 0.02575 at $q_{\mathrm{route}} = 0$ to 0.01928 at 0.3, 0.00870 at 0.7, and zero at 1.

Audit information answers a different question: how accurately can that fixed-route discrepancy be estimated from finite evidence? At a 10% evidence rate, ambiguity-selective unweighted sampling has bias 0.03163 and RMSE 0.03193, while Hájek IPW using the known simulated inclusion probabilities reduces them to 0.00018 and 0.00262. This design matches Theorem E.12, but the estimand here is pair-average discrepancy; these results do not instantiate the worst-pair structural-regret certificate in Corollary E.11.

Calibration is different again. For the affine-linked control, out-of-fold calibration reduces pairwise discrepancy from 0.241 to 0.046, giving recoverability 0.809. After unit-level correspondence is destroyed within temporal blocks, calibration still reduces discrepancy from 0.396 to 0.293, while recoverability remains 0.260. Lower calibrated discrepancy is therefore not, by itself, evidence of structural identification.

Experiment 4 realizes the separation in Section E directly: route information changes the population route, audit information changes finite-evidence estimation, and calibration changes restricted recoverability. These operations can all improve a reported quantity without providing interchangeable evidence of validity.

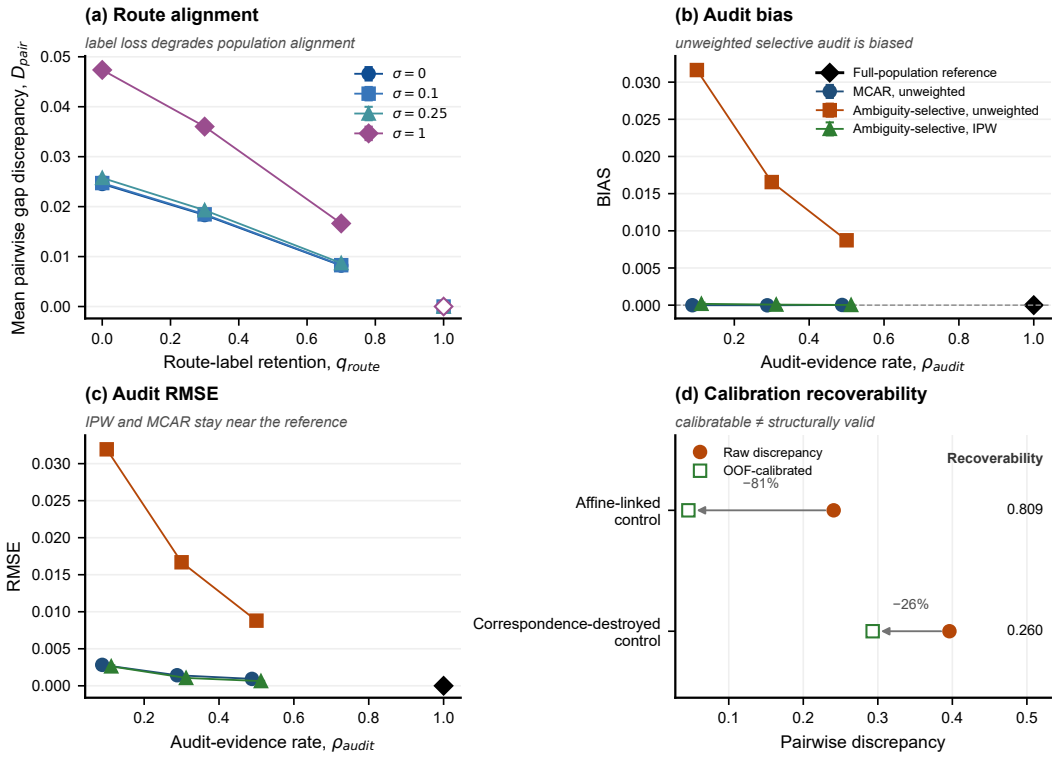

Figure 8: **Experiment 4: route information, audit information, and calibration change different empirical objects.** Panel (a) changes information used to construct the route and therefore changes the population discrepancy D^r_{pair}. Panels (b)–(c) freeze that route and vary audit evidence, separating estimation bias and RMSE from the underlying route discrepancy; filled black diamonds mark the deterministic full-population reference. Panel (d) changes only the calibration map: discrepancy falls strongly for the affine-linked control but also falls after unit-level correspondence is destroyed, showing that recoverability is not an identification certificate. Marker shape in Panel (a) encodes σ_{proxy}, open markers are the deterministic retention endpoint, and audit intervals are estimate ± 1.96 MCSE; finite-audit Monte Carlo details are reported in Appendix I.5.

F.6 Empirical Synthesis

Across the four experiments, the common empirical object is preservation of decision-time action comparisons, but the evidence supporting each conclusion is not interchangeable. Controlled complete maps show that similar delay can induce different benchmark damage and directly realize the cancellation, margin-threshold, and persistent-conflict regimes. Logged conversion data localize route non-uniqueness to source ambiguity, while the semi-synthetic KuaiRand study separates lower full-sample predictive MAE from gap and choice recovery, with Ridge's decision-side disadvantage recurring in most user resamples. Controlled audit interventions finally separate route alignment, audit reliability, and calibration recoverability. The controlled experiments expose structural comparison information by construction; the delayed-conversion log supports source-route non-uniqueness under ambiguous source correspondence, and the semi-synthetic study supports a narrower diagnostic of predictive fit versus action-gap/choice recovery on common logged support.

G Timing and Notation Support

Figure 9 illustrates the distinction between structural decision time and observation clock time used throughout the paper.

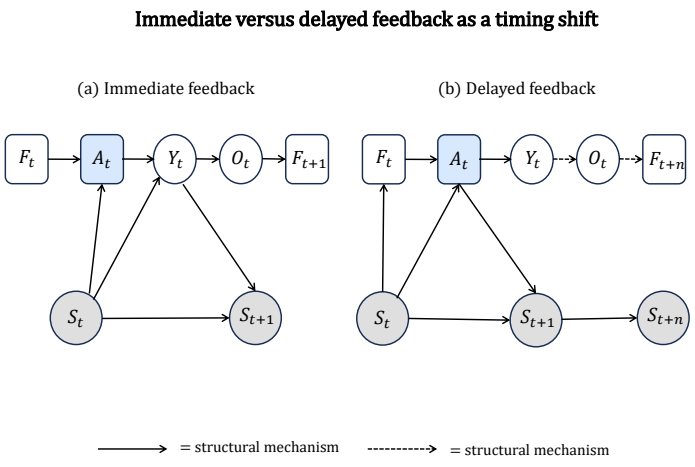

Figure 9: Timing distinction underlying source-time evaluation. With immediate feedback, A_t, Y_t, and O_t share the same round index. With delayed feedback, Y_t is still generated by A_t under the structural state S_t, but becomes observable only at a later clock time.

H Proofs and Technical Derivations

This section proves the results in Sections D and E and gives the supporting route-specific derivations. Four proof devices recur throughout: oscillation control on loss maps modulo action-independent shifts; margin arguments for set-valued structural and route optima; secant-slope control for nonlinear transformations; and observational-equivalence constructions that remain valid under adaptive policies. The short route-specific bounds are then obtained as direct specializations of these generic arguments.

Proof of Proposition C.1. By admissibility, conditional on $\mathcal{F}_t$, A_t has law $\pi_t(\cdot \mid \mathcal{F}_t)$, so

$$\mathbb{E}[L_t^c(A_t) \mid \mathcal{F}_t] = \sum_{a \in \mathcal{A}} \pi_t(a \mid \mathcal{F}_t)\mathbb{E}[L_t^c(a) \mid \mathcal{F}_t] = \sum_{a \in \mathcal{A}} \pi_t(a \mid \mathcal{F}_t) L_t^{\mathcal{F}}(a).$$

The same conditioning yields

$$\mathbb{E}[r_t^c \mid \mathcal{F}_t] = \sum_a \pi_t(a \mid \mathcal{F}_t) L_t^{\mathcal{F}}(a) - \mathbb{E}\Big[\min_a L_t^c(a) \Big| \mathcal{F}_t\Big] = d_t^{\mathcal{F}}(\pi) + \iota_t.$$

Moreover, for every fixed a, $\mathbb{E}[\min_b L_t^c(b) \mid \mathcal{F}_t] \le \mathbb{E}[L_t^c(a) \mid \mathcal{F}_t] = L_t^{\mathcal{F}}(a)$. Taking the minimum over a proves $\iota_t \ge 0$. Finally, the tower property and finite summation give the horizon identity. □

H.1 Validity Geometry and the Preservation Hierarchy

Recall $e_t^r(a) = \widetilde{L}_t^r(a) - L_t^c(a)$. For any $a, b \in \mathcal{A}$, $G_t^r(a,b) - G_t^c(a,b) = e_t^r(a) - e_t^r(b)$. Hence

$$\begin{aligned}\delta_t^r &= \max_{a,b\in\mathcal{A}} |e_t^r(a) - e_t^r(b)| \\ &= \max_{a\in\mathcal{A}} e_t^r(a) - \min_{a\in\mathcal{A}} e_t^r(a) = \operatorname{osc}(e_t^r).\end{aligned}$$

Proof of Proposition D.3. If $\delta_t^r = 0$, then $G_t^r(a,b) = G_t^c(a,b), \qquad \forall a, b \in \mathcal{A}$. Therefore every pair has the same exact sign, including ties, and hence $\rho_t^r = 0$.

Now suppose $\rho_t^r = 0$. Let $b \in \mathcal{A}_t^{r,\star}$. If $b \notin \mathcal{A}_t^{c,\star}$, choose any $a^\star \in \mathcal{A}_t^{c,\star}$. Then $G_t^c(b, a^\star) > 0$. Exact ranking preservation gives $G_t^r(b, a^\star) > 0$, which contradicts the route optimality of b, since route optimality requires $\widetilde{L}_t^r(b) \le \widetilde{L}_t^r(a^\star)$. Thus every route-optimal action is structurally optimal: $\mathcal{A}_t^{r,\star} \subseteq \mathcal{A}_t^{c,\star}$, and therefore $\chi_t^r = 0$.

At the horizon level, $\mathfrak{A}_T^r/T = 0$ implies $\delta_t^r = 0$ for every t, since the defects are nonnegative. Likewise, $\overline{\rho}_T^r = 0$ implies $\rho_t^r = 0$ for every t. The horizon hierarchy follows.

□

H.2 Sharp Regret Stability and Uniform Regret Maps

Proof of Theorem D.4. Fix a round t and write $e(a) = \widetilde{L}_t^r(a) - L_t^c(a)$. Define

$$m_t := \min_{a\in\mathcal{A}} \{L_t^c(a) + e(a)\} - \min_{a\in\mathcal{A}} L_t^c(a).$$

Because $\min_a L_t^c(a) + \min_a e(a) \le \min_a\{L_t^c(a) + e(a)\}$, we have $m_t \ge \min_a e(a)$. Conversely, evaluating the route loss at any structural minimizer gives $m_t \le \max_a e(a)$. Therefore $m_t \in [\min_a e(a), \max_a e(a)]$.

Using $\widetilde{L}_t^r(a) = L_t^c(a) + e(a)$,

$$r_t^r - r_t^c = e(A_t) - m_t.$$

Both $e(A_t)$ and m_t lie in the interval $[\min_a e(a), \max_a e(a)]$, so $|r_t^r - r_t^c| \le \max_a e(a) - \min_a e(a) = \delta_t^r$. Thus

$$|r_t^c - r_t^r| \le \delta_t^r.$$

Summing and applying the triangle inequality yields $|R_T^c - R_T^r| \le \sum_{t=1}^T \delta_t^r = \mathfrak{A}_T^r$. For a common admissible policy π,

$$\begin{aligned}|\mathcal{R}_T^c(\pi) - \mathcal{R}_T^r(\pi)| &= |\mathbb{E}_\pi[R_T^c - R_T^r]| \\ &\le \mathbb{E}_\pi\left[|R_T^c - R_T^r|\right] \\ &\le \mathbb{E}_\pi[\mathfrak{A}_T^r].\end{aligned}$$

The constant 1 cannot be improved. With two actions,

$$L^c = (0,1), \qquad \widetilde{L}^r = \left(0, \frac{1}{2}\right), \qquad A = a_2,$$

we have $r^c = 1, \qquad r^r = \frac{1}{2}, \qquad \delta^r = \frac{1}{2}$, and therefore $|r^c - r^r| = \delta^r$.

