DECISION-RELEVANT REPRESENTATIONS UNDER RARE OBSERVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

World models and predictive representations decide what to remember by how well the past predicts the future. This objective has a blind spot: a cue that is rare, unpredictable from the past, and yet decides the action—a traffic light that turns red—contributes almost nothing to an average predictive loss and is the first distinction compression discards. We show that the right currency for such cues is evidence, not frequency. We introduce *coverage-calibrated Bellman quotients* (CBQ), a latent-state learner that merges two histories only when their counts certify that no action consequence separates them, and returns the most compact code that passes this certificate. CBQ comes with a complete finite-sample account. Its supported Q-error splits into approximation bias $\mathfrak{A}_t$ and statistical error $\mathfrak{S}_t$; it keeps every action-flipping cue whenever $2\mathfrak{S}_t < \Delta_t - 2\mathfrak{A}_t$, recovers the exact Bellman quotient—discarding all decision-null nuisance—under a separation condition, and turns both into fresh-rollout regret bounds. A matching obstruction shows that no learner can identify a rare decisive cue when $nq\Delta^2 = O(1)$, so in the one-stage case the evidence requirement of CBQ is tight up to logarithms. Across 288 random POMDPs, predictive objectives keep the rare cue in at most 5% of instances at any sample size, while CBQ keeps it in all of them with a code exactly as compact as the Bellman quotient; all 4,320 runs switch on within one decade of the theory's coverage statistic $nq\Delta^2/\Gamma$, and the certificate is never violated. From pixels, next-frame and self-predictive world models lose a rare traffic light in 107 of 108 runs, whereas CBQ on a learned patch tokenizer keeps exactly the light's patch and discards the watermark entirely.

1 INTRODUCTION

A watermark in the corner of a dashcam frame and a traffic light that has just turned red are equally unpredictable from the preceding video. A representation trained to predict the next observation has no reason to prefer one over the other; if anything it prefers the watermark, whose persistence makes it easy to forecast. Yet forgetting the watermark is harmless, while forgetting the light changes the safe action. This is the regime in which predictive learning is least aligned with control: the decisive cue is *exogenous*, so the past does not predict it, and *rare*, so it barely moves an average loss—while its decision cost, conditional on occurring, stays large.

Modern agents compress history through exactly such averages. World models are trained to reconstruct or forecast observations (Ha & Schmidhuber, 2018; Hafner et al., 2023), value-equivalent models to predict returns (Schrittwieser et al., 2020), and self-predictive representations to forecast their own latent codes (Schwarzer et al., 2021; Ni et al., 2024). Each objective weights an event by how often it occurs. Figure 2a makes the consequence concrete: a budgeted compressor that mixes next-observation log-loss with a reward term of weight λ keeps a decisive cue of rate ρ only when $\lambda\rho$ exceeds a constant. Every tenfold drop in rarity demands a tenfold larger reward weight, and more data does not help (Figure 2b).

Task-grounded abstractions point in the right direction. Bisimulation, DeepMDP, and approximate information states preserve reward and transition consequences (Zhang et al., 2021; Gelada et al., 2019; Subramanian et al., 2022); task-informed and denoised world models separate distractors from task dynamics (Fu et al., 2021; Wang et al., 2022; Zhu et al., 2023). What these methods leave open is the statistical question that rarity raises: *how much evidence is needed before a learner may safely*

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

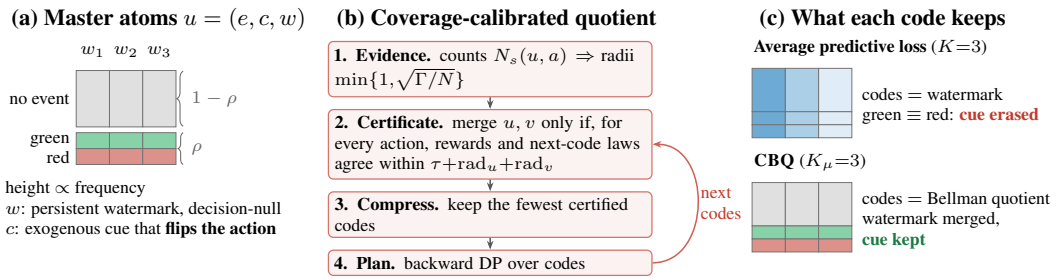

Figure 1: **Rare decisive cues and how CBQ keeps them.** (a) Master atoms cross a frequent non-event row with rare green/red cue rows; a persistent watermark w is decision-null. (b) CBQ converts counts into radii, certifies merges pairwise for every action, keeps the fewest certified codes, and plans backward. (c) A budgeted average predictive loss spends its codes on the easy-to-forecast watermark and merges green with red; CBQ merges the watermark instead and recovers the Bellman quotient.

merge two histories, and how much before it must keep them apart? A rare cue is observed rarely, so any empirical test of its consequences is noisy exactly where the stakes are highest.

We answer this question by making evidence the organizing principle of the representation. CBQ maps histories to a fixed finite set of master atoms and asks, for every pair of atoms and every action, whether the observed rewards and next-code laws could be equal given how often that atom–action row was seen. Two atoms share a latent code only when the counts certify this; among all certified codes, CBQ returns the most compact one, and plans on it by backward dynamic programming (Figure 1). Because the certificate is calibrated row by row, a rare cue is not penalized for being rare: it is kept the moment its own data can tell it apart.

Contributions.

1. **A learner that is calibrated to evidence rather than frequency.** CBQ selects the minimum-rate latent partition among those certified by count-calibrated reward and transition tests, and is well defined on every dataset, including rows that were never observed (Section 4).
2. **A finite-sample guarantee that separates bias from noise.** On a single high-probability event, the supported Q-error is at most $\mathfrak{A}_t + \mathfrak{S}_t$, fresh-rollout regret is at most $2\sum_{s \geq t}(\mathfrak{A}_s + \mathfrak{S}_s)$, every action-flipping cue is retained when $2\mathfrak{S}_t < \Delta_t - 2\mathfrak{A}_t$, and the exact supported Bellman quotient is recovered—with all decision-null nuisance merged—under a $4\mathfrak{B}_s$ separation (Section 5).
3. **A matching obstruction.** Every learner errs with constant probability when $nq\Delta^2 \leq 3/64$, and for CBQ insufficient evidence forces a merge, refinement, or action error; in the one-stage case the upper and lower conditions meet at the scale $nq\Delta^2$ (Section 5).
4. **Evidence at scale and from pixels.** On 288 random POMDPs, predictive objectives keep the cue in at most 5% of instances while CBQ keeps it in all, recovers the exact quotient in 95%, and never violates its certificate; from pixels, world-model latents lose a rare traffic light in 107 of 108 runs, while CBQ on a learned tokenizer keeps exactly the light's patch (Section 6).

2 Related Work

Predictive objectives and what they forget. Predictive-state representations and the information bottleneck compress history with respect to a chosen future or relevance variable (Littman et al., 2001; Tishby et al., 1999); energy-based and self-predictive variants scale the idea to deep partially observed RL (Zhang et al., 2023; Schwarzer et al., 2021; Ni et al., 2024). When reward is left out, even complete observation prediction can be insufficient for control (Baisero & Amato, 2021), state-only and inverse mutual-information objectives admit optima that alias different optimal actions (Rakelly et al., 2021), and controlled experiments show reward-free predictive objectives discarding exogenous control-relevant features (Pendharkar, 2026). Analyses of reconstruction and latent self-prediction

Table 1: **What each family of representations guarantees for a rare, exogenous, action-flipping cue.** "$\propto \rho$": kept only when its frequency-weighted loss contribution is large enough.

Representation family	Keeps rare decisive cue	Merges decision-null nuisance	Finite-sample guarantee	Matching lower bound
Predictive / world-model losses	$\propto \rho$	✗	✗	✗
Bisimulation, DeepMDP, AIS	✓ (population)	✓	✗	✗
Control-endogenous state	✗ (by design)	✓	✓	✗
CBQ (ours)	✓ once $2\mathfrak{S}_t < \Delta_{t,\mathrm{eff}}$	✓	✓	✓ ($nq\Delta^2$)

characterize how these losses interact with distractors (Voelcker et al., 2024). We isolate the statistical mechanism behind the failure—low atom–action coverage against a finite action gap—and give the learner that neutralizes it.

Task-grounded abstraction. Model minimization and bisimulation define the coarsest partition that preserves rewards and transitions (Givan et al., 2003; Li et al., 2006); bisimulation metrics, approximate abstractions, DeepMDP, and approximate information states make this robust and learnable (Ferns et al., 2004; Abel et al., 2016; Gelada et al., 2019; Zhang et al., 2021; Subramanian et al., 2022). Task-informed, denoised, and posterior-predictability-regularized world models separate distractor processes from task dynamics (Fu et al., 2021; Wang et al., 2022; Zhu et al., 2023). These works motivate our target, the Bellman quotient. CBQ adds what a rare cue requires: merge decisions calibrated to each row's own evidence, a separation of approximation bias from sampling error, and a guarantee stated on the behavior support where data exist. Abstraction selection for model-based RL (Jiang et al., 2015) likewise trades approximation against estimation; we specialize that trade-off to rare decisive rows and prove it is tight.

Exogenous noise and rich observations. Block-MDP and exogenous-MDP methods learn provably compact states from rich observations (Du et al., 2019; Misra et al., 2020; Efroni et al., 2022), and multi-step inverse models recover the control-endogenous latent state (Lamb et al., 2023). They discard exogenous processes by design. Our object of study is the complementary case: an exogenous variable that cannot be controlled or predicted, yet determines the optimal action.

Coverage and lower bounds. Offline RL ties finite-sample guarantees to how well the behavior data cover the comparator (Rashidinejad et al., 2021; Xie et al., 2021), and agnostic system identification makes the same point for model learning (Ross & Bagnell, 2012). POMDP theory shows how additional structure enables efficient learning and where lower bounds persist (Smallwood & Sondik, 1973; Lee et al., 2023; Chen et al., 2023). Our obstruction follows Le Cam's two-point method (Le Cam, 1964; Tsybakov, 2009); its contribution is to place the rare row's information budget $nq\Delta^2$ exactly where the upper theorem needs it.

3 Which Distinctions Must a Latent State Keep?

Finite master model. We consider a finite-horizon POMDP with stages $t = 1, \dots, H$, finite actions $\mathcal{A}$, rewards in $[0, 1]$, and pre-action histories $H_t = (O_{1:t}, A_{1:t-1})$ that include the current observation. A finite map χ_t, fixed before data are collected, sends each history to a master atom $U_t = \chi_t(H_t) \in \mathcal{U}_t$. Histories in one atom share the reward mean $r_t(u, a)$ and next-atom law $P_t^\chi(\cdot \mid u, a)$, so the optimal values $Q_t^\star(u, a) = r_t(u, a) + \mathbb{E}_{U' \sim P_t^\chi(\cdot|u,a)}[V_{t+1}^\star(U')]$ are the full-history optimal values; we write $B_t = H - t + 1$ for the value range. The master map plays the role of a fine tokenizer; we ask which of its distinctions a decision-sufficient latent must keep, and how much data certifies that choice.

Events, cues, and nuisance. A master-measurable event $E_t = e_t(U_t)$ marks the rare decision region and a cue $C_t = c_t(U_t)$, observed before the action, carries the information that selects the action there. The atom may also contain a nuisance N_t with arbitrary entropy and dynamics. Nuisance is *Bellman-null* when reward and the next task coordinate $T_{t+1} \ni (E_{t+1}, C_{t+1})$ depend on

the history only through (T_t, A_t). In the running example of Figure 1, e indicates an intersection, $c \in \{\text{green}, \text{red}\}$, and the watermark w is Bellman-null even though it is highly predictable.

The target: the supported Bellman quotient. Let μ be the behavior policy, d_t^μ its atom occupancy, and $S_{\mu,t} = \{u : d_t^\mu(u) > 0\}$ the behavior support. Backward from a single terminal class, two supported atoms are Bellman-equivalent, $u \sim_t^{\mathrm{B}} v$, when for every action their rewards agree and their next-atom laws agree after pushing through the stage-$(t{+}1)$ quotient. The resulting *supported Bellman quotient* is the coarsest code that preserves every action consequence. It merges every Bellman-null nuisance fiber and separates any two atoms with different optimal actions. Its separation

$$\zeta_t = \min_{u \not\sim_t^{\mathrm{B}} v} \max_{a \in \mathcal{A}} \Big\{ |r_t(u,a) - r_t(v,a)| + B_{t+1}\, \mathrm{TV}\big((\varphi_{t+1}^{\mathrm{B}})_{\#} P_t^\chi(\cdot \mid u,a), (\varphi_{t+1}^{\mathrm{B}})_{\#} P_t^\chi(\cdot \mid v,a)\big) \Big\} \quad (1)$$

measures how far apart its blocks are, where $\varphi_{t+1}^{\mathrm{B}}$ is the next-stage quotient map.

Rarity is coverage, not frequency. What limits a learner is not how often the event occurs but how often each decision-relevant *atom–action row* is observed. For the event and non-event strata $D \in \{E, \bar{E}\}$,

$$q_{D,t} = \min_{u \in S_{\mu,t} \cap \mathcal{U}_{D,t},\ a \in \mathcal{A}} d_t^\mu(u,a), \qquad q_{E,t} \ \geq\ \rho_t^\mu\, \lambda_t\, \alpha_t\, \beta_t, \quad (2)$$

where ρ_t^μ is the event rate, λ_t the smallest cue probability given the event, α_t the smallest atom share within a cue cell, and β_t the smallest behavior probability of an action at an event atom (an empty stratum has $q_{D,t} = +\infty$). Event rarity, cue imbalance, nuisance fragmentation, and unexplored actions each shrink coverage multiplicatively. The appendix writes $q_{D,t}$ as $q_{D,t}^{\mathrm{audit}}$.

The action gap. The decisive cue matters through its supported active gap $\Delta_t = \min_{u \in S_{\mu,t} \cap \mathcal{U}_{E,t}} \{Q_t^\star(u, a_t^\star(u)) - \max_{a \neq a_t^\star(u)} Q_t^\star(u,a)\}$, and we assume that at least two supported cue atoms have different optimal actions—the cue flips the action. Target policies π are evaluated on their own occupancy d_t^π, with event rate ρ_t^π and event-conditional regret $\mathcal{R}_{E,t}^\pi / \rho_t^\pi$; we require every target-reachable atom–action row to be behavior-supported, with finite density ratio $C_{D,t}^{\mathrm{atom}}$.

4 Coverage-Calibrated Bellman Quotients

CBQ learns from n independent behavior episodes. For each supported row it forms the count $N_t(u,a)$, the empirical reward $\widehat{r}_t(u,a)$, and the empirical next-atom law $\widehat{P}_t^\chi(\cdot \mid u,a)$; an unobserved row receives a fixed bounded placeholder (reward 0, a point mass on the first next atom) together with the maximal radius, so that every dataset yields a well-defined model. Evidence enters through the radius

$$\mathrm{rad}(N) = \min\Big\{1, \sqrt{\Gamma_{\mathrm{dev}}/N}\Big\}, \qquad \mathrm{rad}(0) = 1, \qquad \Gamma_{\mathrm{dev}} = \max_s |\mathcal{U}_{s+1}| + \log \frac{8H|\mathcal{A}||\Phi| \sum_s |\mathcal{U}_s|}{\delta}, \quad (3)$$

where Φ is the finite candidate class of coarsenings of χ and δ the confidence level. A rarely seen row has a wide radius and therefore cannot be separated from its neighbors by accident; a well-covered row has a narrow radius and cannot be merged with a row whose consequences differ.

Certified merges. Working backward from the terminal stage, suppose the stage-$(s{+}1)$ code $\widehat{\phi}_{s+1}$ has been chosen. Two supported atoms u, v are *compatible* at stage s if, for every action a,

$$|\widehat{r}_s(u,a) - \widehat{r}_s(v,a)| \leq \tau_s^r(u,v,a) + \mathrm{rad}(N_s(u,a)) + \mathrm{rad}(N_s(v,a)), \quad (4)$$

$$\mathrm{TV}\big((\widehat{\phi}_{s+1})_{\#} \widehat{P}_s^\chi(\cdot \mid u,a), (\widehat{\phi}_{s+1})_{\#} \widehat{P}_s^\chi(\cdot \mid v,a)\big) \leq \tau_s^p(u,v,a) + \mathrm{rad}(N_s(u,a)) + \mathrm{rad}(N_s(v,a)), \quad (5)$$

where the tolerances $\tau^r, \tau^p \geq 0$ encode the approximation bias of a declared reference coarsening ($\tau \equiv 0$ in the exact case). A candidate partition is *certified* if all pairs inside each block are compatible. The transition test equation 5 compares next-stage *codes*, not raw atoms, so a consequence that surfaces only several stages later still separates two atoms today.

Algorithm 1 Coverage-calibrated Bellman quotient (CBQ)

Require: episodes $\{(U^{(i)}_{1:H}, A^{(i)}_{1:H}, R^{(i)}_{1:H})\}_{i=1}^{n}$; supports $S_{\mu,s}$; tolerances τ^r, τ^p; confidence δ

$\widehat{\phi}_{H+1} \leftarrow$ single terminal code; $\widehat{V}_{H+1} \leftarrow 0$

for $s = H, H-1, \dots, 1$ **do**

  compute $N_s(u,a)$, $\widehat{r}_s(u,a)$, $\widehat{P}^{\chi}_s(\cdot \mid u,a)$ (fixed placeholder when $N_s(u,a)=0$) and radii equation 3

  push next-atom laws through $\widehat{\phi}_{s+1}$; mark $u, v \in S_{\mu,s}$ compatible iff equation 4–equation 5 hold for all a

  $\widehat{\phi}_s \leftarrow$ first certified partition of $S_{\mu,s}$ with minimum $K_{\mu,s}$ ▷ min. clique cover if Φ holds all partitions

  for each block b: representative $u_b \leftarrow$ first supported atom of b

  $\widehat{Q}_s(b,a) \leftarrow \widehat{r}_s(u_b,a) + \sum_{b'} \big[(\widehat{\phi}_{s+1})_{\#} \widehat{P}^{\chi}_s(\cdot \mid u_b, a)\big](b')\, \widehat{V}_{s+1}(b')$; $\widehat{V}_s(b) \leftarrow \max_a \widehat{Q}_s(b,a)$

end for

return codes $\widehat{\phi}_{1:H}$ and greedy policy $\widehat{g}_s(u) = \arg\max_a \widehat{Q}_s(\widehat{\phi}_s(u), a)$

Table 2: **Quantities that decide whether a rare decisive cue survives compression.**

Quantity	Meaning	Enters as	Improved by
$q_{E,t} \geq \rho\lambda\alpha\beta$	weakest event row coverage	$\mathfrak{S}_t \propto (nq)^{-1/2}$	balanced events, cues, actions
$\mathfrak{A}_t$	reference-abstraction bias	$\Delta_{t,\mathrm{eff}} = \Delta_t - 2\mathfrak{A}_t$	better abstraction, not data
Δ_t	decisive action gap	margin to protect	property of the task
ζ_t	quotient separation	recovery if $\zeta_s > 4\mathfrak{B}_s$	property of the task
Γ_{dev}	alphabet + log candidates	$\mathfrak{S}_t \propto \Gamma^{1/2}$	smaller candidate class Φ

Minimum-rate selection and planning. Among certified candidates, CBQ selects the one with the fewest supported codes, $K_{\mu,s}(\phi) = |\{\phi(u) : u \in S_{\mu,s}\}|$, breaking ties by a fixed order. Minimum rate is what removes nuisance: any distinction that the certificate does not demand is merged. Each block is represented by its first supported atom, and backward dynamic programming on the representative rows yields $\widehat{Q}_s$ and the greedy policy $\widehat{g}_s(u) = \arg\max_a \widehat{Q}_s(\widehat{\phi}_s(u), a)$. Atoms outside the behavior support keep their own singleton codes, so the learner never extrapolates a merge to states it has not seen. Algorithm 1 summarizes the procedure.

The three quantities that govern CBQ. The analysis tracks the deterministic and statistical charges accumulated from stage t to the horizon:

$$\mathfrak{A}_t = \sum_{s=t}^{H} \max_{D,u,a} \big(\varepsilon^r_s(u,a) + B_{s+1}\varepsilon^p_s(u,a)\big), \qquad \mathfrak{S}_t = 4\sqrt{2}\sum_{s=t}^{H} \max_{D,u,a}(1+B_{s+1})\sqrt{\frac{\Gamma_{\mathrm{dev}}}{n\, d^{\mu}_s(u,a)}}, \tag{6}$$

where $\varepsilon^r_s(u,a) = \max_v \tau^r_s(u,v,a)$ and $\varepsilon^p_s(u,a) = \max_v \tau^p_s(u,v,a)$, and $\mathfrak{B}_t = \mathfrak{A}_t + \mathfrak{S}_t$. Bias shrinks the usable gap to the *effective gap* $\Delta_{t,\mathrm{eff}} = \Delta_t - 2\mathfrak{A}_t$; sampling error must then fit inside it. Table 2 collects the roles of these quantities.

5 Guarantees and a Matching Obstruction

All statements below are specializations of a single theorem, Theorem 16 in Appendix A.3, whose constants are explicit and universal. The upper results hold simultaneously on one event, and the lower results hold for every learner.

5.1 Value, regret, and retention

Theorem 1 (Supported value and regret)**.** *Fix a start stage t and $\delta \in (0,1)$.* CBQ *returns a unique code, model, and policy on every dataset. If every nonempty stratum at stages $s \geq t$ satisfies $nq_{D,s} \geq 8\Gamma_{\mathrm{cnt}}$ with $\Gamma_{\mathrm{cnt}} = \log(4H|\mathcal{A}|\sum_s |\mathcal{U}_s|/\delta)$, then there is an event $\mathcal{H}_\delta$ of probability at least $1-\delta$ on which*

$$\sup_{u \in S_{\mu,t},\, a \in \mathcal{A}} \big|Q^{\star}_t(u,a) - \widehat{Q}_t(\widehat{\phi}_t(u), a)\big| \leq \mathfrak{B}_t = \mathfrak{A}_t + \mathfrak{S}_t, \tag{7}$$

and, for every start law supported on $S_{\mu,t}$—including the start marginal of any target policy—the learned policy's fresh-rollout regret is at most $2\sum_{s=t}^{H} \mathfrak{B}_s$.

The event $\mathcal{H}_\delta$ intersects a deviation event, which controls every observed row at its own radius, with a count event guaranteeing $N_s(u,a) \geq \frac{1}{2} n d_s^\mu(u,a)$ on every row the theorem uses; this split lets the guarantee hold on every dataset while stating the error through population coverage. Regret is measured in the true environment.

Corollary 2 (Rare decisive cues are retained)**.** *Under the assumptions of Theorem 1, suppose every supported event atom has a unique optimal action with gap at least Δ_t. If*

$$2\mathfrak{S}_t < \Delta_{t,\mathrm{eff}} = \Delta_t - 2\mathfrak{A}_t, \tag{8}$$

then on $\mathcal{H}_\delta$ CBQ *acts optimally on every target-reachable supported event atom, its event-conditional regret is zero for every target policy with $\rho_t^\pi > 0$, and any two supported cue atoms with different optimal actions receive different codes.*

Condition equation 8 is the central message of the paper in one line. The margin that protects the decisive action is the raw gap minus twice the bias; sampling error must fit inside what remains. Two consequences follow. First, *bias cannot be bought back with data*: $\mathfrak{S}_t \to 0$ as $n \to \infty$, but $\mathfrak{A}_t$ does not, so a reference abstraction that blurs the cue by more than $\Delta_t/2$ loses it at every sample size. Second, *rarity is paid in coverage*: $\mathfrak{S}_t \propto \sqrt{\Gamma_{\mathrm{dev}}/(nq)}$, and the event rate enters only through $q_{E,t} \geq \rho\lambda\alpha\beta$, on equal footing with cue balance, nuisance fragmentation, and exploration of the counterfactual action. A rare event with balanced cues and explored actions can be cheaper to certify than a frequent one observed under a nearly deterministic behavior policy.

5.2 Exact recovery of the Bellman quotient

Theorem 3 (Supported quotient recovery)**.** *Suppose the supported Bellman quotient belongs to the candidate class and its separation equation 1 satisfies $\zeta_s > 4\mathfrak{B}_s$ for every $s \geq t$. Then on $\mathcal{H}_\delta$, simultaneously for all $s \geq t$ and all $u, v \in S_{\mu,s}$, $\widehat{\phi}_s(u) = \widehat{\phi}_s(v) \iff u \sim_s^{\mathrm{B}} v$. In particular* CBQ *merges every supported Bellman-null nuisance fiber and preserves every supported difference in optimal action.*

The proof runs backward from the terminal class. Once the learned and population codes agree at stage $s{+}1$, the population quotient is itself a certified candidate at stage s; a cross-block merge would force $d_s^{\mathrm{B}}(u,v) \leq 4\mathfrak{B}_s < \zeta_s$, and minimum rate excludes any finer certified partition. Recovery is thus the conjunction of two properties that average objectives trade against each other: the certificate forbids merging what matters, and minimum rate forbids keeping what does not.

5.3 No learner can do better

Theorem 4 (Rare-row obstruction)**.** *For every learner and every $0 < q \leq 1/6$, $0 < \Delta \leq 1/4$, there is a pair of bounded one-stage POMDPs with hard atom–action mass q and gap Δ whose data laws satisfy $\mathrm{KL} \leq \frac{16}{3} nq\Delta^2$. If $nq\Delta^2 \leq 3/64$, then on one member the learner picks the wrong action at the hard atom with probability at least $1/8$, its expected event-conditional regret is at least $\Delta/8$, and its expected target regret is at least $\rho^\pi \Delta/32$. For* CBQ*, the same budget forces a wrong merge, an over-refinement, or a wrong action with probability at least $1/4$.*

Corollary 5 (The evidence scale is tight)**.** *Let $H = 1$ with zero tolerances, so that $\mathfrak{A}_1 = 0$ and $\Delta_{1,\mathrm{eff}} = \Delta_1$.* CBQ *retains every decisive cue with probability at least $1 - \delta$ whenever $nq \geq 8\Gamma_{\mathrm{cnt}}$ and $nq\Delta^2 > 128\,\Gamma_{\mathrm{dev}}$, while every learner fails with constant probability on a member of the lower family when $nq\Delta^2 \leq 3/64$. The upper and lower conditions therefore meet at the scale $nq\Delta^2$, up to the logarithmic factor Γ_{dev}.*

Theorem 4 and Corollary 5 close the loop opened in the introduction. The obstruction depends on the data only through $nq\Delta^2$: the event's frequency, the decisive gap, and the coverage of each counterfactual action combine into a single information budget. An average objective spends that budget in proportion to ρ; CBQ spends it where the certificate needs it, and matches what any learner needs up to constants and a logarithmic factor.

Proof architecture. The appendix proves the theorem in thirteen self-contained steps (Appendix A.4): concentration and a multiplicative Chernoff bound give $\mathcal{H}_\delta$; a pairwise endpoint envelope controls every representative against its whole block; a Bellman telescope separates $\mathfrak{A}_t$ from $\mathfrak{S}_t$; and the lower bounds are two-point Le Cam arguments with KL divergence $2nq\Delta \log \frac{1+\Delta}{1-\Delta}$.

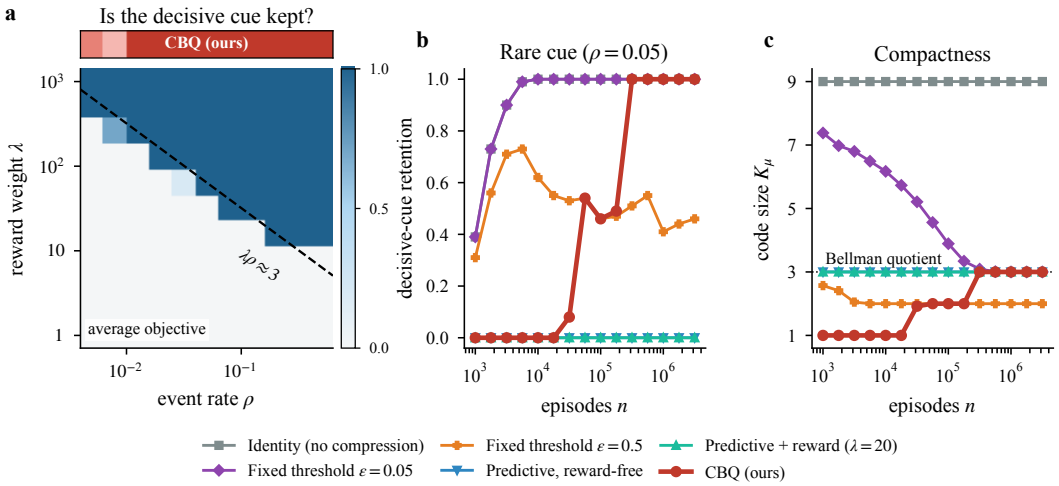

Figure 2: **Average objectives erase the rare cue; CBQ keeps it and compresses** (traffic light). (a) At $n = 10^6$, a budgeted predictive + reward compressor keeps the cue only above $\lambda\rho \approx 3$; the strip shows CBQ at the same budget, which retains it for every $\rho \geq 0.013$. (b) At $\rho = 0.05$, both predictive objectives stay at zero retention up to 3.2×10^6 episodes and a loose fixed threshold plateaus near 0.45. (c) CBQ converges to the three-block Bellman quotient; identity keeps all nine atoms.

6 Numerical Study

We test the theory on a controlled traffic light, 288 random POMDPs, and a pixel task where CBQ runs on a learned tokenizer against deep world-model representations.

Learners and protocol. CBQ uses the radii of equation 3 exactly as analyzed, with $\delta = 0.05$. *Identity* keeps every atom. *Predictive* selects, among candidates with at most as many codes as the Bellman quotient, the one minimizing occupancy-weighted next-observation log-loss, optionally plus λ times the reward squared error (*predictive + reward*)—a budgeted world-model objective. *Fixed threshold* replaces the count-calibrated radius by a constant $\epsilon/2$; *myopic* drops the transition test equation 5. *Retention* means that every event atom receives its optimal action and no two cue atoms with different optimal actions share a code. All runs use seed 42 as the root of every random stream; specifications, per-point replicate counts, and complete tables are in Appendix B.

6.1 Average objectives erase rare cues; certified compression does not

The traffic light ($H = 2$) instantiates Figure 1: {no event, green, red} crossed with $m = 3$ watermarks that persist to the next frame with probability 0.9; green and red flip the optimal action with gap $\Delta = 0.4$. A reward-free predictive compressor spends all three of its codes on the watermark, which carries about 0.8 nats about the next frame, and merges green with red at every sample size. Adding a reward term restores the cue only when its weight compensates for rarity: the boundary in Figure 2a is $\lambda\rho \approx 3$, so a $10\times$ rarer event needs a $10\times$ larger reward weight. CBQ has no such weight and keeps the cue across $\rho \in [0.013, 0.5]$, because it separates green from red by the counts of those rows rather than by their share of the loss. From $n \approx 5.6 \times 10^5$ it returns exactly the three-block quotient in every replicate: nine atoms collapse onto {no event, green, red} and the watermark leaves the representation.

6.2 One coverage law across 288 random POMDPs

Each random instance draws a horizon $H \in \{2, 3, 4\}$, two or three actions, two or three cues, two nuisance factors with three or four values, an event rate $\rho \in \{0.02, 0.05, 0.1\}$, and a random task kernel through which the cue acts on future stages; rewards are planted so that every event atom has

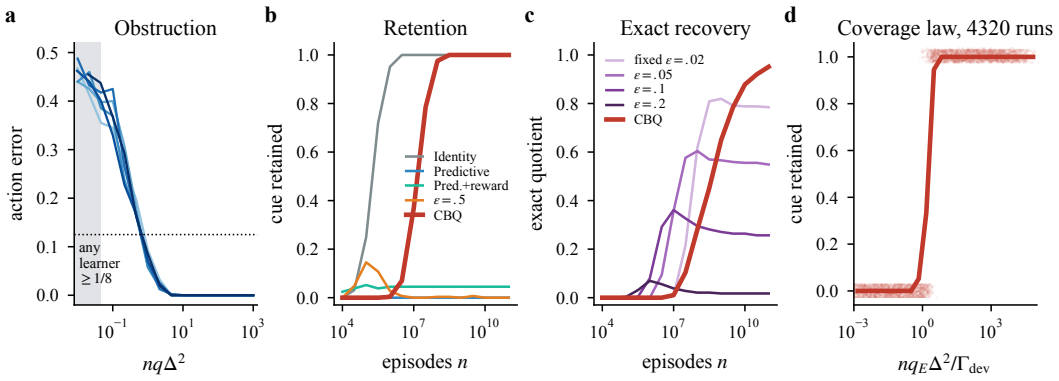

Figure 3: **One coverage law, from the obstruction to 288 random POMDPs.** (a) On the hard pair of Theorem 4, action error for six (q, Δ) configurations collapses in $nq\Delta^2$; shading marks where every learner errs with probability $\geq 1/8$. (b–c) Random POMDPs ($H \leq 4$, up to 3 actions, 128 atoms per stage): predictive objectives keep the cue in at most 5% of instances at any n; fixed thresholds plateau in exact recovery at a level set by ϵ, while CBQ retains the cue in every instance and recovers the exact quotient in 95% at $n = 10^{11}$. (d) All 4,320 CBQ runs switch from failure to retention within one decade of normalized coverage.

gap $\Delta = 0.25$. CBQ uses a factored candidate class—partitions of the decision values times subsets of kept nuisance factors—which contains the Bellman quotient and keeps Γ_{dev} between 79 and 181.

The results hold at scale (Figure 3b–d). *Predictive objectives fail structurally*: reward-free compression keeps the cue in 0% of instances and predictive + reward in at most 5%, at every n up to 10^{11}. *Calibration removes the threshold dilemma*: every fixed ϵ plateaus in exact recovery—at 78%, 55%, 26%, and 2% for $\epsilon = 0.02, 0.05, 0.1, 0.2$—because any constant is too small for some nuisance pairs and too large for some decisive ones; CBQ keeps rising to 95% and returns codes of exactly the quotient size, against $12.5\times$ for identity. *The coverage law is universal*: pooled over all instances, retention is 0% below normalized coverage 0.3 and 99.95% above 3. *The certificate is valid*: in all 2,393 stage-cases where $2\mathfrak{S}_t < \Delta$ and the count condition hold, CBQ retains the cue, and in all 3,237 runs meeting the count condition its exact regret lies below $2\sum_s \mathfrak{B}_s$. Appendix B.5 adds the complementary test of Corollary 2: once injected approximation bias exceeds $\Delta/2$, event regret stops improving with n from 10^6 to 10^9.

6.3 From pixels: certified codes on a learned tokenizer

Frames are 64×64 road scenes with one of eight large textured watermarks, which persist to the next frame with probability 0.9, and, with rate ρ, a small traffic light whose red or green lamp flips the optimal action; pixel noise and brightness jitter make every frame unique. The master map is learned without labels: per-patch k-means on a separate pool of 20k frames, with a fixed resolution below which centroids merge, recovers the 24 underlying scene states exactly. CBQ then selects which of the 16 patch factors to keep ($|\Phi| = 2^{16}$). The deep baselines use the same offline data and a 16-dimensional convolutional latent trained for 4,000 steps: an autoencoder, a next-frame world model, a world model with a reward head (Dreamer-style likelihood weighting), a self-predictive latent with an EMA target, and end-to-end Q regression; frozen latents are probed with a Q-head.

Figure 4 reproduces the tabular picture in pixels. Every objective that predicts the future loses the light: across all event rates and sample sizes, the world model and the self-predictive latent never retain it, and the world model with a reward head does so in one run of 36. CBQ retains the cue in every seed once coverage suffices—from $n = 3 \times 10^5$ at $\rho = 0.1$, 10^6 at $\rho = 0.03$, and 3×10^6 at $\rho = 0.01$—and in every such run it keeps exactly one patch, the traffic light, returning the three-code Bellman quotient. Its code carries no watermark information (probe accuracy 0.125, chance), whereas every deep latent encodes the watermark (0.79–1.00), including end-to-end Q regression, which also keeps the cue. CBQ is the only learner whose pixel representation is simultaneously

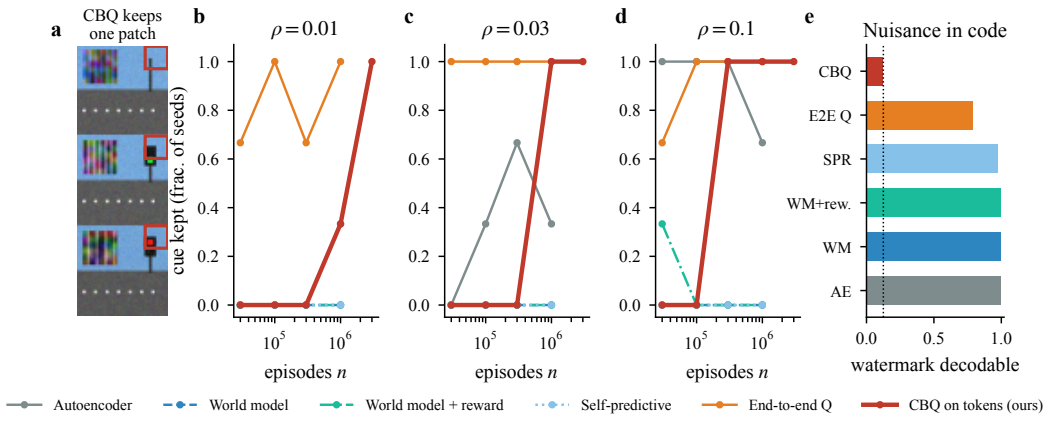

Figure 4: **From pixels.** (a) 64×64 frames with a large persistent watermark and a rare traffic light; the red box is the single patch whose codes CBQ keeps. (b–d) Fraction of three seeds in which the cue is fully retained. Latents trained to predict the future—world model, world model + reward head, self-predictive—lose the cue in 107 of their 108 runs. (e) Linear-probe accuracy of the watermark identity from each code (chance $= 1/8$, dotted).

Table 3: **Summary.** Traffic light: $\rho = 0.05$, $n = 10^6$, 100 replicates. Random POMDPs: 288 instances at $n = 10^{11}$; $K/K^\star$ is the learned-to-quotient code-size ratio. Delayed: three-stage task with reward only at stage three, $n = 10^5$ (Appendix B).

	Traffic light			288 random POMDPs			Delayed	No tuned
Learner	Retain	Nuisance merged	K	Retain	Exact quotient	$K/K^\star$	Retain	threshold
Identity	1.00	0.00	9	1.00	0.00	12.5	1.00	✓
Predictive, reward-free	0.00	0.00	3	0.00	0.00	0.60	0.00	✓
Predictive + reward	0.00	0.00	3	0.05	0.00	0.91	0.00	✗
Fixed threshold $\epsilon = 0.05$	1.00	1.00	3	1.00	0.55	0.96	1.00	✗
Fixed threshold $\epsilon = 0.5$	0.41	1.00	2	0.00	0.00	0.30	0.00	✗
Myopic reward quotient	1.00	1.00	3	1.00	0.94	1.00	0.00	✓
CBQ (ours)	**1.00**	**1.00**	**3**	**1.00**	**0.95**	**1.00**	**1.00**	✓

decision-sufficient, free of nuisance, interpretable down to the single patch that matters, and backed by a certificate.

7 Discussion and Conclusion

What the theory tells a practitioner. *Collect coverage, not volume*: the budget that matters is nq_E, and $q_E \geq \rho\lambda\alpha\beta$ rewards balanced cues and explored counterfactual actions as much as frequent events. *Fix bias before scaling data*: bias subtracts from the gap at every sample size.

Scope. The master partition can be given or, as in Section 6, learned by an unsupervised tokenizer; the certificate applies unchanged on top of any such map, which makes CBQ a drop-in decision layer for discrete world-model tokenizers.

Conclusion. Rare does not mean irrelevant. A latent state should keep a distinction when the evidence says it changes the action, however seldom it occurs, and discard it otherwise, however predictable it is. CBQ implements this with one certificate: it keeps every action-flipping cue once $2\mathfrak{S}_t < \Delta_t - 2\mathfrak{A}_t$, recovers the Bellman quotient under separation, and needs only the unavoidable budget $nq\Delta^2$.

AI USE STATEMENT

Generative AI tools assisted literature triage, formalization, proof drafting and checking, simulation code, and manuscript drafting and editing. The authors checked cited sources against primary proceedings or official metadata, reviewed every AI-assisted proof, and verified every reported number against the released simulation outputs. The authors take responsibility for the final content.

ETHICS STATEMENT

The work is theoretical and uses synthetic simulations only; it involves no human participants, personal data, or physical deployment.

REPRODUCIBILITY STATEMENT

Appendix A.1–A.5 state every assumption and give complete proofs. Appendix B specifies every environment, baseline, metric, and sample size of the numerical study. All random streams descend from seed 42; the tabular and random-POMDP studies run on CPUs in minutes, and the pixel study takes about 45 minutes on one GPU. The supplement contains the simulation code and raw result files.

REFERENCES

David Abel, D. Ellis Hershkowitz, and Michael L. Littman. Near optimal behavior via approximate state abstraction. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2915–2923, 2016.

Andrea Baisero and Christopher Amato. Reconciling rewards with predictive state representations. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 2170–2176, 2021. doi: 10.24963/ijcai.2021/299.

Fan Chen, Huan Wang, Caiming Xiong, Song Mei, and Yu Bai. Lower bounds for learning in revealing POMDPs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 5104–5161, 2023. URL https://proceedings.mlr.press/v202/chen23ae.html.

Simon S. Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudík, and John Langford. Provably efficient RL with rich observations via latent state decoding. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1665–1674, 2019.

Yonathan Efroni, Dylan J. Foster, Dipendra Misra, Akshay Krishnamurthy, and John Langford. Sample-efficient reinforcement learning in the presence of exogenous information. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 5062–5127, 2022.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 162–169, 2004.

Xiang Fu, Ge Yang, Pulkit Agrawal, and Tommi Jaakkola. Learning task informed abstractions. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 2021. URL https://arxiv.org/abs/2106.15612.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. DeepMDP: Learning continuous latent space models for representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2019. URL https://arxiv.org/abs/1906.02736.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Nan Jiang, Alex Kulesza, and Satinder Singh. Abstraction selection in model-based reinforcement learning. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 179–188, 2015.

Alex Lamb, Riashat Islam, Yonathan Efroni, Aniket Rajiv Didolkar, Dipendra Misra, Dylan J. Foster, Lekan P. Molu, Rajan Chari, Akshay Krishnamurthy, and John Langford. Guaranteed discovery of control-endogenous latent states with multi-step inverse models. *Transactions on Machine Learning Research*, 2023.

Lucien Le Cam. Sufficiency and approximate sufficiency. *The Annals of Mathematical Statistics*, 35 (4):1419–1455, 1964. doi: 10.1214/aoms/1177700372.

Jonathan Lee, Alekh Agarwal, Christoph Dann, and Tong Zhang. Learning in POMDPs is sample-efficient with hindsight observability. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 18733–18773, 2023. URL https://proceedings.mlr.press/v202/lee23a.html.

Lihong Li, Thomas J. Walsh, and Michael L. Littman. Towards a unified theory of state abstraction for MDPs. In *International Symposium on Artificial Intelligence and Mathematics*, 2006.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems*, volume 14, 2001. URL https://papers.nips.cc/paper/2001/hash/1e4d36177d71bbb3558e43af9577d70e-Abstract.html.

Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6961–6971, 2020.

Tianwei Ni, Benjamin Eysenbach, Erfan SeyedSalehi, Michel Ma, Clement Gehring, Aditya Mahajan, and Pierre-Luc Bacon. Bridging state and history representations: Understanding self-predictive RL. In *International Conference on Learning Representations*, 2024.

Ayan Pendharkar. Predictive objectives discard exogenous control-relevant features: A controlled mechanistic study. *arXiv preprint arXiv:2606.30068*, 2026. URL https://arxiv.org/abs/2606.30068.

Kate Rakelly, Abhishek Gupta, Carlos Florensa, and Sergey Levine. Which mutual-information representation learning objectives are sufficient for control? In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2106.07278.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Stéphane Ross and J. Andrew Bagnell. Agnostic system identification for model-based reinforcement learning. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 1905–1912, 2012. URL https://arxiv.org/abs/1203.1007.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588: 604–609, 2020.

Max Schwarzer, Ankesh Anand, Rishab Goel, R Devon Hjelm, Aaron Courville, and Philip Bachman. Data-efficient reinforcement learning with self-predictive representations. In *International Conference on Learning Representations*, 2021.

Richard D. Smallwood and Edward J. Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973. doi: 10.1287/opre.21.5.1071.

Jayakumar Subramanian, Amit Sinha, Raihan Seraj, and Aditya Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems. *Journal of Machine Learning Research*, 23, 2022. URL https://arxiv.org/abs/2010.08843.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pp. 368–377, 1999. URL https://mlanthology.org/misc/1999/tishby1999misc-information/.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Claas A. Voelcker, Tyler Kastner, Igor Gilitschenski, and Amir-massoud Farahmand. When does self-prediction help? understanding auxiliary tasks in reinforcement learning. *Reinforcement Learning Journal*, 4:1567–1597, 2024. URL https://rlj.cs.umass.edu/2024/papers/Paper197.html.

Tongzhou Wang, Simon S. Du, Antonio Torralba, Phillip Isola, Amy Zhang, and Yuandong Tian. Denoised MDPs: Learning world models better than the world itself. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 2022. URL https://arxiv.org/abs/2206.15477.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2006.10742.

Tianjun Zhang, Tongzheng Ren, Chenjun Xiao, Wenli Xiao, Joseph E. Gonzalez, Dale Schuurmans, and Bo Dai. Energy-based predictive representations for partially observed reinforcement learning. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2477–2487, 2023. URL https://proceedings.mlr.press/v216/zhang23b.html.

Chuning Zhu, Max Simchowitz, Siri Gadipudi, and Abhishek Gupta. RePo: Resilient model-based reinforcement learning by regularizing posterior predictability. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2309.00082.

A Complete theoretical statements and proofs

A.1 Theoretical Setup

A.1.1 Finite master model and decision target

Fix a finite horizon $H \in \mathbb{N}$, finite action and observation alphabets, and rewards $R_t \in [0,1]$. At stage $t \in [H]$, the pre-action history is

$$H_t = (O_{1:t}, A_{1:t-1}) \in \mathcal{H}_t.$$

A finite, sample-independent master partition $\chi_t : \mathcal{H}_t \to \mathcal{U}_t$ produces $U_t := \chi_t(H_t)$; $\mathcal{U}_{H+1}$ is a singleton. The master reward and transition rows are

$$r_t(u,a) := \mathbb{E}[R_t \mid H_t = h, A_t = a], \qquad P_t^{\chi}(\cdot \mid u,a) := \mathcal{L}(U_{t+1} \mid H_t = h, A_t = a),$$

where the right-hand sides are identical for all $h \in \chi_t^{-1}(u)$. Put

$$B_t := H - t + 1, \qquad B_{H+1} := 0, \qquad V_{H+1}^\star \equiv 0,$$

and define

$$Q_t^\star(u,a) = r_t(u,a) + \mathbb{E}_{U' \sim P_t^\chi(\cdot|u,a)}[V_{t+1}^\star(U')], \qquad V_t^\star(u) := \max_{a \in \mathcal{A}} Q_t^\star(u,a).$$

Thus $0 \le Q_t^\star, V_t^\star \le B_t$, and these quantities equal the full-history optimal values on each master atom.

The active event and observable cue are master-measurable:

$$E_t = e_t(U_t) \in \{0,1\}, \qquad C_t = c_t(U_t), \qquad \mathcal{U}_{E,t} := \{u : e_t(u) = 1\}, \quad \mathcal{U}_{\bar{E},t} := \{u : e_t(u) = 0\}.$$

The raw history may also contain a finite nuisance N_t. A finite nuisance-free task coordinate $T_t = \tau_t(U_t)$, containing E_t and C_t, obeys

$$\mathcal{L}(R_t, T_{t+1} \mid H_t = h, A_t = a) = K_t(\cdot \mid T_t(h), a).$$

Nuisance entropy and nuisance-only dynamics are unrestricted. A future shock $W_{t+1} \notin \sigma(H_t)$ enters only through its action-conditioned law in the reward/transition kernel. A decision-time code may retain that law or posterior, but not the unrealized shock unless its posterior is a point mass.

Let Π_{tar} contain a full-history optimal policy and all learned latent-planning policies to be evaluated. For $\nu \in \{\mu\} \cup \Pi_{\mathrm{tar}}$, write

$$d_t^\nu(u) := \Pr_\nu(U_t = u), \qquad d_t^\nu(u,a) := \Pr_\nu(U_t = u, A_t = a), \qquad \rho_t^\nu := \Pr_\nu(E_t = 1).$$

Fix a total order $\prec_{\mathcal{A}}$ before sampling and let $a_t^\star(u)$ be its first maximizer of $Q_t^\star(u,\cdot)$. With

$$S_{\mu,t} := \{u : d_t^\mu(u) > 0\},$$

the supported active gap is

$$\Delta_t := \min_{u \in S_{\mu,t} \cap \mathcal{U}_{E,t}} \left\{ Q_t^\star(u, a_t^\star(u)) - \max_{a \neq a_t^\star(u)} Q_t^\star(u,a) \right\}.$$

For a selector $g_t : \mathcal{U}_t \to \mathcal{A}$, define

$$\mathcal{R}_{E,t}^\pi(g_t) := \mathbb{E}_{U_t \sim d_t^\pi}\left[\mathbf{1}\{E_t = 1\} \big(Q_t^\star(U_t, a_t^\star(U_t)) - Q_t^\star(U_t, g_t(U_t)) \big) \right].$$

The conditional ratio is reported only when $\rho_t^\pi > 0$; the unnormalized event contribution is defined as zero when $\rho_t^\pi = 0$.

A.1.2 Coverage, candidate partitions, and the learner

The learner observes n independent length-H episodes from μ. For $D \in \{E, \bar{E}\}$, define

$$S_{D,t} := S_{\mu,t} \cap \mathcal{U}_{D,t}, \qquad q_{D,t}^{\mathrm{audit}} := \min_{\substack{u \in S_{D,t} \\ a \in \mathcal{A}}} d_t^\mu(u,a)$$

when $S_{D,t} \neq \varnothing$. If the stratum is empty, set $q_{D,t}^{\mathrm{audit}} = +\infty$ and all of its behavior sums and error envelopes to zero. Also set

$$q_t^{\mathrm{audit}} := \min_{D: S_{D,t} \neq \varnothing} q_{D,t}^{\mathrm{audit}}.$$

On the active cue support, let

$$\lambda_t := \min_c \Pr_\mu(C_t = c \mid E_t = 1), \quad \alpha_t^{\mathrm{audit}} := \min_{c,u: c_t(u)=c} \Pr_\mu(U_t = u \mid E_t = 1, C_t = c),$$

$$\beta_t := \min_{u \in S_{E,t},\, a \in \mathcal{A}} \Pr_\mu(A_t = a \mid U_t = u), \qquad q_{E,t}^{\mathrm{audit}} \geq \rho_t^\mu \lambda_t \alpha_t^{\mathrm{audit}} \beta_t,$$

where each minimum is over its behavior-supported domain.

For a target occupancy and queried action, put

$$d_t^{\pi \rightsquigarrow a}(u,a) := d_t^\pi(u), \qquad C_{D,t}^{\mathrm{atom}} := \sup_{\substack{\pi \in \Pi_{\mathrm{tar}},\, u \in \mathcal{U}_{D,t} \\ a \in \mathcal{A}}} \frac{d_t^{\pi \rightsquigarrow a}(u,a)}{d_t^\mu(u,a)}.$$

Here a positive numerator over a zero denominator is $+\infty$, $0/0 := 0$, and the supremum over an empty master stratum is zero. Hence a finite ratio on every target-relevant branch excludes target-positive, behavior-zero queried rows.

At each stage, $\mathcal{G}_t$ is a fixed finite family of maps on $S_{\mu,t}$, containing the supported identity. Every candidate has the common canonical extension

$$\phi_t(h) = \widetilde{g}_t(\chi_t(h)), \qquad \widetilde{g}_t(u) = \begin{cases} g_t(u), & u \in S_{\mu,t}, \\ \bot_{t,u}, & u \notin S_{\mu,t}, \end{cases}$$

where off-support labels are fixed, injective, candidate-independent, and disjoint from supported labels. Write

$$\Phi := \prod_{t=1}^{H} \{\widetilde{g}_t \circ \chi_t : g_t \in \mathcal{G}_t\}, \qquad K_{\mu,t}(\phi_t) := |\{\phi_t(u) : u \in S_{\mu,t}\}|.$$

Let $\sim_{H+1}^{\mathrm{B}}$ have one class. Recursively, $u \sim_t^{\mathrm{B}} v$ when, for every action, their rewards agree and the pushforwards of their next-master laws through the next quotient map agree. Denoting that quotient map by φ_t^{B}, define

$$d_t^{\mathrm{B}}(u,v) := \max_{a \in \mathcal{A}} \left\{ |r_t(u,a) - r_t(v,a)| + B_{t+1} \, \mathrm{TV}\big((\varphi_{t+1}^{\mathrm{B}})_\# P_t^\chi(\cdot \mid u,a), (\varphi_{t+1}^{\mathrm{B}})_\# P_t^\chi(\cdot \mid v,a)\big) \right\},$$

$$\zeta_t := \min_{\substack{u,v \in S_{\mu,t} \\ u \not\sim_t^{\mathrm{B}} v}} d_t^{\mathrm{B}}(u,v),$$

with $\zeta_t = +\infty$ for one supported quotient block. Total variation is $\mathrm{TV}(P,Q) = \frac{1}{2}\|P - Q\|_1$.

For supported u, v, fixed symmetric tolerances $\tau_t^r(u,v,a), \tau_t^p(u,v,a) \geq 0$ bound the reward and next-code discrepancies of a declared nuisance-invariant reference coarsening $\phi^\circ \in \Phi$: whenever $\phi_t^\circ(u) = \phi_t^\circ(v)$,

$$|r_t(u,a) - r_t(v,a)| \leq \tau_t^r(u,v,a),$$

$$\mathrm{TV}\big((\phi_{t+1}^\circ)_\# P_t^\chi(\cdot \mid u,a), (\phi_{t+1}^\circ)_\# P_t^\chi(\cdot \mid v,a)\big) \leq \tau_t^p(u,v,a) \qquad (a \in \mathcal{A}).$$

The exact specialization takes $\phi^\circ = \varphi^{\mathrm{B}}$ on support and all within-block tolerances equal to zero.

For each row, let

$$N_t(u,a) := \sum_{i=1}^{n} \mathbf{1}\{U_t^{(i)} = u, A_t^{(i)} = a\}.$$

Let u_{t+1}^0 be the first atom of the fixed order on $\mathcal{U}_{t+1}$, with the terminal singleton convention, and set

$$r_t^0(u,a) := 0, \qquad P_t^0(\cdot \mid u,a) := \delta_{u_{t+1}^0}.$$

The empirical-or-default rows used in every result below are

$$\widehat{r}_t(u,a) := \begin{cases} \dfrac{1}{N_t(u,a)} \displaystyle\sum_{i=1}^{n} \mathbf{1}\{U_t^{(i)} = u, A_t^{(i)} = a\} R_t^{(i)}, & N_t(u,a) > 0, \\ r_t^0(u,a), & N_t(u,a) = 0, \end{cases}$$

and, for $w \in \mathcal{U}_{t+1}$,

$$\widehat{P}_t^{\chi}(w \mid u,a) := \begin{cases} \dfrac{1}{N_t(u,a)} \displaystyle\sum_{i=1}^{n} \mathbf{1}\{U_t^{(i)} = u, A_t^{(i)} = a, U_{t+1}^{(i)} = w\}, & N_t(u,a) > 0, \\ P_t^0(w \mid u,a), & N_t(u,a) = 0. \end{cases}$$

At zero count both reward and transition radii equal one. Backward from the terminal singleton, a candidate may merge supported u, v only if, for every action, its empirical reward and selected-next-code discrepancies do not exceed the corresponding tolerance plus the two endpoint radii. The learner chooses the first minimum-$K_{\mu,t}$ feasible candidate, the first supported atom in each selected block as representative, and the prescribed singleton representative off support. Its representative empirical-or-default model is solved by backward dynamic programming, producing $\widehat{Q}_t, \widehat{V}_t$ and

$$\widehat{g}_t(u) := \min_{\prec_{\mathcal{A}}} \arg\max_a \widehat{Q}_t(\widehat{\phi}_t(u), a).$$

All atom, action, candidate, representative, and tie orders are fixed before sampling.

A.1.3 Assumptions

Assumption 6 (Fixed homogeneous master)**.** *The finite, sample-independent master maps have exactly homogeneous reward and next-master laws on every atom–action row, with the finite horizon, finite alphabets, and reward range stated above.*

Assumption 7 (Pre-action cue boundary)**.** *The event and cue are master-measurable before action. No nondegenerate realization of a future shock outside $\sigma(H_t)$ is supplied to the learner at decision time.*

Assumption 8 (Nuisance-free task kernel)**.** *Reward and next task-coordinate laws factor through (T_t, A_t). Nuisance entropy and nuisance-only dynamics are otherwise unrestricted.*

Assumption 9 (Independent fixed-master episodes)**.** *The n episodes are iid. The master partitions, behavior supports, candidate families, tolerances, defaults, and total orders are fixed before sampling.*

Assumption 10 (Population audit coverage)**.** *Every nonempty stratum consumed by an upper or recovery conclusion has $q_{D,t}^{\mathrm{audit}} > 0$. On a nonempty active stratum, $\rho_t^{\mu}, \lambda_t, \alpha_t^{\mathrm{audit}}, \beta_t > 0$. Realized count lower bounds are not assumed.*

Assumption 11 (Full counterfactual target support)**.** *For every theorem start t, every consumed stage $s \geq t$, and every target-relevant stratum, $C_{D,s}^{\mathrm{atom}} < \infty$. In particular, if $S_{D,s} = \varnothing$, then $d_s^{\pi}(\mathcal{U}_{D,s}) = 0$ for every $\pi \in \Pi_{\mathrm{tar}}$.*

Assumption 12 (Support-aligned coarsening class)**.** *Every candidate coarsens the common master map, uses the same injective canonical off-support extension, and is ranked by $K_{\mu,t}$; the class contains the supported identity. In the recovery specialization it also contains the supported Bellman quotient.*

Assumption 13 (Primitive approximation bias)**.** *The declared reference coarsening obeys the fixed reward and next-code pair inequalities specified above with τ_t^r, τ_t^p. These primitive tolerances generate approximation bias; they do not assume Bellman accuracy, correct actions, or cue retention.*

Assumption 14 (Fixed learner specification)**.** *Zero-count rewards, next-master laws, endpoint radii, atom/action/candidate orders, representative rules, model rows, dynamic-programming initialization, action and partition ties, canonical off-support labels, and empty-stratum outputs are fixed exactly as above before sampling.*

Assumption 15 (Supported active action margin)**.** *Every supported active atom has a unique optimal action with raw gap at least $\Delta_t > 0$, and at least two supported active cue atoms have different optimal actions. Effective-gap positivity, learned correctness, and cue retention are conclusions rather than assumptions.*

A.1.4 OBJECTIVE

The goal is a single theorem for the learner defined above. It separates deterministic approximation bias from statistical error, proves a joint-event finite-sample Q-certificate and target-occupancy decision bounds, retains observable action-relevant cues, recovers the supported Bellman quotient under separation, and pairs these upper conclusions with action-identification and estimator-specific structural lower bounds. In the one-stage zero-tolerance case, the upper and lower conditions meet at the scale $nq\Delta^2$.

A.2 PRELIMINARIES

A.2.1 CONFIDENCE EVENT AND ACCUMULATED ERROR

Fix a theorem start $t \in [H]$ and confidence $\delta \in (0,1)$. The two finite-union complexities appearing in the main theorem are

$$\Gamma_{\mathrm{dev}}(\delta/2) := \max_{s\in[H]} |\mathcal{U}_{s+1}| + \log\left(\frac{8H|\mathcal{A}||\Phi|\sum_{s=1}^{H}|\mathcal{U}_s|}{\delta}\right),$$

$$\Gamma_{\mathrm{cnt}}(\delta/2) := \log\left(\frac{4H|\mathcal{A}|\sum_{s=1}^{H}|\mathcal{U}_s|}{\delta}\right).$$

For $N > 0$, set

$$\mathrm{rad}_r(N) := \min\left\{1, c_r\sqrt{\frac{\Gamma_{\mathrm{dev}}(\delta/2)}{N}}\right\}, \qquad \mathrm{rad}_p(N) := \min\left\{1, c_p\sqrt{\frac{\Gamma_{\mathrm{dev}}(\delta/2)}{N}}\right\};$$

at $N = 0$, both radii are one. These radii define the learner's pairwise feasibility tests. Specifically, a candidate may merge supported u, v at stage s only if, for every action,

$$|\widehat{r}_s(u,a) - \widehat{r}_s(v,a)| \le \tau_s^r(u,v,a) + \mathrm{rad}_r(N_s(u,a)) + \mathrm{rad}_r(N_s(v,a))$$

and

$$\begin{aligned}&\mathrm{TV}\Big((\widehat{\phi}_{s+1})_{\#}\widehat{P}_s^{\chi}(\cdot \mid u,a), (\widehat{\phi}_{s+1})_{\#}\widehat{P}_s^{\chi}(\cdot \mid v,a)\Big)\\ &\qquad \le \tau_s^p(u,v,a) + \mathrm{rad}_p(N_s(u,a)) + \mathrm{rad}_p(N_s(v,a)).\end{aligned}$$

They also define the simultaneous fixed-master deviation event

$$\begin{aligned}\mathcal{G}_{\delta/2}^{\mathrm{dev}} := \bigcap_{\substack{s,\, u\in S_{\mu,s},\, a\\ N_s(u,a)>0}} \Big\{&|\widehat{r}_s(u,a) - r_s(u,a)| \le \mathrm{rad}_r(N_s(u,a)),\\ &\mathrm{TV}\Big(\widehat{P}_s^{\chi}(\cdot \mid u,a), P_s^{\chi}(\cdot \mid u,a)\Big) \le \mathrm{rad}_p(N_s(u,a))\Big\}.\end{aligned}$$

The rows consumed from start t are

$$\mathcal{I}_t := \{(s,u,a) : s \ge t,\ u \in S_{\mu,s},\ a \in \mathcal{A}\}.$$

The realized-count event and sole upper/recovery event are

$$\mathcal{C}_{\delta/2}^{\mathrm{cnt}} := \bigcap_{(s,u,a)\in\mathcal{I}_t} \left\{N_s(u,a) \ge \tfrac{1}{2}nd_s^{\mu}(u,a)\right\}, \qquad \mathcal{H}_{\delta} := \mathcal{G}_{\delta/2}^{\mathrm{dev}} \cap \mathcal{C}_{\delta/2}^{\mathrm{cnt}}.$$

The events are conclusions of concentration, not assumptions.

To expose approximation and sampling effects separately, fix a consumed supported row (u,a). Assumption 10 gives $d_s^{\mu}(u,a) > 0$ on this domain, and we define

$$\varepsilon_s^r(u,a) := \max_{v\in S_{\mu,s}} \tau_s^r(u,v,a), \qquad \varepsilon_s^p(u,a) := \max_{v\in S_{\mu,s}} \tau_s^p(u,v,a),$$

and, on a nonempty stratum, define

$$\varepsilon_{D,s}^r := \max_{\substack{u\in S_{D,s}\\ a\in\mathcal{A}}} \varepsilon_s^r(u,a), \qquad \varepsilon_{D,s}^p := \max_{\substack{u\in S_{D,s}\\ a\in\mathcal{A}}} \varepsilon_s^p(u,a).$$

Both stratum maxima are defined as zero when $S_{D,s} = \varnothing$. The stratum quantities expose the public bias rate, while the pointwise quantities remain available for rowwise transfer.

$$a_s(u,a) := c_b\big(\varepsilon_s^r(u,a) + B_{s+1}\varepsilon_s^p(u,a)\big),$$

$$s_s(u,a) := c_s(1+B_{s+1})\sqrt{\frac{\Gamma_{\mathrm{dev}}(\delta/2)}{n\, d_s^\mu(u,a)}}, \qquad \ell_s(u,a) := a_s(u,a) + s_s(u,a).$$

On a nonempty stratum,

$$a_{D,s} := \max_{\substack{u\in S_{D,s}\\ a\in\mathcal{A}}} a_s(u,a), \qquad s_{D,s} := \max_{\substack{u\in S_{D,s}\\ a\in\mathcal{A}}} s_s(u,a), \qquad e_{D,s} := a_{D,s} + s_{D,s};$$

all three are zero on an empty stratum. The accumulated quantities used below are

$$\mathfrak{A}_t := \sum_{s=t}^{H} \max_{D\in\{E,\bar{E}\}} a_{D,s}, \qquad \mathfrak{S}_t := \sum_{s=t}^{H} \max_{D\in\{E,\bar{E}\}} s_{D,s}, \qquad \mathfrak{B}_t := \mathfrak{A}_t + \mathfrak{S}_t,$$

and the bias-adjusted margin is

$$\Delta_{t,\mathrm{eff}} := \Delta_t - 2\mathfrak{A}_t.$$

These symbols display, rather than hide, the downstream deterministic and statistical charges entering the Q-bound, retention condition, and recovery condition. Their setting-derived envelopes are

$$a_{D,s} \le c_b\big(\varepsilon_{D,s}^r + B_{s+1}\varepsilon_{D,s}^p\big), \qquad s_{D,s} \le c_s(1+B_{s+1})\sqrt{\frac{\Gamma_{\mathrm{dev}}(\delta/2)}{n q_{D,s}^{\mathrm{audit}}}},$$

where the right-hand side is invoked only for a nonempty stratum.