□

Proof of Corollary D.5. The one-sided part of Theorem D.4 gives $0 \le R_T^c \le R_T^r + \mathfrak{A}_T^r$. Dividing by T proves the pathwise statement. Taking expectations in the same inequality gives $0 \le \mathcal{R}_T^c(\pi) \le \mathcal{R}_T^r(\pi) + \mathbb{E}_\pi[\mathfrak{A}_T^r]$, which proves the expected statement. □

Proof of Proposition D.6. Retain the notation from the preceding proof. For every action a,

$$q_t^r(a) - q_t^c(a) = e(a) - m_t.$$

Since both $e(a)$ and m_t lie between $\min e$ and $\max e$, $|q_t^r(a) - q_t^c(a)| \le \operatorname{osc}(e) = \delta_t^r$. Taking the maximum over a gives $d_{t,\mathrm{reg}}^r \le \delta_t^r$.

For the reverse inequality, choose a_+ and a_- satisfying $e(a_+) = \max_a e(a)$, $e(a_-) = \min_a e(a)$. Then

$$\begin{aligned}\delta_t^r &= e(a_+) - e(a_-)\\ &= \big[q_t^r(a_+) - q_t^c(a_+)\big] - \big[q_t^r(a_-) - q_t^c(a_-)\big]\\ &\le 2d_{t,\mathrm{reg}}^r.\end{aligned}$$

Hence $\frac{1}{2}\delta_t^r \le d_{t,\mathrm{reg}}^r \le \delta_t^r$.

Summing over t gives

$$\frac{1}{2}\mathfrak{A}_T^r \le \sum_{t=1}^T d_{t,\mathrm{reg}}^r \le \mathfrak{A}_T^r.$$

□

H.3 Margins, Choice Preservation, and Persistent Route Conflict

Proof of Proposition D.7. For every ordered pair (a, b), Definition D.1 gives

$$|G_t^r(a,b) - G_t^c(a,b)| \le \delta_t^r.$$

If $G_t^c(a,b) > \delta_t^r$, then $G_t^r(a,b) > 0$; if $G_t^c(a,b) < -\delta_t^r$, then $G_t^r(a,b) < 0$. This proves part (1).

For part (2), suppose that $b \in \mathcal{A}_t^{r,\star}$ but $b \notin \mathcal{A}_t^{c,\star}$, and choose $a^\star \in \mathcal{A}_t^{c,\star}$. Then $G_t^c(b,a^\star) \ge \mu_t^c > \delta_t^r$. Part (1) implies $G_t^r(b,a^\star) > 0$, contradicting route optimality of b. Thus every route minimizer is a structural minimizer.

For part (3), choose $b^\star \in \mathcal{A}_t^{r,\star}$ satisfying $q_t^c(b^\star) = \gamma_t^r$, and choose any $a^\star \in \mathcal{A}_t^{c,\star}$. The two optimal sets are disjoint, so $a^\star \notin \mathcal{A}_t^{r,\star}$ and $q_t^r(a^\star) \ge \eta_t^r$. Consequently,

$$G_t^c(b^\star,a^\star) = \gamma_t^r, \qquad G_t^r(b^\star,a^\star) = -q_t^r(a^\star) \le -\eta_t^r,$$

which yields $\delta_t^r \ge \gamma_t^r + \eta_t^r$. Equality is attained with two actions and loss vectors $L_t^c = (0,\gamma)$ and $\widetilde{L}_t^r = (\eta, 0)$, for arbitrary $\gamma, \eta > 0$. □

Proof of Proposition D.10. Fix $\kappa > 0$. For each horizon $T \ge 2$, construct a separate two-action instance $\mathcal{I}_T$ and set

$$d_T := T^{-1/\kappa}.$$

On exactly one round of $\mathcal{I}_T$, use

$$L_t^c = (0, d_T), \qquad \widetilde{L}_t^r = (d_T, 0),$$

and on each of the remaining $T-1$ rounds use

$$L_t^c = \widetilde{L}_t^r = (0,1).$$

The reversed round has $\delta_t^r = 2d_T$ and $\rho_t^r = \chi_t^r = 1$; all other rounds have zero defect and no disagreement. Therefore

$$\frac{\mathfrak{A}_T^r}{T} = 2T^{-(1+1/\kappa)}, \qquad \overline{\rho}_T^r = \overline{\chi}_T^r = \frac{1}{T}.$$

The structural pairwise gap and structural optimality margin equal d_T on the reversed round and 1 otherwise. Hence, for either $\ell \in \{\mathrm{ord}, \mathrm{ch}\}$,

$$H_{T,\ell}(\varepsilon) = \begin{cases} 0, & 0 < \varepsilon < d_T,\\ 1/T, & d_T \le \varepsilon < 1,\\ 1, & \varepsilon \ge 1.\end{cases}$$

Because $1/T = d_T^\kappa$,

$$H_{T,\ell}(\varepsilon) \le \varepsilon^\kappa \qquad \text{for every } \varepsilon > 0.$$

Thus the margin-mass condition holds for every horizon T with the same constant $C_\ell = 1$ and exponent $\kappa_\ell = \kappa$.

Finally,

$$\frac{1}{T} = 2^{-\kappa/(\kappa+1)} \left(\frac{\mathfrak{A}_T^r}{T}\right)^{\kappa/(\kappa+1)}.$$

Therefore the exponent is attained by every member of the horizon-indexed family $\{\mathcal{I}_T : T \ge 2\}$. These instances are constructed separately for each horizon and are not claimed to be restrictions of a single nested infinite trajectory. □

Proof of Corollary D.8. Under the uniform margin condition $\mu_t^c \ge \underline{\mu} > 0$, Part (2) of Proposition D.7 implies $\chi_t^r > 0 \Rightarrow \delta_t^r \ge \underline{\mu}$. Since $\chi_t^r \le \mathbf{1}\{\chi_t^r > 0\}$, we obtain $\chi_t^r \le \frac{\delta_t^r}{\underline{\mu}}$. Averaging over t gives $\overline{\chi}_T^r \le \mathfrak{A}_T^r/(\underline{\mu}T)$.

□

Proof of Proposition D.9. For an audited pair (a, b), Proposition D.7(1) implies that a ranking disagreement can occur only if $|G_t^c(a, b)| \le \delta_t^r$. Fix $\varepsilon > 0$. Therefore

$$\mathbf{1}\{\operatorname{sgn} G_t^r(a, b) \ne \operatorname{sgn} G_t^c(a, b)\} \le \mathbf{1}\{|G_t^c(a, b)| \le \varepsilon\} + \mathbf{1}\{\delta_t^r \ge \varepsilon\}.$$

Averaging first over $\mathcal{P}$ and then over t, and using $\mathbf{1}\{\delta_t^r \ge \varepsilon\} \le \delta_t^r/\varepsilon$, gives

$$\overline{\rho}_T^r \le H_{T,\mathrm{ord}}(\varepsilon) + \frac{1}{T}\sum_{t=1}^T \frac{\delta_t^r}{\varepsilon} = H_{T,\mathrm{ord}}(\varepsilon) + \frac{\mathfrak{A}_T^r}{\varepsilon T}.$$

For choice disagreement, Proposition D.7(2) implies that $\chi_t^r > 0$ and $\mu_t^c > \varepsilon$ together require $\delta_t^r \ge \mu_t^c > \varepsilon$. Since $0 \le \chi_t^r \le 1$,

$$\chi_t^r \le \mathbf{1}\{\mu_t^c \le \varepsilon\} + \mathbf{1}\{\delta_t^r \ge \varepsilon\}.$$

Averaging and applying the same bound on the second indicator yields

$$\overline{\chi}_T^r \le H_{T,\mathrm{ch}}(\varepsilon) + \frac{\mathfrak{A}_T^r}{\varepsilon T}.$$

Now suppose $H_{T,\ell}(\varepsilon) \le C_\ell \varepsilon^{\kappa_\ell}$. Writing $x_T = \mathfrak{A}_T^r/T$, the preceding display gives $\overline{d}_{T,\ell}^r \le C_\ell \varepsilon^{\kappa_\ell} + x_T/\varepsilon$. If $x_T = 0$, nonnegativity implies $\delta_t^r = 0$ for every round, and Proposition D.3 gives zero ordinal and choice discrepancy. If $x_T > 0$, choosing $\varepsilon \asymp x_T^{1/(\kappa_\ell+1)}$ gives

$$\overline{d}_{T,\ell}^r = O\Big(x_T^{\kappa_\ell/(\kappa_\ell+1)}\Big),$$

which proves the stated rate.

□

Proof of Theorem D.11. Fix $t \in \mathcal{T}_0$.

If $A_t \in \mathcal{A}_t^{r,\star}$, complete route conflict implies that A_t is structurally suboptimal. By definition of γ_t^r, $r_t^c = q_t^c(A_t) \ge \gamma_t^r \ge \gamma$, and therefore $\frac{r_t^c}{\gamma} + \frac{r_t^r}{\eta} \ge 1$.

If instead $A_t \notin \mathcal{A}_t^{r,\star}$, then by definition of η_t^r, $r_t^r = q_t^r(A_t) \ge \eta_t^r \ge \eta$, so again $\frac{r_t^c}{\gamma} + \frac{r_t^r}{\eta} \ge 1$.

Summing over all $t \in \mathcal{T}_0$ gives

$$\frac{R_T^c}{\gamma} + \frac{R_T^r}{\eta} \ge |\mathcal{T}_0|.$$

Rearranging yields $R_T^c \ge \gamma|\mathcal{T}_0| - \frac{\gamma}{\eta}R_T^r$. Along a sequence of horizons on which $|\mathcal{T}_0| \ge cT$ for some $c > 0$ and the same constants $\gamma, \eta > 0$ lower-bound the conflict margins uniformly, $R_T^r = o(T)$ therefore implies $R_T^c = \Omega(T)$.

□

Proof of Corollary D.12. If $A_t \in \mathcal{A}_t^{r,\star}$ on every $t \in \mathcal{T}_0$, then $r_t^r = 0$ and, by complete conflict and the structural margin condition, $r_t^c \geq \gamma$. Thus $R_T^c \geq \gamma|\mathcal{T}_0|$.

□

H.4 Decision-Preserving Transformation Classes

Proof of Theorem D.13. Fix t and write

$$\widetilde{L}_t^r(a) = \alpha_t L_t^c(a) + \beta_t + \zeta_t^r(a), \qquad \alpha_t > 0.$$

The action-independent term β_t cancels from route regret.

Define

$$m_t^\zeta := \min_a \{\alpha_t L_t^c(a) + \zeta_t^r(a)\} - \alpha_t \min_a L_t^c(a).$$

As in the proof of Theorem D.4, $m_t^\zeta \in [\min_a \zeta_t^r(a), \max_a \zeta_t^r(a)]$. Therefore

$$r_t^r - \alpha_t r_t^c = \zeta_t^r(A_t) - m_t^\zeta,$$

and hence

$$|r_t^r - \alpha_t r_t^c| \leq \mathrm{osc}(\zeta_t^r) = \varepsilon_t^r.$$

Summing gives

$$\left| R_T^r - \sum_{t=1}^T \alpha_t r_t^c \right| \leq \sum_{t=1}^T \varepsilon_t^r.$$

If $0 < \underline{\alpha} \leq \alpha_t \leq \overline{\alpha}$, then

$$\underline{\alpha} R_T^c \leq \sum_t \alpha_t r_t^c \leq \overline{\alpha} R_T^c.$$

Combining these inequalities with the preceding bound yields

$$\underline{\alpha} R_T^c - \sum_t \varepsilon_t^r \leq R_T^r \leq \overline{\alpha} R_T^c + \sum_t \varepsilon_t^r,$$

and therefore

$$R_T^c \leq \frac{R_T^r + \sum_t \varepsilon_t^r}{\underline{\alpha}}.$$

□

Proof of Proposition D.14. Define the transformation-only loss map $L_t^h(a) = h_t(L_t^c(a))$ and its regret $r_t^h = L_t^h(A_t) - \min_a L_t^h(a)$. Because every secant slope of h_t on I_t is strictly positive, h_t is strictly increasing on I_t. Hence every minimizer of L_t^c is also a minimizer of L_t^h.