A.2.2 Target-complete change of measure

The atom-level target shift is defined before summation. For $\pi \in \Pi_{\mathrm{tar}}$, set

$$\mathfrak{v}_{D,s}^\pi := \begin{cases} C_{D,s}^{\mathrm{atom}} \displaystyle\sum_{\substack{u\in S_{D,s},\, a\in\mathcal{A}\\ d_s^\mu(u,a)>0}} d_s^\mu(u,a)\ell_s(u,a), & S_{D,s}\neq\varnothing, \\ 0, & S_{D,s}=\varnothing \text{ and hence } d_s^\pi(\mathcal{U}_{D,s})=0, \end{cases}$$

and

$$\mathfrak{V}_t^\pi := \sum_{s=t}^{H}\sum_{D\in\{E,\bar{E}\}} \mathfrak{v}_{D,s}^\pi.$$

This is the target-complete piecewise definition used in the main theorem. It never evaluates ℓ_s off behavior support and never forms $+\infty\cdot 0$. It is an algebraic selector-loss transport quantity, not a row-local substitute for the selected representative-model error.

The learned deterministic policy is

$$\widehat{\pi}_s(a \mid u) := \mathbf{1}\{a = \widehat{g}_s(u)\}.$$

Call a stage-t start law ξ admissible when $\mathrm{supp}(\xi) \subseteq S_{\mu,t}$. The declared atom-level absolute continuity makes every target-policy start marginal admissible. For such a law, let $V_t^{\widehat{\pi}}(\xi)$ be its actual fresh-rollout cumulative reward under this policy and let $V_t^\star(\xi)$ be the corresponding optimal value. These quantities state the full-policy regret conclusion without introducing an auxiliary learned-model value.

A.2.3 Lower experiments

The universal lower family is a terminal binary-action experiment with a hard observable pre-action atom $u_\star$ of behavior mass $2q$ and uniform actions. The two reward-mean vectors at that atom are

$$\big(\tfrac{1}{2}+\tfrac{\Delta}{2}, \tfrac{1}{2}-\tfrac{\Delta}{2}\big) \quad \text{and} \quad \big(\tfrac{1}{2}-\tfrac{\Delta}{2}, \tfrac{1}{2}+\tfrac{\Delta}{2}\big).$$

All other laws are common, every supported atom–action mass is at least q, $q^{\mathrm{audit}}_{\bar{E},t} \geq q^{\mathrm{audit}}_{E,t} = q$, and two supported anchors have fixed opposite optimal actions. An arbitrary finite independent nuisance and a common zero-reward prefix may be appended without changing the information calculation. The prefix fixes

$$\eta^{\pi}_t := d^{\pi}_t(u_\star) \geq \tfrac{1}{4}\rho^{\pi}_t$$

before the training sample is observed. This family tests action identification and regret only. For any learner, $\widehat{a}(u_\star)$ denotes its possibly randomized reported action at the hard atom, and $a^\star_\theta(u_\star)$ denotes the optimal hard-atom action in paired member θ.

For the named minimum-K_μ estimator, the common-protocol trilemma uses supported terminal atoms A, B, C. Its two members have respective supported Bellman quotients

$$\{A\} \mid \{B, C\}, \qquad \{A, B\} \mid \{C\},$$

while sharing the master map, support, candidate family, tolerances, zero-count defaults, orders, labels, radii, and tie rules. For member $\theta \in \{0, 1\}$, define

$$\mathsf{MergeErr}_\theta := \{\exists u \not\sim^{\mathrm{B}}_{\theta,t} v : \widehat{\phi}_t(u) = \widehat{\phi}_t(v)\},$$

$$\mathsf{RefineErr}_\theta := \{\exists u \sim^{\mathrm{B}}_{\theta,t} v : \widehat{\phi}_t(u) \neq \widehat{\phi}_t(v)\}, \qquad \mathsf{ActionErr}_\theta := \{\widehat{g}_t(B) \neq a^\star_{\theta,t}(B)\}.$$

These definitions make the structural lower statement a relation-testing claim.

A.3 Main Theorem

Theorem 16 (Finite-sample guarantee for coverage-calibrated Bellman quotients)**.** *Under Assumptions 6–15, the universal numerical constants may be chosen simultaneously as*

$$c_r = c_p = c_b = 1, \qquad c_{\mathrm{cnt}} = 8, \qquad c_s = 4\sqrt{2},$$

$$c_{\mathrm{KL}} = c_{\mathrm{tri}} = \frac{16}{3}, \qquad c_0 = \frac{3}{64}, \qquad c_1 = c_2 = \frac{1}{8}, \qquad \kappa = \frac{1}{4}.$$

For every start $t \in [H]$ and $\delta \in (0, 1)$, the following conclusions hold with the conditions and probability spaces stated in their respective clauses.

(i) Well-definedness. *Every dataset, including one with all row counts equal to zero, produces a unique measurable supported partition with its canonical off-support extension, a representative map, a valid bounded latent model, arrays $0 \leq \widehat{Q}_s, \widehat{V}_s \leq B_s$, an ordered action, and the deterministic policy $\widehat{\pi}$. The supported identity supplies partition feasibility; the fixed defaults and orders supply numerical and decision totality.*

(ii) Joint-event upper bounds and target regret. *Suppose that, for every $s \geq t$ and every nonempty consumed stratum,*

$$n q^{\mathrm{audit}}_{D,s} \geq 8\Gamma_{\mathrm{cnt}}(\delta/2).$$

Then

$$\Pr(\mathcal{G}^{\mathrm{dev}}_{\delta/2}) \geq 1 - \frac{\delta}{2}, \qquad \Pr(\mathcal{C}^{\mathrm{cnt}}_{\delta/2}) \geq 1 - \frac{\delta}{2}, \qquad \Pr(\mathcal{H}_\delta) \geq 1 - \delta.$$

On the single event $\mathcal{H}_\delta$ for this fixed start t,

$$\sup_{\substack{u \in S_{\mu,t} \\ a \in \mathcal{A}}} \left| Q^\star_t(u, a) - \widehat{Q}_t(\widehat{\phi}_t(u), a) \right| \leq \mathfrak{A}_t + \mathfrak{S}_t = \mathfrak{B}_t,$$

where

$$\mathfrak{A}_t \leq \sum_{s=t}^{H} \max_{D \in \{E, \bar{E}\}} \left(\varepsilon^r_{D,s} + B_{s+1} \varepsilon^p_{D,s} \right),$$

$$\mathfrak{S}_t \leq 4\sqrt{2} \sum_{s=t}^{H} \max_{\substack{D \in \{E, \bar{E}\}: \\ S_{D,s} \neq \varnothing}} (1 + B_{s+1}) \sqrt{\frac{\Gamma_{\mathrm{dev}}(\delta/2)}{n q^{\mathrm{audit}}_{D,s}}}.$$

Independently of this event, Assumption 11 gives, for every $\pi \in \Pi_{\mathrm{tar}}$ and selector sequence g,

$$\sum_{s=t}^{H} \sum_{D \in \{E,\bar{E}\}} \mathbb{E}_{d_s^\pi}\left[\mathbf{1}\{U_s \in \mathcal{U}_{D,s}\}\ell_s(U_s, g_s(U_s))\right] \le \mathfrak{V}_t^\pi,$$

where $\mathfrak{V}_t^\pi$ is the finite piecewise sum defined above. Each expectation denotes its finite positive-target-mass sum; every queried row in that sum is behavior positive by Assumption 11, and an empty stratum uses the zero branch. On $\mathcal{H}_\delta$, for every $u \in S_{\mu,t}$,

$$Q_t^\star(u, a_t^\star(u)) - Q_t^\star(u, \widehat{g}_t(u)) \le \min\{B_t, 2\mathfrak{B}_t\}.$$

Consequently, for every target policy,

$$\frac{\mathcal{R}_{E,t}^\pi(\widehat{g}_t)}{\rho_t^\pi} \le \min\{B_t, 2\mathfrak{B}_t\} \quad \textit{when } \rho_t^\pi > 0,$$

and the unnormalized event contribution is zero when $\rho_t^\pi = 0$. For every admissible start law ξ, the actual fresh-rollout value satisfies

$$V_t^\star(\xi) - V_t^{\widehat{\pi}}(\xi) \le 2\sum_{s=t}^{H} \mathfrak{B}_s.$$

(iii) Bias-aware cue retention. *If*

$$\Delta_{t,\mathrm{eff}} > 0, \qquad 2\mathfrak{S}_t < \Delta_{t,\mathrm{eff}},$$

then, on $\mathcal{H}_\delta$, every supported active atom is acted on optimally, $\mathcal{R}_{E,t}^\pi(\widehat{g}_t) = 0$ for every $\pi \in \Pi_{\mathrm{tar}}$, and any two supported active atoms with different ordered optimal actions receive different learned codes. Equivalently, the surviving learned action margin is at least $\Delta_{t,\mathrm{eff}} - 2\mathfrak{S}_t > 0$.

(iv) Exact supported quotient recovery. *If the supported Bellman quotient belongs to the candidate class and*

$$4\mathfrak{B}_s < \zeta_s \qquad \textit{for every } s = t, \dots, H,$$

then, on $\mathcal{H}_\delta$, simultaneously for every $s \ge t$,

$$\widehat{\phi}_s(u) = \widehat{\phi}_s(v) \iff u \sim_s^{\mathrm{B}} v, \qquad u, v \in S_{\mu,s}.$$

Thus the learned partition equals the supported master Bellman quotient up to relabeling, merges every supported Bellman-null nuisance fiber, and preserves supported ordered-action differences. Off-support atoms retain only their fixed canonical singleton labels.

(v) Universal raw-gap action and regret obstruction. *For every learner and every $0 < q \le 1/6$, $0 < \Delta \le 1/4$, there is a bounded two-action pair with $q_{E,t}^{\mathrm{audit}} = q$, $q_{\bar{E},t}^{\mathrm{audit}} \ge q$, raw gap $\Delta_t = \Delta$, and*

$$D_{\mathrm{KL}}\big(\mathsf{P}_+^{n,\mu} \| \mathsf{P}_-^{n,\mu}\big) \le \frac{16}{3} nq\Delta^2.$$

For this pair,

$$nq\Delta^2 \le \frac{3}{64} \implies \sup_{\theta \in \{+,-\}} \Pr_\theta(\widehat{a}(u_\star) \ne a_\theta^\star(u_\star)) \ge \frac{1}{8}.$$

Under one member, the probability that $\mathcal{R}_{E,t}^\pi(\widehat{g}_t) \ge \eta_t^\pi \Delta$ is at least $1/8$, the expected hard-atom conditional regret is at least $\Delta/8$, and

$$\mathbb{E}_\theta[\mathcal{R}_{E,t}^\pi(\widehat{g}_t)] \ge \frac{1}{8}\eta_t^\pi \Delta \ge \frac{1}{32}\rho_t^\pi \Delta.$$

(vi) Estimator-specific common-protocol trilemma. *For the named minimum-K_μ estimator and every $0 < q \le 1/6$, $0 < \Delta \le 1/4$, there is a pair sharing every learner-visible protocol element, with opposite supported A, B quotient relations. In this common protocol, atom B*

has mass $2q$, the two binary behavior actions are uniform, and every supported atom–action row has mass at least q. All supported atoms are active, so

$$q_{E,t}^{\mathrm{audit}} = q, \qquad S_{\bar{E},t} = \varnothing, \qquad q_{\bar{E},t}^{\mathrm{audit}} = +\infty$$

under the empty non-event convention. Moreover,

$$D_{\mathrm{KL}}(\mathsf{T}_0^{n,\mu} \| \mathsf{T}_1^{n,\mu}) \le \frac{16}{3} nq\Delta^2.$$

Whenever $nq\Delta^2 \le 3/64$,

$$\max_{\theta \in \{0,1\}} \Pr_{\theta}(\mathsf{MergeErr}_\theta \cup \mathsf{RefineErr}_\theta \cup \mathsf{ActionErr}_\theta) \ge \frac{1}{4}.$$

(vii) One-stage positive-tolerance specialization. *When $H = 1$, if*

$$\Delta_{1,\mathrm{eff}} > 0, \qquad nq_1^{\mathrm{audit}} \ge 8\Gamma_{\mathrm{cnt}}(\delta/2),$$
$$nq_1^{\mathrm{audit}} \Delta_{1,\mathrm{eff}}^2 > 4c_s^2 \Gamma_{\mathrm{dev}}(\delta/2),$$

then all target-reachable supported active actions are correct with probability at least $1-\delta$. No correctness conclusion follows from the squared inequality alone when $\Delta_{1,\mathrm{eff}} \le 0$.

Rate, mode, and dependence declaration. *The exposed upper variables are*

$$H, t, n, \delta, |\mathcal{A}|, |\Phi|, \{|\mathcal{U}_s|\}_{s=1}^{H+1}, \{B_s\}_{s=t}^{H+1}, \Gamma_{\mathrm{dev}}, \Gamma_{\mathrm{cnt}},$$
$$\{\tau_s^r, \tau_s^p, \varepsilon_{D,s}^r, \varepsilon_{D,s}^p, q_{D,s}^{\mathrm{audit}}, \rho_s^\mu, \lambda_s, \alpha_s^{\mathrm{audit}}, \beta_s, d_s^\pi, \rho_s^\pi, C_{D,s}^{\mathrm{atom}}, a_{D,s}, s_{D,s}\}_{s=t}^{H},$$
$$\{K_{\mu,s}, \Delta_s, \Delta_{s,\mathrm{eff}}, \zeta_s\}_{s=t}^{H}, \{\mathfrak{A}_s, \mathfrak{S}_s, \mathfrak{B}_s\}_{s=t}^{H}, \mathfrak{V}_t^\pi,$$

and the lower side additionally exposes $q, \Delta, \eta_t^\pi, \rho_t^\pi$. The displayed universal constants depend on none of these quantities; no other hidden constant is used. As n varies, the POMDP, master audit, support masks, candidate class, raw tolerances, defaults, total orders, target class, and finite horizon remain fixed. Upper conclusions are simultaneous high-probability statements on $\mathcal{H}_\delta$; lower conclusions use their own binary-testing probabilities and expectations. The horizon mode is fixed finite. The Q-norm is the supported atom–action supremum, policy regret uses target occupancy and actual fresh rollouts, recovery is equality of supported equivalence relations up to relabeling, and total variation is half the ℓ_1 distance.

Boundary cases. *At zero bias and statistical charge the supported Q-values are exact, and with quotient membership the recovery, nuisance-merging, and action-separation statements hold with no remainder. Recovery concerns the behavior support; off-support atoms keep their fixed singleton codes. Regret refers to the learned policy's true rollout, and $\ell_s, \mathfrak{V}_t^\pi$ enter only the change-of-measure display. Retention is stated for $q_{E,t}^{\mathrm{audit}} > 0$ and $\Delta_{t,\mathrm{eff}} > 0$, and the lower regret vanishes as $\Delta \to 0$. The positive-tolerance upper condition uses the effective gap $\Delta_{1,\mathrm{eff}}$, whereas the lower scale uses the raw gap.*

Corollary 17 (One-stage zero-tolerance decision scale)**.** *Under the assumptions of Theorem 16, suppose $H = 1$, every supported stage-one raw reward and next-code tolerance is zero, and*

$$q_{\bar{E},1}^{\mathrm{audit}} \ge q_{E,1}^{\mathrm{audit}} =: q < +\infty.$$

Then

$$\mathfrak{A}_1 = 0, \qquad \Delta_{1,\mathrm{eff}} = \Delta_1, \qquad q_1^{\mathrm{audit}} = q,$$

and, on $\mathcal{H}_\delta$,

$$\sup_{\substack{u \in S_{\mu,1} \\ a \in \mathcal{A}}} \left| Q_1^\star(u,a) - \widehat{Q}_1(\widehat{\phi}_1(u), a) \right| \le 4\sqrt{2} \sqrt{\frac{\Gamma_{\mathrm{dev}}(\delta/2)}{nq}}.$$

Consequently, the specified total learner succeeds simultaneously on all target-reachable supported active atoms with probability at least $1-\delta$ whenever

$$nq \ge 8\Gamma_{\mathrm{cnt}}(\delta/2), \qquad nq\Delta_1^2 > 4c_s^2 \Gamma_{\mathrm{dev}}(\delta/2).$$

On the common range $0 < q \le 1/6$ and $0 < \Delta = \Delta_1 \le 1/4$, every learner also faces a member of the zero-tolerance one-stage lower family for which

$$nq\Delta^2 \le \frac{3}{64} \quad \Longrightarrow \quad \Pr(\textit{hard-action error}) \ge \frac{1}{8},$$

with expected hard-atom conditional regret at least $\Delta/8$ and expected target-occupancy regret at least $\rho_1^\pi \Delta/32$. Thus the upper decision term and lower obstruction use the same bare scalar $nq\Delta^2$, up to the displayed finite-class/confidence logarithm and the separate upper count prerequisite. The two implications use separate experiments and probability spaces. There are no hidden constants; the fixed quantities, one-stage horizon mode, supported Q-supremum norm, and testing/expectation lower modes are those declared in Theorem 16.

Proof. Proposition 63 verifies $B_2 = 0$, the count condition, the strict two-error action margin, the conversion from $\mathcal{H}_\delta$ to unconditional success probability, the identities $\mathfrak{A}_1 = 0$, $\Delta_{1,\mathrm{eff}} = \Delta_1$, and $q_1^{\mathrm{audit}} = q$, and the model-common zero-tolerance completion of the lower pair. It then compares the two displayed implications while preserving their separate probability spaces and modes. □

A.4 Proof Sketch

The proof has three branches: a joint-event upper bound, a structural recovery argument, and two binary-testing lower bounds.

First, Proposition 21 totalizes every zero-count and tie branch before any probabilistic argument. Reward concentration and finite-alphabet total-variation concentration give $\mathcal{G}_{\delta/2}^{\mathrm{dev}}$, while an independent count calculation gives $\mathcal{C}_{\delta/2}^{\mathrm{cnt}}$. Their intersection is the only event used for empirical upper and recovery conclusions; Proposition 28 supplies its probability without treating a realized count inequality as an assumption.

The structural induction begins with support closure in Proposition 30. On the joint event, feasibility of the selected pair and the two endpoint deviations control the selected representative row. The larger endpoint radius is retained, which is essential when a block representative is rarer than the queried atom. This yields the one-stage envelope of Proposition 35. Comparing the true and representative Bellman recursions gives

$$D_s \le \max_{D \in \{E, \bar{E}\}} a_{D,s} + \max_{D \in \{E, \bar{E}\}} s_{D,s} + D_{s+1}, \qquad D_{H+1} = 0.$$

Unrolling this recurrence separately for deterministic and statistical terms proves Proposition 38 and the $\mathfrak{A}_t + \mathfrak{S}_t$ supported Q-bound.

Target shift is handled atom by atom. A target-relevant behavior-empty stratum is excluded by full counterfactual support and contributes zero before any loss is evaluated; otherwise the target atom and queried action are dominated by the corresponding behavior atom–action row. This proves the piecewise transfer in Proposition 39. The total learned policy is then evaluated in the true environment: master homogeneity identifies its finite master recursion with its fresh-rollout return in Proposition 40. Each optimal advantage is at most $2\mathfrak{B}_s$, and the finite performance-difference telescope gives Proposition 42.

The raw action gap loses two Q-error terms. Hence the remaining margin is

$$\Delta_t - 2\mathfrak{B}_t = \Delta_{t,\mathrm{eff}} - 2\mathfrak{S}_t.$$

When it is positive, the learned action is correct and two atoms requiring different actions cannot share a learned code; these are the conclusions of Proposition 43 and Proposition 45. No sample size is used to remove the fixed bias $\mathfrak{A}_t$.

For exact supported recovery, the task-factorized kernel first makes every task-equal nuisance fiber a population Bellman-quotient block. The proof then runs backward from the terminal singleton. Once the downstream learned and population quotient codes agree up to relabeling, the population quotient is a feasible current candidate. A cross-block merge would force $d_s^{\mathrm{B}}(u, v) \le 4\mathfrak{B}_s < \zeta_s$, which is impossible. Minimum supported cardinality then rules out within-block refinement. This proves the supported equivalence-relation equality through quotient-specific compatibility.

For the universal lower bound, only the hard atom's two Bernoulli reward rows differ. Direct calculation gives

$$D_{\mathrm{KL}}(\mathsf{P}_{+}^{n,\mu}\|\mathsf{P}_{-}^{n,\mu}) = 2nq\Delta\log\frac{1+\Delta}{1-\Delta} \le \frac{16}{3}nq\Delta^2.$$

A binary overlap inequality then forces constant hard-action error, and the model-common occupancy converts it to expected event regret. The estimator-specific construction is restricted to $0 < q \le 1/6$, $0 < \Delta \le 1/4$. Atom B has mass $2q$, binary behavior actions are uniform, and every supported atom–action row has mass at least q; since all supported atoms are active, this gives $q_{E,t}^{\mathrm{audit}} = q$ and, by the empty non-event convention, $q_{\bar{E},t}^{\mathrm{audit}} = +\infty$. The construction tests whether A and B share a learned block. Because their true relation is opposite in the two common-protocol members, a testing error is a merge error in one member or a refinement error in the other, giving the trilemma of Proposition 58.

Finally, Proposition 63 sets $H = 1$, checks the count prerequisite and strict two-error margin, converts the joint event to unconditional upper success probability, and then sets all raw tolerances to zero. Only after

$$\mathfrak{A}_1 = 0, \qquad \Delta_{1,\mathrm{eff}} = \Delta_1, \qquad q_1^{\mathrm{audit}} = q_{E,1}^{\mathrm{audit}}$$

does it compare the upper and lower $nq\Delta^2$ scales. Their probability spaces remain separate, and the positive-tolerance upper continues to use the effective rather than raw gap.

A.5 Proof Details

A.5.1 Well-Definedness of the Empirical Latent Planner

Lemma 18 (Total empirical-or-default rows)**.** *Under Assumption 14 and the bounded finite-master setup, for every data set, stage $t \in [H]$, master atom $u \in \mathcal{U}_t$, and action $a \in \mathcal{A}$, the row*

$$\big(\widehat{r}_t(u,a), \widehat{P}_t^{\chi}(\cdot \mid u,a)\big)$$

is defined and satisfies

$$0 \le \widehat{r}_t(u,a) \le 1, \qquad \widehat{P}_t^{\chi}(\cdot \mid u,a) \in \Delta(\mathcal{U}_{t+1}).$$

Its reward and transition radii lie in $[0,1]$. If $\psi : \mathcal{U}_{t+1} \to \mathcal{Z}$ is any total map into a finite set, then $\psi_{\#}\widehat{P}_t^{\chi}(\cdot \mid u,a)$ is a probability law on $\mathcal{Z}$. The same conclusions hold for every canonical off-support default row.

Proof. Fix an arbitrary data set and abbreviate

$$I_i^{t,u,a} := \mathbf{1}\{U_t^{(i)} = u, A_t^{(i)} = a\}, \qquad N := \sum_{i=1}^{n} I_i^{t,u,a}.$$

When $N > 0$, the empirical formulas are

$$\widehat{r}_t(u,a) = \frac{1}{N}\sum_{i=1}^{n} I_i^{t,u,a} R_t^{(i)}$$

and, for $w \in \mathcal{U}_{t+1}$,

$$\widehat{P}_t^{\chi}(w \mid u,a) = \frac{1}{N}\sum_{i=1}^{n} I_i^{t,u,a}\mathbf{1}\{U_{t+1}^{(i)} = w\}.$$

The reward is an average of numbers in $[0,1]$. The transition coordinates are nonnegative and sum to one, including at $t = H$, where the next alphabet is the terminal singleton.

When $N = 0$, Assumption 14 gives

$$\widehat{r}_t(u,a) = r_t^0(u,a) = 0, \qquad \widehat{P}_t^{\chi}(\cdot \mid u,a) = P_t^0(\cdot \mid u,a) = \delta_{u_{t+1}^0},$$

together with $\mathrm{rad}_r(0) = \mathrm{rad}_p(0) = 1$. At positive counts, the clipped definitions also put both radii in $[0,1]$. Consequently, every reward difference, total-variation distance, and right-hand side in a feasibility comparison is a finite real number, even when one or both endpoint counts vanish.

For a total finite code ψ, define

$$(\psi_{\#}\widehat{P}_t^{\chi})(z \mid u, a) := \sum_{w:\, \psi(w)=z} \widehat{P}_t^{\chi}(w \mid u, a).$$

These coordinates are nonnegative and their sum over $z \in \mathcal{Z}$ is one. A canonical off-support block uses the atom itself with its fixed default row, so the zero-count argument applies verbatim. This covers positive, zero, and mixed endpoint counts without restricting to any event. □

Lemma 19 (Unique finite candidate and representative selection)**.** *Under Assumptions 12 and 14, fix $t \in [H]$, an arbitrary data set, and an already-defined total downstream code $\psi : \mathcal{U}_{t+1} \to \mathcal{Z}_{t+1}$. Then the set of candidates satisfying the reward and transition feasibility inequalities with downstream code ψ is finite and nonempty. The rule that first minimizes $K_{\mu,t}$ and then uses $\prec_{\Phi,t}$ selects exactly one candidate. Each supported block has exactly one prescribed representative, while every off-support atom has exactly one candidate-independent canonical singleton label and representative.*

Proof. Lemma 18 makes every empirical-or-default reward row and every pushforward through ψ well-defined. Thus each feasibility comparison has a definite truth value. Let $\mathcal{F}_t(\psi) \subseteq \mathcal{G}_t$ be the candidates passing all comparisons for every supported pair they merge and every action. The family $\mathcal{G}_t$ is fixed and finite.

The supported identity assigns distinct labels to distinct supported atoms, so it merges no distinct supported pair. If self-pairs are listed, both left-hand sides are zero and the inequalities hold because tolerances and radii are nonnegative. Hence

$$g_t^{\mathrm{id}} \in \mathcal{F}_t(\psi), \qquad \mathcal{F}_t(\psi) \neq \varnothing.$$

The identity is used only for feasibility; it supplies neither a numerical default row nor an action.

Because $\mathcal{F}_t(\psi)$ is finite and nonempty, the integer $K_{\mu,t}$ attains a minimum. The fixed total order $\prec_{\Phi,t}$ has a unique first element among the minimizers, so the rule selects one exact map $\widehat{\phi}_t$, rather than merely an unlabeled partition class. For a supported selected label z, the fiber

$$\{u \in S_{\mu,t} : \widehat{\phi}_t(u) = z\}$$

is a nonempty finite set, and its first atom in the fixed order on $\mathcal{U}_t$ is the unique representative $\widehat{v}_t(z)$. If $u \notin S_{\mu,t}$, every candidate uses the same injective label $\perp_{t,u}$, disjoint from the supported labels, and the protocol makes $\{u\}$ its block with representative u and the fixed default row. Thus all labels and representatives are total and unambiguous, including singleton fibers and complete candidate-rate ties. □

Proposition 20 (Bounded backward learner construction)**.** *Under Assumptions 12 and 14, and under Lemmas 18 and 19, the backward construction produces, for every $t \in [H+1]$, unique total codes $\widehat{\phi}_s$, representative maps $\widehat{v}_s$, reward and transition model rows, action values $\widehat{Q}_s$, values $\widehat{V}_s$, ordered greedy selectors $\widehat{g}_s$, and deterministic master policies*

$$\widehat{\pi}_s(\cdot \mid u) := \delta_{\widehat{g}_s(u)}, \qquad u \in \mathcal{U}_s,$$

for every $s = t, \ldots, H$. The model rewards lie in $[0,1]$, the transition rows are probability laws, and

$$0 \le \widehat{Q}_s(z,a) \le B_s, \qquad 0 \le \widehat{V}_s(z) \le B_s$$

for every selected latent label z and action a. At $t = H+1$, the conclusion is the prescribed singleton code and $\widehat{V}_{H+1} = 0$.

Proof. At $H+1$, the code has the prescribed singleton label, $B_{H+1} = 0$, and $\widehat{V}_{H+1} = 0$. Suppose the conclusion has been established from stage $t+1$ onward. In particular, $\widehat{\phi}_{t+1}$ is total and $0 \le \widehat{V}_{t+1} \le B_{t+1}$. Applying Lemma 19 with $\psi = \widehat{\phi}_{t+1}$ gives a unique $\widehat{\phi}_t$, one representative $\widehat{v}_t(z)$ for each supported label, and the prescribed representative for each canonical off-support label.

For a supported label z, form the model row from the empirical-or-default row of $\widehat{v}_t(z)$; for an off-support label use its fixed default row. Push the corresponding next-master law through $\widehat{\phi}_{t+1}$, and denote the resulting rows by

$$\widehat{r}_t^{\,\mathrm{mod}}(z,a), \qquad \widehat{P}_t^{\,\mathrm{mod}}(\cdot \mid z, a).$$

Lemma 18 yields

$$0 \le \widehat{r}_t^{\mathrm{mod}}(z,a) \le 1, \qquad \widehat{P}_t^{\mathrm{mod}}(\cdot \mid z,a) \text{ is a probability law.}$$

Therefore

$$\widehat{Q}_t(z,a) := \widehat{r}_t^{\mathrm{mod}}(z,a) + \sum_{z'} \widehat{P}_t^{\mathrm{mod}}(z' \mid z,a)\widehat{V}_{t+1}(z')$$

is a finite sum and satisfies

$$0 \le \widehat{Q}_t(z,a) \le 1 + B_{t+1} = B_t.$$

The finite nonempty action set gives

$$\widehat{V}_t(z) := \max_{a \in \mathcal{A}} \widehat{Q}_t(z,a) \in [0, B_t].$$

The first maximizer in $\prec_{\mathcal{A}}$ is unique, so

$$\widehat{g}_t(u) := \min_{\prec_{\mathcal{A}}} \arg\max_{a \in \mathcal{A}} \widehat{Q}_t(\widehat{\phi}_t(u), a)$$

is defined on every supported and off-support atom, and $\widehat{\pi}_t(\cdot \mid u) = \delta_{\widehat{g}_t(u)}$ is a probability law. Equivalently, after the training data are fixed,

$$\widehat{\pi}_t(\cdot \mid h_t) := \delta_{\widehat{g}_t(\chi_t(h_t))}$$

is a total deterministic master-measurable policy. This constructs the policy; its equality with a fresh-rollout master policy-evaluation recursion is proved later from master homogeneity.

The induction also resolves the boundary cases. At $t = H$, the downstream code is the terminal singleton and the continuation value is zero, so $\widehat{Q}_H = \widehat{r}_H^{\mathrm{mod}}$. If all counts are zero, all model rewards and hence all learned values are zero; the fixed orders still resolve every candidate and action tie. When $H = 1$, one has $t = H = 1$, so this reward-only update is the entire backward construction. Mixed counts, including a zero-count representative paired with a positive-count atom, remain total by Lemma 18. A canonical off-support atom always uses its singleton label, itself as representative, and its default row; this is a totality statement, not an off-support recovery claim. Finite backward induction completes the proof. □

Proposition 21 (Measurability of the deterministic learner)**.** *Under Assumptions 12 and 14, the tuple and induced point-mass policy in Proposition 20 are deterministic measurable functions of the data set, with every protocol parameter fixed before sampling.*

Proof. Give the finite observation, action, and master coordinates their discrete sigma-fields and the reward coordinates their Borel sigma-fields. The n-episode data space carries the corresponding finite-product sigma-field. For fixed t, u, a, the count $N_t(u,a)$ is a finite sum of indicators and is measurable. On $\{N_t(u,a) > 0\}$, every empirical reward and transition coordinate is a ratio of finite measurable sums by the positive integer $N_t(u,a)$; on $\{N_t(u,a) = 0\}$, it is constant. The empirical-or-default rows and clipped radii are therefore measurable piecewise functions.

The terminal code and value are constant. Suppose inductively that $\widehat{\phi}_{t+1}$ and $\widehat{V}_{t+1}$ are measurable. The union of labels from the fixed finite candidate family and canonical extension is finite. Extend $\widehat{V}_{t+1}$ by zero on ambient labels outside the selected range; every pushforward assigns those labels zero mass, so this bookkeeping extension leaves the Bellman update unchanged. For a fixed candidate, each coordinate of a pushed law is a finite sum of terms

$$\widehat{P}_t^{\chi}(w \mid u,a)\mathbf{1}\{\widehat{\phi}_{t+1}(w) = z\},$$

and is measurable. Total variation is one half of a finite sum of absolute coordinate differences, so every feasibility inequality defines a measurable set.

Order the candidates by the fixed lexicographic pair consisting of $K_{\mu,t}$ and their rank under $\prec_{\Phi,t}$. If candidate j has rank $r(j)$ in this composite order, its selection event is

$$\{j \text{ is feasible}\} \cap \bigcap_{k:\, r(k) < r(j)} \{k \text{ is not feasible}\}.$$

These events are measurable and, by Lemma 19, form a measurable partition. Representatives and canonical labels are fixed finite lookups once the candidate is known. Each selected model coordinate is a finite piecewise selection followed by a finite pushforward; each action value is a finite sum of products of measurable functions; and each value is a finite maximum.

If $a_1 \prec_{\mathcal{A}} \cdots \prec_{\mathcal{A}} a_m$, the event that a_j is the first maximizer at z is

$$\bigcap_{k<j} \{\widehat{Q}_t(z, a_j) > \widehat{Q}_t(z, a_k)\} \cap \bigcap_{k>j} \{\widehat{Q}_t(z, a_j) \geq \widehat{Q}_t(z, a_k)\},$$

a measurable finite intersection that includes exact ties. Finally,

$$\widehat{\pi}_t(a \mid u) = \mathbf{1}\{a = \widehat{g}_t(u)\}$$

is measurable, as is its composition with the fixed master map. Finite backward induction proves measurability of every tuple and policy component. Every apparent choice is an ordered singleton selection, so the data map is deterministic and defined everywhere. □

Proposition 22 (Total bounded empirical latent planner)**.** *Under Assumptions 12 and 14, every data set, including an all-zero-count data set, produces a unique measurable partition, representative map, valid latent model, bounded value and action-value arrays, ordered greedy action at every master atom, and deterministic point-mass policy. In particular,*

$$0 \leq \widehat{Q}_t, \widehat{V}_t \leq B_t \qquad (t \in [H]),$$

and the conclusion holds for mixed counts, all candidate and action ties, singleton blocks, the terminal singleton, and canonical off-support atoms, without conditioning on a statistical event.

Proof. Lemma 18 closes every positive-count, zero-count, mixed-count, terminal, supported, and off-support numerical row. Lemma 19 uses the supported identity only to make the finite feasible set nonempty and then uses the fixed orders to choose one candidate and one representative per block. Proposition 20 composes those facts from the terminal singleton through all H stages and proves the displayed bounds and total policy. Proposition 21 proves deterministic measurability, including equality boundaries. These conclusions are all unconditional; the separate true-rollout semantics of the learned policy require master homogeneity and are not asserted here. □

A.5.2 Fixed-Master Deviation Control

Lemma 23 (Random-count concentration for fixed master rewards)**.** *Under Assumptions 6 and 9, fix $j = (s, u, a)$ with $u \in S_{\mu,s}$, and set*

$$I_i^j := \mathbf{1}\{U_s^{(i)} = u, A_s^{(i)} = a\}, \quad \mathscr{I}_j := \{i \in [n] : I_i^j = 1\}, \quad N_j := |\mathscr{I}_j|.$$

For either choice of selected row mark

$$X_i = R_s^{(i)}, \quad \mu = r_s(u, a), \qquad \text{or} \qquad X_i = \mathbf{1}\{U_{s+1}^{(i)} \in A\}, \quad \mu = P_s^{\chi}(A \mid u, a),$$

where $A \subseteq \mathcal{U}_{s+1}$ is fixed, every deterministic $I \subseteq [n]$ with $|I| = m > 0$ and $\Pr(\mathscr{I}_j = I) > 0$ satisfies, for $x > 0$,

$$\Pr\left(\left|\frac{1}{m}\sum_{i\in I} X_i - \mu\right| > x \,\middle|\, \mathscr{I}_j = I\right) \leq 2e^{-2mx^2}. \tag{9}$$

Let

$$K := \max_{r\in[H]} |\mathcal{U}_{r+1}|, \quad S := \sum_{r=1}^{H} |\mathcal{U}_r|, \quad J_{\text{row}} := |\mathcal{A}| \sum_{r=1}^{H} |S_{\mu,r}| \leq |\mathcal{A}|S,$$

$$L := \frac{8H|\mathcal{A}||\Phi|S}{\delta}, \qquad \Gamma := K + \log L = \Gamma_{\text{dev}}(\delta/2).$$

With the universal choice $c_r = 1$, the simultaneous positive-count reward failure probability is bounded by

$$\Pr\left(\begin{array}{c} \exists\, r, u \in S_{\mu,r}, a :\ N_r(u,a) > 0, \\ |\widehat{r}_r(u,a) - r_r(u,a)| > \text{rad}_r(N_r(u,a)) \end{array} \right) \leq 2J_{\text{row}} e^{-2\Gamma} \leq \frac{\delta}{32}. \tag{10}$$

Proof. The event $\{\mathscr{I}_j = I\}$ factors episode by episode: episodes in I satisfy the fixed row-membership event and episodes outside I do not. Because the complete episodes are iid, conditioning on this product event preserves independence among the selected episode records. Master homogeneity gives their common conditional reward mean $r_s(u,a)$ and common next-master law $P_s^\chi(\cdot \mid u,a)$, even after histories inside the master atom are mixed. The subset indicators therefore have the displayed mean $P_s^\chi(A \mid u,a)$.

It remains to derive the bounded-mark tail. If $X \in [0,1]$ has mean μ, convexity of $x \mapsto e^{\lambda x}$ on $[0,1]$ gives

$$e^{\lambda x} \le (1-x) + xe^{\lambda}$$

and hence

$$\mathbb{E}e^{\lambda(X-\mu)} \le \exp\bigl(\log(1-\mu+\mu e^{\lambda}) - \mu\lambda\bigr).$$

For

$$h_\mu(\lambda) := \log(1-\mu+\mu e^{\lambda}) - \mu\lambda, \qquad p_\lambda := \frac{\mu e^{\lambda}}{1-\mu+\mu e^{\lambda}},$$

one has $h_\mu(0) = h'_\mu(0) = 0$ and

$$h''_\mu(\lambda) = p_\lambda(1-p_\lambda) \le \frac{1}{4}.$$

Taylor's formula on the segment joining 0 and λ, for either sign of λ, yields

$$h_\mu(\lambda) = \lambda^2 \int_0^1 (1-v) h''_\mu(v\lambda)\, dv \le \frac{\lambda^2}{8}. \tag{11}$$

For the m conditionally independent selected marks,

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i\in I}(X_i-\mu)\right) \,\middle|\, \mathscr{I}_j = I\right] \le e^{m\lambda^2/8}.$$

Markov's inequality with $\lambda = 4x$ gives the upper-tail bound e^{-2mx^2}, and the same calculation with $\lambda = -4x$ gives the lower-tail bound. Their union proves equation 9.

The estimate is uniform over all membership sets I of size m, so mixing over these sets gives the same bound conditional on $N_j = m$. If $\mathrm{rad}_r(m) = 1$, boundedness makes the strict failure event empty. Otherwise, $\mathrm{rad}_r(m) = \sqrt{\Gamma/m} < 1$, and equation 9 gives conditional failure probability at most $2e^{-2\Gamma}$. Total probability therefore gives, without a factor of n,

$$\begin{aligned}
&\Pr(N_j > 0,\ |\widehat{r}_s(u,a) - r_s(u,a)| > \mathrm{rad}_r(N_j)) \\
&\quad = \sum_{m=1}^{n} \Pr(N_j = m) \Pr(|\widehat{r}_s(u,a) - r_s(u,a)| > \mathrm{rad}_r(m) \mid N_j = m) \\
&\quad \le 2e^{-2\Gamma}.
\end{aligned} \tag{12}$$

If $n = 0$, the sum is empty. A union bound over the J_{row} supported rows proves the first inequality in equation 10. Finally, using $\Gamma = K + \log L$, $J_{\mathrm{row}} \le |\mathcal{A}|S$, and $\delta^2 \le \delta$,

$$\begin{aligned}
2J_{\mathrm{row}}e^{-2\Gamma} &= 2J_{\mathrm{row}}e^{-2K}\left(\frac{\delta}{8H|\mathcal{A}||\Phi|S}\right)^2 \\
&\le \frac{\delta^2 e^{-2K}}{32H^2|\mathcal{A}||\Phi|^2 S} \le \frac{\delta}{32}.
\end{aligned}$$

□

Proposition 24 (Finite-subset concentration for empirical total variation)**.** *Under Assumptions 6 and 9, fix a supported row $j = (s,u,a)$, condition on $N_j = m > 0$, and write $k := |\mathcal{U}_{s+1}|$. Then, for every $x > 0$,*

$$\Pr\Bigl(\mathrm{TV}\Bigl(\widehat{P}_s^\chi(\cdot \mid u,a), P_s^\chi(\cdot \mid u,a)\Bigr) > x \,\Big|\, N_j = m\Bigr) \le (2^k - 2)e^{-2mx^2}. \tag{13}$$

For $k = 1$, the deviation is identically zero. With $K, S, J_{\mathrm{row}}, \Gamma$ as in Lemma 23 and the universal choice $c_p = 1$,

$$\Pr\left(\begin{array}{c} \exists\, s,u \in S_{\mu,s}, a:\ N_s(u,a) > 0, \\ \mathrm{TV}(\widehat{P}_s^\chi(\cdot \mid u,a), P_s^\chi(\cdot \mid u,a)) > \mathrm{rad}_p(N_s(u,a)) \end{array}\right) \le J_{\mathrm{row}}2^K e^{-2\Gamma} \le \frac{\delta}{64}. \tag{14}$$

Proof. For probability laws $\widehat{P}, P$ on a finite alphabet $\mathcal{X}$, put $q(x) := \widehat{P}(x) - P(x)$. Since $\sum_x q(x) = 0$, the positive set $A_+ := \{x : q(x) \geq 0\}$ satisfies

$$\frac{1}{2}\sum_{x\in\mathcal{X}} |q(x)| = \sum_{x\in A_+} q(x).$$

For every $A \subseteq \mathcal{X}$, both $q(A)$ and $-q(A) = q(A^c)$ are at most this positive mass. Hence

$$\mathrm{TV}(\widehat{P}, P) = \max_{A\subseteq\mathcal{X}} |\widehat{P}(A) - P(A)|. \tag{15}$$

Complementary subsets have the same absolute discrepancy. The 2^k subsets form 2^{k-1} complementary pairs, and $\{\varnothing, \mathcal{X}\}$ has zero discrepancy. It suffices to union over $2^{k-1} - 1$ fixed nontrivial representatives. For each one, apply Lemma 23 to

$$X_i = \mathbf{1}\{U_{s+1}^{(i)} \in A\}, \qquad \mu = P_s^{\chi}(A \mid u, a).$$

The factor two for the two tail signs and the number of representatives give

$$2(2^{k-1} - 1)e^{-2mx^2} = (2^k - 2)e^{-2mx^2},$$

which proves equation 13, including the singleton case.

If $\mathrm{rad}_p(m) = 1$, total variation is at most one and the strict failure event is empty. Otherwise, $\mathrm{rad}_p(m) = \sqrt{\Gamma/m}$, so equation 13 gives conditional failure probability at most $(2^k - 2)e^{-2\Gamma}$. Mixing over the random count exactly as in equation 12, unioning over rows, and using $k \leq K$ gives

$$\Pr(F_p) \leq \sum_{s=1}^{H} \sum_{u\in S_{\mu,s}} \sum_{a\in\mathcal{A}} (2^{|\mathcal{U}_{s+1}|} - 2)e^{-2\Gamma} \leq J_{\mathrm{row}} 2^K e^{-2\Gamma},$$

where F_p is the transition failure event in equation 14. Finally,

$$\begin{aligned} J_{\mathrm{row}} 2^K e^{-2\Gamma} &= J_{\mathrm{row}} \left(\frac{2}{e^2}\right)^K \left(\frac{\delta}{8H|\mathcal{A}||\Phi|S}\right)^2 \\ &\leq \frac{\delta^2}{64H^2|\mathcal{A}||\Phi|^2 S} \leq \frac{\delta}{64}, \end{aligned}$$

because $(2/e^2)^K \leq 1$. □

Proposition 25 (Simultaneous fixed-master deviation event)**.** *Under Assumptions 6 and 9, let $\delta \in (0, 1)$, use the setting-defined $\Gamma_{\mathrm{dev}}(\delta/2)$, and choose the universal constants $c_r = c_p = 1$ in the clipped positive-count radii. Then*

$$\Pr(\mathcal{G}_{\delta/2}^{\mathrm{dev}}) \geq 1 - \frac{3\delta}{64} \geq 1 - \frac{\delta}{2}.$$

The assertion is unconditional and simultaneous over all fixed supported master atom-action rows, and it imposes no realized-count or population-mass lower bound.

Proof. Let F_r and F_p denote, respectively, the reward and transition failure events in equation 10 and equation 14. By the definition of $\mathcal{G}_{\delta/2}^{\mathrm{dev}}$, rows with zero count contribute no conjunct, and every positive-count supported row contributes exactly the two inequalities above. Thus

$$(\mathcal{G}_{\delta/2}^{\mathrm{dev}})^c = F_r \cup F_p.$$

Lemma 23 and Proposition 24, followed by a union bound, yield

$$\Pr\Big((\mathcal{G}_{\delta/2}^{\mathrm{dev}})^c\Big) \leq \frac{\delta}{32} + \frac{\delta}{64} = \frac{3\delta}{64} \leq \frac{\delta}{2}.$$

No independence between the two failure families is used. A singleton next alphabet makes the transition term zero; a zero count makes the corresponding event conjunct vacuous; and clipping at one makes strict low-positive-count failure impossible. The result therefore supplies only the stated deviation event and no count conclusion. □

A.5.3 Simultaneous Count Activation and the Joint Event

Lemma 26 (Multiplicative lower tail for one consumed row)**.** *Under Assumption 9, fix a deterministic stage–atom–action row $j = (s, u, a)$, and set*

$$Z_i^j := \mathbf{1}\{U_s^{(i)} = u, A_s^{(i)} = a\}, \qquad p_j := d_s^{\mu}(u, a), \qquad N_j := \sum_{i=1}^{n} Z_i^j.$$

Then, for every $p_j \in [0, 1]$,

$$\Pr\Big(N_j < \frac{np_j}{2}\Big) \le \exp\Big(-\frac{np_j}{8}\Big). \tag{16}$$

Proof. The variables $Z_1^j, \dots, Z_n^j$ are iid Bernoulli(p_j) by Assumption 9. If $np_j = 0$, the event in equation 16 is $\{N_j < 0\}$ and hence empty. Assume $np_j > 0$. For $\theta > 0$, monotonicity of $x \mapsto e^{-\theta x}$, followed by Markov's inequality, gives

$$\Pr(Y \ge b) \le \frac{\mathbb{E}Y}{b} \qquad (Y \ge 0,\ b > 0),$$

because $Y \ge b\,\mathbf{1}\{Y \ge b\}$. Applying this with $Y = e^{-\theta N_j}$ and $b = e^{-\theta np_j/2}$ yields

$$\begin{aligned} \Pr\Big(N_j < \frac{np_j}{2}\Big) &\le e^{\theta np_j/2}\,\mathbb{E}e^{-\theta N_j} \\ &= e^{\theta np_j/2}\left(1 - p_j + p_j e^{-\theta}\right)^n \\ &\le \exp\left\{np_j\left(\frac{\theta}{2} + e^{-\theta} - 1\right)\right\}. \end{aligned} \tag{17}$$

The last inequality uses $1 + x \le e^x$ with $x = p_j(e^{-\theta} - 1) \in [-1, 0]$. For completeness, if $g(x) := e^x - 1 - x$, then $g'(0) = 0$, $g''(x) = e^x > 0$, and the global minimum is $g(0) = 0$.

The other elementary envelope is also direct. Let

$$f(x) := 1 - x + \frac{x^2}{2} - e^{-x}, \qquad x \ge 0.$$

Then $f(0) = f'(0) = 0$ and $f''(x) = 1 - e^{-x} \ge 0$; hence $f'(x) \ge 0$ and $f(x) \ge 0$. Therefore

$$e^{-x} \le 1 - x + \frac{x^2}{2}, \qquad x \ge 0.$$

Substituting this in equation 17 yields

$$\frac{\theta}{2} + e^{-\theta} - 1 \le -\frac{\theta}{2} + \frac{\theta^2}{2}.$$

Taking $\theta = 1/2$ makes the right-hand side $-1/8$, proving equation 16. No relationship among counts for different rows or stages is used. □

Proposition 27 (Simultaneous realized-count activation)**.** *Under Assumptions 9 and 10, let $t \in [H]$, $\delta \in (0, 1)$, and define*

$$S_{\mathrm{mast}} := \sum_{r=1}^{H} |\mathcal{U}_r|, \qquad J_t := |\mathcal{I}_t| = |\mathcal{A}| \sum_{s=t}^{H} |S_{\mu,s}| \le |\mathcal{A}| S_{\mathrm{mast}},$$

$$\Gamma := \Gamma_{\mathrm{cnt}}(\delta/2) = \log \frac{4H|\mathcal{A}|S_{\mathrm{mast}}}{\delta}.$$

Choose $c_{\mathrm{cnt}} = 8$. If, for every $s \ge t$ and $D \in \{E, \bar{E}\}$ with $S_{D,s} \ne \varnothing$,

$$nq_{D,s}^{\mathrm{audit}} \ge 8\Gamma, \tag{18}$$

then

$$\Pr\Big((\mathcal{C}_{\delta/2}^{\mathrm{cnt}})^c\Big) \le \frac{\delta}{4H} \le \frac{\delta}{2}, \qquad \Pr(\mathcal{C}_{\delta/2}^{\mathrm{cnt}}) \ge 1 - \frac{\delta}{2}. \tag{19}$$

Moreover, every consumed row has strictly positive realized count on $\mathcal{C}_{\delta/2}^{\mathrm{cnt}}$. Empty strata contribute no row and require no threshold.

Proof. For each $j = (s, u, a) \in \mathcal{I}_t$, the binary event label places u in a unique stratum $D(j) \in \{E, \bar{E}\}$. This stratum is nonempty. By the audit-mass definition and Assumption 10,

$$p_j = d_s^{\mu}(u, a) \geq q^{\mathrm{audit}}_{D(j),s} > 0. \tag{20}$$

The threshold equation 18 implies

$$np_j \geq 8\Gamma.$$

Lemma 26 therefore gives

$$\Pr\Big(N_j < \frac{np_j}{2}\Big) \leq e^{-np_j/8} \leq e^{-\Gamma}. \tag{21}$$

The complement of the count event is exactly the union of these fixed row shortfalls. A union bound, without any independence assumption among rows or stages, gives

$$\begin{aligned}\Pr\Big((\mathcal{C}^{\mathrm{cnt}}_{\delta/2})^c\Big) &\leq J_t e^{-\Gamma} \\ &\leq |\mathcal{A}| S_{\mathrm{mast}} \frac{\delta}{4H|\mathcal{A}|S_{\mathrm{mast}}} \\ &= \frac{\delta}{4H} \leq \frac{\delta}{2},\end{aligned}$$

which proves equation 19.

If $S_{D,s} = \varnothing$, no u from that stratum occurs in $\mathcal{I}_t$. The stratum contributes no union term, and the convention $q^{\mathrm{audit}}_{D,s} = +\infty$ is never multiplied by n or inserted into a tail bound. If all of $\mathcal{I}_t$ is empty, its defining intersection is the sure event and the conclusion holds with $J_t = 0$.

On $\mathcal{C}^{\mathrm{cnt}}_{\delta/2}$, every consumed row satisfies, using $\Gamma > 0$,

$$N_j \geq \frac{np_j}{2} \geq 4\Gamma > 0. \tag{22}$$

Thus count-one and all-zero-count data sets are not silently certified: whenever they violate this display, they lie in the explicitly controlled complement. □

Proposition 28 (Joint split-confidence event)**.** *Under Assumptions 9 and 10, let $\delta \in (0, 1)$, fix $t \in [H]$, and suppose that every $s \geq t$ and every $D \in \{E, \bar{E}\}$ with $S_{D,s} \neq \varnothing$ satisfies*

$$nq^{\mathrm{audit}}_{D,s} \geq 8\Gamma_{\mathrm{cnt}}(\delta/2).$$

Together with Proposition 25, these conditions imply that the event

$$\mathcal{H}_\delta = \mathcal{G}^{\mathrm{dev}}_{\delta/2} \cap \mathcal{C}^{\mathrm{cnt}}_{\delta/2}$$

satisfies

$$\Pr(\mathcal{H}_\delta) \geq 1 - \delta. \tag{23}$$

No independence between the deviation and count events is required.

Proof. De Morgan's identity gives

$$\mathcal{H}_\delta^c = (\mathcal{G}^{\mathrm{dev}}_{\delta/2})^c \cup (\mathcal{C}^{\mathrm{cnt}}_{\delta/2})^c.$$

Propositions 25 and 27, followed by a union bound, imply

$$\begin{aligned}\Pr(\mathcal{H}_\delta^c) &\leq \Pr\Big((\mathcal{G}^{\mathrm{dev}}_{\delta/2})^c\Big) + \Pr\Big((\mathcal{C}^{\mathrm{cnt}}_{\delta/2})^c\Big) \\ &\leq \frac{\delta}{2} + \frac{\delta}{4H} \leq \delta.\end{aligned}$$

Taking complements proves equation 23. The argument allows arbitrary dependence between the event halves. □

Corollary 29 (Count and joint-event guarantees)**.** *Under Assumptions 6, 9, and 10, let $\delta \in (0,1)$, fix $t \in [H]$, choose $c_r = c_p = 1$ and $c_{\mathrm{cnt}} = 8$, and suppose that every nonempty consumed stratum satisfies*

$$nq_{D,s}^{\mathrm{audit}} \geq 8\Gamma_{\mathrm{cnt}}(\delta/2).$$

Then the three simultaneous probability statements

$$\Pr(\mathcal{G}_{\delta/2}^{\mathrm{dev}}) \geq 1 - \frac{\delta}{2}, \qquad \Pr(\mathcal{C}_{\delta/2}^{\mathrm{cnt}}) \geq 1 - \frac{\delta}{2}, \qquad \Pr(\mathcal{H}_{\delta}) \geq 1 - \delta$$

hold. On the count event, every row in $\mathcal{I}_t$ satisfies

$$N_s(u,a) \geq \frac{1}{2} n d_s^{\mu}(u,a) > 0.$$

Proof. The deviation probability is Proposition 25; the count probability and positive-count conclusion are Proposition 27; and their joint conclusion is Proposition 28. Empty strata add neither a row nor an arithmetic operation involving their $+\infty$ audit-mass convention. □

A.5.4 Audited Support Closure and Nuisance-Invariant Quotients

Proposition 30 (All-action closure of the audited support)**.** *Under Assumptions 6 and 10, fix a theorem start $t \in [H]$. For every $s \in \{t, \dots, H\}$, $u \in S_{\mu,s}$, and $a \in \mathcal{A}$,*

$$P_s^{\chi}(S_{\mu,s+1} \mid u, a) = 1,$$

where $S_{\mu,H+1} := \mathcal{U}_{H+1}$ is the singleton terminal alphabet. Equivalently,

$$P_s^{\chi}(u' \mid u, a) > 0 \quad \Longrightarrow \quad u' \in S_{\mu,s+1}.$$

Consequently, a finite true rollout that is in $S_{\mu,s}$ at one stage remains in the successive support masks almost surely under arbitrary action selection.

Proof. Fix $s \in \{t, \dots, H\}$, $u \in S_{\mu,s}$, and $a \in \mathcal{A}$. Let $D \in \{E, \bar{E}\}$ be the stratum containing u. It is nonempty and consumed from stage t, so Assumption 10 gives

$$d_s^{\mu}(u,a) \geq q_{D,s}^{\mathrm{audit}} > 0. \tag{24}$$

The all-action audit is essential here: this inequality holds for the particular queried action, rather than only for a behavior-policy favorite. If a stratum is empty, it contains no supported u, so the assertion is vacuous and the $+\infty$ convention is not used as a positive-mass argument.

By Assumption 6, conditioning on $(U_s, A_s) = (w, b)$ gives the row $P_s^{\chi}(\cdot \mid w, b)$. The finite law of total probability therefore yields, for $u' \in \mathcal{U}_{s+1}$,

$$\begin{aligned} d_{s+1}^{\mu}(u') &= \Pr_{\mu}(U_{s+1} = u') \\ &= \sum_{w \in \mathcal{U}_s} \sum_{b \in \mathcal{A}} \Pr_{\mu}(U_s = w, A_s = b) \Pr_{\mu}(U_{s+1} = u' \mid U_s = w, A_s = b) \\ &= \sum_{w \in \mathcal{U}_s} \sum_{b \in \mathcal{A}} d_s^{\mu}(w,b) P_s^{\chi}(u' \mid w, b) \\ &\geq d_s^{\mu}(u,a) P_s^{\chi}(u' \mid u, a). \end{aligned} \tag{25}$$

Thus a positive transition probability implies $d_{s+1}^{\mu}(u') > 0$, which is exactly $u' \in S_{\mu,s+1}$. Since the alphabet is finite,

$$P_s^{\chi}(\mathcal{U}_{s+1} \setminus S_{\mu,s+1} \mid u, a) = 0,$$

and the asserted mass-one identity follows.

The conclusion is pointwise for every action, so it composes along any finite action sequence. Once a true rollout is in $S_{\mu,s}$, its next atom belongs to $S_{\mu,s+1}$ almost surely regardless of the selected action. Induction over the remaining finite horizon proves the rollout statement. At $s = H$, the transition assigns mass one to the terminal singleton, agreeing with the convention above. □

Proposition 31 (Task-fiber invariance of the supported Bellman quotient)**.** *Under Assumptions 6, 8, and 10, and under Proposition 30, for every $s \in \{t, \dots, H\}$,*

$$u, v \in S_{\mu,s}, \qquad \tau_s(u) = \tau_s(v) \quad \Longrightarrow \quad u \sim_s^{\mathrm{B}} v. \tag{26}$$

Equivalently, there is a map ψ_s on the supported task values $\tau_s(S_{\mu,s})$ such that

$$\varphi_s^{\mathrm{B}}(u) = \psi_s(\tau_s(u)), \qquad u \in S_{\mu,s}.$$

Thus every supported task-equal nuisance fiber is contained in a single Bellman-quotient block.

Proof. We argue by backward induction. At stage $H+1$, both $\mathcal{U}_{H+1}$ and its Bellman quotient consist of one class, so the factorization is immediate.

Suppose the assertion holds at stage $s+1$. Then a well-defined map ψ_{s+1} on $\tau_{s+1}(S_{\mu,s+1})$ satisfies

$$\varphi_{s+1}^{\mathrm{B}}(w) = \psi_{s+1}(\tau_{s+1}(w)), \qquad w \in S_{\mu,s+1}. \tag{27}$$

Indeed, two supported atoms with the same task value are equivalent by the induction hypothesis and hence carry the same quotient label.

Fix $u, v \in S_{\mu,s}$ with $\tau_s(u) = \tau_s(v) =: \theta$, and fix $a \in \mathcal{A}$. Define the next-task marginal of the setting kernel by

$$K_s^T(z \mid \theta, a) := K_s([0,1] \times \{z\} \mid \theta, a).$$

Assumption 8 says that every history in either master atom has the same conditional joint law $K_s(\cdot \mid \theta, a)$ for (R_s, T_{s+1}). Averaging within each master atom and using Assumption 6 gives

$$r_s(u,a) = \int r \, K_s(dr, dz \mid \theta, a) = r_s(v,a) \tag{28}$$

and

$$(\tau_{s+1})_{\#} P_s^{\chi}(\cdot \mid u, a) = K_s^T(\cdot \mid \theta, a) = (\tau_{s+1})_{\#} P_s^{\chi}(\cdot \mid v, a). \tag{29}$$

These equations identify the reward mean and next-task law; they do not identify the full next-master laws, which may differ through nuisance-only dynamics.

Proposition 30 gives

$$P_s^{\chi}(S_{\mu,s+1} \mid u, a) = P_s^{\chi}(S_{\mu,s+1} \mid v, a) = 1.$$

For any set B of stage-$s+1$ quotient labels,

$$\begin{aligned}
&[(\varphi_{s+1}^{\mathrm{B}})_{\#} P_s^{\chi}(\cdot \mid u, a)](B) \\
&\quad = \sum_{w \in S_{\mu,s+1}} \mathbf{1}\{\varphi_{s+1}^{\mathrm{B}}(w) \in B\} P_s^{\chi}(w \mid u, a) \\
&\quad = \sum_{w \in S_{\mu,s+1}} \mathbf{1}\{\psi_{s+1}(\tau_{s+1}(w)) \in B\} P_s^{\chi}(w \mid u, a) \\
&\quad = [(\psi_{s+1})_{\#}((\tau_{s+1})_{\#} P_s^{\chi}(\cdot \mid u, a))](B).
\end{aligned} \tag{30}$$

The identical calculation holds with v in place of u. Equation equation 29 therefore implies

$$(\varphi_{s+1}^{\mathrm{B}})_{\#} P_s^{\chi}(\cdot \mid u, a) = (\varphi_{s+1}^{\mathrm{B}})_{\#} P_s^{\chi}(\cdot \mid v, a).$$

Together with equation 28, this proves equality of the two Bellman-signature components for the arbitrary action a. Since a was arbitrary, the recursive definition of $\sim_s^{\mathrm{B}}$ yields $u \sim_s^{\mathrm{B}} v$.

For each task value in $\tau_s(S_{\mu,s})$, choose a supported atom with that value and let ψ_s be its quotient label. The equivalence just proved makes this definition independent of the choice, closing the induction. Only supported atoms occur: support closure gives zero mass to off-support next atoms under every row used above. Thus arbitrary off-support quotient or learner labels cannot affect the pushed laws. The conclusion is population nuisance invariance on support, not equality of full master kernels or empirical quotient recovery. If the supported quotient has one block, the same argument gives the immediate one-block conclusion. □

Corollary 32 (Support and nuisance-quotient structure)**.** *Under Assumptions 6, 8, and 10, every action-conditioned continuation of a supported master atom stays in the next support, and every pair of supported atoms with a common task coordinate lies in one population Bellman-quotient block. These conclusions hold through the terminal singleton and concern behavior support only.*

Proof. The all-action support conclusion, including its finite-rollout and terminal consequences, is Proposition 30. With that closed continuation domain, the backward task-kernel argument in Proposition 31 proves equation 26. Neither proposition assigns structural meaning to off-support labels or asserts learned partition recovery. □

A.5.5 Selected-Block Representative Accuracy

Lemma 33 (Pathwise contraction through the realized selected code)**.** *Under Proposition 22, fix an arbitrary data set and a stage s. For any probability laws P, Q on $\mathcal{U}_{s+1}$, the realized finite map $\widehat{\phi}_{s+1}$ satisfies*

$$\mathrm{TV}\Big((\widehat{\phi}_{s+1})_{\#}P, (\widehat{\phi}_{s+1})_{\#}Q\Big) \le \mathrm{TV}(P, Q). \tag{31}$$

The inequality holds pointwise on every event despite the data dependence of the selected code.