Let $x_t = \min_a L_t^c(a)$. Then $r_t^h = h_t(L_t^c(A_t)) - h_t(x_t)$. When $r_t^c > 0$, the secant-slope bounds give $m_t r_t^c \leq r_t^h \leq M_t r_t^c$. When $r_t^c = 0$, all three quantities are zero, so the same inequalities hold.

Now add the residual $\zeta_t^r(a)$. Applying the same oscillation argument as in Theorem D.4 to L_t^h and $L_t^h + \zeta_t^r$ yields $|r_t^r - r_t^h| \leq \varepsilon_t^r$. Consequently,

$$m_t r_t^c - \varepsilon_t^r \leq r_t^r \leq M_t r_t^c + \varepsilon_t^r.$$

If $m_t \geq \underline{m} > 0$, then $\underline{m} R_T^c \leq R_T^r + \sum_{t=1}^T \varepsilon_t^r$, which gives

$$R_T^c \leq \frac{R_T^r + \sum_{t=1}^T \varepsilon_t^r}{\underline{m}}.$$

□

Proof of Corollary D.15. If $\widetilde{L}_t^r(a) = h_t(L_t^c(a)) + c_t$ with h_t strictly increasing, then for every pair a, b,

$$\operatorname{sgn}\left(\widetilde{L}_t^r(a) - \widetilde{L}_t^r(b)\right) = \operatorname{sgn}\left(L_t^c(a) - L_t^c(b)\right).$$

Thus $\rho_t^r = 0$. Strict monotonicity also preserves the complete structural minimizer set, so $\mathcal{A}_t^{r,\star} = \mathcal{A}_t^{c,\star}$ and hence $\chi_t^r = 0$. The numerical magnitudes of the gaps need not be preserved, so δ_t^r may remain positive.

□

H.5 Arrival, Batch, and Proxy Specializations

Proof of Proposition D.16. Let $z_a = \Psi_a(S_t') - \Psi_a(S_t)$. For any actions a, b,

$$\begin{aligned} G_t^r(a,b) - G_t^c(a,b) &= [L^c(a; S_t') - L^c(a; S_t)] \\ &\quad - [L^c(b; S_t') - L^c(b; S_t)] \\ &= z_a - z_b. \end{aligned}$$

Therefore

$$\begin{aligned} \delta_t^r &= \max_{a,b} |z_a - z_b| \\ &= \max_a z_a - \min_a z_a \\ &= \operatorname{osc}\left(\Psi(S_t') - \Psi(S_t)\right) \\ &= d_{\mathrm{dec}}(S_t, S_t'). \end{aligned}$$

□

Proof of Theorem D.17. By definition, $\delta_t^{\mathrm{arr}} = 0$ if and only if

$$G_t^{\mathrm{arr}}(a,b) = G_t^c(a,b), \qquad \forall a, b \in \mathcal{A}.$$

Suppose these gaps are equal. Fix any reference action $a_0 \in \mathcal{A}$. Then for every a, $L_{\kappa(t)}^c(a) - L_{\kappa(t)}^c(a_0) = L_t^c(a) - L_t^c(a_0)$. Hence $L_{\kappa(t)}^c(a) - L_t^c(a) = L_{\kappa(t)}^c(a_0) - L_t^c(a_0)$, which is independent of a. Defining $c_t = L_{\kappa(t)}^c(a_0) - L_t^c(a_0)$ gives $L_{\kappa(t)}^c(a) = L_t^c(a) + c_t, \qquad \forall a$.

Conversely, if $L_{\kappa(t)}^c(a) = L_t^c(a) + c_t$ for every a, then the common shift cancels from every pairwise action gap, so $G_t^{\mathrm{arr}}(a,b) = G_t^c(a,b)$ for all pairs and hence $\delta_t^{\mathrm{arr}} = 0$.

□

Proof of Proposition D.18. From Proposition D.16,

$$\delta_t^{\mathrm{arr}} = d_{\mathrm{dec}}\left(S_t, S_{\kappa(t)}\right).$$

Therefore, under $d_{\mathrm{dec}}(s, s') \le L_\Psi \|s - s'\|$,

$$\delta_t^{\mathrm{arr}} \le L_\Psi \left\|S_t - S_{\kappa(t)}\right\|.$$

Summing yields

$$\mathfrak{A}_T^{\mathrm{arr}} \le L_\Psi \sum_{t=1}^{T} \left\|S_t - S_{\kappa(t)}\right\|.$$

Under the stronger action-wise Lipschitz condition, $|L^c(a; s) - L^c(a; s')| \le L_S \|s - s'\|$, for any a, b,

$$\begin{aligned} &|G^c(a,b; s) - G^c(a,b; s')| \\ &\le |L^c(a; s) - L^c(a; s')| + |L^c(b; s) - L^c(b; s')| \\ &\le 2L_S \|s - s'\|. \end{aligned}$$

Taking the maximum over a, b gives $d_{\mathrm{dec}}(s, s') \le 2L_S \|s - s'\|$, and the stated stronger bound follows.

□

Batch-arrival derivation. Suppose

$$\widetilde{L}_t^{\mathrm{batch}}(a) = \sum_{s \in \mathcal{B}_t} w_{t,s} L^c(a; S_s), \qquad w_{t,s} \geq 0, \qquad \sum_{s \in \mathcal{B}_t} w_{t,s} = 1.$$

For every pair a, b,

$$G_t^{\mathrm{batch}}(a,b) = \sum_{s \in \mathcal{B}_t} w_{t,s} G^c(a,b; S_s).$$

Hence

$$\begin{aligned} &\left| G_t^{\mathrm{batch}}(a,b) - G_t^c(a,b) \right| \\ &\leq \sum_{s \in \mathcal{B}_t} w_{t,s} \left| G^c(a,b;S_s) - G^c(a,b;S_t) \right| \\ &\leq \sum_{s \in \mathcal{B}_t} w_{t,s} d_{\mathrm{dec}}(S_t, S_s). \end{aligned}$$

Taking the maximum over a, b gives

$$\delta_t^{\mathrm{batch}} \leq \sum_{s \in \mathcal{B}_t} w_{t,s} d_{\mathrm{dec}}(S_t, S_s).$$

The state-space Lipschitz version follows immediately.

Proof of Proposition D.19. By Proposition D.16, $\delta_t^{\mathrm{proxy}} = d_{\mathrm{dec}}(S_t, \widehat{S}_t)$. The assumed decision-space Lipschitz condition therefore gives $\delta_t^{\mathrm{proxy}} \leq L_\Psi \|S_t - \widehat{S}_t\|$. Summing over t yields

$$\mathfrak{A}_T^{\mathrm{proxy}} \leq L_\Psi \sum_{t=1}^{T} \|S_t - \widehat{S}_t\|.$$

The stronger action-wise Lipschitz bound follows from the same two-term triangle inequality used in the proof of Proposition D.18: $d_{\mathrm{dec}}(s, s') \leq 2L_S \|s - s'\|$.

□

Proof of Corollary D.20. For the state-valued proxy route, $\delta_t^{\mathrm{proxy}} = d_{\mathrm{dec}}(S_t, \widehat{S}_t)$. If $d_{\mathrm{dec}}(S_t, \widehat{S}_t) < \mu_t^c$, then Proposition D.7 implies $\mathcal{A}_t^{\mathrm{proxy},\star} \subseteq \mathcal{A}_t^{c,\star}$. The stronger state-space condition follows from the action-wise Lipschitz comparison bound established above.

□

Proof of Proposition D.21. Write the implemented route error vector as

$$\widetilde{L}_t^r - L_t^c = \left[L^c(\,\cdot\,; \widehat{S}_t) - L^c(\,\cdot\,; S_t) \right] + \left[\widehat{L}(\,\cdot\,; \widehat{S}_t) - L^c(\,\cdot\,; \widehat{S}_t) \right].$$

The gap defect is the oscillation of this vector. By the triangle inequality for oscillation,

$$\begin{aligned} \delta_t^r &= \mathrm{osc}\Big(\widetilde{L}_t^r - L_t^c \Big) \\ &\leq \mathrm{osc}\Big(L^c(\,\cdot\,; \widehat{S}_t) - L^c(\,\cdot\,; S_t) \Big) + \mathrm{osc}\Big(\widehat{L}(\,\cdot\,; \widehat{S}_t) - L^c(\,\cdot\,; \widehat{S}_t) \Big) \\ &= d_{\mathrm{dec}}(S_t, \widehat{S}_t) + \varepsilon_{t,\mathrm{ker}}^r. \end{aligned}$$

Summing over t gives the cumulative bound. The inequality need not be an equality because the two actionwise error vectors can offset one another in their pairwise contrasts. If the fitted kernel equals the structural kernel at $\widehat{S}_t$, then $\varepsilon_{t,\mathrm{ker}}^r = 0$, recovering the oracle proxy-route identity.

□

H.6 Identification and Fully Unlabelled Feedback

Proof of Proposition E.3. Under the proposition's edge-only cardinal-information contract, the compatible loss maps are constrained by the exact audited edge gaps and by no additional restrictions linking disconnected components. Suppose first that $\mathcal{G}_i^{\mathrm{aud}}$ is connected. Fix a reference action a_0. For any $a \in \mathcal{A}$, choose a path $a_0 = v_0, v_1, \ldots, v_m = a$. The exact edge gaps determine $L(v_j) - L(v_{j-1})$ for every edge on this path. Summing along the path determines $L(a) - L(a_0)$. Hence every action loss relative to a_0 is determined, and therefore the complete quotient class

$$[L] \in \mathbb{R}^{|\mathcal{A}|} / \operatorname{span}\{\mathbf{1}\}$$

is uniquely determined. Thus all structural action gaps are identified.

Conversely, suppose $\mathcal{G}_i^{\mathrm{aud}}$ is disconnected. Let C be one connected component and define

$$L'(a) = \begin{cases} L(a) + c, & a \in C, \\ L(a), & a \notin C, \end{cases}$$

for some $c \neq 0$. Every observed edge lies within a connected component, so every audited edge gap is unchanged. However, every gap between an action in C and an action outside C changes by c. Thus the complete cardinal gap geometry is not uniquely determined.

□

Proof of Theorem E.4. Take $T = 2m$, $\mathcal{A} = \{0, 1\}$, $X_t \equiv \varnothing$, $S_{t+1} = 1 - S_t$, $Y_t(a) = S_t$, and $\ell(a, y) = \mathbf{1}\{a \neq y\}$. The worlds differ only in the initial state: $\mathcal{M}_0 : S_1 = 0$ and $\mathcal{M}_1 : S_1 = 1$. Let $\tau_{2k-1} = 1$ and $\tau_{2k} = 0$. Under either world, odd clocks have no arrival and every even clock contains the unlabelled multiset $\{\{0, 1\}\}$.

Because the state and outcome processes are action-independent, identical observable histories induce identical conditional action distributions under any common admissible policy. Induction over decision rounds therefore gives $P_{\mathcal{M}_0}^{\pi, I_{\mathrm{unlab}}} = P_{\mathcal{M}_1}^{\pi, I_{\mathrm{unlab}}}$ for every common admissible π.