Proof. After the data set is fixed, $\widehat{\phi}_{s+1}$ is a definite total map on the finite next-master alphabet. Put $q(w) := P(w) - Q(w)$. Under the convention $\mathrm{TV} = \frac{1}{2}\ell_1$,

$$\begin{aligned} 2\,\mathrm{TV}\Big((\widehat{\phi}_{s+1})_{\#}P, (\widehat{\phi}_{s+1})_{\#}Q\Big) &= \sum_{z} \left| \sum_{w:\, \widehat{\phi}_{s+1}(w)=z} q(w) \right| \\ &\le \sum_{z} \sum_{w:\, \widehat{\phi}_{s+1}(w)=z} |q(w)| \\ &= \sum_{w} |q(w)| = 2\,\mathrm{TV}(P, Q). \end{aligned}$$

This deterministic fiber calculation needs neither conditioning on the selected code nor a union over candidate maps. At $s = H$, both pushforwards lie on the terminal singleton and the distance is zero. □

Lemma 34 (Selected-pair population and representative triangles)**.** *Under Assumptions 13 and 14, Proposition 22, Proposition 25, and Corollary 29, work on $\mathcal{H}_\delta$. For a consumed row write*

$$R_s^r(w, a) := \mathrm{rad}_r(N_s(w, a)), \qquad R_s^p(w, a) := \mathrm{rad}_p(N_s(w, a)). \tag{32}$$

Fix $s \ge t$, $a \in \mathcal{A}$, and $u, v \in S_{\mu,s}$ with $\widehat{\phi}_s(u) = \widehat{\phi}_s(v)$. Then

$$|r_s(u, a) - r_s(v, a)| \le \tau_s^r(u, v, a) + 2R_s^r(u, a) + 2R_s^r(v, a), \tag{33}$$

$$\begin{aligned} &\mathrm{TV}\Big((\widehat{\phi}_{s+1})_{\#}P_s^{\mathcal{X}}(\cdot \mid u, a), (\widehat{\phi}_{s+1})_{\#}P_s^{\mathcal{X}}(\cdot \mid v, a)\Big) \\ &\quad \le \tau_s^p(u, v, a) + 2R_s^p(u, a) + 2R_s^p(v, a). \end{aligned} \tag{34}$$

If v is the prescribed representative of the selected block containing u, then the exact selected representative-model row satisfies

$$|r_s(u, a) - \widehat{r}_s(v, a)| \le \tau_s^r(u, v, a) + 2R_s^r(u, a) + R_s^r(v, a), \tag{35}$$

$$\begin{aligned} &\mathrm{TV}\Big((\widehat{\phi}_{s+1})_{\#}P_s^{\mathcal{X}}(\cdot \mid u, a), (\widehat{\phi}_{s+1})_{\#}\widehat{P}_s^{\mathcal{X}}(\cdot \mid v, a)\Big) \\ &\quad \le \tau_s^p(u, v, a) + 2R_s^p(u, a) + R_s^p(v, a). \end{aligned} \tag{36}$$

Proof. Because $\mathcal{H}_\delta \subseteq \mathcal{C}_{\delta/2}^{\mathrm{cnt}}$, Corollary 29 gives positive counts at both (s, u, a) and (s, v, a). Because $\mathcal{H}_\delta \subseteq \mathcal{G}_{\delta/2}^{\mathrm{dev}}$, Proposition 25 gives, for $w \in \{u, v\}$,

$$|r_s(w, a) - \widehat{r}_s(w, a)| \le R_s^r(w, a) \tag{37}$$

and

$$\mathrm{TV}\Big(P_s^{\chi}(\cdot \mid w,a), \widehat{P}_s^{\chi}(\cdot \mid w,a)\Big) \le R_s^p(w,a). \tag{38}$$

The selected candidate supplied by Proposition 22 is feasible. Since it merges u and v, its reward comparison reads

$$|\widehat{r}_s(u,a) - \widehat{r}_s(v,a)| \le \tau_s^r(u,v,a) + R_s^r(u,a) + R_s^r(v,a). \tag{39}$$

Combining equation 37 at both endpoints with equation 39 proves equation 33. Combining it only at the queried endpoint u proves equation 35. Under Assumption 14, $\widehat{r}_s(v,a)$ is exactly the reward row of the selected model at $\widehat{\phi}_s(u)$.

The transition feasibility comparison uses the same realized downstream code:

$$\begin{aligned}&\mathrm{TV}\Big((\widehat{\phi}_{s+1})_{\#}\widehat{P}_s^{\chi}(\cdot \mid u,a), (\widehat{\phi}_{s+1})_{\#}\widehat{P}_s^{\chi}(\cdot \mid v,a)\Big)\\ &\qquad \le \tau_s^p(u,v,a) + R_s^p(u,a) + R_s^p(v,a).\end{aligned} \tag{40}$$

Lemma 33 and equation 38 imply, for each $w \in \{u,v\}$,

$$\mathrm{TV}\Big((\widehat{\phi}_{s+1})_{\#}P_s^{\chi}(\cdot \mid w,a), (\widehat{\phi}_{s+1})_{\#}\widehat{P}_s^{\chi}(\cdot \mid w,a)\Big) \le R_s^p(w,a). \tag{41}$$

Adding equation 41 at both endpoints to equation 40 proves equation 34; adding it only at u proves equation 36. The final empirical pushed law is precisely the selected representative-model transition row.

This argument uses only the selected candidate's own feasibility. A selected block may cross the declared reference coarsening ϕ°; no feasibility of that reference under the crossing selected downstream code is asserted. □

Proposition 35 (Global endpoint envelope for representative and pair signatures)**.** *Under Assumptions 13 and 14, Propositions 22, 25, and 28, and Lemmas 33 and 34, choose universal constants satisfying*

$$c_b \ge 1, \qquad c_s \ge 4\sqrt{2}\max\{c_r, c_p\}. \tag{42}$$

On $\mathcal{H}_\delta$, for every $s \ge t$, action a, supported queried atom u, and prescribed representative v of its selected block,

$$\begin{aligned}&|r_s(u,a) - \widehat{r}_s(v,a)|\\ &\quad + B_{s+1}\,\mathrm{TV}\Big((\widehat{\phi}_{s+1})_{\#}P_s^{\chi}(\cdot \mid u,a), (\widehat{\phi}_{s+1})_{\#}\widehat{P}_s^{\chi}(\cdot \mid v,a)\Big)\\ &\quad \le \max_{D\in\{E,\bar{E}\}} a_{D,s} + \max_{D\in\{E,\bar{E}\}} s_{D,s}.\end{aligned} \tag{43}$$

Every supported pair u, v merged by the selected code also satisfies

$$\begin{aligned}&|r_s(u,a) - r_s(v,a)|\\ &\quad + B_{s+1}\,\mathrm{TV}\Big((\widehat{\phi}_{s+1})_{\#}P_s^{\chi}(\cdot \mid u,a), (\widehat{\phi}_{s+1})_{\#}P_s^{\chi}(\cdot \mid v,a)\Big)\\ &\quad \le 4\left(\max_{D\in\{E,\bar{E}\}} a_{D,s} + \max_{D\in\{E,\bar{E}\}} s_{D,s}\right).\end{aligned} \tag{44}$$

Proof. On $\mathcal{H}_\delta$, the count conclusion in Corollary 29 and the positive-count radius definitions give, for every supported consumed endpoint w,

$$\begin{aligned}R_s^r(w,a) &\le c_r\sqrt{\frac{\Gamma_{\mathrm{dev}}(\delta/2)}{N_s(w,a)}} \le \sqrt{2}c_r\sqrt{\frac{\Gamma_{\mathrm{dev}}(\delta/2)}{nd_s^{\mu}(w,a)}},\\ R_s^p(w,a) &\le c_p\sqrt{\frac{\Gamma_{\mathrm{dev}}(\delta/2)}{N_s(w,a)}} \le \sqrt{2}c_p\sqrt{\frac{\Gamma_{\mathrm{dev}}(\delta/2)}{nd_s^{\mu}(w,a)}}.\end{aligned} \tag{45}$$

The first inequalities remain true under clipping because $\min\{1, x\} \le x$. Define the proof-local endpoint scale

$$x_s(w, a) := \sqrt{\frac{\Gamma_{\mathrm{dev}}(\delta/2)}{n d_s^{\mu}(w, a)}}.$$

Choose $w_\star \in \{u, v\}$ for which

$$x_s(w_\star, a) = \max\{x_s(u, a), x_s(v, a)\},$$

and let $D_\star$ be its event or non-event stratum. Symmetry in Assumption 13 permits orienting the pair from $w_\star$, and hence

$$\begin{aligned} \tau_s^r(u, v, a) + B_{s+1}\tau_s^p(u, v, a) &\le \varepsilon_s^r(w_\star, a) + B_{s+1}\varepsilon_s^p(w_\star, a) \\ &= \frac{1}{c_b} a_s(w_\star, a) \\ &\le \frac{1}{c_b} \max_{D \in \{E, \bar{E}\}} a_{D,s}. \end{aligned} \tag{46}$$

For either $w \in \{u, v\}$,

$$\begin{aligned} R_s^r(w, a) + B_{s+1}R_s^p(w, a) &\le \sqrt{2}\max\{c_r, c_p\}(1 + B_{s+1})x_s(w, a) \\ &\le \frac{\sqrt{2}\max\{c_r, c_p\}}{c_s} s_s(w_\star, a) \\ &\le \frac{\sqrt{2}\max\{c_r, c_p\}}{c_s} \max_{D \in \{E, \bar{E}\}} s_{D,s}. \end{aligned} \tag{47}$$

If the representative is rarer than the query, it becomes $w_\star$; its population scale, rather than the query's scale, is therefore charged. If the endpoints lie in different strata, the displayed maximum includes $D_\star$.

Weighting equation 35 and equation 36 by 1 and B_{s+1}, respectively, and applying equation 46–equation 47 gives

$$\begin{aligned} \text{representative weighted error} &\le \frac{1}{c_b}\max_D a_{D,s} + \frac{3\sqrt{2}\max\{c_r, c_p\}}{c_s}\max_D s_{D,s} \\ &\le \max_D a_{D,s} + \frac{3}{4}\max_D s_{D,s} \\ &\le \max_D a_{D,s} + \max_D s_{D,s}, \end{aligned} \tag{48}$$

where both maxima are over $D \in \{E, \bar{E}\}$. This proves equation 43.

Likewise, weighting equation 33 and equation 34 gives

$$\begin{aligned} \text{true selected-code pair signature} &\le \frac{1}{c_b}\max_D a_{D,s} + \frac{4\sqrt{2}\max\{c_r, c_p\}}{c_s}\max_D s_{D,s} \\ &\le \max_D a_{D,s} + \max_D s_{D,s} \\ &\le 4\left(\max_D a_{D,s} + \max_D s_{D,s}\right). \end{aligned} \tag{49}$$

The final deliberately conservative form is equation 44. Before the maxima are taken, the four endpoint-radius copies, the reward and transition tolerance classes, both atom-action masses, and the factor B_{s+1} all remain explicit.

At $s = H$, $B_{H+1} = 0$ and the selected next-code alphabet is the terminal singleton, so both conclusions reduce to reward-only statements. If the relevant tolerances and endpoint deviations vanish, the representative model row and the selected-code population signatures are exact. Thus equation 43 is the complete one-step model interface for the Bellman recursion, while equation 44 is the supported separation interface; neither is replaced by the rowwise quantity $\ell_s(u, a)$. □

A.5.6 Backward Supported Bellman Error Propagation

For $s = t, \dots, H$, define the appendix-local supported error

$$D_s := \sup_{\substack{u \in S_{\mu,s} \\ a \in \mathcal{A}}} \left| Q_s^\star(u,a) - \widehat{Q}_s(\widehat{\phi}_s(u), a) \right|, \qquad D_{H+1} := 0. \tag{50}$$

The finite supremum is a maximum. The terminal convention records zero terminal value error and does not introduce a stage-$H+1$ action-value function.

Lemma 36 (Finite-range Bellman comparison primitives)**.** *Under Proposition 22, the following statements hold on every data set.*

First, for every $s \in \{1, \dots, H+1\}$ and every selected latent label z,

$$0 \le \widehat{V}_s(z) \le B_s, \qquad B_{H+1} = 0. \tag{51}$$

Second, for every $s \in \{t, \dots, H\}$ and $u \in S_{\mu,s}$,

$$\left| V_s^\star(u) - \widehat{V}_s(\widehat{\phi}_s(u)) \right| \le D_s. \tag{52}$$

At the terminal stage, both values and the corresponding error are zero. Third, if P, Q are probability laws on a selected stage-s latent alphabet and f takes values in $[0, B_s]$, then

$$|\mathbb{E}_P f - \mathbb{E}_Q f| \le B_s \, \mathrm{TV}(P, Q). \tag{53}$$

Proof. Every empirical reward mean lies in $[0,1]$, the fixed zero-count reward is zero, and every empirical or default transition row is a probability law. The selected model uses precisely such representative rows. Starting from $\widehat{V}_{H+1} = 0$, if equation 51 holds at $s+1$, then

$$0 \le \widehat{Q}_s(z,a) \le 1 + B_{s+1} = B_s.$$

Taking a maximum over the finite action set proves equation 51 by backward induction for every selected supported and canonical off-support label.

For finite vectors $(x_a)_a$ and $(y_a)_a$,

$$\left| \max_a x_a - \max_a y_a \right| \le \max_a |x_a - y_a|. \tag{54}$$

Indeed, $x_a \le y_a + \max_b |x_b - y_b|$ for every a; maximizing gives one direction, and exchanging x and y gives the other. Apply equation 54 with

$$x_a = Q_s^\star(u,a), \qquad y_a = \widehat{Q}_s(\widehat{\phi}_s(u), a).$$

Definition equation 50 then proves equation 52.

Finally, put $m(z) := P(z) - Q(z)$. Since P and Q are probability laws,

$$\sum_{z:\, m(z) > 0} m(z) = \sum_{z:\, m(z) < 0} (-m(z)) = \frac{1}{2} \sum_z |m(z)| = \mathrm{TV}(P, Q).$$

For $0 \le f \le B_s$,

$$-B_s \, \mathrm{TV}(P,Q) \le \sum_z f(z) m(z) \le B_s \, \mathrm{TV}(P,Q),$$

which proves equation 53 with no missing factor under the half-ℓ_1 convention. □

Proposition 37 (One-step supported Bellman error recurrence)**.** *Under Assumption 6, Proposition 30, Proposition 35, and Lemma 36, on $\mathcal{H}_\delta$, for every $s \in \{t, \dots, H\}$,*

$$D_s \le \max_{D \in \{E, \bar{E}\}} a_{D,s} + \max_{D \in \{E, \bar{E}\}} s_{D,s} + D_{s+1}. \tag{55}$$

Proof. Work on $\mathcal{H}_\delta$, and fix $s \in \{t, \dots, H\}$, $u \in S_{\mu,s}$, and $a \in \mathcal{A}$. Let v be the prescribed representative of the selected block containing u. Assumption 6 and the representative model construction give

$$Q_s^\star(u,a) = r_s(u,a) + \mathbb{E}_{U' \sim P_s^\chi(\cdot|u,a)} V_{s+1}^\star(U') \tag{56}$$

and

$$\widehat{Q}_s(\widehat{\phi}_s(u), a) = \widehat{r}_s(v,a) + \mathbb{E}_{U' \sim \widehat{P}_s^\chi(\cdot|v,a)} \widehat{V}_{s+1}(\widehat{\phi}_{s+1}(U')). \tag{57}$$

Insert the learned continuation evaluated under the true next-master law and through the same realized selected code. The triangle inequality yields

$$\begin{aligned} &\left| Q_s^\star(u,a) - \widehat{Q}_s(\widehat{\phi}_s(u), a) \right| \\ &\le |r_s(u,a) - \widehat{r}_s(v,a)| \\ &\quad + \left| \mathbb{E}_{U' \sim P_s^\chi(\cdot|u,a)} \left[V_{s+1}^\star(U') - \widehat{V}_{s+1}(\widehat{\phi}_{s+1}(U')) \right] \right| \\ &\quad + \left| \mathbb{E}_{(\widehat{\phi}_{s+1})_\# P_s^\chi(\cdot|u,a)} \widehat{V}_{s+1} - \mathbb{E}_{(\widehat{\phi}_{s+1})_\# \widehat{P}_s^\chi(\cdot|v,a)} \widehat{V}_{s+1} \right|. \end{aligned} \tag{58}$$

For $s < H$, Proposition 30 makes the true law in the middle line concentrate on $S_{\mu,s+1}$. Therefore equation 52 gives

$$\begin{aligned} &\left| \mathbb{E}_{P_s^\chi(\cdot|u,a)} \left[V_{s+1}^\star(U') - \widehat{V}_{s+1}(\widehat{\phi}_{s+1}(U')) \right] \right| \\ &\qquad \le \mathbb{E}_{P_s^\chi(\cdot|u,a)} \left| V_{s+1}^\star(U') - \widehat{V}_{s+1}(\widehat{\phi}_{s+1}(U')) \right| \le D_{s+1}. \end{aligned} \tag{59}$$

At $s = H$, both terminal values are zero, so the same display holds with $D_{H+1} = 0$.

By equation 51, $\widehat{V}_{s+1}$ takes values in $[0, B_{s+1}]$. Applying equation 53 to the last two laws in equation 58 gives

$$\begin{aligned} &\left| \mathbb{E}_{(\widehat{\phi}_{s+1})_\# P_s^\chi(\cdot|u,a)} \widehat{V}_{s+1} - \mathbb{E}_{(\widehat{\phi}_{s+1})_\# \widehat{P}_s^\chi(\cdot|v,a)} \widehat{V}_{s+1} \right| \\ &\qquad \le B_{s+1} \operatorname{TV}\Big((\widehat{\phi}_{s+1})_\# P_s^\chi(\cdot \mid u,a), (\widehat{\phi}_{s+1})_\# \widehat{P}_s^\chi(\cdot \mid v,a) \Big). \end{aligned} \tag{60}$$

At $s = H$, $B_{H+1} = 0$, so this contribution vanishes exactly.

Substituting equation 59–equation 60 into equation 58 and applying equation 43 yields

$$\left| Q_s^\star(u,a) - \widehat{Q}_s(\widehat{\phi}_s(u), a) \right| \le \max_D a_{D,s} + \max_D s_{D,s} + D_{s+1}, \tag{61}$$

where both maxima are over $D \in \{E, \bar{E}\}$. Taking the supremum over supported u and all actions proves equation 55. Every transition law in this comparison is pushed through the identical realized $\widehat{\phi}_{s+1}$, and no event or non-event charge is replaced by a row-local representative quantity. □

Proposition 38 (Separated finite-horizon bias and statistical telescope)**.** *Under Assumption 6, Propositions 30, 35, and 37, on $\mathcal{H}_\delta$,*

$$\begin{aligned} D_t &\le \sum_{s=t}^H \max_{D \in \{E,\bar{E}\}} a_{D,s} + \sum_{s=t}^H \max_{D \in \{E,\bar{E}\}} s_{D,s} \\ &= \mathfrak{A}_t + \mathfrak{S}_t = \mathfrak{B}_t. \end{aligned} \tag{62}$$

Furthermore,

$$\mathfrak{A}_t \le c_b \sum_{s=t}^H \max_{D \in \{E,\bar{E}\}} \left(\varepsilon_{D,s}^r + B_{s+1} \varepsilon_{D,s}^p \right) \tag{63}$$

and

$$\mathfrak{S}_t \le c_s \sum_{s=t}^H \max_{D \in \{E,\bar{E}\}} \left[(1 + B_{s+1}) \sqrt{\frac{\Gamma_{\mathrm{dev}}(\delta/2)}{n q_{D,s}^{\mathrm{audit}}}} \right], \tag{64}$$

where the bracketed contribution is defined to be zero for an empty stratum, without evaluating an expression involving its $+\infty$ audit-mass convention. Under $H = 1$, zero raw tolerances, and $q^{\mathrm{audit}}_{\bar{E},1} \geq q^{\mathrm{audit}}_{E,1} < +\infty$,

$$D_1 \leq \mathfrak{S}_1 \leq c_s \sqrt{\frac{\Gamma_{\mathrm{dev}}(\delta/2)}{n q^{\mathrm{audit}}_{E,1}}}. \tag{65}$$

Proof. Iterate equation 55:

$$\begin{aligned} D_t &\leq \max_D a_{D,t} + \max_D s_{D,t} + D_{t+1} \\ &\leq \sum_{s=t}^{H} \left(\max_D a_{D,s} + \max_D s_{D,s} \right) + D_{H+1}, \end{aligned} \tag{66}$$

where every maximum is over $D \in \{E, \bar{E}\}$. Because $D_{H+1} = 0$ and the horizon is finite, distributivity separates the last expression exactly into the two sums defining $\mathfrak{A}_t$ and $\mathfrak{S}_t$. Every persistent local bias charge and every persistent local statistical charge therefore appears once for each downstream stage; there is no contraction, cancellation, or asymptotic absorption.

For each stratum and stage, the setting gives

$$a_{D,s} \leq c_b \left(\varepsilon^r_{D,s} + B_{s+1} \varepsilon^p_{D,s} \right)$$

and

$$s_{D,s} \leq c_s (1 + B_{s+1}) \sqrt{\frac{\Gamma_{\mathrm{dev}}(\delta/2)}{n q^{\mathrm{audit}}_{D,s}}}.$$

Taking the maximum over the same two strata separately at each stage and summing proves equation 63–equation 64. The event and non-event audit masses are not replaced by an unstated coverage parameter. For an empty stratum, the setting assigns $a_{D,s} = s_{D,s} = 0$; this direct zero branch is used, so no quotient or product involving $q^{\mathrm{audit}}_{D,s} = +\infty$ is formed.

If all local bias and statistical charges vanish, equation 62 gives $D_t = 0$, hence exact supported action values. At $H = 1$, $B_2 = 0$ and there is only one stage charge:

$$D_1 \leq \max_D a_{D,1} + \max_D s_{D,1} = \mathfrak{A}_1 + \mathfrak{S}_1.$$

Under the further zero-tolerance assumptions, $\mathfrak{A}_1 = 0$. Since the event stratum has the smaller audit mass and an empty non-event stratum contributes zero, equation 64 becomes equation 65.

Finally, by equation 50, the principal bound is exactly

$$\sup_{\substack{u \in S_{\mu,t} \\ a \in \mathcal{A}}} \left| Q^\star_t(u,a) - \widehat{Q}_t(\widehat{\phi}_t(u), a) \right| \leq \mathfrak{A}_t + \mathfrak{S}_t = \mathfrak{B}_t.$$

This is a supported supremum statement on the single joint event. It makes no target-shift, action-correctness, recovery, or upper–lower matching claim; the one-stage display above is only the supported action-value input to the later rate specialization. □

A.5.7 Target Transfer and True-Rollout Planning Regret

Proposition 39 (Piecewise atom-level change of measure)**.** *Under Assumption 11, fix $s \in \{t, \ldots, H\}$, $D \in \{E, \bar{E}\}$, $\pi \in \Pi_{\mathrm{tar}}$, and a selector $g_s : \mathcal{U}_s \to \mathcal{A}$. Define*

$$\mathfrak{l}^{\pi,g}_{D,s} := \begin{cases} \displaystyle\sum_{\substack{u \in S_{D,s} \\ d^\pi_s(u) > 0}} d^\pi_s(u) \ell_s(u, g_s(u)), & S_{D,s} \neq \varnothing, \\ 0, & S_{D,s} = \varnothing, \end{cases} \tag{67}$$

and

$$\mathfrak{v}^\pi_{D,s} := \begin{cases} C^{\mathrm{atom}}_{D,s} \displaystyle\sum_{\substack{u \in S_{D,s},\, a \in \mathcal{A} \\ d^\mu_s(u,a) > 0}} d^\mu_s(u,a) \ell_s(u,a), & S_{D,s} \neq \varnothing, \\ 0, & S_{D,s} = \varnothing. \end{cases} \tag{68}$$

Then every target-positive atom is behavior supported for every queried action, and

$$\mathfrak{l}_{D,s}^{\pi,g} \le \mathfrak{v}_{D,s}^{\pi}. \tag{69}$$

The finite quantity

$$\mathfrak{V}_t^{\pi} := \sum_{s=t}^{H} \sum_{D\in\{E,\bar{E}\}} \mathfrak{v}_{D,s}^{\pi} \tag{70}$$

satisfies, for every selector sequence g,

$$\sum_{s=t}^{H} \sum_{D\in\{E,\bar{E}\}} \mathfrak{l}_{D,s}^{\pi,g} \le \mathfrak{V}_t^{\pi}. \tag{71}$$

On nonempty strata, the restricted sums equal their intended probability expectations because only zero-mass atoms or rows are omitted. On empty strata, both contributions are defined as zero after target emptiness is established and before any loss or density product is evaluated. The assertion is deterministic.

Proof. First suppose $S_{D,s} = \varnothing$. If some $\pi_0 \in \Pi_{\mathrm{tar}}$ assigned positive mass to $\mathcal{U}_{D,s}$, finiteness of the master alphabet would give $u \in \mathcal{U}_{D,s}$ with $d_s^{\pi_0}(u) > 0$. Behavior emptiness gives $d_s^{\mu}(u,a) = 0$ for every action, so the counterfactual ratio in $C_{D,s}^{\mathrm{atom}}$ would be $+\infty$, contradicting Assumption 11. Hence every policy in Π_{tar} assigns this stratum zero mass. The zero branches in equation 67 and equation 68 are taken before evaluating ℓ_s or multiplying by $C_{D,s}^{\mathrm{atom}}$. Thus equation 69 becomes $0 \le 0$, with neither an off-support loss nor a $+\infty \cdot 0$ expression.

Now suppose $S_{D,s} \neq \varnothing$. The finite-support condition gives $C_{D,s}^{\mathrm{atom}} < \infty$. If $d_s^{\pi}(u) > 0$, then $d_s^{\mu}(u,a) > 0$ for every $a \in \mathcal{A}$; otherwise the corresponding counterfactual ratio would be infinite. All target mass in the stratum therefore lies in $S_{D,s}$, and every loss in equation 67 is evaluated on a positive behavior row at its queried action.

For each $u \in S_{D,s}$ with $d_s^{\pi}(u) > 0$, instantiate the counterfactual coefficient at $a = g_s(u)$:

$$d_s^{\pi}(u) \le C_{D,s}^{\mathrm{atom}} d_s^{\mu}(u, g_s(u)). \tag{72}$$

Since $\ell_s = a_s + s_s \ge 0$, multiplication and finite summation give

$$\begin{aligned} \mathfrak{l}_{D,s}^{\pi,g} &\le C_{D,s}^{\mathrm{atom}} \sum_{\substack{u\in S_{D,s}\\ d_s^{\pi}(u)>0}} d_s^{\mu}(u, g_s(u))\ell_s(u, g_s(u)) \\ &\le C_{D,s}^{\mathrm{atom}} \sum_{\substack{u\in S_{D,s},\, a\in\mathcal{A}\\ d_s^{\mu}(u,a)>0}} d_s^{\mu}(u,a)\ell_s(u,a) \\ &= \mathfrak{v}_{D,s}^{\pi}. \end{aligned} \tag{73}$$

Every selected row in the first sum is behavior positive, and the second nonnegative sum contains it. No factor $|\mathcal{A}|$ is introduced. Behavior-null rows never appear and are not assigned a numerical loss. The finite horizon and alphabets, finite ratios on nonempty branches, and finite positive-row losses make $\mathfrak{V}_t^{\pi}$ finite. Summing equation 69 proves equation 71. This is a rowwise-loss transfer only; it is not a representative-model error certificate. □

Proposition 40 (Homogeneous evaluation of the total learned policy)**.** *Under Assumptions 6, 12, and 14, and under Propositions 20 and 21, fix any realized training data set d. Let $\widehat{g}_s^d$ and $\widehat{\pi}^d$ be the exact selector and deterministic policy produced from d, and define*

$$V_{H+1}^{\widehat{\pi}^d,\chi} \equiv 0,$$

$$V_s^{\widehat{\pi}^d,\chi}(u) := r_s(u, \widehat{g}_s^d(u)) + \mathbb{E}_{U'\sim P_s^{\chi}(\cdot|u,\widehat{g}_s^d(u))}\left[V_{s+1}^{\widehat{\pi}^d,\chi}(U')\right]. \tag{74}$$

For a fresh true-environment rollout initialized at any history h_s, set

$$J_s^d(h_s) := \mathbb{E}_{s,h_s}^{\widehat{\pi}^d}\left[\sum_{j=s}^{H} R_j\right], \qquad J_{H+1}^d \equiv 0. \tag{75}$$

Then, on every data set, at every stage and history,

$$J_s^d(h_s) = V_s^{\widehat{\pi}^d,\chi}(\chi_s(h_s)), \qquad 0 \le J_s^d(h_s), V_s^{\widehat{\pi}^d,\chi}(\chi_s(h_s)) \le B_s. \tag{76}$$

Thus the master recursion is the actual conditional fresh-rollout value of the exact learned policy, rather than an auxiliary model value.

Proof. Propositions 20 and 21 provide, for every data set,

$$\widehat{\pi}_s^d(\cdot \mid h_s) = \delta_{\widehat{g}_s^d(\chi_s(h_s))} \tag{77}$$

as a total deterministic measurable policy, including on all-zero counts. No statistical event is needed for this definition.

We prove equation 76 by backward induction. At $H+1$, both sides are zero. Suppose the identity holds at $s+1$. Fix h_s, put $u := \chi_s(h_s)$, and let $a := \widehat{g}_s^d(u)$. One-step conditioning and linearity give

$$\begin{aligned} J_s^d(h_s) &= \mathbb{E}[R_s \mid H_s = h_s, A_s = a] \\ &\quad + \mathbb{E}\big[J_{s+1}^d(H_{s+1}) \mid H_s = h_s, A_s = a\big]. \end{aligned} \tag{78}$$

These conditional quantities use the setting-provided finite kernel for a rollout initialized at the specified history; the history need not have positive mass under a previously generated trajectory. The induction hypothesis turns the continuation integrand into $V_{s+1}^{\widehat{\pi}^d,\chi}(U_{s+1})$. Assumption 6 identifies the first term with $r_s(u,a)$ and the next-master marginal with $P_s^\chi(\cdot \mid u,a)$. Consequently,

$$\begin{aligned} J_s^d(h_s) &= r_s(u,a) + \mathbb{E}_{P_s^\chi(\cdot|u,a)}\left[V_{s+1}^{\widehat{\pi}^d,\chi}(U')\right] \\ &= V_s^{\widehat{\pi}^d,\chi}(u). \end{aligned} \tag{79}$$

No conditional independence between reward and next master atom is used: expected current reward and expected continuation separate by linearity, and only the next-master marginal is needed. Rewards lie in $[0,1]$, so the initialized return lies in $[0, B_s]$; equivalently, the same bound follows directly from equation 74 by backward induction. □

Lemma 41 (Stagewise greedy advantage and event regret)**.** *Under Proposition 39, Proposition 37, and the exact learned selector in Proposition 22, fix a training data set $d \in \mathcal{H}_\delta$. For every $s \in \{t, \dots, H\}$ and $u \in S_{\mu,s}$,*

$$\begin{aligned} 0 &\le V_s^\star(u) - Q_s^\star(u, \widehat{g}_s^d(u)) \\ &= Q_s^\star(u, a_s^\star(u)) - Q_s^\star(u, \widehat{g}_s^d(u)) \\ &\le \min\{B_s, 2\mathfrak{B}_s\}. \end{aligned} \tag{80}$$

In particular,

$$Q_t^\star(u, a_t^\star(u)) - Q_t^\star(u, \widehat{g}_t^d(u)) \le \min\{B_t, 2\mathfrak{B}_t\}. \tag{81}$$

For every $\pi \in \Pi_{\mathrm{tar}}$,

$$\mathcal{R}_{E,t}^\pi(\widehat{g}_t^d) \le \rho_t^\pi \min\{B_t, 2\mathfrak{B}_t\}. \tag{82}$$

Thus the normalized event regret obeys the same bound when $\rho_t^\pi > 0$, while the unnormalized contribution is exactly zero when $\rho_t^\pi = 0$.

Proof. Retelescope Proposition 37 from an arbitrary s:

$$\begin{aligned} D_s &\le \sum_{j=s}^{H} \max_{D \in \{E,\bar{E}\}} a_{D,j} + \sum_{j=s}^{H} \max_{D \in \{E,\bar{E}\}} s_{D,j} + D_{H+1} \\ &= \mathfrak{A}_s + \mathfrak{S}_s = \mathfrak{B}_s. \end{aligned} \tag{83}$$

For $u \in S_{\mu,s}$, learned greediness gives

$$\widehat{Q}_s(\widehat{\phi}_s(u), \widehat{g}_s^d(u)) \ge \widehat{Q}_s(\widehat{\phi}_s(u), a_s^\star(u)). \tag{84}$$

Add and subtract these two learned values. Their middle difference has the favorable sign, so

$$\begin{aligned} &Q_s^\star(u, a_s^\star(u)) - Q_s^\star(u, \widehat{g}_s^d(u)) \\ &\quad \le \left| Q_s^\star(u, a_s^\star(u)) - \widehat{Q}_s(\widehat{\phi}_s(u), a_s^\star(u)) \right| \\ &\qquad + \left| Q_s^\star(u, \widehat{g}_s^d(u)) - \widehat{Q}_s(\widehat{\phi}_s(u), \widehat{g}_s^d(u)) \right| \\ &\quad \le 2D_s \le 2\mathfrak{B}_s. \end{aligned} \tag{85}$$

True optimality makes the left side nonnegative, and the range $0 \le Q_s^\star \le B_s$ supplies the B_s cap.

Proposition 39 puts every target-positive atom on behavior support. Hence

$$\begin{aligned} \mathcal{R}_{E,t}^\pi(\widehat{g}_t^d) &= \sum_{u \in S_{E,t}} d_t^\pi(u) \left[Q_t^\star(u, a_t^\star(u)) - Q_t^\star(u, \widehat{g}_t^d(u)) \right] \\ &\le \rho_t^\pi \min\{B_t, 2\mathfrak{B}_t\}, \end{aligned}$$

including the exact zero branch when $S_{E,t} = \varnothing$ or $\rho_t^\pi = 0$. Division is performed only on the positive-occupancy branch. □

Proposition 42 (True-rollout performance difference and repeated-error accounting)**.** *Under Proposition 40, Lemma 41, Proposition 30, and target support from Proposition 39, fix $d \in \mathcal{H}_\delta$. For every history h_t with $u := \chi_t(h_t) \in S_{\mu,t}$,*

$$V_t^\star(u) - J_t^d(h_t) = \mathbb{E}_{t,h_t}^{\widehat{\pi}^d} \left[\sum_{s=t}^{H} \left(V_s^\star(U_s) - Q_s^\star(U_s, \widehat{g}_s^d(U_s)) \right) \right]. \tag{86}$$

Consequently,

$$0 \le V_t^\star(u) - J_t^d(h_t) \le \sum_{s=t}^{H} \min\{B_s, 2\mathfrak{B}_s\} \le 2 \sum_{s=t}^{H} \mathfrak{B}_s. \tag{87}$$

The conservative final term has the exact expansion

$$2 \sum_{s=t}^{H} \mathfrak{B}_s = 2 \sum_{j=t}^{H} (j - t + 1) \left(\max_{D \in \{E, \bar{E}\}} a_{D,j} + \max_{D \in \{E, \bar{E}\}} s_{D,j} \right). \tag{88}$$

The same bound holds after averaging over any supported fresh-rollout start law. In particular, for the learned policy's own occupancies,

$$\begin{aligned} &\mathbb{E}_{U_t \sim d_t^{\widehat{\pi}^d}} \left[V_t^\star(U_t) - V_t^{\widehat{\pi}^d, \chi}(U_t) \right] \\ &\quad = \sum_{s=t}^{H} \mathbb{E}_{U_s \sim d_s^{\widehat{\pi}^d}} \left[V_s^\star(U_s) - Q_s^\star(U_s, \widehat{g}_s^d(U_s)) \right] \\ &\quad \le 2 \sum_{s=t}^{H} \mathfrak{B}_s. \end{aligned} \tag{89}$$

The first line is the actual expected fresh-rollout return gap.

Proof. Define the proof-local value gap and optimal advantage

$$\delta_s^d(u) := V_s^\star(u) - V_s^{\widehat{\pi}^d, \chi}(u), \qquad A_s^d(u) := V_s^\star(u) - Q_s^\star(u, \widehat{g}_s^d(u)). \tag{90}$$

Subtract equation 74 from the optimal Bellman recursion. For every master atom,

$$\delta_s^d(u) = A_s^d(u) + \mathbb{E}_{U' \sim P_s^\chi(\cdot | u, \widehat{g}_s^d(u))} \left[\delta_{s+1}^d(U') \right], \qquad \delta_{H+1}^d \equiv 0. \tag{91}$$

Proposition 40 identifies every policy value in this recursion with the actual conditional fresh-rollout value and identifies the transition with the true next-master transition under the exact learned policy. Finite iteration and the tower property therefore give equation 86 from any initialized history; no auxiliary master process replaces the rollout.

Starting from support, Proposition 30 keeps every visited U_s in $S_{\mu,s}$. Lemma 41 then gives

$$0 \le A_s^d(U_s) \le \min\{B_s, 2\mathfrak{B}_s\}$$

at every stage, proving equation 87. Event and non-event probabilities sum to one, so no additional factor for the two strata occurs.

Expanding the definitions and interchanging two finite sums gives

$$\begin{aligned} 2\sum_{s=t}^{H} \mathfrak{B}_s &= 2\sum_{s=t}^{H}\left[\sum_{j=s}^{H} \max_D a_{D,j} + \sum_{j=s}^{H} \max_D s_{D,j}\right] \\ &= 2\sum_{j=t}^{H}(j-t+1)\left(\max_D a_{D,j} + \max_D s_{D,j}\right), \end{aligned} \tag{92}$$

where the two maxima are separately over $D \in \{E, \bar{E}\}$. They cannot be replaced by $\max_D(a_{D,j} + s_{D,j})$, because their maximizing strata may differ.

By the preliminary declaration that Π_{tar} contains every fixed-data learned policy, the total policy constructed here belongs to Π_{tar}. Target support from Proposition 39 places the learned policy's target occupancy on $S_{\mu,t}$. Averaging the exact historywise identity and using the learned policy's own future occupancies proves equation 89; occupancies of an unrelated reference policy do not appear.

The inherited stagewise envelopes make every dependence in the full-regret bound explicit:

$$\begin{aligned} 2\sum_{s=t}^{H} \mathfrak{B}_s \le\ & 2c_b \sum_{j=t}^{H}(j-t+1) \max_{D\in\{E,\bar{E}\}} \left(\varepsilon^r_{D,j} + B_{j+1}\varepsilon^p_{D,j}\right) \\ & + 2c_s \sum_{j=t}^{H}(j-t+1) \max_{D\in\{E,\bar{E}\}} \left[(1+B_{j+1})\sqrt{\frac{\Gamma_{\mathrm{dev}}(\delta/2)}{n q^{\mathrm{audit}}_{D,j}}}\right]. \end{aligned} \tag{93}$$

An empty stratum takes its setting-defined zero branch. On a nonempty active stratum,

$$s_{E,j} \le c_s(1+B_{j+1})\sqrt{\frac{\Gamma_{\mathrm{dev}}(\delta/2)}{n\rho^{\mu}_j \lambda_j \alpha^{\mathrm{audit}}_j \beta_j}}. \tag{94}$$

Finally, expanding the finite piecewise target-transfer quantity gives

$$\begin{aligned} \mathfrak{V}^{\pi}_t = \sum_{s=t}^{H} & \sum_{\substack{D\in\{E,\bar{E}\}\\ S_{D,s}\neq\varnothing}} C^{\mathrm{atom}}_{D,s} \sum_{\substack{u\in S_{D,s},\, a\in\mathcal{A}\\ d^{\mu}_s(u,a)>0}} d^{\mu}_s(u,a) \\ & \cdot \left[c_b\left(\varepsilon^r_s(u,a) + B_{s+1}\varepsilon^p_s(u,a)\right) + c_s(1+B_{s+1})\sqrt{\frac{\Gamma_{\mathrm{dev}}(\delta/2)}{n d^{\mu}_s(u,a)}}\right]. \end{aligned} \tag{95}$$

Every omitted empty-stratum contribution is the already-defined zero branch, and each omitted row within a nonempty stratum has zero behavior mass. Thus no null-row loss is evaluated. The target-shift coefficient appears only in this formal rowwise transfer; it does not multiply the uniform-on-support performance-difference bound. The four named results in this subsection therefore supply, respectively, typed target transfer, exact learned-policy rollout semantics, pointwise and event regret, and full true-rollout regret with its exact repeated local charges. □

A.5.8 Effective-Gap Cue Retention

Proposition 43 (Strict effective-gap dominance and greedy correctness)**.** *Under Assumption 15 and Proposition 38, suppose*

$$\Delta_{t,\mathrm{eff}} := \Delta_t - 2\mathfrak{A}_t > 0, \qquad 2\mathfrak{S}_t < \Delta_{t,\mathrm{eff}}. \tag{96}$$

Then, on $\mathcal{H}_\delta$, for every $u \in S_{\mu,t} \cap \mathcal{U}_{E,t}$ and every $a \neq a^\star_t(u)$,

$$\begin{aligned} &\widehat{Q}_t(\widehat{\phi}_t(u), a^\star_t(u)) - \widehat{Q}_t(\widehat{\phi}_t(u), a) \\ &\qquad \ge \Delta_t - 2\mathfrak{B}_t = \Delta_{t,\mathrm{eff}} - 2\mathfrak{S}_t > 0. \end{aligned} \tag{97}$$

Consequently, $a_t^\star(u)$ is the unique maximizer of the learned action-value vector at $\widehat{\phi}_t(u)$, and

$$\widehat{g}_t(u) = a_t^\star(u). \tag{98}$$

In particular, this holds at every target-reachable supported active atom.

Proof. Work on $\mathcal{H}_\delta$, and fix a supported active atom u and $a \neq a_t^\star(u)$. With D_t defined in equation 50, Proposition 38 gives

$$D_t \leq \mathfrak{A}_t + \mathfrak{S}_t = \mathfrak{B}_t. \tag{99}$$

Applying this error once at the true optimal action and once at the competitor gives

$$\widehat{Q}_t(\widehat{\phi}_t(u), a_t^\star(u)) \geq Q_t^\star(u, a_t^\star(u)) - D_t \tag{100}$$

and

$$\widehat{Q}_t(\widehat{\phi}_t(u), a) \leq Q_t^\star(u, a) + D_t. \tag{101}$$

Subtracting, applying the raw supported active gap from Assumption 15, and then using equation 99, we obtain

$$\begin{aligned} &\widehat{Q}_t(\widehat{\phi}_t(u), a_t^\star(u)) - \widehat{Q}_t(\widehat{\phi}_t(u), a) \\ &\quad \geq Q_t^\star(u, a_t^\star(u)) - Q_t^\star(u, a) - 2D_t \\ &\quad \geq \Delta_t - 2D_t \\ &\quad \geq \Delta_t - 2(\mathfrak{A}_t + \mathfrak{S}_t) \\ &\quad = \Delta_{t,\mathrm{eff}} - 2\mathfrak{S}_t. \end{aligned} \tag{102}$$

The last expression is strictly positive by equation 96. This applies to every action other than $a_t^\star(u)$, so that action is the unique learned maximizer. The exact learned selector therefore satisfies equation 98; its tie rule is not invoked because the learned inequalities are strict. □

Proposition 44 (Zero target event regret under effective-gap retention)**.** *Under Propositions 43 and 39, on $\mathcal{H}_\delta$, every $\pi \in \Pi_{\mathrm{tar}}$ satisfies*

$$\mathcal{R}_{E,t}^\pi(\widehat{g}_t) = 0. \tag{103}$$

If $\rho_t^\pi > 0$, then

$$\frac{\mathcal{R}_{E,t}^\pi(\widehat{g}_t)}{\rho_t^\pi} = 0.$$

If $\rho_t^\pi = 0$, the unnormalized event contribution is exactly zero and no ratio is formed.

Proof. Proposition 39 puts every atom with $d_t^\pi(u) > 0$ on behavior support. If the active behavior stratum is empty, the same proposition proves target emptiness before any loss is evaluated. Thus event regret is the finite supported sum

$$\mathcal{R}_{E,t}^\pi(\widehat{g}_t) = \sum_{\substack{u \in S_{\mu,t} \cap \mathcal{U}_{E,t} \\ d_t^\pi(u) > 0}} d_t^\pi(u) \left[Q_t^\star(u, a_t^\star(u)) - Q_t^\star(u, \widehat{g}_t(u)) \right]. \tag{104}$$

Proposition 43 gives $\widehat{g}_t(u) = a_t^\star(u)$ at every atom in the sum, so every bracket is zero. This proves equation 103. Division occurs only when the event occupancy is positive; the zero-occupancy branch uses the unnormalized-zero convention. Because correctness is pointwise and uniform on supported active atoms, no target-shift coefficient enters the result. □

Proposition 45 (Non-aliasing of opposite-action supported cues)**.** *Under Proposition 43, on $\mathcal{H}_\delta$, if*

$$u, v \in S_{\mu,t} \cap \mathcal{U}_{E,t}, \qquad a_t^\star(u) \neq a_t^\star(v),$$

then

$$\widehat{\phi}_t(u) \neq \widehat{\phi}_t(v). \tag{105}$$

Thus every target-reachable supported action-flipping cue pair is retained. The conclusion is support restricted and makes no statement about off-support atoms or pairs with the same optimal action.

Proof. Suppose instead that $\widehat{\phi}_t(u) = \widehat{\phi}_t(v) = z$. Both atoms use the identical learned action-value vector

$$(\widehat{Q}_t(z, a))_{a\in\mathcal{A}}.$$

Proposition 43 says that $a_t^\star(u)$ is the unique maximizer of this vector and, when applied to v, that $a_t^\star(v)$ is its unique maximizer. A finite vector cannot have two distinct unique maximizers, contradicting $a_t^\star(u) \neq a_t^\star(v)$. Hence equation 105 holds.

The three conclusions in this subsection use the exact identity

$$\Delta_t - 2\mathfrak{B}_t = \Delta_t - 2\mathfrak{A}_t - 2\mathfrak{S}_t = \Delta_{t,\mathrm{eff}} - 2\mathfrak{S}_t. \tag{106}$$

Since $\mathfrak{S}_t \geq 0$, the margin conditions are equivalent to $2\mathfrak{B}_t < \Delta_t$, but the bias-adjusted form keeps the irreducible tolerance bias explicit. At $\Delta_{t,\mathrm{eff}} \leq 0$, or at equality $2\mathfrak{S}_t = \Delta_{t,\mathrm{eff}}$, the strict argument does not activate. When raw tolerances vanish, $\mathfrak{A}_t = 0$ and the effective gap reduces to the raw gap; if statistical error also vanishes, correct action, zero event regret, and opposite-action non-aliasing hold exactly. □

A.5.9 Exact Recovery of the Supported Bellman Quotient

Use the terminal convention $S_{\mu,H+1} := \mathcal{U}_{H+1}$, and define the supported label sets

$$\mathcal{Z}_s^{\mathrm{B}} := \varphi_s^{\mathrm{B}}(S_{\mu,s}), \qquad \widehat{\mathcal{Z}}_s := \widehat{\phi}_s(S_{\mu,s}). \tag{107}$$

For bookkeeping within this subsection, put

$$\mathfrak{b}_s^{\mathrm{loc}} := \max_{D\in\{E,\bar{E}\}} a_{D,s} + \max_{D\in\{E,\bar{E}\}} s_{D,s}. \tag{108}$$

The accumulated-sum definitions give

$$\mathfrak{b}_s^{\mathrm{loc}} \leq \mathfrak{A}_s + \mathfrak{S}_s = \mathfrak{B}_s. \tag{109}$$

For a positive consumed count, abbreviate

$$R_s^r(w, a) := \mathrm{rad}_r(N_s(w, a)), \qquad R_s^p(w, a) := \mathrm{rad}_p(N_s(w, a)). \tag{110}$$

Assumption 12 supplies, at each recovery stage, a candidate whose fibers on $S_{\mu,s}$ are exactly the classes of $\sim_s^{\mathrm{B}}$. Denote one such proof-local witness by ϕ_s^{quo}. Its supported labels are immaterial and its off-support extension is the common canonical one, so

$$K_{\mu,s}(\phi_s^{\mathrm{quo}}) = |\mathcal{Z}_s^{\mathrm{B}}|. \tag{111}$$

Lemma 46 (Support-restricted transfer through a quotient-block relabeling)**.** *Under Proposition 30, fix $s \in \{t, \dots, H\}$. Suppose there is a bijection*

$$b_{s+1} : \mathcal{Z}_{s+1}^{\mathrm{B}} \longrightarrow \widehat{\mathcal{Z}}_{s+1} \tag{112}$$

such that

$$\widehat{\phi}_{s+1}(w) = b_{s+1}(\varphi_{s+1}^{\mathrm{B}}(w)), \qquad w \in S_{\mu,s+1}. \tag{113}$$

Then, for every $u \in S_{\mu,s}$ and $a \in \mathcal{A}$,

$$(\widehat{\phi}_{s+1})_{\#}P_s^\chi(\cdot \mid u, a) = (b_{s+1})_{\#}\left[(\varphi_{s+1}^{\mathrm{B}})_{\#}P_s^\chi(\cdot \mid u, a)\right]. \tag{114}$$

Consequently, for $u, v \in S_{\mu,s}$ and every action,

$$\begin{aligned}&\mathrm{TV}\Big((\widehat{\phi}_{s+1})_{\#}P_s^\chi(\cdot \mid u, a), (\widehat{\phi}_{s+1})_{\#}P_s^\chi(\cdot \mid v, a)\Big)\\ &\quad = \mathrm{TV}\big((\varphi_{s+1}^{\mathrm{B}})_{\#}P_s^\chi(\cdot \mid u, a), (\varphi_{s+1}^{\mathrm{B}})_{\#}P_s^\chi(\cdot \mid v, a)\big).\end{aligned} \tag{115}$$

Proof. Proposition 30 gives $P_s^\chi(S_{\mu,s+1} \mid u, a) = 1$. Thus, for every $\widehat{z} \in \widehat{\mathcal{Z}}_{s+1}$,

$$\begin{aligned}&[(\widehat{\phi}_{s+1})_{\#}P_s^\chi(\cdot \mid u, a)](\widehat{z})\\ &\quad = \sum_{w\in S_{\mu,s+1}} \mathbf{1}\{\widehat{\phi}_{s+1}(w) = \widehat{z}\}P_s^\chi(w \mid u, a)\\ &\quad = \sum_{w\in S_{\mu,s+1}} \mathbf{1}\{b_{s+1}(\varphi_{s+1}^{\mathrm{B}}(w)) = \widehat{z}\}P_s^\chi(w \mid u, a),\end{aligned}$$

which is equation 114.

If μ_1, μ_2 are probability laws on $\mathcal{Z}^{\mathrm{B}}_{s+1}$, bijectivity of b_{s+1} and the half-ℓ_1 convention give

$$\begin{aligned} &2\,\mathrm{TV}((b_{s+1})_{\#}\mu_1, (b_{s+1})_{\#}\mu_2) \\ &\quad= \sum_{\widehat{z}\in\widehat{\mathcal{Z}}_{s+1}} \left|\mu_1(b^{-1}_{s+1}(\widehat{z})) - \mu_2(b^{-1}_{s+1}(\widehat{z}))\right| \\ &\quad= \sum_{z\in\mathcal{Z}^{\mathrm{B}}_{s+1}} |\mu_1(z) - \mu_2(z)|. \end{aligned}$$

Apply this identity to the two quotient pushforwards in equation 114 to obtain equation 115. At $s = H$, the stage-$H+1$ label sets are singletons and their unique bijection supplies the hypothesis. Arbitrary off-support labels carry no true mass and do not enter the calculation. □

Proposition 47 (Feasibility of the supported quotient under recovered downstream blocks)**.** *Under Assumptions 12 and 13, Corollary 29, Lemma 33, and Lemma 46, work on $\mathcal{H}_\delta$. Fix $s \in \{t, \dots, H\}$ and suppose equation 112–equation 113 hold. Then ϕ^{quo}_s is feasible for the learner's stage-s comparison with the realized downstream code $\widehat{\phi}_{s+1}$. Consequently,*

$$K_{\mu,s}(\widehat{\phi}_s) \le K_{\mu,s}(\phi^{\mathrm{quo}}_s) = |\mathcal{Z}^{\mathrm{B}}_s|. \tag{116}$$

Proof. Let $u, v \in S_{\mu,s}$ lie in one supported quotient block, and fix $a \in \mathcal{A}$. By the definition of $u \sim^{\mathrm{B}}_s v$,

$$r_s(u,a) = r_s(v,a) \tag{117}$$

and

$$(\varphi^{\mathrm{B}}_{s+1})_{\#}P^{\chi}_s(\cdot \mid u,a) = (\varphi^{\mathrm{B}}_{s+1})_{\#}P^{\chi}_s(\cdot \mid v,a). \tag{118}$$

Lemma 46 converts the latter identity into

$$(\widehat{\phi}_{s+1})_{\#}P^{\chi}_s(\cdot \mid u,a) = (\widehat{\phi}_{s+1})_{\#}P^{\chi}_s(\cdot \mid v,a). \tag{119}$$

On $\mathcal{H}_\delta$, Corollary 29 makes both endpoint counts positive, and the deviation component of the same event gives, for $w \in \{u, v\}$,

$$|\widehat{r}_s(w,a) - r_s(w,a)| \le R^r_s(w,a) \tag{120}$$

and

$$\mathrm{TV}\Big(\widehat{P}^{\chi}_s(\cdot \mid w,a), P^{\chi}_s(\cdot \mid w,a)\Big) \le R^p_s(w,a). \tag{121}$$

Equation equation 117 and the scalar triangle inequality give

$$\begin{aligned} |\widehat{r}_s(u,a) - \widehat{r}_s(v,a)| &\le R^r_s(u,a) + R^r_s(v,a) \\ &\le \tau^r_s(u,v,a) + R^r_s(u,a) + R^r_s(v,a), \end{aligned} \tag{122}$$

where the final inequality uses only nonnegativity of the primitive tolerance.

Lemma 33 applies to the same realized downstream code and equation 121. Each empirical pushforward is within $R^p_s(w,a)$ of its true counterpart. Combining the two endpoint bounds with equation 119 gives

$$\begin{aligned} &\mathrm{TV}\Big((\widehat{\phi}_{s+1})_{\#}\widehat{P}^{\chi}_s(\cdot \mid u,a), (\widehat{\phi}_{s+1})_{\#}\widehat{P}^{\chi}_s(\cdot \mid v,a)\Big) \\ &\quad\le R^p_s(u,a) + R^p_s(v,a) \\ &\quad\le \tau^p_s(u,v,a) + R^p_s(u,a) + R^p_s(v,a). \end{aligned} \tag{123}$$

Every action comparison for a pair merged by ϕ^{quo}_s therefore passes the learner's exact feasibility threshold. No off-support pair is tested, so the common canonical extension adds no condition. The selected candidate minimizes $K_{\mu,s}$ over the nonempty feasible set, which proves equation 116. At $s = H$, the pushed laws are singletons and the same argument has no continuation residual. □

Proposition 48 (Separation excludes every cross-quotient merge)**.** *Under Proposition 35, Proposition 38, and Lemma 46, work on $\mathcal{H}_\delta$. Fix $s \in \{t, \dots, H\}$, suppose equation 112–equation 113 hold, and assume*

$$4\mathfrak{B}_s < \zeta_s. \tag{124}$$

Then, for $u, v \in S_{\mu,s}$,

$$\widehat{\phi}_s(u) = \widehat{\phi}_s(v) \quad \Longrightarrow \quad u \sim_s^{\mathrm{B}} v. \tag{125}$$

Thus the learned supported partition refines the supported Bellman quotient.

Proof. If the supported quotient has one block, there is no cross-quotient pair and equation 125 is immediate; this is the branch $\zeta_s = +\infty$. Otherwise, suppose for contradiction that $\widehat{\phi}_s$ merges supported u, v with $u \not\sim_s^{\mathrm{B}} v$. For every action, equation 44 and equation 108 give

$$\begin{aligned} &|r_s(u,a) - r_s(v,a)| \\ &\quad + B_{s+1}\,\mathrm{TV}\Big((\widehat{\phi}_{s+1})_{\#}P_s^{\chi}(\cdot \mid u,a), (\widehat{\phi}_{s+1})_{\#}P_s^{\chi}(\cdot \mid v,a)\Big) \\ &\quad \le 4\mathfrak{b}_s^{\mathrm{loc}}. \end{aligned} \tag{126}$$

Lemma 46 replaces the TV term exactly by the corresponding $\varphi_{s+1}^{\mathrm{B}}$-pushforward TV. Maximizing over actions and applying equation 109 gives

$$d_s^{\mathrm{B}}(u,v) \le 4\mathfrak{b}_s^{\mathrm{loc}} \le 4\mathfrak{B}_s < \zeta_s. \tag{127}$$

The definition of ζ_s gives $d_s^{\mathrm{B}}(u,v) \ge \zeta_s$ for every cross-quotient pair, a contradiction. The factor four is exactly the conservative factor already present in the selected-pair signature; it is not duplicated. □

Proposition 49 (Minimum supported cardinality closes stagewise recovery)**.** *Under Propositions 47 and 48, at stage s,*

$$\widehat{\phi}_s(u) = \widehat{\phi}_s(v) \quad \Longleftrightarrow \quad u \sim_s^{\mathrm{B}} v, \qquad u, v \in S_{\mu,s}, \tag{128}$$

and

$$K_{\mu,s}(\widehat{\phi}_s) = |\mathcal{Z}_s^{\mathrm{B}}|. \tag{129}$$

Moreover, there is a unique bijection

$$b_s : \mathcal{Z}_s^{\mathrm{B}} \longrightarrow \widehat{\mathcal{Z}}_s \tag{130}$$

satisfying

$$\widehat{\phi}_s(u) = b_s(\varphi_s^{\mathrm{B}}(u)), \qquad u \in S_{\mu,s}. \tag{131}$$

Proof. Proposition 48 places each learned supported block inside a unique quotient block. Mapping each learned block to its containing quotient block is a surjection between finite block sets: every quotient block contains a supported atom, and that atom belongs to a learned block. Therefore

$$K_{\mu,s}(\widehat{\phi}_s) \ge |\mathcal{Z}_s^{\mathrm{B}}|. \tag{132}$$

Proposition 47 gives the reverse inequality, so equation 129 holds. The finite surjection is consequently a bijection, and every quotient block contains exactly one learned block. This proves the reverse implication in equation 128; the forward implication is the established refinement.

In the one-quotient-block case, the same argument is explicitly

$$1 \le K_{\mu,s}(\widehat{\phi}_s) \le K_{\mu,s}(\phi_s^{\mathrm{quo}}) = 1,$$

because $S_{\mu,s}$ is nonempty. Thus the convention $\zeta_s = +\infty$ leaves no hidden cardinality branch.

For $z \in \mathcal{Z}_s^{\mathrm{B}}$, choose any $u \in S_{\mu,s}$ with $\varphi_s^{\mathrm{B}}(u) = z$ and set

$$b_s(z) := \widehat{\phi}_s(u).$$

Relation equality makes this independent of the chosen atom. Distinct quotient labels correspond to distinct learned blocks, so b_s is injective; every learned supported label has a supported preimage, so it is surjective. Equation equation 131 follows, and it uniquely determines b_s on $\mathcal{Z}_s^{\mathrm{B}}$. □

Lemma 50 (Bellman values factor through the supported quotient)**.** *Under Assumption 6, Proposition 30, and the setting-defined Bellman quotient, for every $s \in \{t, \dots, H\}$,*

$$u, v \in S_{\mu,s}, \quad u \sim_s^{\mathrm{B}} v \quad \Longrightarrow \quad Q_s^\star(u,a) = Q_s^\star(v,a) \quad \text{for every } a \in \mathcal{A}. \tag{133}$$

For the same supported pair,

$$V_s^\star(u) = V_s^\star(v), \qquad a_s^\star(u) = a_s^\star(v). \tag{134}$$

Proof. We argue backward on supported atoms. At $H+1$, the supported quotient has one block and $V_{H+1}^\star = 0$, so the value is constant on that block. Suppose $V_{s+1}^\star$ is constant on every supported $\sim_{s+1}^{\mathrm{B}}$-class. Define

$$\overline{V}_{s+1}^{\mathrm{B}}(z) := V_{s+1}^\star(w) \quad \text{for any } w \in S_{\mu,s+1} \text{ with } \varphi_{s+1}^{\mathrm{B}}(w) = z. \tag{135}$$

The induction hypothesis makes this well-defined. Proposition 30 then gives, for supported u and every action,

$$\mathbb{E}_{P_s^{\mathrm{X}}(\cdot|u,a)} V_{s+1}^\star(U') = \mathbb{E}_{(\varphi_{s+1}^{\mathrm{B}})_\# P_s^{\mathrm{X}}(\cdot|u,a)} \overline{V}_{s+1}^{\mathrm{B}}. \tag{136}$$

If $u, v \in S_{\mu,s}$ and $u \sim_s^{\mathrm{B}} v$, the quotient definition supplies equal rewards and equal next-quotient laws for every action. Substitution into the finite-horizon Bellman recursion, using equation 136, proves equation 133. Maximizing the identical action-value vectors proves equality of $V_s^\star$, closing the induction. Applying the same fixed total action order to the identical vectors selects the same first maximizer, proving equation 134, including true ties. No assertion is made off behavior support. □

Proposition 51 (Exact supported Bellman-quotient recovery)**.** *Under Assumptions 12 and 13, and under Corollary 29, Propositions 30, 31, 35, and 38, suppose the candidate class contains the supported Bellman quotient at every recovered stage and*

$$4\mathfrak{B}_s < \zeta_s, \qquad s = t, \dots, H. \tag{137}$$

Then, on $\mathcal{H}_\delta$, for every $s \geq t$ there is a unique bijection

$$b_s : \mathcal{Z}_s^{\mathrm{B}} \longrightarrow \widehat{\mathcal{Z}}_s$$

such that

$$\widehat{\phi}_s(u) = b_s(\varphi_s^{\mathrm{B}}(u)), \qquad u \in S_{\mu,s},$$

or, equivalently,

$$\widehat{\phi}_s(u) = \widehat{\phi}_s(v) \quad \Longleftrightarrow \quad u \sim_s^{\mathrm{B}} v \qquad (u, v \in S_{\mu,s}). \tag{138}$$

In particular,

$$K_{\mu,s}(\widehat{\phi}_s) = |\mathcal{Z}_s^{\mathrm{B}}|.$$

The learned partition therefore merges every supported task-equal Bellman-null nuisance fiber and never merges two supported atoms with different ordered optimal actions. For every data set and every $u \notin S_{\mu,s}$,

$$\widehat{\phi}_s(u) = \perp_{s,u}. \tag{139}$$

This final identity is canonical protocol preservation, not off-support statistical recovery.

Proof. At stage $H+1$, the learned code and population quotient both have one supported terminal block, so there is a unique b_{H+1} satisfying equation 113. Proceed backward from $s = H$ to $s = t$. Given b_{s+1}, Lemma 46 aligns the learned and quotient next-law interfaces exactly. Proposition 47 makes the quotient candidate feasible and yields the upper supported block-count bound. Proposition 48 combines

$$4\mathfrak{b}_s^{\mathrm{loc}} \leq 4\mathfrak{B}_s < \zeta_s$$

with the selected-pair signature to exclude every cross-quotient merge. Proposition 49 then combines refinement with minimum supported cardinality, eliminates every within-quotient split, and produces b_s. This closes the induction and proves equation 138 simultaneously for all $s \geq t$.

Proposition 31 and relation equality show that every supported task-equal nuisance fiber is merged by the learned partition. Lemma 50 shows that supported quotient-equivalent atoms have the same ordered optimal action; hence atoms with different ordered optimal actions cannot be merged. This

is a partition-preservation conclusion from the exact quotient signature and does not assert that the learned greedy action is correct.

Finally, Assumption 12 makes equation 139 deterministic and candidate-independent. The labels $\perp_{s,u}$ are injective, disjoint from supported labels, and absent from $K_{\mu,s}$. Therefore they neither finance nor obstruct supported recovery, and they carry no learned population-quotient meaning.

If every relevant tolerance and statistical charge vanishes, then $\mathfrak{B}_s = 0$. With more than one supported quotient block, finiteness makes $\zeta_s > 0$; with one block, $\zeta_s = +\infty$. Thus the exact/noiseless specialization satisfies equation 137, recovers the supported quotient exactly, merges supported nuisance fibers, preserves distinct ordered actions, and leaves every canonical off-support label unchanged. □

A.5.10 Universal Raw-Gap Action and Regret Obstruction

Proposition 52 (Feasible binary lower family with fixed occupancy)**.** *Under Assumption 7, for every*

$$0 < q \le \frac{1}{6}, \qquad 0 < \Delta \le \frac{1}{4},$$

there are two bounded terminal experiments P_+ and P_-, with common action set $\{a_0, a_1\}$, satisfying all of the following.