Now fix the clock-only route $\widetilde{L}_t^r = (0, 1)$ on odd rounds and $(1, 0)$ on even rounds. Under $\mathcal{M}_0$, this equals the structural map at every round, so $\mathsf{Val}_T^r(\mathcal{M}_0) = (0, 0, 0)$. Under $\mathcal{M}_1$, every comparison is reversed, giving $\delta_t^r = 2$, $\rho_t^r = \chi_t^r = 1$, and hence $\mathsf{Val}_T^r(\mathcal{M}_1) = (2, 1, 1)$. This proves non-identification. □

Proof of Corollary E.5. The route-greedy action is the structural optimum in $\mathcal{M}_0$ and the unique structurally worse action in $\mathcal{M}_1$ at every round. The claim follows by summation. □

Proof of Corollary E.6. For n independent fully unlabelled trajectories, the observable laws remain identical under the two worlds; denote their common law by P_n. For any estimator $Z = \widehat{\Theta}_n$, the triangle inequality gives $|Z - \theta_0| + |Z - \theta_1| \geq \Delta_\Theta$. Taking expectation under P_n shows that the larger of the two absolute-error risks is at least $\Delta_\Theta/2$. The constant midpoint estimator $(\theta_0 + \theta_1)/2$ attains this value in both worlds, so the minimax risk equals $\Delta_\Theta/2$. □

Proof of Corollary E.7. In the two-world construction, under $\mathcal{M}_0$ odd sources have outcome 0 and even sources outcome 1; under $\mathcal{M}_1$ the parity pattern is reversed. Thus every observed source–outcome pair (s, Y_s) is world-separating. Since $T = 2m$ and the delay pattern is $\tau_{2k-1} = 1$, $\tau_{2k} = 0$, every one of the T source outcomes arrives within the declared observation window, so $N_T = T$.

Let E_0 be the event that no source label is retained across the nT opportunities. Independent world-independent retention gives $\Pr(E_0) = (1 - p_{\mathrm{src}})^{nT}$. On E_0^c, at least one retained source–outcome pair identifies the world. On E_0, the record reduces to the fully unlabelled record and the conditional observable laws coincide. Hence the conditional minimax absolute-error risk on E_0 is $\Delta_\Theta/2$.

Multiplying by $\Pr(E_0)$ gives the lower bound $\Delta_\Theta(1 - p_{\mathrm{src}})^{nT}/2$. A rule that outputs the identified θ_j on E_0^c and the midpoint on E_0 attains the same risk, proving equality. □

Proof of Corollary E.8. Applying the same Markov kernel to equal input laws preserves equality of the pushforward laws. Deterministic aggregation is the special case of a degenerate kernel. □

H.7 Finite-Evidence Audit and Structural-Regret Certificates

Proof of Proposition E.10. Condition on the fixed implemented route, the fixed finite audit population, and the fixed unit-level contributions $\{d_{i,\ell}^{r,\mathrm{obs}}\}_{i=1}^{M}$ before drawing the simple random audit sample. This conditioning includes all realized validation references and any fitting or calibration used in the contributions. By assumption, the audit sample remains simple random sampling without replacement conditional on these objects. All probabilities in the following argument concern only this conditional audit draw. Suppose $0 \le d_{i,\ell}^{r,\mathrm{obs}} \le B_\ell$.

For a simple random sample without replacement of size n, the conservative Hoeffding bound for finite-population sampling (Hoeffding, 1963) gives

$$\Pr\left(\left|\widehat{D}_{M,\ell}^{r,\mathrm{obs}} - D_{M,\ell}^{r,\mathrm{obs}}\right| \ge \varepsilon\right) \le 2\exp\left(-\frac{2n\varepsilon^2}{B_\ell^2}\right).$$

Setting $\varepsilon = B_\ell\sqrt{\frac{\log(2/\alpha)}{2n}}$ makes the right-hand side equal to α. Hence, with probability at least $1-\alpha$,

$$D_{M,\ell}^{r,\mathrm{obs}} \le \widehat{D}_{M,\ell}^{r,\mathrm{obs}} + B_\ell\sqrt{\frac{\log(2/\alpha)}{2n}}.$$

The proof depends only on the prespecified bounded unit-level audit contribution. Accordingly, the same argument applies to a covered contribution or, when the validation design supports a complete reconstructed cardinal reference, to the full-map contribution used below, provided the same conditional sampling assumptions hold. If the frozen pre-audit fit is random and the assumptions hold almost surely, the conditional failure probability is at most α for almost every realization. Averaging over this randomness preserves the failure bound α; it does not bound the fitted reference's error. If references are refitted as a function of the audit sample, conditioning on their outputs need not preserve the required sampling law, and this argument does not justify that procedure.

□

Proof of Theorem E.12. Write

$$X_i := \frac{d_i}{M}\left(\frac{R_i}{\pi_i} - 1\right), \qquad Z_i := \frac{1}{M}\left(\frac{R_i}{\pi_i} - 1\right).$$

The variables within each row are independent and centered. Direct calculation gives

$$\mathbb{E}[\widehat{D}_{\mathrm{HT}}] = D_M, \qquad \sum_i \mathbb{E}[X_i^2] = \frac{1}{M^2}\sum_i \frac{1-\pi_i}{\pi_i} d_i^2 = V_D,$$

and similarly $\sum_i \mathbb{E}[Z_i^2] = V_W$. Since $d_i^2 \le B^2$, one also has $V_D \le B^2 V_W$, which justifies the design-computable replacement $\epsilon_D \le \bar{\epsilon}_D$. Moreover, $|X_i| \le B/(M\pi_{\min})$ and $|Z_i| \le 1/(M\pi_{\min})$. Bernstein's inequality therefore implies, for every $x > 0$,

$$\mathbb{P}\left(|\widehat{D}_{\mathrm{HT}} - D_M| > \sqrt{2V_D x} + \frac{Bx}{3M\pi_{\min}}\right) \le 2e^{-x},$$

and the analogous bound for $|\widehat{W} - 1|$ with V_W and $B = 1$. Setting $x = \log(4/\alpha)$ and taking a union bound yields both events with probability at least $1-\alpha$. On that event, $\widehat{W} \ge 1 - \epsilon_W > 0$, and

$$\begin{aligned}|\widehat{D}_{\mathrm{H}} - D_M| &= \frac{|\widehat{D}_{\mathrm{HT}} - D_M\widehat{W}|}{\widehat{W}} \\ &\le \frac{|\widehat{D}_{\mathrm{HT}} - D_M| + D_M|1-\widehat{W}|}{1-\epsilon_W} \le \frac{\epsilon_D + B\epsilon_W}{1-\epsilon_W} \le \frac{\bar{\epsilon}_D + B\epsilon_W}{1-\epsilon_W}.\end{aligned}$$

Independence also gives $\mathbb{P}(\widehat{W} = 0) = \mathbb{P}(R_1 = \cdots = R_M = 0) = \prod_i(1-\pi_i) \le \exp(-\sum_i \pi_i)$.

For the final claim, consider the centered triangular-array summands X_i. Their total variance is V_D, and the stated condition implies $\max_i |X_i|/\sqrt{V_D} \to 0$. Hence, for every $\varepsilon > 0$, the Lindeberg sum

$$\frac{1}{V_D}\sum_i \mathbb{E}\Big[X_i^2 \mathbf{1}\{|X_i| > \varepsilon\sqrt{V_D}\}\Big]$$

is eventually zero. The Lindeberg–Feller central limit theorem gives the asserted convergence in distribution. □

Covered reference-quality bridge. For a cardinal-supported audit unit, define

$$x_{a,b} = \Delta_i^r(a,b) - \phi_i^{a,b}, \qquad y_{a,b} = \Delta_i^r(a,b) - \Delta_i^c(a,b), \qquad (a,b) \in \mathcal{P}_i^{\mathrm{aud}}.$$

Then $x_{a,b} - y_{a,b} = \Delta_i^c(a,b) - \phi_i^{a,b}$. Using

$$\left| \max_j |x_j| - \max_j |y_j| \right| \le \max_j |x_j - y_j|,$$

we obtain

$$\left| d_{i,\delta}^{r,\mathrm{obs}} - d_{i,\delta}^{r,\mathrm{cov}} \right| \le v_{i,\delta}^{\mathrm{cov}}.$$

Averaging over the finite population gives

$$D_{M,\delta}^{r,\mathrm{cov}} \le D_{M,\delta}^{r,\mathrm{obs}} + V_{M,\delta}^{\mathrm{cov}}.$$

Thus, on the joint event consisting of the sampling bound from Proposition E.10 and the reference-quality bound $V_{M,\delta}^{\mathrm{cov}} \le \overline{V}_{\delta,\beta}^{\mathrm{cov}}$, we have

$$D_{M,\delta}^{r,\mathrm{cov}} \le \widehat{D}_{M,\delta}^{r,\mathrm{obs}} + B_\delta \sqrt{\frac{\log(2/\alpha)}{2n}} + \overline{V}_{\delta,\beta}^{\mathrm{cov}}.$$

By the union bound, this event has probability at least $1 - \alpha - \beta$.

Complete reconstructed cardinal reference. Suppose now that the validation evidence is cardinal comparison-identifying and that the validation protocol constructs a complete reference gap map $\phi_{i,\mathrm{full}}^{a,b}, \qquad a,b \in \mathcal{A}$. For exact mutually compatible edge references on a connected comparison graph, Proposition E.3 implies that this complete gap map is uniquely determined.

Define

$$x_{a,b}^{\mathrm{full}} = \Delta_i^r(a,b) - \phi_{i,\mathrm{full}}^{a,b},$$

and $y_{a,b}^{\mathrm{full}} = \Delta_i^r(a,b) - \Delta_i^c(a,b)$. Then

$$x_{a,b}^{\mathrm{full}} - y_{a,b}^{\mathrm{full}} = \Delta_i^c(a,b) - \phi_{i,\mathrm{full}}^{a,b}.$$

By definition,

$$d_{i,\delta}^{r,\mathrm{full,obs}} = \max_{a,b \in \mathcal{A}} \left| x_{a,b}^{\mathrm{full}} \right|,$$

whereas

$$d_{i,\delta}^{r,\mathrm{full}} = \max_{a,b \in \mathcal{A}} \left| y_{a,b}^{\mathrm{full}} \right|.$$

Applying the same max-norm inequality gives

$$\left| d_{i,\delta}^{r,\mathrm{full,obs}} - d_{i,\delta}^{r,\mathrm{full}} \right| \le v_{i,\delta}^{\mathrm{full}},$$

where

$$v_{i,\delta}^{\mathrm{full}} = \max_{a,b \in \mathcal{A}} \left| \phi_{i,\mathrm{full}}^{a,b} - \Delta_i^c(a,b) \right|.$$

Averaging over $i = 1, \dots, M$ yields

$$\begin{aligned} D_{M,\delta}^{r,\mathrm{full}} &\le \frac{1}{M} \sum_{i=1}^{M} d_{i,\delta}^{r,\mathrm{full,obs}} + \frac{1}{M} \sum_{i=1}^{M} v_{i,\delta}^{\mathrm{full}} \\ &= D_{M,\delta}^{r,\mathrm{full,obs}} + V_{M,\delta}^{\mathrm{full}}. \end{aligned}$$

If

$$0 \le d_{i,\delta}^{r,\mathrm{full,obs}} \le B_\delta^{\mathrm{full}},$$

then Proposition E.10, applied to these full-map unit-level contributions, gives an event $E^{\mathrm{full}}_{\mathrm{samp}}$ with probability at least $1-\alpha$ on which

$$D^{r,\mathrm{full,obs}}_{M,\delta} \le \widehat{D}^{r,\mathrm{full,obs}}_{M,\delta} + B^{\mathrm{full}}_{\delta}\sqrt{\frac{\log(2/\alpha)}{2n}}.$$

Suppose additionally that the reconstructed reference satisfies $V^{\mathrm{full}}_{M,\delta} \le \overline{V}^{\mathrm{full}}_{\delta,\beta}$ on an event $E^{\mathrm{full}}_{\mathrm{ref}}$ with probability at least $1-\beta$. Then

$$\Pr\left(E^{\mathrm{full}}_{\mathrm{samp}} \cap E^{\mathrm{full}}_{\mathrm{ref}}\right) \ge 1-\alpha-\beta.$$

On this joint event,

$$D^{r,\mathrm{full}}_{M,\delta} \le \widehat{D}^{r,\mathrm{full,obs}}_{M,\delta} + B^{\mathrm{full}}_{\delta}\sqrt{\frac{\log(2/\alpha)}{2n}} + \overline{V}^{\mathrm{full}}_{\delta,\beta}.$$

If the cardinal references are exact and comparison-identifying, then $\phi^{a,b}_{i,\mathrm{full}} = \Delta^c_i(a,b)$ for every action pair, so $v^{\mathrm{full}}_{i,\delta} = 0$ and hence $V^{\mathrm{full}}_{M,\delta} = 0$.