(a) *Their action and observation alphabets, master/event/cue map, action labels and order, state–action law, prefix, nonhard reward laws, and any appended nuisance law are identical. Only the hard atom's two action-conditioned reward laws depend on the model.*

(b) *Their audit masses obey*

$$q_E^{\mathrm{audit}} = q, \qquad q_{\bar{E}}^{\mathrm{audit}} \ge q,$$

where an empty non-event stratum has its prescribed value $+\infty$.

(c) *A pre-action hard atom $u_\star$ has mass $2q$, uniform behavior actions, and raw gap Δ. Its optimal action is a_0 under P_+ and a_1 under P_-.*

(d) *Two active anchors have common, fixed, opposite optimal actions.*

(e) *A model-common action-independent prefix fixes, before training,*

$$\eta_t^\pi := d_t^\pi(u_\star) \ge \frac{1}{4}\rho_t^\pi$$

for every evaluation policy π.

The range $q \le 1/6$ is the full feasible range for this binary construction with one hard atom and two required active anchors when every active atom–action mass is at least q.

Proof. Put

$$p_\Delta := \frac{1+\Delta}{2}, \qquad r_\Delta := \frac{1-\Delta}{2}.$$

Then $p_\Delta, r_\Delta \in [0,1]$ and $p_\Delta - r_\Delta = \Delta$. If the hard atom $u_\star$ and two anchors u_+, u_- each have two audited actions of probability at least q, necessarily

$$1 \ge \sum_{u \in \{u_\star, u_+, u_-\}} \sum_{a \in \{a_0, a_1\}} d_t^\mu(u,a) \ge 6q. \tag{140}$$

Thus $q \le 1/6$ is necessary.

Use the same finite alphabets, master map, event/cue map, action labels, and action order in both experiments. The behavior action is uniform at every supported atom. Declare $u_\star, u_+, u_-$ active, take the cue to be the atom label, and use the terminal Bernoulli reward means

	$u_\star$	u_+	u_-
P_+	(p_Δ, r_Δ)	(p_Δ, r_Δ)	(r_Δ, p_Δ)
P_-	(r_Δ, p_Δ)	(p_Δ, r_Δ)	(r_Δ, p_Δ).

(141)

The hard optimal action is swapped, the two anchors retain the common opposite optimal actions a_0 and a_1, and every active raw gap is Δ. All terminal dynamics are identical and trivial.

For $0 < q \le 1/8$, take

$$d_t(u_\star) = d_t(u_+) = d_t(u_-) = 2q, \qquad d_t(v_0) = 1 - 6q,$$

where v_0 is the sole non-event atom and has the same zero reward law under both models. Uniform actions give

$$q_E^{\mathrm{audit}} = q, \qquad q_{\bar{E}}^{\mathrm{audit}} = \frac{1-6q}{2} \ge q,$$

and the model-common occupancy satisfies

$$\eta_t^\pi = 2q, \qquad \rho_t^\pi = 6q, \qquad \frac{\eta_t^\pi}{\rho_t^\pi} = \frac{1}{3}. \tag{142}$$

For $1/8 < q \le 1/6$, omit the non-event atom and take

$$d_t(u_\star) = d_t(u_+) = 2q, \qquad d_t(u_-) = 1 - 4q.$$

The last atom–action mass is $(1-4q)/2 \ge q$ exactly when $q \le 1/6$, while each action at the first two atoms has mass q. Hence

$$q_E^{\mathrm{audit}} = q, \qquad q_{\bar{E}}^{\mathrm{audit}} = +\infty, \qquad \eta_t^\pi = 2q, \qquad \rho_t^\pi = 1,$$

and $\eta_t^\pi/\rho_t^\pi = 2q > 1/4$. Together with equation 142, this proves the uniform choice $\kappa = 1/4$, including $q = 1/6$.

At $t = 1$, the displayed atom law is the common initial law. At a later terminal decision stage, prepend a model-common, action-independent, zero-reward process that draws the terminal atom independently of every earlier history and action. Thus $C_t = U_t$ may be unpredictable from H_{t-1} while remaining observed before A_t, as required by Assumption 7. The stage-t occupancy is common to both models and all evaluation policies and is fixed before the training sample or learner output.

Finally, append any finite nuisance variable whose common law is independent of the prefix, task atom, action, and reward. The master coordinate retains the task atom and ignores the nuisance. Master homogeneity and every audit mass above are unchanged, and the nuisance contains no model information. This attains every q permitted by equation 140. □

Lemma 53 (Exact divergence of the signed Bernoulli swap)**.** *Under Assumption 7 and Proposition 52, the laws of n independent training episodes satisfy*

$$\begin{aligned} D_{\mathrm{KL}}\big(\mathsf{P}_+^{n,\mu} \big\| \mathsf{P}_-^{n,\mu}\big) &= 2nq\Delta \log \frac{1+\Delta}{1-\Delta} \\ &\le \frac{16}{3} nq\Delta^2. \end{aligned} \tag{143}$$

Every model-common prefix variable, nonhard row, and independently appended nuisance variable contributes exactly zero divergence.

Proof. For one episode, collect the common prefix, terminal atom, behavior action, and nuisance into X, and denote the terminal reward by Y. Proposition 52 makes the marginal law of X common. Conditional reward laws differ only at $(U_t, A_t) = (u_\star, a_0)$ and $(U_t, A_t) = (u_\star, a_1)$, each of probability q. Expanding the finite likelihood ratio gives

$$D_{\mathrm{KL}}(\mathsf{P}_+^\mu \| \mathsf{P}_-^\mu) = q\ \mathrm{kl}(p_\Delta \| r_\Delta) + q\ \mathrm{kl}(r_\Delta \| p_\Delta).$$

Since $1 - p_\Delta = r_\Delta$, $1 - r_\Delta = p_\Delta$, and $p_\Delta - r_\Delta = \Delta$,

$$\mathrm{kl}(p_\Delta \| r_\Delta) = p_\Delta \log \frac{p_\Delta}{r_\Delta} + r_\Delta \log \frac{r_\Delta}{p_\Delta} = \Delta \log \frac{1+\Delta}{1-\Delta}.$$

The reverse divergence has the same value. The product likelihood ratio over independent episodes is the product of the episodewise ratios, so expectation of its logarithm yields the exact factor n. Every common coordinate has likelihood ratio one and contributes zero.

Finally,

$$\log \frac{1+\Delta}{1-\Delta} = \int_{-\Delta}^{\Delta} \frac{dx}{1+x} \le \frac{2\Delta}{1-\Delta} \le \frac{8}{3}\Delta, \qquad 0 < \Delta \le \frac{1}{4}.$$

Substitution proves equation 143. □

Lemma 54 (Explicit binary testing overlap)**.** *Let P, Q be finite probability laws, and let a possibly randomized binary decision have error probabilities e_P, e_Q under the two opposite labels. Then*

$$e_P + e_Q \geq 1 - \|P - Q\|_{\mathrm{TV}}, \qquad \|P - Q\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P\|Q)}. \tag{144}$$

Consequently,

$$\max\{e_P, e_Q\} \geq \frac{1 - \sqrt{D_{\mathrm{KL}}(P\|Q)/2}}{2}$$

whenever the right-hand side is nonnegative. In particular, $D_{\mathrm{KL}}(P\|Q) \leq 1/4$ implies $\max\{e_P, e_Q\} \geq 1/4$, and therefore also the conservative bound $1/8$.

Proof. Represent the randomized decision by $\varphi : \Omega \to [0, 1]$, where $\varphi(z)$ is the probability of choosing the label that is wrong under P and correct under Q. Then

$$e_P = \mathbb{E}_P[\varphi], \qquad e_Q = \mathbb{E}_Q[1 - \varphi],$$

and hence

$$e_P + e_Q = 1 + \mathbb{E}_P[\varphi] - \mathbb{E}_Q[\varphi] \geq 1 - \|P - Q\|_{\mathrm{TV}}.$$

The last inequality follows directly from the finite signed-mass definition of total variation: the expectation difference of every $[0, 1]$-valued function has magnitude at most total variation.

We also derive the KL-to-TV bound. Let

$$B := \{z : P(z) \geq Q(z)\}, \qquad x := P(B), \quad y := Q(B).$$

Then $x - y = \|P - Q\|_{\mathrm{TV}}$. If $x, y > 0$, define on B

$$\alpha_z := \frac{P(z)}{x}, \qquad \beta_z := \frac{Q(z)}{y}.$$

Direct expansion gives

$$\sum_{z \in B} P(z) \log \frac{P(z)}{Q(z)} = x \log \frac{x}{y} + x \sum_{z \in B} \alpha_z \log \frac{\alpha_z}{\beta_z} \geq x \log \frac{x}{y}.$$

Indeed, $-\log s \geq 1 - s$ implies

$$\sum_z \alpha_z \log \frac{\alpha_z}{\beta_z} \geq \sum_z \alpha_z \left(1 - \frac{\beta_z}{\alpha_z}\right) = 0,$$

and a positive α_z with zero β_z gives $+\infty$. Applying the same calculation on B^c, with boundary cases obtained by continuity, yields

$$D_{\mathrm{KL}}(P\|Q) \geq x \log \frac{x}{y} + (1 - x) \log \frac{1 - x}{1 - y} =: \mathrm{kl}(x\|y). \tag{145}$$

For fixed $y \in (0, 1)$, set

$$f_y(x) := \mathrm{kl}(x\|y) - 2(x - y)^2.$$

One has $f_y(y) = f_y'(y) = 0$ and

$$f_y''(x) = \frac{1}{x(1 - x)} - 4 \geq 0$$

because $x(1 - x) \leq 1/4$. Thus $f_y(x) \geq 0$; boundary cases follow by continuity, while positive P-mass on zero Q-mass gives infinite KL. Combining this with equation 145 gives

$$D_{\mathrm{KL}}(P\|Q) \geq 2\|P - Q\|_{\mathrm{TV}}^2.$$

This proves equation 144. If the KL is at most $1/4$, then total variation is at most $1/(2\sqrt{2}) \leq 1/2$; the error sum is at least $1/2$, so one error is at least $1/4$. □

Proposition 55 (Raw-gap action and fixed-occupancy regret lower bound)**.** *Under Assumption 7, Proposition 52, and Lemmas 53 and 54, let any possibly randomized learner use n training episodes to report $\widehat{g}_t$, and write*

$$\widehat{a}(u_\star) := \widehat{g}_t(u_\star) \in \{a_0, a_1\}.$$

Here $\Pr_\theta$ and $\mathbb{E}_\theta$ are joint over the n-episode sample under $\mathsf{P}_\theta^{n,\mu}$, learner randomization, and any fresh model-common prefix or nuisance randomness used in evaluation. If

$$0 < q \le \frac{1}{6}, \qquad 0 < \Delta \le \frac{1}{4}, \qquad nq\Delta^2 \le \frac{3}{64}, \tag{146}$$

then

$$\sup_{\theta \in \{+,-\}} \Pr_\theta(\widehat{a}(u_\star) \neq a_\theta^\star(u_\star)) \ge \frac{1}{8}. \tag{147}$$

For at least one member θ, the expected hard-atom conditional regret is at least $\Delta/8$, and

$$\Pr_\theta\big(\mathcal{R}^\pi_{E,t}(\widehat{g}_t) \ge \eta_t^\pi \Delta\big) \ge \frac{1}{8}, \tag{148}$$

$$\mathbb{E}_\theta\big[\mathcal{R}^\pi_{E,t}(\widehat{g}_t)\big] \ge \frac{1}{8}\eta_t^\pi \Delta \ge \frac{1}{32}\rho_t^\pi \Delta. \tag{149}$$

The conclusion concerns action identification and regret only; it does not force an encoder to alias, merge, or refine any atoms.

Proof. Adding fresh common prefix or nuisance randomness multiplies both paired laws by the same conditional factor. It therefore leaves the zero-divergence coordinates in Lemma 53 and the testing interface in Lemma 54 unchanged. From equation 143 and equation 146,

$$D_{\mathrm{KL}}\big(\mathsf{P}_+^{n,\mu} \big\| \mathsf{P}_-^{n,\mu}\big) \le \frac{16}{3} nq\Delta^2 \le \frac{16}{3}\frac{3}{64} = \frac{1}{4}.$$

Map the learner's action to a binary test that outputs $+$ after a_0 and $-$ after a_1. These are the correct hard actions under P_+ and P_-, respectively. Lemma 54 gives the stronger conclusion that one action-error probability is at least $1/4$, proving equation 147 with the conservative constant $c_1 = 1/8$.

At $u_\star$, the two actions have raw value gap exactly Δ. Thus, for either model,

$$\begin{aligned} &Q^\star_{t,\theta}(u_\star, a^\star_\theta(u_\star)) - Q^\star_{t,\theta}(u_\star, \widehat{a}(u_\star)) \\ &\qquad = \Delta\, \mathbf{1}\{\widehat{a}(u_\star) \neq a^\star_\theta(u_\star)\}\,. \end{aligned} \tag{150}$$

For the member whose error probability is at least $1/4$, the expected hard regret is at least $\Delta/4$, hence at least $c_2\Delta = \Delta/8$.

The prefix in Proposition 52 fixes η_t^π and ρ_t^π before training and identically in both models. All optimal-action advantages are nonnegative, so every realized learner output satisfies

$$\mathcal{R}^\pi_{E,t}(\widehat{g}_t) \ge \eta_t^\pi \Delta\, \mathbf{1}\{\widehat{a}(u_\star) \neq a^\star_\theta(u_\star)\}\,. \tag{151}$$

The hard-error event is therefore contained in the event in equation 148. Taking expectation and using the stronger $1/4$ testing probability gives

$$\mathbb{E}_\theta[\mathcal{R}^\pi_{E,t}(\widehat{g}_t)] \ge \frac{1}{4}\eta_t^\pi \Delta \ge \frac{1}{8}\eta_t^\pi \Delta \ge \frac{1}{32}\rho_t^\pi \Delta,$$

where the last inequality uses $\eta_t^\pi \ge \rho_t^\pi/4$. This proves equation 148–equation 149.

The construction and the four results above establish the complete lower interface with the compatible constants

$$c_{\mathrm{KL}} = \frac{16}{3}, \qquad c_0 = \frac{3}{64}, \qquad c_1 = c_2 = \frac{1}{8}, \qquad \kappa = \frac{1}{4}.$$

The probability mode is worst-member binary testing and expectation, the horizon is one terminal decision possibly after a fixed finite common prefix, and the metric is hard-action error and target-occupancy regret. At $q = 0$ no positive identification claim is made, and at $\Delta = 0$ the paired laws and lower regret coincide at zero. Independent nuisance and the common prefix add no divergence. An encoder may retain every atom and nuisance value, so the obstruction cannot be strengthened to universal representation aliasing. □

A.5.11 MERGE–REFINE–ACTION TRILEMMA UNDER A SHARED SPECIFICATION

Proposition 56 (Common protocol with opposite supported quotient relations)**.** *Under Assumptions 12 and 14, for every*

$$0 < q \le \frac{1}{6}, \qquad 0 < \Delta \le \frac{1}{4},$$

there are two bounded terminal experiments $\mathsf{T}_0, \mathsf{T}_1$*, with common binary actions and common supported master atoms* $S_{\mu,t} = \{A, B, C\}$*, such that:*

(a) *B has mass $2q$, behavior is uniform on actions, every supported atom–action mass is at least q, and all atom/action, prefix, transition, and optional nuisance laws are common;*

(b) *with*

$$p_\Delta := \frac{1+\Delta}{2}, \qquad r_\Delta := \frac{1-\Delta}{2},$$

the reward-mean vectors, ordered as (a_0, a_1), are

	A	B	C
T_0	(p_Δ, r_Δ)	(r_Δ, p_Δ)	(r_Δ, p_Δ)
T_1	(p_Δ, r_Δ)	(p_Δ, r_Δ)	(r_Δ, p_Δ);

(152)

(c) *their true supported terminal Bellman quotients are*

$$\mathsf{T}_0 : \{A\} \mid \{B, C\}, \qquad \mathsf{T}_1 : \{A, B\} \mid \{C\}; \tag{153}$$

(d) *they use the same master, support, behavior law, candidate family, tolerances, reference candidate, default rows, radii, atom, action and candidate orders, labels, representatives, model construction, terminal initialization, and partition/action tie rules; and*

(e) *both quotient candidates and the identity belong to the common class, so neither the class nor a protocol field reveals the model.*

Proof. Use a one-stage terminal problem, or append it after a model-common, action-independent, zero-reward prefix. The observation reveals $U_t \in \{A, B, C\}$ before action and χ_t returns that atom. Take $E_t = 1$, $C_t = U_t$, $T_t = U_t$, and the terminal next-master singleton. An arbitrary finite independent nuisance may be appended with the same law under both members and omitted by χ_t.

Use the common atom masses

$$d_t^\mu(A) = 2q, \qquad d_t^\mu(B) = 2q, \qquad d_t^\mu(C) = 1 - 4q,$$

and choose a_0, a_1 uniformly at every atom. Then

$$d_t^\mu(A, a_j) = d_t^\mu(B, a_j) = q, \qquad d_t^\mu(C, a_j) = \frac{1-4q}{2}, \quad j \in \{0, 1\}.$$

The condition $q \le 1/6$ is exactly $(1-4q)/2 \ge q$. Hence every row has mass at least q, the least audit mass is exactly q, and the endpoint is included. Conversely, three atoms with two actions and row mass at least q require $6q \le 1$, so the construction covers its entire feasible range. Every supported atom is active, giving

$$q_{E,t}^{\mathrm{audit}} = q, \qquad q_{\bar{E},t}^{\mathrm{audit}} = +\infty.$$

Give the rewards Bernoulli laws with means equation 152. Since $0 < \Delta \le 1/4$,

$$\frac{3}{8} \le r_\Delta < p_\Delta \le \frac{5}{8},$$

so the laws are bounded and mutually absolutely continuous. The terminal transition is the same point mass for every row and model. All raw action gaps equal Δ; A prefers a_0, C prefers a_1, and B prefers a_1 in T_0 and a_0 in T_1. The pre-action cue, fixed atom rows, independent episodes, and positive row masses satisfy the corresponding master, cue, sampling, coverage, and target-support conditions.

At the terminal stage, two master atoms are Bellman-equivalent exactly when their reward means agree under every action, because the next quotient is the singleton. Thus B, C agree under T_0 and A, B agree under T_1, whereas the remaining pairs differ by Δ. In particular,

$$A \not\sim^{\mathrm{B}}_{t;0} B, \qquad A \sim^{\mathrm{B}}_{t;1} B, \tag{154}$$

where the semicolon indicates which experiment supplies the true relation.

For the common candidate family, use one fixed labeled map for each of the five partitions of $\{A, B, C\}$: the identity, $A \mid BC$, $AB \mid C$, the all-merged partition, and $AC \mid B$. Extend a larger master alphabet by common injective canonical off-support labels. Fix one atom order, action order, candidate order, block-label scheme, first-atom representative rule, and every tie rule. Use the prescribed zero reward and terminal point-mass default, unit zero-count radii, and the same positive-count radius calibration. Set every pair tolerance to zero and declare the supported identity as the common reference; since it merges no distinct supported pair, its reference conditions are vacuous. Both model-specific quotients belong to this same class, but neither is marked as the true quotient. The only model-indexed random laws seen by the learner are the two reward rows at B. □

Lemma 57 (Information budget of the common-protocol relation pair)**.** *Under Proposition 56, the laws of n independent training episodes satisfy*

$$\begin{aligned} D_{\mathrm{KL}}(\mathsf{T}_0^{n,\mu} \| \mathsf{T}_1^{n,\mu}) &= 2nq\Delta \log \frac{1+\Delta}{1-\Delta} \\ &\le \frac{16}{3} nq\Delta^2. \end{aligned} \tag{155}$$

Every common protocol field, non-B row, prefix, and independent nuisance contributes exactly zero divergence.

Proof. Protocol fields are fixed constants, so including them in the input adds no likelihood ratio. The atom/action law is common and all conditional reward laws agree except at (B, a_0) and (B, a_1), each of probability q. The one-episode likelihood-ratio expansion therefore gives

$$D_{\mathrm{KL}}(\mathsf{T}_0^{\mu} \| \mathsf{T}_1^{\mu}) = q \ \mathrm{kl}(r_\Delta \| p_\Delta) + q \ \mathrm{kl}(p_\Delta \| r_\Delta).$$

The two Bernoulli divergences both equal

$$\Delta \log \frac{p_\Delta}{r_\Delta} = \Delta \log \frac{1+\Delta}{1-\Delta}.$$

Independence multiplies the likelihood ratios, so expectation of their logarithm multiplies the one-episode divergence by n. For $0 < \Delta \le 1/4$,

$$\log \frac{1+\Delta}{1-\Delta} = \int_{-\Delta}^{\Delta} \frac{dx}{1+x} \le \frac{2\Delta}{1-\Delta} \le \frac{8}{3}\Delta.$$

This proves equation 155. Every common factor has likelihood ratio one and adds zero. □

Proposition 58 (Learned-relation testing trilemma)**.** *Under Assumptions 12 and 14, Propositions 21 and 56, and Lemmas 54 and 57, run the specified minimum-$K_{\mu,t}$ estimator on n iid episodes and define*

$$Z := \mathbf{1}\{\widehat{\phi}_t(A) = \widehat{\phi}_t(B)\}.$$

Let the model-indexed error events use the true quotient and optimal action of T_θ. Then

$$\{Z = 1\} \subseteq \mathsf{MergeErr}_0, \qquad \{Z = 0\} \subseteq \mathsf{RefineErr}_1, \tag{156}$$

and

$$\begin{aligned} &\max_{\theta \in \{0,1\}} \Pr_\theta(\mathsf{MergeErr}_\theta \cup \mathsf{RefineErr}_\theta \cup \mathsf{ActionErr}_\theta) \\ &\qquad \ge \frac{1 - \sqrt{D_{\mathrm{KL}}(\mathsf{T}_0^{n,\mu} \| \mathsf{T}_1^{n,\mu})/2}}{2}. \end{aligned} \tag{157}$$

In particular,

$$nq\Delta^2 \le \frac{3}{64} \quad \Longrightarrow \quad \max_{\theta \in \{0,1\}} \Pr_\theta(\mathsf{MergeErr}_\theta \cup \mathsf{RefineErr}_\theta \cup \mathsf{ActionErr}_\theta) \ge \frac{1}{4}. \tag{158}$$

Proof. Proposition 21, instantiated with the common protocol, makes $\widehat{\phi}_t$ a total measurable data map. Since all protocol inputs are identical under the two experiments, the same sample-to-output map defines Z under both laws.

Under T_0, equation 154 gives $A \not\sim^{\mathrm{B}}_{t;0} B$. Thus $Z = 1$ is a witness that the learner merged two true blocks. Under T_1, the same display gives $A \sim^{\mathrm{B}}_{t;1} B$, so $Z = 0$ is a witness that the learner refined one true block. This proves equation 156 on the same supported atoms; no off-support or surrogate relation is involved.

Interpret Z as the test that reports T_1 when $Z = 1$ and T_0 when $Z = 0$. Its errors are

$$e_0 := \Pr_0(Z = 1), \qquad e_1 := \Pr_1(Z = 0).$$

Lemma 54 gives

$$\begin{aligned} e_0 + e_1 &\geq 1 - \|\mathsf{T}_0^{n,\mu} - \mathsf{T}_1^{n,\mu}\|_{\mathrm{TV}} \\ &\geq 1 - \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mathsf{T}_0^{n,\mu} \| \mathsf{T}_1^{n,\mu})}. \end{aligned}$$

For

$$\mathcal{U}_\theta := \mathsf{MergeErr}_\theta \cup \mathsf{RefineErr}_\theta \cup \mathsf{ActionErr}_\theta,$$

the inclusions give

$$\Pr_0(\mathcal{U}_0) \geq e_0, \qquad \Pr_1(\mathcal{U}_1) \geq e_1.$$

Therefore

$$\max_\theta \Pr_\theta(\mathcal{U}_\theta) \geq \max\{e_0, e_1\} \geq \frac{e_0 + e_1}{2},$$

which proves equation 157. When $nq\Delta^2 \leq 3/64$, Lemma 57 gives

$$D_{\mathrm{KL}}(\mathsf{T}_0^{n,\mu} \| \mathsf{T}_1^{n,\mu}) \leq \frac{16}{3} \frac{3}{64} = \frac{1}{4}.$$

The testing lemma then yields $\max\{e_0, e_1\} \geq 1/4$, proving equation 158.

The action event is not needed: adding it only enlarges the union. An identity output has $Z = 0$ and refines the true A, B block under T_1; an all-merged output has $Z = 1$ and merges two blocks under T_0; a crossing partition is likewise classified by its A, B relation. Correct action at B cannot remove the relation error. Thus the common pair, exact KL, binary testing interface, and same-object inclusions prove the complete trilemma in raw $nq\Delta^2$ mode, with $c_{\mathrm{tri}} = 16/3$, small-information coefficient $3/64$, and probability $1/4$. At $q = 0$ or $\Delta = 0$ no positive trilemma is asserted; at $q = 1/6$ all six atom–action rows have mass $1/6$ and the calculation remains valid. □

A.5.12 One-Stage Rate Specialization

For this subsection only, define the target-reachable supported active set and simultaneous-correctness event

$$\begin{aligned} \mathcal{S}_{E,1}^{\mathrm{tar}} &:= \{u \in S_{\mu,1} \cap \mathcal{U}_{E,1} : d_1^\pi(u) > 0 \text{ for some } \pi \in \Pi_{\mathrm{tar}}\}, \\ \mathcal{K}_1 &:= \left\{\widehat{g}_1(u) = a_1^\star(u) \text{ for every } u \in \mathcal{S}_{E,1}^{\mathrm{tar}}\right\}. \end{aligned} \tag{159}$$

Recall that

$$q_1^{\mathrm{audit}} := \min_{D : S_{D,1} \neq \varnothing} q_{D,1}^{\mathrm{audit}}. \tag{160}$$

This quantity is finite and positive in the active-gap regime; empty strata are omitted and have zero error charge.

Lemma 59 (Direct two-error action-margin transfer)**.** *Under Assumption 15 and Proposition 38, let $H = 1$. On $\mathcal{H}_\delta$, for every $u \in \mathcal{S}_{E,1}^{\mathrm{tar}}$ and every $a \neq a_1^\star(u)$,*

$$\begin{aligned} &\widehat{Q}_1(\widehat{\phi}_1(u), a_1^\star(u)) - \widehat{Q}_1(\widehat{\phi}_1(u), a) \\ &\quad \geq \Delta_1 - 2D_1 \\ &\quad \geq \Delta_1 - 2\mathfrak{A}_1 - 2\mathfrak{S}_1 \\ &\quad = \Delta_{1,\mathrm{eff}} - 2\mathfrak{S}_1. \end{aligned} \tag{161}$$

If the last expression is positive, then $\widehat{g}_1(u) = a_1^\star(u)$ simultaneously for all $u \in \mathcal{S}_{E,1}^{\mathrm{tar}}$.

Proof. Assumption 15 gives

$$Q_1^\star(u, a_1^\star(u)) - Q_1^\star(u, a) \geq \Delta_1.$$

The definition of D_1 in equation 50 controls the learned error at both actions. Hence

$$\begin{aligned}
&\widehat{Q}_1(\widehat{\phi}_1(u), a_1^\star(u)) - \widehat{Q}_1(\widehat{\phi}_1(u), a) \\
&= \big[\widehat{Q}_1(\widehat{\phi}_1(u), a_1^\star(u)) - Q_1^\star(u, a_1^\star(u))\big] \\
&\quad + \big[Q_1^\star(u, a_1^\star(u)) - Q_1^\star(u, a)\big] \\
&\quad + \big[Q_1^\star(u, a) - \widehat{Q}_1(\widehat{\phi}_1(u), a)\big] \\
&\geq -D_1 + \Delta_1 - D_1.
\end{aligned}$$

Proposition 38 gives $D_1 \leq \mathfrak{A}_1 + \mathfrak{S}_1$, and the definition $\Delta_{1,\text{eff}} = \Delta_1 - 2\mathfrak{A}_1$ gives the last line of equation 161. If it is positive, the true ordered optimum has strictly larger learned value than every competitor. It is therefore the unique learned maximizer, so the ordered greedy rule selects it. The conclusion is simultaneous because D_1 is the supported atom–action supremum. □

Proposition 60 (Bias-aware one-stage upper specialization)**.** *Under Assumptions 13 and 15, Propositions 28 and 38, and Lemma 59, let $H = 1$. If*

$$\Delta_{1,\text{eff}} > 0, \tag{162}$$

$$nq_1^{\text{audit}} \geq 8\Gamma_{\text{cnt}}(\delta/2), \tag{163}$$

and

$$nq_1^{\text{audit}}\Delta_{1,\text{eff}}^2 > 4c_s^2\Gamma_{\text{dev}}(\delta/2), \tag{164}$$

then

$$\Pr(\mathcal{K}_1) \geq 1 - \delta. \tag{165}$$

No correctness conclusion is made when $\Delta_{1,\text{eff}} \leq 0$, regardless of its square.

Proof. Since $H = 1$, $B_2 = 0$ and the statistical envelope has one stage. For every nonempty stratum, $q_{D,1}^{\text{audit}} \geq q_1^{\text{audit}}$. Hence

$$\begin{aligned}
\mathfrak{S}_1 &= \max_{D \in \{E, \bar{E}\}} s_{D,1} \\
&\leq c_s \max_{D: S_{D,1} \neq \varnothing} \sqrt{\frac{\Gamma_{\text{dev}}(\delta/2)}{nq_{D,1}^{\text{audit}}}} \\
&\leq c_s \sqrt{\frac{\Gamma_{\text{dev}}(\delta/2)}{nq_1^{\text{audit}}}}.
\end{aligned} \tag{166}$$

An empty stratum contributes its prescribed zero.

The count condition is checked separately. For every nonempty D,

$$nq_{D,1}^{\text{audit}} \geq nq_1^{\text{audit}} \geq 8\Gamma_{\text{cnt}}(\delta/2). \tag{167}$$

Proposition 28 therefore gives $\Pr(\mathcal{H}_\delta) \geq 1 - \delta$.

Since $\Delta_{1,\text{eff}} > 0$, the strict scalar inequality equation 164 is equivalent to

$$c_s \sqrt{\frac{\Gamma_{\text{dev}}(\delta/2)}{nq_1^{\text{audit}}}} < \frac{\Delta_{1,\text{eff}}}{2}. \tag{168}$$

Combining this with equation 166 gives

$$2\mathfrak{S}_1 < \Delta_{1,\text{eff}}.$$

Lemma 59 now proves $\mathcal{H}_\delta \subseteq \mathcal{K}_1$, and therefore

$$\Pr(\mathcal{K}_1) \geq \Pr(\mathcal{H}_\delta) \geq 1 - \delta.$$

If $\Delta_{1,\text{eff}} \leq 0$, the signed lower source in equation 161 is absent; the nonnegative square cannot restore it. □

Proposition 61 (All-raw-zero one-stage baseline reduction)**.** *Under Assumptions 13 and 15, Propositions 38 and 60, let $H = 1$. Suppose every supported stage-one raw reward and next-code pair tolerance vanishes and*

$$q_{\bar{E},1}^{\mathrm{audit}} \geq q_{E,1}^{\mathrm{audit}} =: q < +\infty. \tag{169}$$

Then

$$q_1^{\mathrm{audit}} = q, \qquad \mathfrak{A}_1 = 0, \qquad \Delta_{1,\mathrm{eff}} = \Delta_1, \tag{170}$$

and, on $\mathcal{H}_\delta$,

$$D_1 \leq \mathfrak{S}_1 \leq c_s \sqrt{\frac{\Gamma_{\mathrm{dev}}(\delta/2)}{nq}}. \tag{171}$$

Consequently,

$$nq \geq 8\Gamma_{\mathrm{cnt}}(\delta/2), \qquad nq\Delta_1^2 > 4c_s^2\Gamma_{\mathrm{dev}}(\delta/2) \tag{172}$$

imply $\Pr(\mathcal{K}_1) \geq 1 - \delta$.