For estimated primitive references, the reference-quality condition is imposed on the complete reconstructed comparison object used by the audit.

Ordinal reference-quality bridge. For ordinal audit, for each supported pair write

$$x = \operatorname{sgn}\Delta^r_i(a,b), \qquad y = \operatorname{sgn}\phi^{a,b}_i, \qquad z = \operatorname{sgn}\Delta^c_i(a,b).$$

The elementary disagreement inequality

$$\mathbf{1}\{x \ne z\} \le \mathbf{1}\{x \ne y\} + \mathbf{1}\{y \ne z\}$$

therefore gives

$$d^{r,\mathrm{cov}}_{i,\rho} \le d^{r,\mathrm{obs}}_{i,\rho} + v_{i,\rho}$$

after averaging over $\mathcal{P}^{\mathrm{aud}}_i$.

At the choice level, structural interpretation requires the validation-supported optimal-action set to identify, or otherwise control its discrepancy from, the structural optimal-action set. No scalar reference-error reduction is required for the results in the main text.

Proof of Corollary E.11. Under the conditions of Section E.6.2, the audit units are the structural decision rounds: $M = T, \quad i = t$. Moreover, the route-side and structural audit comparisons coincide with the Section D objects, $\Delta^r_t(a,b) = G^r_t(a,b)$ and $\Delta^c_t(a,b) = G^c_t(a,b)$, so that

$$d^{r,\mathrm{full}}_{t,\delta} = \delta^r_t, \qquad D^{r,\mathrm{full}}_{T,\delta} = \frac{1}{T}\sum_{t=1}^{T}\delta^r_t.$$

The validation protocol additionally provides a complete reconstructed cardinal reference.

Applying the preceding full-map finite-evidence result gives, with probability at least $1-\alpha-\beta$,

$$D^{r,\mathrm{full}}_{T,\delta} \le \widehat{D}^{r,\mathrm{full,obs}}_{T,\delta} + B^{\mathrm{full}}_{\delta}\sqrt{\frac{\log(2/\alpha)}{2n}} + \overline{V}^{\mathrm{full}}_{\delta,\beta}.$$

By Definition D.1, $D^{r,\mathrm{full}}_{T,\delta} = \frac{\mathfrak{A}^r_T}{T}$. Hence

$$\frac{\mathfrak{A}^r_T}{T} \le \widehat{D}^{r,\mathrm{full,obs}}_{T,\delta} + B^{\mathrm{full}}_{\delta}\sqrt{\frac{\log(2/\alpha)}{2n}} + \overline{V}^{\mathrm{full}}_{\delta,\beta}.$$

Finally, Theorem D.4 gives the pathwise stability relation $R^c_T \le R^r_T + \mathfrak{A}^r_T$. Substituting the preceding upper bound on the alignment budget yields

$$R^c_T \le R^r_T + T\left[\widehat{D}^{r,\mathrm{full,obs}}_{T,\delta} + B^{\mathrm{full}}_{\delta}\sqrt{\frac{\log(2/\alpha)}{2n}} + \overline{V}^{\mathrm{full}}_{\delta,\beta}\right].$$

If the available comparisons are not cardinal comparison-identifying, the complete reconstructed reference required in this proof is unavailable and the audit stops at the covered-validity result above.

□

I Experimental Protocols and Supplementary Results

This section records the experiment-specific protocols supporting Section F. The accompanying anonymized code archive contains the implementations and computational materials described in Section I.8.

I.1 Reporting, Uncertainty, and Provenance Conventions

A *full-sample estimate* is the statistic computed from the prespecified analysis sample or the complete set of controlled evaluation seeds specified by the experiment. Exp1 uncertainty is a 95% seed-bootstrap interval over 30 shared evaluation seeds with 2000 bootstrap repetitions. Exp2 and Exp3 report empirical 2.5%–97.5% cluster-resampling sensitivity ranges; these ranges recompute the relevant analysis pipeline and are not treated as conventional confidence intervals. Exp4 Panel (a) uses paired-seed intervals over 100 shared structural seeds, while finite-audit bias and RMSE use estimate ± 1.96 Monte Carlo standard errors from 1000 independent audit replications. Deterministic controlled endpoints are reported without artificial uncertainty intervals.

Frozen derived results are distributed with the anonymized code archive alongside analysis and per-experiment plotting scripts. Regenerating figures from these saved results is distinct from rerunning the experiments.

I.2 Experiment 1: Controlled Alignment and Regret Transfer

I.2.1 Prespecified mechanisms and simulation protocol

Experiment 1 uses $K = 10$ actions, primary horizon $T = 5000$, maximum delay $d_{\max} = 100$, and 30 shared evaluation seeds. The simulator generates 100 prehistory rounds for the arrival-assigned complete-map route and uses an AR burn-in of 500 rounds. Twenty calibration seeds, disjoint from the 30 evaluation seeds, determine the state-process and delay parameters before evaluation.

For the smooth mechanisms, the latent process is

$$Z_t = \rho Z_{t-1} + \xi_t, \qquad \xi_t \sim N(0, \sigma_\xi^2), \qquad S_t = \tanh(Z_t),$$

with $\rho = 0.98$ and $\sigma_\xi = 0.15$. The public context is $X_t = \tanh(Z_t + \varepsilon_t)$, with $\varepsilon_t \sim N(0, 0.35^2)$. Action locations are $\mu_a = -1 + 2(a-1)/(K-1)$, and the smooth structural loss is $L_t^c(a) = ((S_t - \mu_a)/2)^2$, so $0 \le L_t^c(a) \le 1$. This AR(1) process provides the controlled latent-state generator for the smooth mechanisms.

The exact-cardinal-valid control uses $L_t^c(a) = g(a) + c_t$, where

$$g(a) = 0.6\,\frac{(a-5)^2}{\max_{j \in \mathcal{A}} (j-5)^2}, \qquad c_t = 0.1(1 + S_t).$$

The action-independent term changes loss levels without changing action gaps. The exact-cardinal-valid and geometric mechanisms share the same realized delay path within each seed.

The systematic-misbinding control alternates between preferred actions $a_L = 3$ and $a_R = 8$ in blocks of length 30, with fixed delay 15. Its structural reversal margin in a fully reversed block is $25/81 \approx 0.308642$. In the primary evaluation, the arrival-assigned route has complete-conflict rate 0.500013, with a 95% seed-bootstrap interval of $[0.499780, 0.500240]$. On conflict rounds the minimum structural and route margins are $25/81 \approx 0.308642$ and $1/81 \approx 0.012346$, respectively. Because the route-greedy diagnostic has zero route regret, Theorem D.11 gives a structural-regret rate of at least $(25/81) \times 0.500013 = 0.154325$, matching the reported 0.1543 up to rounding. The remaining stochastic mechanisms use calibrated truncated geometric, mixture, or state-coupled delay laws, all with generated mean delay near 15.

Table 3 summarizes the prespecified mechanisms.

Table 3: Experiment 1 prespecified mechanism summary.

Mechanism	Structural purpose	Delay construction	Primary role
Zero delay	Identity control	$\tau_t = 0$	No observation displacement.
Exact-cardinal-valid shift	Action gaps invariant over time	Same path as geometric	Positive-delay exact-validity control.
Geometric delay	Smooth structural state	Calibrated truncated geometric	Matched-mean misalignment.
Mixture delay	Smooth structural state	Fast/slow geometric mixture	Matched mean with different delay shape.
State-coupled delay	Smooth structural state	State-dependent truncated geometric	Matched mean with state-dependent timing.
Systematic misbinding	Alternating preferred actions	Fixed delay 15	Persistent route conflict.

I.2.2 Route maps and learner construction

Let $\mathcal{B}_t = \{s : s+\tau_s = t\}$ be the source rounds whose feedback arrives at clock t. For complete-map diagnostics, the arrival-assigned route uses

$$\widetilde{L}_t^{\mathrm{arrival}}(a) = \frac{1}{|\mathcal{B}_t|} \sum_{s \in \mathcal{B}_t} L_s^c(a)$$

when $\mathcal{B}_t \neq \varnothing$, and carries forward the previous route map when no new feedback arrives. The prehistory guarantees that the carried-forward map is defined at the start of the evaluation horizon. The controlled source-bound reference is $\widetilde{L}_t^{\mathrm{source}}(a) = L_t^c(a)$. Because the simulator exposes all action losses, it provides the complete counterfactual action-loss map required for this reference. Logged source identifiers instead provide factual source correspondence.

For numerical evaluation, define $e_t^r(a) = \widetilde{L}_t^r(a) - L_t^c(a)$. The worst-pair action-gap defect is computed by the equivalent range expression

$$\delta_t^r = \max_a e_t^r(a) - \min_a e_t^r(a), \qquad \mathfrak{A}_T^r = \sum_{t=1}^{T} \delta_t^r.$$

The route-greedy diagnostic chooses the smallest element of $\arg\min_a \widetilde{L}_t^r(a)$. Its route regret is therefore zero, and the reported structural regret evaluates that same action sequence against L_t^c.

The factual-feedback component uses $B = 10$ context cells. Cell boundaries are equal-probability quantiles estimated from calibration paths and then fixed before evaluation. Each cell maintains a K-action EXP3 log-weight vector. With $T_{\mathrm{cell}} = \lceil T/B \rceil$,

$$\gamma = \min\left\{1, \sqrt{\frac{K \log K}{(e-1)T_{\mathrm{cell}}}}\right\}, \qquad \eta = \frac{\gamma}{K},$$

giving $\gamma \approx 0.163710$ and $\eta \approx 0.016371$ for the primary run. For cell b, the action probability is

$$p_{b,t}(a) = (1-\gamma)\frac{\exp(q_{b,t}(a))}{\sum_j \exp(q_{b,t}(j))} + \frac{\gamma}{K}.$$

For factual scalar loss y, selected action a, and stored selection probability p, the update is

$$\widehat{\ell}(a) = \frac{y}{p}, \qquad q_b(a) \leftarrow q_b(a) - \eta\widehat{\ell}(a).$$

No complete loss vector is supplied to the learner.

At each clock, the learner observes the current public context, selects the action, stores the factual source event, schedules its arrival, and only then processes feedback arriving at that clock. Hence feedback observed at clock t cannot affect the action chosen at the same clock. For an event generated at source round s, the arrival-clock branch updates the current context/action cell using the arriving factual loss, whereas the source-round branch uses the stored source-time context, action, and selection probability. The source-round branch updates the current learner state; it does not rewind the learner trajectory.

I.2.3 Complete primary results and targeted validation

Table 4 reports the complete primary results.

All complete-map transfer checks are performed seed by seed before aggregation. The alignment budget is interpreted as a stability control term, not as an estimator of the realized structural regret.

Targeted validation varies the geometric target mean delay over 5, 15, and 30, and the systematic-misbinding horizon over $T \in \{1000, 5000, 10000\}$. These checks assess finite-horizon sensitivity without changing the primary estimand. Additional diagnostics verify generated delay behavior, route-map age, multi-arrival frequency, state coupling, and reversal margins.

The exact-cardinal-valid control separates map-level validity from learner-level sensitivity. Its complete route map has zero action-gap defect. At the learner level, however, the arriving common shift updates only the selected action–context cell, so the scalar feedback path can alter subsequent learning dynamics even when the complete route map preserves all action gaps. The resulting arrival–source learner contrast is reported as a separate algorithmic diagnostic.

I.2.4 Matched cancellation, margin threshold, and realized stability utilization

The targeted diagnostic separates a component shared by all actions in a round from a residual that varies by action:

$$\widetilde{L}_t^r(a) = L_t^c(a) + \alpha_{\mathrm{shared}} c_t + u_t(a).$$

The shared component is either static, $c_t = 1$, or state-varying, $c_t = 1 + S_t$, and is added identically to every action in round t. The action-dependent residual perturbs the unique structural optimum and its deterministic nearest competitor by $+d_t/2$ and $-d_t/2$, respectively, with $d_t = \alpha_{\mathrm{dep}} \mu_t^c$. The grid uses $\alpha_{\mathrm{shared}} \in \{0, 0.05, 0.10, 0.20\}$, $\alpha_{\mathrm{dep}} \in \{0, 0.50, 0.99, 1, 1.01, 1.50\}$, both shared profiles, and 30 seeds, for 1440 cells. Figure 10 shows the resulting sweep and the realized stability utilization.

Holding α_{dep} fixed, changing the shared amplitude or profile leaves all decision diagnostics unchanged up to 2.22×10^{-16}. These diagnostics include the action-gap defect, sign and choice disagreement, complete conflict, optimal-action masks, and the all-action regret map. The mean absolute action-wise level error, by contrast, increases with α_{shared}. Thus the sweep isolates within-round cancellation of action-independent shifts, including state-varying shared components.

The action-dependent residual confronts the margin boundary directly. On the eligible-round support, the native identity $\delta_t^r = \alpha_{\mathrm{dep}} \mu_t^c$ holds within numerical tolerance. Table 5 reports the resulting directed choice disagreement; all eight shared-profile–amplitude combinations give the same value at each α_{dep}.

The same controlled route maps yield a descriptive realized stability-bound utilization,

$$U_T^r = \frac{|R_T^c - R_T^r|}{\mathfrak{A}_T^r},$$

defined only when $\mathfrak{A}_T^r$ is above the existing numerical stability tolerance. Zero-budget cases remain undefined rather than being assigned a value of zero. Table 6 reports seed-level-ratio means and 95% seed-bootstrap intervals. This statistic measures the realized fraction of the available same-sequence alignment budget; it is not an empirical estimator of the sharp constant in Theorem D.4.

For systematic misbinding, the route-map horizon diagnostic reuses the same shared-prefix construction at $T \in \{1000, 5000, 10000\}$. Route regret is zero at every horizon, while the structural-regret rate, alignment-budget rate, conflict rate, and utilization remain stable, as shown in Table 7. Across these horizons, the stable rates exhibit persistent margin-separated conflict.

Finally, the route-map and learner diagnostics are compared rather than added. No delay and the exact-cardinal-valid delayed shift both have numerically zero alignment budget and zero route-greedy structural regret, while the latter has an arrival-minus-source learner contrast of 0.0675. Systematic misbinding instead has $\mathfrak{A}_T^r/T = 0.5556$, route-greedy $R_T^c/T = 0.1543$, and learner contrast 0.0714. The learner contrast is an algorithm-level sensitivity diagnostic, not an additive component of route-map structural regret.

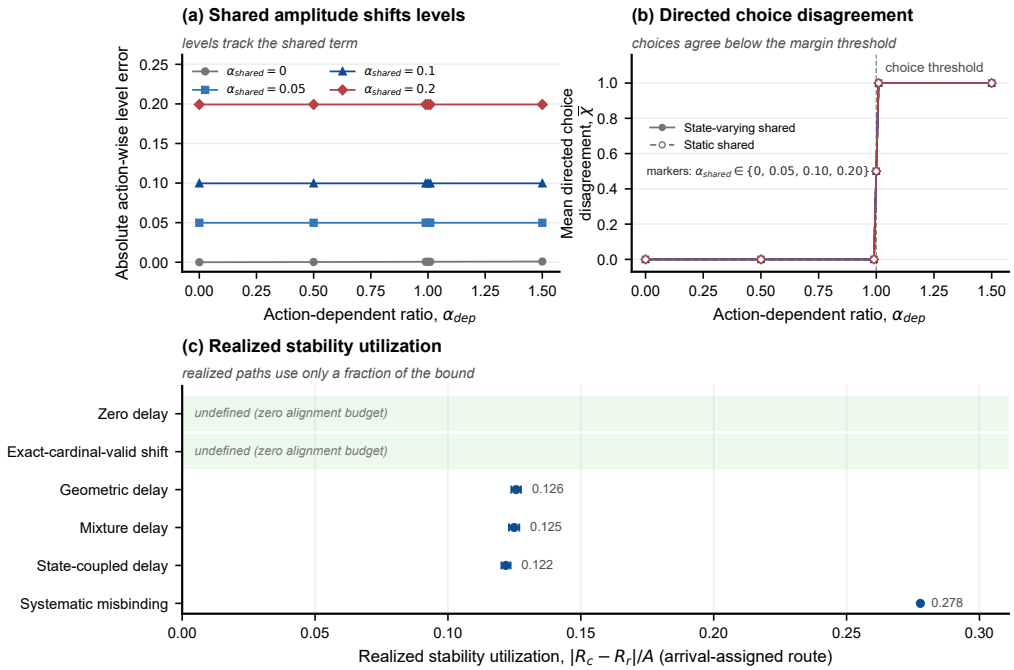

Figure 10: Experiment 1 targeted cancellation sweep and realized stability utilization. Panel (a): the absolute action-wise level error tracks the shared amplitude α_{shared} and is essentially flat in α_{dep}; the shared component moves loss levels. Panel (b): the mean directed choice disagreement $\overline{\chi}$ is invariant to the shared profile and amplitude. In this construction it is zero for $\alpha_{\text{dep}} < 1$, equals 0.5 at $\alpha_{\text{dep}} = 1$, and equals 1 above the threshold. Panel (c): realized stability utilization $|R_T^c - R_T^r|/\mathfrak{A}_T^r$ on the arrival-assigned route, with 95% seed-bootstrap intervals. Mechanisms with numerically zero alignment budget are typed as undefined and are not assigned a value of zero.

Table 4: Experiment 1 primary controlled results. Alignment and regret columns are rates.

Mechanism	Mean delay	$\mathfrak{A}_T^{\mathrm{arr}}/T$	Route-greedy R_T^c/T	Transfer bound	Learner arrival–source
Zero delay	0.0000	0.0000	0.0000	0.0000	0.0000
Exact-cardinal-valid shift	14.9709	≈ 0	0.0000	≈ 0	0.0675
Geometric delay	14.9709	0.2588	0.0326	0.2588	0.0544
Mixture delay	14.9706	0.2191	0.0274	0.2191	0.0437
State-coupled delay	14.9020	0.2481	0.0303	0.2481	0.0516
Systematic misbinding	15.0000	0.5556	0.1543	0.5556	0.0714

Table 5: Experiment 1 targeted margin-threshold diagnostic. The directed choice-disagreement rate is invariant to the shared amplitude and to whether the shared component is static or state-varying.

α_{dep}	Mean directed choice disagreement
0.00	0.000
0.50	0.000
0.99	0.000
1.00	0.500
1.01	1.000
1.50	1.000

Table 6: Experiment 1 realized stability-bound utilization on the arrival-assigned route.

Mechanism	Mean utilization	95% seed-bootstrap interval
Zero delay	undefined	–
Exact-cardinal-valid shift	undefined	–
Geometric delay	0.1257	[0.1239, 0.1275]
Mixture delay	0.1249	[0.1230, 0.1268]
State-coupled delay	0.1218	[0.1201, 0.1236]
Systematic misbinding	0.2778	[0.2778, 0.2778]

Table 7: Experiment 1 systematic-misbinding route-map diagnostics across horizon.

T	R_T^c/T	$\mathfrak{A}_T^r/T$	Complete-conflict rate	U_T^r
1000	0.15425	0.55530	0.49977	0.27778
5000	0.15433	0.55557	0.50001	0.27778
10000	0.15431	0.55553	0.49998	0.27778

I.3 Experiment 2: Attribution Sensitivity in Delayed-Conversion Logs

I.3.1 Cohort and candidate-source construction

The analysis uses the Criteo delayed-conversion log (Diemert et al., 2017). Impressions are potential source events. The primary candidate set contains eligible source events within 7 days before conversion, and only conversion journeys with complete lookback are retained. The source-time decision cell is $c = (\mathrm{campaign}, \mathrm{source\ calendar\ day})$.

Table 8 reports the primary cohort and support.

All source-time routes use the same retained journeys, journey-specific candidate sets, and fixed aggregate cell universe. Cells receiving zero route-specific credit remain in that universe so that normalized allocation comparisons do not change support across routes.

I.3.2 Attribution routes, metrics, and uncertainty

For journey j with candidate cells $\mathcal{C}_j$, route weights satisfy $w_{jc}^{(r)} \geq 0$ and $\sum_{c\in\mathcal{C}_j} w_{jc}^{(r)} = 1$. First-click-or-touch and Last-click-or-touch assign all credit to the source cell containing the earliest and latest eligible clicked impression, respectively; if a journey contains no eligible clicked impression,

Table 8: Experiment 2 primary cohort and support.

Quantity	Value
Raw log rows	16,468,027
Candidate conversion journeys after temporal filters	308,605
Final retained complete-lookback journeys	240,981
Retained UIDs	183,533
Eligible campaign–source-day cells	19,307
Journeys with one candidate cell	193,131
Journeys with two candidate cells	32,736
Journeys with three or more candidate cells	15,114
Attribution-ambiguous journeys	47,850
Minimum impressions per eligible cell	50
UID-cluster resampling repetitions	1,000

they use the earliest and latest eligible impression instead. Linear credit distributes mass equally across eligible source cells. Time-decay uses

$$w_{jc}^{(\mathrm{decay})} = \frac{\exp(-\lambda \Delta_{jc})}{\sum_{c' \in \mathcal{C}_j} \exp(-\lambda \Delta_{jc'})},$$

with the primary half-life fixed at 1.386 days. The constructed arrival-time anchor allocates conversion credit according to the arrival-side accounting rule and is used only as a diagnostic reference.

For route r, let

$$Q_r(c) = \frac{\sum_j w_{jc}^{(r)}}{\sum_{c'} \sum_j w_{jc'}^{(r)}}.$$

Allocation displacement is

$$\mathrm{TV}(Q_r, Q_{r'}) = \frac{1}{2} \sum_c |Q_r(c) - Q_{r'}(c)|.$$

Kendall's τ_b is computed over the fixed common active-cell support. Top-k overlap is retained as a supplementary ranking-depth diagnostic rather than a main-figure metric.

UID-cluster resampling uses 1000 repetitions and holds the full-sample support universe fixed. The empirical 2.5%–97.5% ranges quantify pipeline sensitivity. Because these high-dimensional normalized statistics are non-smooth, the ranges are not interpreted as conventional confidence intervals.

I.3.3 Complete pairwise results and robustness

Table 9 reports the arrival-anchored comparisons.

Table 10 reports the pairwise comparisons among source-time routes.

Route disagreement is zero by construction for singleton-candidate journeys. The ambiguity analysis stratifies journeys by candidate-cell multiplicity. Singleton-candidate journeys number $193{,}131$, or 80.14% of the retained cohort, and have zero pairwise source-route allocation TV. The two-candidate stratum contains $32{,}736$ journeys (13.58%) with mean allocation TV 0.179 across the six source-route pairs. The $15{,}114$ journeys with three or more candidate cells (6.27%) have mean pairwise allocation TV 0.275.

Robustness checks vary the candidate window, the eligible-cell impression threshold, the time-decay half-life, and ranking depth. Across these specifications, the source-route sensitivity pattern remains stable, with disagreement concentrated where multiple candidate source cells are available. Experiment 2 therefore characterizes source-route non-uniqueness on the retained log support.

Table 9: Experiment 2 source-time routes versus the arrival-time accounting anchor.

Route	Allocation TV	Kendall's τ_b
First-click-or-touch	0.138	0.482
Last-click-or-touch	0.103	0.579
Linear credit	0.119	0.560
Time-decay credit	0.109	0.589

Table 10: Experiment 2 pairwise comparisons among source-time attribution routes.

Route pair	Allocation TV	Kendall's τ_b
First-click-or-touch vs. Last-click-or-touch	0.062	0.717
First-click-or-touch vs. Linear	0.032	0.848
First-click-or-touch vs. Time-decay	0.050	0.774
Last-click-or-touch vs. Linear	0.033	0.855
Last-click-or-touch vs. Time-decay	0.017	0.925
Linear vs. Time-decay	0.019	0.917

I.4 Experiment 3: Logged-Support Decision Recovery

I.4.1 Temporal split, target, and action support

Experiment 3 uses KuaiRand-1K (Gao et al., 2022). The history period is April 8–21, 2022, and the evaluation period is April 22–May 8, 2022. Observations outside the prespecified temporal boundaries are excluded from the corresponding fitting or evaluation stage.

For a user–video exposure at source time t, the constructed six-hour target is

$$Y_{u,t}^{(6h)} = \log\Big(1 + 0.5\,\mathrm{LongView}_{u,t:t+6h} + \mathrm{Like}_{u,t:t+6h} + \mathrm{Comment}_{u,t:t+6h} + \mathrm{Forward}_{u,t:t+6h} + 1.5\,\mathrm{Follow}_{u,t:t+6h}\Big).$$

The delayed target is assigned a deterministic pseudo-random arrival lag between 6 and 10 hours for the delayed-feedback diagnostic. This construction retains the source event for each delayed target, enabling direct evaluation of source misbinding on the logged support.

The candidate action space contains the 20 most frequent history-side video tags and covers 85.0% of evaluation exposure mass. These tag categories are the decision actions used in this diagnostic, not complete recommender policies. An action is supported within a fold–audit-unit cell only when at least 500 source events are available. The primary analysis contains 85 audit units; all 85 are retained, mean supported-action count is 18.49, action coverage is 0.925, and reference-pair coverage is 0.921.

I.4.2 Route construction, cross-fitting, and metrics

Arrival carrier is a deliberate source-misbinding baseline: an arriving target is associated with the most recent eligible exposure of the same user before arrival. Historical mean predicts using matured history-side action averages. Ridge proxy uses action indicators and admissible history-side proxy summaries; its penalty is fixed at $\alpha = 30$ after rolling temporal validation on the history split.

Reference comparisons are cross-fitted over two deterministic user folds. An action selected using one fold is evaluated on the opposite fold, so the selected reference action can change across folds or resamples. We therefore report the signed reference-minus-route value difference as a paired held-out contrast, whose sign records the relative held-out value of the two selected actions.

The primary diagnostic chain is:

1. score recovery: supported-cell Spearman correlation and MAE;
2. gap recovery: maximum held-out reference-pair gap error and pairwise sign agreement;

3. decision recovery: top-action agreement with the cross-fitted reference and the paired Ridge-minus-Historical value contrast.

User-cluster resampling uses 1000 repetitions. Supported action sets and cross-fitted selections are recomputed inside each resample. The resulting 2.5%–97.5% ranges are pipeline-sensitivity ranges, not validated confidence intervals. Because support sets and cross-fitted reference selections are reconstructed within each resample, these ranges are reported as pipeline-sensitivity diagnostics rather than as an application of the fixed-finite-population concentration proposition.

I.4.3 Complete primary results and resampling diagnostics

Table 11 reports the full-sample logged-support diagnostics.

Paired resampling contrasts. For metric m and user-cluster resample b, define

$$\Delta_m^{(b)} := m_{\text{Ridge}}^{(b)} - m_{\text{Historical}}^{(b)}.$$

Both routes use the same resampled user multiplicities and the same within-resample support and reference-construction protocol. The fitted routes remain fixed; supported action sets and cross-fitted reference selections are reconstructed within each resample. Table 12 summarizes the differences computed within each of the 1000 paired resamples, rather than subtracting the endpoints of separate route-wise ranges. Negative differences favor Ridge for MAE and maximum gap error; positive differences favor Ridge for the two agreement metrics.

All four paired ranges cross zero. Ridge has lower MAE in 54.6% of the resamples, higher maximum gap error in 74.2%, lower sign agreement in 91.0%, and lower top-action agreement in 84.9%. Top-action agreement ties in 3.5% of the resamples. Across the paired resampling pipeline, the decision-side disadvantage is directionally prevalent, although all four empirical sensitivity ranges cross zero.

The full-sample Ridge-minus-Historical paired value contrast is -0.0149, with empirical user-resampling sensitivity range $[-0.0347,\ 0.0075]$. Positive values are defined to favor Ridge. The corresponding sensitivity range crosses zero.

The deliberate Arrival-carrier route provides the main source-misbinding control. Its poorer gap and ranking diagnostics show that source misbinding can degrade decision comparisons even when all evaluation remains on common logged support. Together, these diagnostics characterize how predictive recovery, action-gap recovery, and decision recovery separate under a common logged-support evaluation design.

I.5 Experiment 4: Controlled Route Alignment and Evidence-Qualified Audit

I.5.1 Population estimand, route grid, and construction

Experiment 4 uses $K = 10$ actions and a three-dimensional structural state. Module A uses horizon $T = 5000$, warm-up $W = 250$, and 100 shared structural seeds. The primary delay–state coupling coefficient is 2.0, target mean delay is 2, maximum candidate delay is 20, and feedback-noise standard deviation is 0.009.

Let $\mathcal{P} = \{(a, b) : a < b,\ a, b \in \mathcal{A}\}$. For controlled evaluation unit i, define

$$d_{i,\text{pair}}^r = \frac{1}{|\mathcal{P}|} \sum_{(a,b)\in\mathcal{P}} \left| G_i^r(a, b) - G_i^c(a, b) \right|,$$

and the population mean pairwise gap discrepancy

$$D_{\text{pair}}^r = \frac{1}{M} \sum_{i=1}^{M} d_{i,\text{pair}}^r.$$

This pair-average statistic is the primary population estimand for the evidence-quality analysis in Section E, whereas the sharp transfer bound uses the worst-pair defect δ_t^r and its cumulative alignment budget.

Table 11: Experiment 3 full-sample logged-support recovery diagnostics.

Route	Score Spearman	Score MAE	Max gap error	Gap-sign agreement	Top-action agreement
Arrival carrier	0.610	0.128	0.642	0.765	0.335
Historical mean	0.714	0.106	0.386	0.867	0.518
Ridge proxy	0.696	0.105	0.400	0.857	0.447

Table 12: Experiment 3 paired Ridge-minus-Historical resampling diagnostics. All four metrics have 1000 valid paired draws. Ranges report the empirical 2.5%–97.5% variation across the paired user-resampling pipeline. The negative-draw share reports the fraction of resamples with $\Delta_m^{(b)} < 0$.

Metric	Full sample	Median	2.5%–97.5% range	Negative draws
Score MAE	−0.00100	−0.00102	[−0.01597, +0.01383]	54.6%
Maximum gap error	+0.01371	+0.00631	[−0.01070, +0.02538]	25.8%
Sign agreement	−0.01079	−0.00661	[−0.01765, +0.00301]	91.0%
Top-action agreement	−0.07059	−0.04118	[−0.11765, +0.02941]	84.9%

The partial-label proxy route first constructs a label-blind source assignment for each arrival. Historical candidate sources are restricted to the maximum-delay window. For candidate s, the unnormalized attribution weight is

$$\widetilde{w}_s = \exp\left(-\frac{\|Z_s^{\text{proxy}} - Z_{\text{arr}}^{\text{sig}}\|^2}{2h^2}\right) p_{\text{delay}}(\tau_s),$$

where h is the calibrated kernel bandwidth fixed before evaluation and p_{delay} is the fixed delay prior; weights are normalized to sum to one. If the true source label is retained, the complete controlled source map is returned to its true round. Otherwise, the source map is distributed over feasible historical candidates using the normalized proxy weights. The route-level label rate is varied over $q_{\text{route}} \in \{0, 0.3, 0.7, 1\}$ and proxy-noise standard deviation over $\{0, 0.1, 0.25, 1\}$.

At $q_{\text{route}} = 1$, the simulator exposes every structural source map and returns it to its generating round, so $D^r_{\text{pair}} = 0$ by construction.

I.5.2 Audit inclusion and finite-evidence diagnostics

Module B fixes $q_{\text{route}} = 0.3$ and $\sigma_{\text{proxy}} = 0.25$, uses horizon 2000 after a 100-round warm-up, and repeats the audit 1000 times. Audit evidence rates are $\rho_{\text{audit}} \in \{0.1, 0.3, 0.5, 1\}$.

Let M denote the number of units in the audit population for a given replication. The endpoint $\rho_{\text{audit}} = 1$ is a deterministic census:

$$\pi_i = 1, \qquad R_i = 1 \qquad (i = 1, \ldots, M).$$

This endpoint is defined separately and does not use the clipped logistic construction.

For the non-census rates $\rho_{\text{audit}} \in \{0.1, 0.3, 0.5\}$, the MCAR design includes each unit with constant probability $\pi_i = \rho_{\text{audit}}$. For ambiguity-selective sampling, let z_i be standardized route-attribution ambiguity and set

$$\pi_i = \text{clip}\left\{\text{logit}^{-1}(\alpha_\rho + 1.5 z_i), 0.05, 0.95\right\}, \qquad \frac{1}{M}\sum_{i=1}^{M} \pi_i = \rho_{\text{audit}}.$$

The intercept α_ρ is solved separately for each non-census rate and audit population. This calibration fixes the expected audit fraction, not the realized fraction in each sampling repetition.

The unweighted selective estimator averages $d^r_{i,\text{pair}}$ over included units. The design-aware weighted estimator is the Hájek IPW mean

$$\widehat{D}_{\text{IPW}} = \frac{\sum_{i:R_i=1} \pi_i^{-1} d^r_{i,\text{pair}}}{\sum_{i:R_i=1} \pi_i^{-1}}.$$

Table 13: Experiment 4 population mean pairwise gap discrepancy D^r_{pair} over the route-label and proxy-noise grid.

	Proxy-noise sd σ_{proxy}			
q_{route}	0	0.1	0.25	1.0
0.0	0.02454	0.02472	0.02575	0.04737
0.3	0.01831	0.01846	0.01928	0.03602
0.7	0.00820	0.00828	0.00870	0.01663
1.0	0.00000	0.00000	0.00000	0.00000

when its denominator is positive. If no unit is selected, the implementation records the estimate as unavailable rather than assigning a numerical discrepancy; Theorem E.12 bounds the probability of this event under the controlled independent-inclusion design. The same selective inclusion mask is used for the unweighted and weighted versions, isolating the effect of weighting.

Table 14 reports the finite-audit performance.

At the census endpoint $\rho_{\text{audit}} = 1$, every unit is included and all inverse-probability weights equal one. The unweighted and Hájek estimators therefore both equal the corresponding population discrepancy D^r_{pair} in every audit replication, giving zero audit bias and RMSE by construction. This eliminates audit-sampling error; it does not imply that the route's population discrepancy is zero. Effective sample size and weight concentration are reported separately from discrepancy. For the inverse-probability weights $w_i = \pi_i^{-1}$, the effective support is

$$n_{\text{eff}} = \frac{\left(\sum_{i:R_i=1} w_i\right)^2}{\sum_{i:R_i=1} w_i^2}.$$

We report n_{eff} as a diagnostic of weight concentration and effective audit support.

The results show that known-probability Hájek weighting substantially reduces both bias and RMSE relative to the corresponding ambiguity-selective unweighted estimator in this controlled design. This comparison isolates the known-propensity setting; observational audits with estimated propensities require additional selection-model analysis.

I.5.3 Calibration controls and validation

Calibration uses five temporal folds. For each supported action pair, an affine map is fitted without using the held-out unit:

$$\widehat{G}^r_{i,\text{cal}}(a,b) = \widehat{\alpha}_{r,a,b} + \widehat{\beta}_{r,a,b}\Delta^r_i(a,b).$$

Let $\widehat{D}^r_{\text{raw}}$ and $\widehat{D}^r_{\text{cal}}$ be the out-of-fold pairwise discrepancies on the same target. When $\widehat{D}^r_{\text{raw}} > 0$,

$$\widehat{\text{Rec}}^r = 1 - \frac{\widehat{D}^r_{\text{cal}}}{\widehat{D}^r_{\text{raw}}}.$$

Table 15 reports the calibration-family controls.

The affine-linked control preserves unit-level correspondence by construction. The correspondence-destroyed control shuffles the relevant correspondence within temporal blocks while retaining a calibration task with the same reporting interface. Recoverability decreases from 0.809 in the affine-linked control to 0.260 after correspondence destruction, while remaining positive in both cases. A nonlinear-monotone control provides a supplementary diagnostic of calibration-family misspecification.

Together, these controls characterize how much of the pairwise comparison discrepancy can be recovered by the prespecified calibration family under preserved, disrupted, and nonlinear correspondence structures.

Table 14: Experiment 4 finite-audit performance. Values are Monte Carlo bias and RMSE for D^r_{pair}.

	$\rho = 0.1$		$\rho = 0.3$		$\rho = 0.5$	
Audit design	Bias	RMSE	Bias	RMSE	Bias	RMSE
MCAR, unweighted	−0.00001	0.00283	−0.00002	0.00143	0.00003	0.00093
Ambiguity-selective, unweighted	0.03163	0.03193	0.01656	0.01668	0.00873	0.00880
Ambiguity-selective, Hájek IPW	0.00018	0.00262	0.00007	0.00106	0.00002	0.00063

Table 15: Experiment 4 calibration-family controls.

Control	Raw discrepancy	OOF calibrated	Recoverability	Estimability
Affine-linked	0.241	0.046	0.809	1.000
Correspondence-destroyed	0.396	0.293	0.260	1.000

I.6 Numerical Checks and Descriptive Diagnostics

This subsection separates numerical checks of explicit constructions and closed-form expressions from a descriptive association in the recorded simulator outputs.

Margin threshold. A fixed two-action construction varies the alignment-defect ratio δ^r_t/μ^c_t across a grid spanning both sides of the threshold in Proposition D.7. The observed choice disagreement matches the closed-form value at every grid point, with maximum deviation 0.0: $\overline{\chi}^r_T = 0$ for $\delta^r_t/\mu^c_t < 1$, $\overline{\chi}^r_T = 0.5$ exactly at the threshold, and $\overline{\chi}^r_T = 1$ above it (Figure 12, Panel (a)).

Descriptive alignment–disagreement association. Across the controlled route grid of Experiment 4, the pairwise sign disagreement $\overline{\rho}^r_T$ and the alignment-budget rate $\mathfrak{A}^r_T/T$ have a fitted log–log slope of 0.89 (Figure 12, Panel (b)). This is a descriptive association across the primary proxy-label route configurations. The fitted slope is reported as a descriptive route-grid association; Proposition D.9 instead supplies an upper bound under a margin-mass condition.

Sharp exponent. The explicit two-action construction of Proposition D.10 is evaluated directly over a grid of $\kappa \in \{0.5, 1, 2, 4, 8\}$ and $T = 2, \ldots, 2000$. For each κ the realized pair $\left(\mathfrak{A}^r_T/T, \overline{\rho}^r_T\right)$ satisfies the claimed identity $\overline{\rho}^r_T = 2^{-\kappa/(\kappa+1)}(\mathfrak{A}^r_T/T)^{\kappa/(\kappa+1)}$ to machine precision (maximum absolute deviation 1.1×10^{-16}), numerically reproducing the matching identity for this explicit construction (Figure 11).

Information boundary. For the paired-arrival construction of Theorem E.4, we restrict the numerical check to even horizons $T = 2, 4, \ldots, 20$, with $n = 1$ and $\Delta_\Theta = 1$. We sample independent source-label retention opportunities and estimate the probability that no world-separating label is retained. Multiplying this probability by $\Delta_\Theta/2$ checks the closed-form risk $\frac{\Delta_\Theta}{2}(1 - p_{\mathrm{src}})^{nT}$ of Corollary E.7. With 200,000 Monte Carlo replications per grid point, the maximum absolute deviation is 7.845×10^{-4} over $p_{\mathrm{src}} \in \{0, 0.1, 0.3, 0.5\}$ (Figure 13).This numerical check targets the closed-form label-retention probability in Corollary E.7.

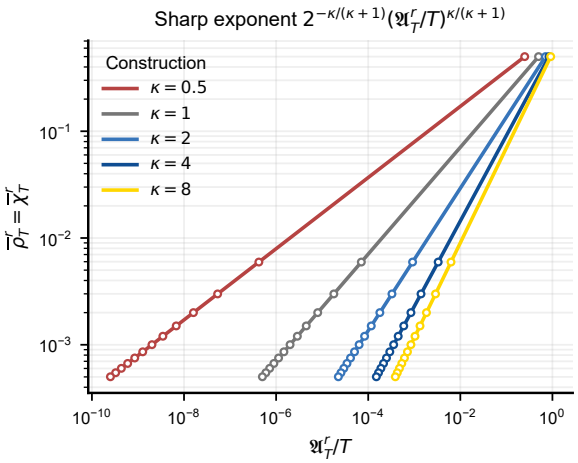

Figure 11: **Sharp-exponent verification.** The two-action construction of Proposition D.10 across $\kappa \in \{0.5, 1, 2, 4, 8\}$; for each κ the realized points (markers) lie on the closed-form curve $2^{-\kappa/(\kappa+1)}(\mathfrak{A}_T^r/T)^{\kappa/(\kappa+1)}$.

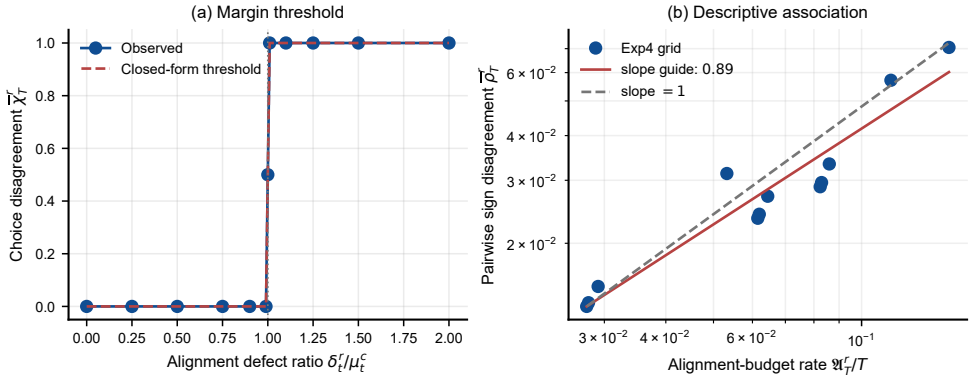

Figure 12: **Margin-threshold check and descriptive alignment–disagreement association.** Panel (a): observed choice disagreement against the alignment-defect ratio, together with the closed-form threshold of Proposition D.7. Panel (b): pairwise sign disagreement against the alignment-budget rate on log–log axes, with a guide using the fitted slope, anchored at the smallest positive budget, and a unit-slope reference. Panel (b) is descriptive and does not verify a soft-margin exponent.

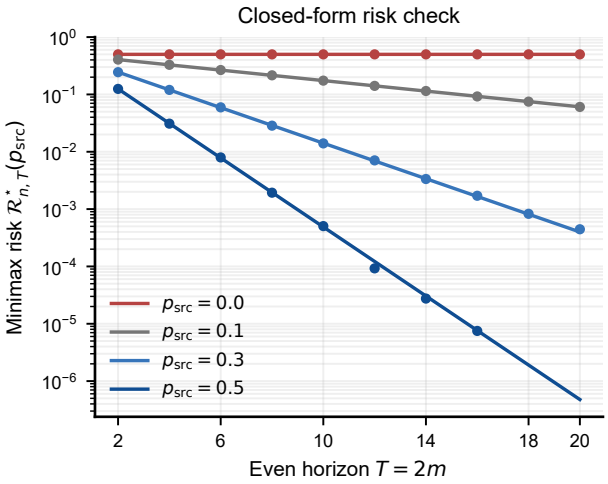

Figure 13: **Closed-form risk check at even horizons.** Lines show the two-world minimax risk in Corollary E.7 for $T = 2, 4, \ldots, 20$, $n = 1$, and $\Delta_\Theta = 1$. Markers estimate the same expression from independent label-retention draws, using 200,000 Monte Carlo replications per grid point; zero Monte Carlo estimates are not visible on the logarithmic axis. The markers provide a Monte Carlo check of the same closed-form retention probability.

I.7 Supplementary Empirical Views

Attribution entropy by candidate multiplicity. Within the delayed-conversion cohort, the credit-dispersion entropy of the source-time routes rises with candidate-source multiplicity: the Linear route attains mean assignment entropy 0.00, 0.69, and 1.25 in the one-, two-, and three-or-more candidate strata, and the Time-decay route 0.00, 0.51, and 0.94, whereas the single-cell assignment routes (First-click-or-touch, Last-click-or-touch, and the arrival anchor) remain at zero entropy by construction. Dispersion in credited mass therefore tracks exactly the candidate ambiguity that drives the allocation disagreement of Experiment 2 (Figure 14, Panel (a)).

Horizon scaling of the learner contrast. The learner-level contrast of Experiment 1 changes qualitatively with the horizon. Under arrival-clock binding the structural-regret rate is essentially flat across $T \in \{10^3, 5 \times 10^3, 10^4\}$ (0.1799, 0.1785, 0.1783), whereas under source-round binding it declines (0.1555, 0.1071, 0.0861). The two bindings are therefore separated by a gap that widens across the evaluated horizons (Figure 14, Panel (b)).

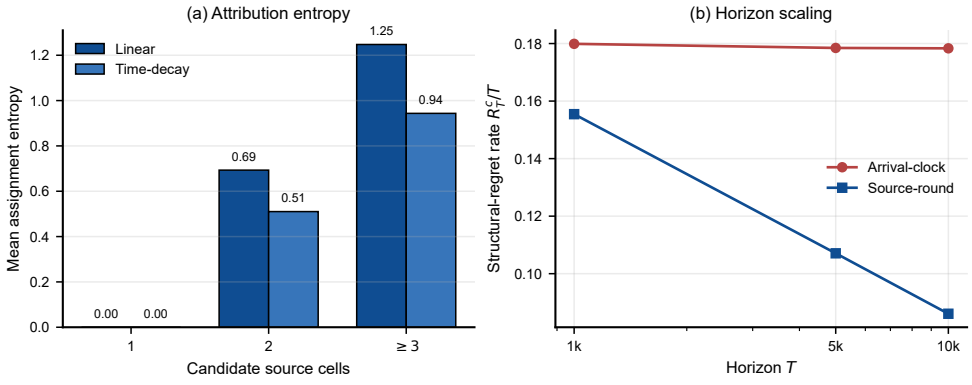

Figure 14: **Supplementary empirical views from the recorded outputs.** Panel (a): mean assignment entropy of the source-time routes within candidate-source strata; the single-cell routes have zero entropy by construction. Panel (b): learner-level structural-regret rate against horizon under arrival-clock and source-round factual-feedback binding.

I.8 Reproducibility and Artifact Lineage

The accompanying anonymized code archive contains experiment implementations, configurations, random seeds, environment specifications, and frozen derived results. `REPRODUCE.md` documents execution, `DATA.md` describes external data requirements, and `ARTIFACT_MANIFEST.json` records the distributed files and their checksums. External raw datasets are obtained separately under their original release conditions.