Proof. Vanishing raw tolerances make every stage-one $\varepsilon_{D,1}^r, \varepsilon_{D,1}^p$ zero on a nonempty stratum; they are already zero on an empty stratum. Since $B_2 = 0$,

$$0 \leq a_{D,1} \leq c_b(0 + B_2 \cdot 0) = 0.$$

Thus $\mathfrak{A}_1 = 0$ and $\Delta_{1,\mathrm{eff}} = \Delta_1$.

The active stratum is nonempty. If the non-event stratum is nonempty, equation 169 makes the minimum in equation 160 equal to q; if it is empty, it is omitted and the same equality holds. This proves equation 170. The one-stage specialization in Proposition 38 is exactly equation 171.

Substituting equation 170 into the sign, count, and statistical conditions of Proposition 60 turns them into the positive raw gap and equation 172. The probability conclusion follows. Equivalently, the strict threshold and equation 171 give, on $\mathcal{H}_\delta$,

$$\widehat{Q}_1(\widehat{\phi}_1(u), a_1^\star(u)) - \widehat{Q}_1(\widehat{\phi}_1(u), a) \geq \Delta_1 - 2\mathfrak{S}_1 > 0.$$

No bias remainder survives. □

Proposition 62 (Zero-tolerance admissibility of the raw lower family)**.** *Under Assumptions 13 and 15 and Proposition 52, each paired lower experiment with*

$$0 < q \leq \frac{1}{6}, \qquad 0 < \Delta \leq \frac{1}{4}$$

has a model-common completion inside the stipulated $H = 1$, all-raw-zero class. In this completion,

$$q_1^{\mathrm{audit}} = q, \qquad \mathfrak{A}_1 = 0, \qquad \Delta_{1,\mathrm{eff}} = \Delta, \tag{173}$$

and neither paired training law nor any conclusion of Proposition 55 changes.

Proof. Use the direct terminal realization in Proposition 52, take the optional nuisance alphabet to be a singleton, and keep its master/event/cue map, binary actions, behavior routing, bounded rewards, and terminal transition. Let the task coordinate equal the master atom. The three active atoms are target reachable under the model-common action-independent initial occupancy and have raw gaps Δ.

Let the supported candidate family consist of the supported identity, with the prescribed candidate-independent canonical extension, and take that identity as the reference coarsening. Fix common total orders and the prescribed zero-count defaults. Set

$$\tau_1^r(u, v, a) = \tau_1^p(u, v, a) = 0$$

for every supported pair and action. The identity merges no distinct supported pair, so the reference inequalities are vacuous. This all-zero profile is therefore admissible, gives $\mathfrak{A}_1 = 0$, and leaves $\Delta_{1,\mathrm{eff}} = \Delta_1 = \Delta$.

Proposition 52 gives $q_E^{\mathrm{audit}} = q$. When $q \leq 1/8$, its non-event stratum is nonempty and $q_{\bar{E}}^{\mathrm{audit}} \geq q$; when $q > 1/8$, that stratum is empty and has value $+\infty$. Thus $q_1^{\mathrm{audit}} = q$ in both branches.

The identity, zero tolerances, singleton nuisance, labels, defaults, and orders are deterministic and common. They add no random coordinate and alter no reward or occupancy law. Hence the paired likelihood ratio, KL, testing error, and regret conclusions remain unchanged. This proves equation 173 within the same stipulated class. □

Proposition 63 (Separate-mode zero-tolerance rate specialization bridge)**.** *Under Propositions 61, 62, and 55, fix*

$$0 < q \le \frac{1}{6}, \qquad 0 < \Delta \le \frac{1}{4}.$$

For $H = 1$, all raw tolerances zero, and

$$q_{\bar{E},1}^{\mathrm{audit}} \ge q_{E,1}^{\mathrm{audit}} = q, \qquad \Delta_1 = \Delta,$$

the following two implications are independently valid.

(i) *For the specified total learner, if*

$$nq \ge 8\Gamma_{\mathrm{cnt}}(\delta/2) \quad \textit{and} \quad nq\Delta^2 > 4c_s^2\Gamma_{\mathrm{dev}}(\delta/2), \tag{174}$$

then every target-reachable supported active action is correct with probability at least $1-\delta$ under the fixed upper experiment.

(ii) *For every learner, the bounded paired family with the same scalar q, Δ satisfies*

$$nq\Delta^2 \le \frac{3}{64} \quad \Longrightarrow \quad \sup_{\theta\in\{+,-\}} \Pr_\theta(\widehat{a}(u_\star) \ne a_\theta^\star(u_\star)) \ge \frac{1}{8}, \tag{175}$$

with expected hard-atom conditional regret at least $\Delta/8$ and expected target-occupancy regret at least $\rho_1^\pi\Delta/32$ under one member.

Thus the one-stage zero-tolerance upper decision term and lower obstruction use the common scalar $nq\Delta^2$, up to the displayed upper finite-class/confidence logarithm and separate count prerequisite. This is a scale comparison, not a joint probability statement or a claim that the two threshold regimes overlap. With positive tolerances, the valid upper term remains $nq_1^{\mathrm{audit}}\Delta_{1,\mathrm{eff}}^2$ under the separate sign, count, and strict statistical conditions in Proposition 60; it is not identified with the raw-gap lower.

Proof. Proposition 61 proves the upper implication after the auxiliary choices $H = 1$, $B_2 = 0$, zero raw tolerances, $q_1^{\mathrm{audit}} = q$, and $\Delta_{1,\mathrm{eff}} = \Delta_1 = \Delta$. Its proof checks every nonempty-stratum count threshold through equation 167, proves

$$\mathfrak{S}_1 \le c_s\sqrt{\frac{\Gamma_{\mathrm{dev}}(\delta/2)}{nq}} < \frac{\Delta}{2},$$

and obtains $\mathcal{H}_\delta \subseteq \mathcal{K}_1$. The probability conversion is

$$\Pr(\mathcal{K}_1) \ge \Pr(\mathcal{H}_\delta) \ge 1-\delta.$$

Thus no count, bias, logarithmic, or confidence term is absorbed.

Proposition 62 places the paired family from Proposition 52 inside the same all-raw-zero one-stage class with the same q, Δ, without changing either data law. Proposition 55 then gives the lower implication and expectation bounds on the common parameter range.

The upper probability is generated by one fixed audit experiment, whereas the lower probability and expectation are generated by a distinct two-member testing experiment. They are not intersected, unioned, or converted into one another. Only the identities

$$\mathfrak{A}_1 = 0, \qquad \Delta_{1,\mathrm{eff}} = \Delta_1 = \Delta, \qquad q_1^{\mathrm{audit}} = q_E^{\mathrm{audit}} = q$$

permit both sides to use the same raw scalar. Outside this specialization, Proposition 60 retains the irreducible bias through $\Delta_{1,\mathrm{eff}} = \Delta_1 - 2\mathfrak{A}_1$ and uses the worst nonempty-stratum audit mass q_1^{audit}, while the lower result retains raw Δ and q_E^{audit}. Identifying those interfaces would change the result. □

A.5.13 CLAUSE-PRESERVING ASSEMBLY

Proposition 64 (Complete master-audit clause conjunction)**.** *Under Assumptions 6, 7, 8, 9, 10, 11, 12, 13, 14, and 15, the universal constants can be chosen compatibly as*

$$c_r = c_p = c_b = 1, \qquad c_{\mathrm{cnt}} = 8, \qquad c_s = 4\sqrt{2},$$
$$c_{\mathrm{KL}} = c_{\mathrm{tri}} = \frac{16}{3}, \qquad c_0 = \frac{3}{64}, \qquad c_1 = c_2 = \frac{1}{8}, \qquad \kappa = \frac{1}{4}. \tag{176}$$

For every $t \in [H]$ and $\delta \in (0,1)$, the following clause-local conclusions hold.

(i) *Every data set produces a unique measurable bounded partition, representative, model, value, action, and deterministic-policy tuple. This includes all-zero counts, mixed counts, every partition and action tie, the terminal singleton, and canonical off-support atoms.*

(ii) *If every nonempty consumed stratum at $s \geq t$ satisfies*

$$n q_{D,s}^{\mathrm{audit}} \geq 8\Gamma_{\mathrm{cnt}}(\delta/2), \tag{177}$$

then

$$\Pr(\mathcal{G}_{\delta/2}^{\mathrm{dev}}) \geq 1 - \frac{\delta}{2}, \qquad \Pr(\mathcal{C}_{\delta/2}^{\mathrm{cnt}}) \geq 1 - \frac{\delta}{2}, \qquad \Pr(\mathcal{H}_\delta) \geq 1 - \delta. \tag{178}$$

On $\mathcal{H}_\delta$,

$$\sup_{\substack{u \in S_{\mu,t} \\ a \in \mathcal{A}}} \left| Q_t^\star(u,a) - \widehat{Q}_t(\widehat{\phi}_t(u), a) \right| \leq \mathfrak{A}_t + \mathfrak{S}_t = \mathfrak{B}_t, \tag{179}$$

with the separate envelopes

$$\mathfrak{A}_t \leq \sum_{s=t}^{H} \max_{D \in \{E, \bar{E}\}} \left(\varepsilon_{D,s}^r + B_{s+1} \varepsilon_{D,s}^p \right), \tag{180}$$

$$\mathfrak{S}_t \leq 4\sqrt{2} \sum_{s=t}^{H} \max_{D \in \{E, \bar{E}\}} \left[(1 + B_{s+1}) \sqrt{\frac{\Gamma_{\mathrm{dev}}(\delta/2)}{n q_{D,s}^{\mathrm{audit}}}} \right], \tag{181}$$

where an empty stratum takes its direct zero branch.

For every target policy and selector sequence, the deterministic piecewise transfer is

$$\sum_{s=t}^{H} \sum_{D \in \{E, \bar{E}\}} \mathbb{E}_{d_s^\pi} \left[\mathbf{1}\{U_s \in \mathcal{U}_{D,s}\} \ell_s(U_s, g_s(U_s)) \right] \leq \mathfrak{V}_t^\pi. \tag{182}$$

On $\mathcal{H}_\delta$,

$$Q_t^\star(u, a_t^\star(u)) - Q_t^\star(u, \widehat{g}_t(u)) \leq \min\{B_t, 2\mathfrak{B}_t\}, \qquad u \in S_{\mu,t}, \tag{183}$$

$$\frac{\mathcal{R}_{E,t}^\pi(\widehat{g}_t)}{\rho_t^\pi} \leq \min\{B_t, 2\mathfrak{B}_t\} \quad (\rho_t^\pi > 0), \tag{184}$$

with zero unnormalized event contribution at $\rho_t^\pi = 0$. For every admissible start law ξ,

$$V_t^\star(\xi) - V_t^{\widehat{\pi}}(\xi) \leq 2 \sum_{s=t}^{H} \mathfrak{B}_s \tag{185}$$

for the actual fresh rollout of the total learned policy.

(iii) *If*

$$\Delta_{t,\mathrm{eff}} > 0, \qquad 2\mathfrak{S}_t < \Delta_{t,\mathrm{eff}}, \tag{186}$$

then, on $\mathcal{H}_\delta$, every supported active atom is assigned its true optimal action, every target event regret is zero, and any two supported active atoms requiring different optimal actions receive different learned codes. The learned action margin is at least $\Delta_{t,\mathrm{eff}} - 2\mathfrak{S}_t > 0$.

(iv) *If the supported Bellman quotient belongs to the candidate class and*

$$4\mathfrak{B}_s < \zeta_s, \qquad s = t, \dots, H, \tag{187}$$

then, on $\mathcal{H}_\delta$,

$$\widehat{\phi}_s(u) = \widehat{\phi}_s(v) \quad \Longleftrightarrow \quad u \sim_s^{\mathrm{B}} v, \qquad u, v \in S_{\mu,s}, \quad s \geq t. \tag{188}$$

The learned partition equals the supported Bellman quotient up to relabeling, merges every supported task-equal Bellman-null nuisance fiber, preserves supported ordered-action differences, and leaves each off-support atom at its fixed canonical singleton label.

(v) *For every learner, $0 < q \leq 1/6$, and $0 < \Delta \leq 1/4$, the bounded raw-gap pair satisfies*

$$D_{\mathrm{KL}}(\mathsf{P}_+^{n,\mu} \| \mathsf{P}_-^{n,\mu}) \leq \frac{16}{3} nq\Delta^2,$$

and

$$nq\Delta^2 \leq \frac{3}{64} \quad \Longrightarrow \quad \sup_{\theta \in \{+,-\}} \Pr_\theta(\widehat{a}(u_\star) \neq a_\theta^\star(u_\star)) \geq \frac{1}{8}. \tag{189}$$

Under one member, occupancy regret is at least $\eta_t^\pi \Delta$ with probability at least $1/8$, the expected hard-atom conditional regret is at least $\Delta/8$, and

$$\mathbb{E}_\theta[\mathcal{R}_{E,t}^\pi(\widehat{g}_t)] \geq \frac{1}{8}\eta_t^\pi \Delta \geq \frac{1}{32}\rho_t^\pi \Delta. \tag{190}$$

This is an action/regret obstruction only.

(vi) *For the specified minimum-K_μ estimator and every $0 < q \leq 1/6$, $0 < \Delta \leq 1/4$, the common-protocol pair has atom B of mass $2q$, uniform binary behavior actions, and every supported atom–action row of mass at least q. All supported atoms are active, so*

$$q_{E,t}^{\mathrm{audit}} = q, \qquad S_{\bar{E},t} = \varnothing, \qquad q_{\bar{E},t}^{\mathrm{audit}} = +\infty$$

under the empty non-event convention. The pair satisfies

$$D_{\mathrm{KL}}(\mathsf{T}_0^{n,\mu} \| \mathsf{T}_1^{n,\mu}) \leq \frac{16}{3} nq\Delta^2$$

and

$$nq\Delta^2 \leq \frac{3}{64} \quad \Longrightarrow \quad \max_{\theta \in \{0,1\}} \Pr_\theta(\mathsf{MergeErr}_\theta \cup \mathsf{RefineErr}_\theta \cup \mathsf{ActionErr}_\theta) \geq \frac{1}{4}. \tag{191}$$

This conclusion is estimator specific and does not force a fixed alias under both members.

(vii) *At $H = 1$, positive tolerances retain the implication*

$$\left.\begin{aligned} \Delta_{1,\mathrm{eff}} &> 0, \\ nq_1^{\mathrm{audit}} &\geq 8\Gamma_{\mathrm{cnt}}(\delta/2), \\ nq_1^{\mathrm{audit}} \Delta_{1,\mathrm{eff}}^2 &> 4c_s^2 \Gamma_{\mathrm{dev}}(\delta/2) \end{aligned}\right\} \Longrightarrow \Pr(\mathcal{K}_1) \geq 1 - \delta. \tag{192}$$

If every stage-one raw tolerance is zero and

$$q_{\bar{E},1}^{\mathrm{audit}} \geq q_{E,1}^{\mathrm{audit}} =: q,$$

then

$$\mathfrak{A}_1 = 0, \qquad \Delta_{1,\mathrm{eff}} = \Delta_1, \qquad q_1^{\mathrm{audit}} = q,$$

so the upper uses the bare scalar $nq\Delta_1^2$, with the separate count prerequisite and displayed finite-class logarithm. The paired lower independently uses the same scalar $nq\Delta^2$ only after the zero-tolerance completion. The upper and lower probability spaces are not combined, and their threshold antecedents need not overlap.

The exposed upper variables are

$$H, t, n, \delta, |\mathcal{A}|, |\Phi|, \{|\mathcal{U}_s|\}_{s=1}^{H+1}, \{B_s\}_{s=t}^{H+1}, \Gamma_{\mathrm{dev}}, \Gamma_{\mathrm{cnt}},$$
$$\{\tau_s^r, \tau_s^p, \varepsilon_{D,s}^r, \varepsilon_{D,s}^p, q_{D,s}^{\mathrm{audit}}, \rho_s^\mu, \lambda_s, \alpha_s^{\mathrm{audit}}, \beta_s, d_s^\pi, \rho_s^\pi, C_{D,s}^{\mathrm{atom}}, a_{D,s}, s_{D,s}\}_{s=t}^{H}, \tag{193}$$
$$\{K_{\mu,s}, \Delta_s, \Delta_{s,\mathrm{eff}}, \zeta_s, \mathfrak{A}_s, \mathfrak{S}_s, \mathfrak{B}_s\}_{s=t}^{H}, \mathfrak{V}_t^\pi.$$

The lower side additionally exposes $q, \Delta, \eta_t^\pi, \rho_t^\pi$. The constants in equation 176 depend on none of these quantities, and there is no other hidden constant. As n varies, the POMDP, master audit, support masks, candidate class, raw tolerances, defaults, orders, target-policy class, and finite horizon are fixed. Upper statements are simultaneous on $\mathcal{H}_\delta$; lower statements use separate worst-member testing probabilities and expectations. Horizon mode is fixed finite. The norms are the supported atom–action Q-supremum, target-occupancy action and policy regret under actual fresh rollouts, supported equivalence-relation equality up to relabeling, hard-action error, and supported relation error. Total variation is half the ℓ_1 distance.

Six scope restrictions are part of the conjunction:

(1) *no general positive-tolerance reference is declared feasible under an arbitrary crossing selected downstream code;*

(2) *ℓ_s and $\mathfrak{V}_t^\pi$ are used only for equation 182, not as representative-model error;*

(3) *a behavior-empty stratum is proved target empty and assigned zero before loss or density evaluation, so no $+\infty \cdot 0$ product is formed;*

(4) *learned-policy regret uses the total data-defined policy and its true homogeneous-master rollout rather than an auxiliary model recursion;*

(5) *quotient recovery is support restricted, and canonical off-support labels are not statistical recovery; and*

(6) *positive-tolerance $nq_1^{\mathrm{audit}}\Delta_{1,\mathrm{eff}}^2$ is not identified with the raw-gap lower scale.*

At zero local bias and statistical charge, supported action values are exact. With quotient membership and zero charges, exact supported quotient recovery, nuisance merging, opposite-action separation, and canonical-label preservation have no remainder. No finite retention claim is made at conceptual $q_{E,t}^{\mathrm{audit}} = 0$ or $\Delta_{t,\mathrm{eff}} \le 0$; lower regret vanishes at raw $\Delta = 0$, and no positive identification lower bound is asserted at $q = 0$. When the supported quotient has one block, $\zeta_s = +\infty$ makes cross-block separation vacuous and supported cardinality still closes recovery.

Proof. The constants in equation 176 satisfy all earlier calibrations simultaneously:

$$c_r = c_p = 1, \qquad c_b = 1, \qquad c_s = 4\sqrt{2}\max\{c_r, c_p\}, \qquad c_{\mathrm{cnt}} = 8.$$

The lower constants are those derived in Lemmas 53, 54, and 57.

Proposition 22 proves clause (i) unconditionally. Proposition 25 and Corollary 29 give equation 178 under equation 177. The count conclusion is not inferred from the deviation event.

Propositions 30 and 31 directly supply the unchanged population structural interfaces

$$P_s^\chi(S_{\mu,s+1} \mid u, a) = 1 \quad (u \in S_{\mu,s},\ a \in \mathcal{A})$$

and

$$u, v \in S_{\mu,s}, \quad \tau_s(u) = \tau_s(v) \quad \Longrightarrow \quad u \sim_s^{\mathrm{B}} v.$$

Thus every true continuation consumed below remains on audited support, and the population quotient merges each supported task-equal nuisance fiber.

Proposition 35 supplies the representative-model interface through the realized selected next code, and Proposition 38 gives equation 179–equation 181. The endpoint envelope, not the rowwise loss, controls a representative that may be rarer than the queried atom. Proposition 39 proves equation 182 with its piecewise empty branch. Lemma 41 gives equation 183–equation 184. Proposition 40 identifies the total learned policy's master recursion with its true fresh-rollout return, and Proposition 42 gives equation 185.

Under equation 186, Propositions 43, 44, and 45 give clause (iii) with residual margin

$$\Delta_t - 2\mathfrak{B}_t = \Delta_{t,\mathrm{eff}} - 2\mathfrak{S}_t > 0.$$

Under equation 187, Proposition 51 gives equation 188; its proof uses quotient-specific downstream relabeling, separation, and equal supported cardinality.

Proposition 55 gives equation 189–equation 190, with the fixed occupancy $\eta_t^\pi \geq \rho_t^\pi/4$. Proposition 58 gives equation 191. The former makes no representation claim; the latter is restricted to the common-protocol minimum-K_μ estimator.

Proposition 60 gives equation 192. Propositions 61 and 62 establish the zero-bias identities and same-class lower completion. Proposition 63 then compares the upper and lower scalar rates without mixing their probability spaces.

Each conclusion has thus been taken under its original local antecedent, event, support, norm, horizon, and probability mode. Conjunction adds no confidence expenditure, rate simplification, support extension, or new mathematical implication. The preceding paragraphs also verify every scope restriction and boundary statement, completing the proposition. □

A.5.14 Proof of the Main Theorem

Proof. Proposition 22 gives unconditional totality. Proposition 25 and Corollary 29 give the three probability statements. Propositions 30, 35, and 38 give the closed supported Bellman domain, the representative interface, and the separated supported action-value bound.

Propositions 39, 40, and 42, together with Lemma 41, give the target-shift, pointwise, event, and actual-rollout policy-regret conclusions. Propositions 43, 44, and 45 give bias-aware retention. Proposition 51 gives exact supported quotient recovery under its clause-local separation conditions.

The raw-gap action and occupancy-regret obstruction is Proposition 55; the common-protocol structural trilemma is Proposition 58; and the positive-tolerance one-stage upper and zero-tolerance scale comparison are Propositions 60 and 63. Proposition 64 verifies that these results admit the common constants displayed in equation 176, preserve every clause-local antecedent and probability space, expose every rate and mode, retain the exact/noiseless baselines, and enforce all six scope restrictions. These conclusions are precisely the clauses of Theorem 16, completing the proof. □

B Numerical Study: Specifications and Complete Results

B.1 Environments

All environments are finite master models with Bernoulli rewards. Episodes are sampled exactly in law through their sufficient statistics: atom counts are multinomial under the stage occupancy, action counts multinomial under the behavior policy, reward sums binomial, and next-atom counts multinomial under P_t^χ.

Traffic light ($H = 2$). Stage-one atoms are (e, c, w) with $(e, c) \in \{(0, \varnothing), (1, \text{green}), (1, \text{red})\}$ and watermark $w \in \{1, \dots, m\}$; $\Pr(e = 1) = \rho$, the cue is uniform given the event, and w is uniform. Rewards are $(0.9, 0.1)$ for (GO,STOP) without an event, $(\frac{1}{2} + \frac{\Delta}{2}, \frac{1}{2} - \frac{\Delta}{2})$ under green, and the reverse under red. The behavior policy is uniform on non-event atoms and plays STOP with probability $\beta = \frac{1}{2}$ on event atoms. The next observation keeps w with probability 0.9 (otherwise resampled uniformly) and draws a fresh event and cue independently of the current atom and action. Stage-two rewards are zero, so the stage-two quotient is a single block and the stage-one quotient is {no event}, {green}, {red}. Default parameters: $m = 3$, $\Delta = 0.4$, $\rho = 0.05$. The one-stage variant ($H = 1$) used for the coverage law drops stage two. The event coverage is $q_E = \rho\beta/(2m)$.

Delayed consequence ($H = 3$). Stage-one atoms are as in the traffic light with $m = 2$ and $\rho = 0.2$; the behavior policy is uniform and all stage-one rewards are zero. The stage-one action enters lane *safe* with probability $\frac{1}{2} + \frac{\Delta}{2}$ if it is the cue-appropriate action (GO without event or under green, STOP under red) and $\frac{1}{2} - \frac{\Delta}{2}$ otherwise, with $\Delta = 0.4$; stage-two and stage-three atoms are (lane, watermark) pairs with fresh uniform watermarks. The lane persists; stage-three reward is 0.75 in the safe lane and 0.25 in the hazard lane regardless of action. The stage-one gap is 0.2, and the Bellman quotient has two blocks at every stage.

Hard pair ($H = 1$). Three atoms: the hard atom $u_\star$ with mass $2q$ and rewards $(\frac{1}{2} \pm \frac{\Delta}{2}, \frac{1}{2} \mp \frac{\Delta}{2})$, and two anchors of mass $\frac{1}{2} - q$ each with rewards $(0.75, 0.25)$ and $(0.25, 0.75)$; actions are uniform. Replicates alternate the sign of the hard atom.

B.2 Learners

CBQ. Radii follow equation 3 with $\delta = 0.05$ and Φ the product over stages of all set partitions of the support, so $\log|\Phi| = \sum_t \log \text{Bell}(|S_{\mu,t}|)$. This gives $\Gamma_{\text{dev}} = 38.3$ for the traffic light ($m = 3$), 13.9 and 24.5 for the one-stage traffic light with $m = 2$ and $m = 4$, 24.2 for the delayed task, and 9.5 for the hard pair. Tolerances are zero. The minimum-K_μ certified partition is the lexicographically first minimum clique cover of the compatibility graph under the atom order, found by exact backtracking. The atom order is a uniformly random permutation drawn before each replicate's data; it fixes compatibility-cover ties and representatives.

Baselines. *Identity*: singleton partition with the same planner. *Myopic*: CBQ with the transition test equation 5 removed. *Fixed threshold*: CBQ with every radius replaced by $\epsilon/2$. *Predictive*: among all partitions of the stage-one support with exactly K blocks (K equal to the Bellman-quotient size: 3 for the traffic light, 2 for the delayed task), minimize $\sum_b N_b \,\mathrm{H}(\widehat{P}_b) + \lambda \sum_{u,a} N(u,a)(\widehat{r}(u,a) - \widehat{r}_{b(u)}(a))^2$, where $\widehat{P}_b$ is the pooled empirical law of the next master atom in block b and $\widehat{r}_b$ the pooled block reward; $\lambda = 0$ for the reward-free variant and $\lambda = 20$ otherwise. It acts greedily on pooled block rewards.

Metrics. *Retention*: every event atom receives its optimal action and no green atom shares a stage-one code with a red atom. *Nuisance merged*: every fiber $\{(e, c, w)\}_w$ lies in one code. *Exact quotient*: the learned supported partition equals the Bellman quotient at every stage. *Event-conditional regret*: $\sum_{u \in \mathcal{U}_E} d(u)(V^\star(u) - Q^\star(u, \widehat{g}(u)))/\rho$.

Replicates and compute. Figure 2b–c and Figure 5: 100 replicates per point; Figure 2a: 10 replicates per cell; Figure 3a: 400; Figure 6: 200. The tabular studies run in under a minute on a laptop CPU; the random-POMDP benchmark takes about 15 minutes on 22 CPU cores; the pixel

study takes about 45 minutes on one RTX PRO 6000 GPU. All random streams descend from seed 42.

B.3 Complete results

Threshold dilemma and delayed consequences. Figure 5 isolates the two ingredients of CBQ. On the traffic light at $n = 3 \times 10^6$, a fixed merge threshold retains the cue and merges the watermark only inside a band of ϵ whose upper edge moves with Δ; below $\epsilon \approx 0.02$ it cannot merge nuisance, above roughly Δ it merges green with red, and from $\epsilon = 0.68$ it merges everything. In the delayed task the cue only selects a lane that pays off two stages later, so a learner certifying merges from immediate rewards merges the cue at every n, while CBQ propagates the {safe, hazard} split backward and recovers the exact three-stage quotient from $n \approx 3 \times 10^4$ with 6 codes instead of identity's 14.

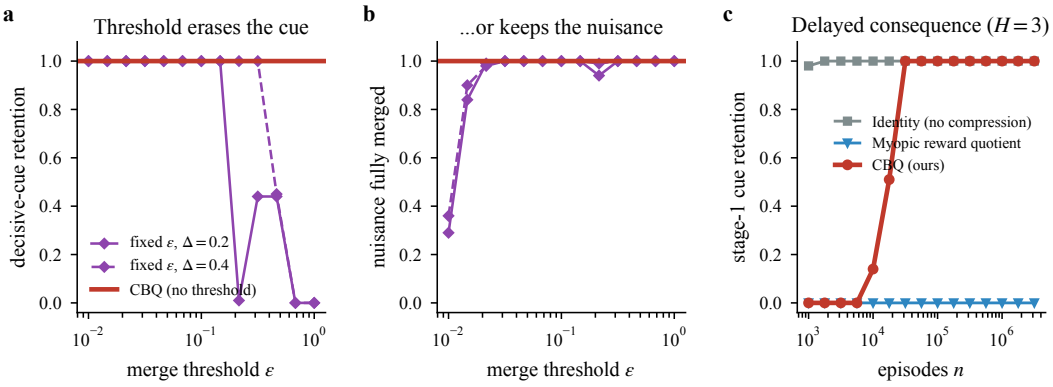

Figure 5: (a–b) Fixed merge thresholds on the traffic light ($n = 3 \times 10^6$, $\rho = 0.05$). (c) Delayed consequence ($H = 3$, $\rho = 0.2$).

Coverage law on the traffic light. Figure 6 repeats the coverage-law test in twelve one-stage traffic-light settings ($\rho \in \{0.02, 0.1, 0.4\}$, $m \in \{2, 4\}$, $\Delta \in \{0.2, 0.4\}$, candidate class of all partitions). The required n ranges from 9.8×10^3 to 2.8×10^6; in coverage units all twelve reach full retention at the same grid point (5.6) and none is fully retained one grid step earlier.

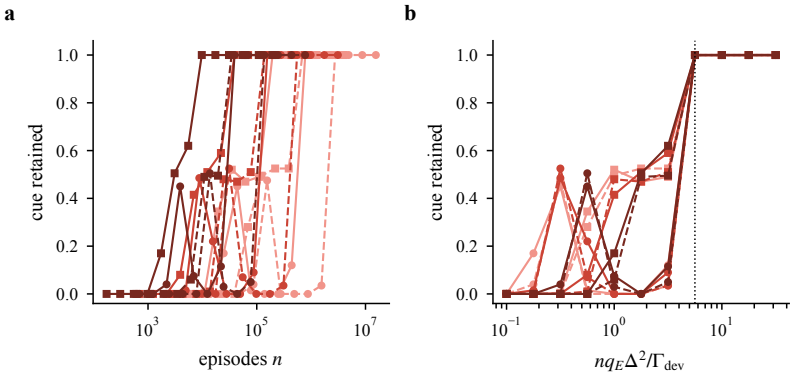

Figure 6: Twelve traffic-light settings: retention against (a) raw sample size and (b) normalized coverage.

Figure 7 reports the remaining views of the same runs. On the hard pair, CBQ and the per-atom plug-in both collapse in $nq\Delta^2$ and both reach zero error within the displayed range; CBQ additionally certifies whether the hard atom may share a code with an anchor, a decision the uncompressed plug-in never makes. Tables 4 and 5 list retention together with code size or exact recovery at representative sample sizes.

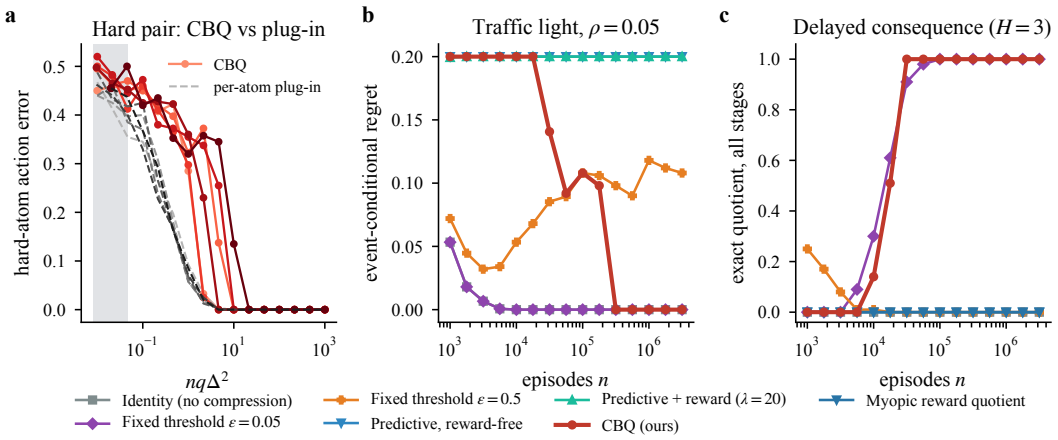

Figure 7: **Additional views.** (a) Hard-pair action error for CBQ (solid) and the per-atom plug-in (dashed), six (q, Δ) configurations. (b) Event-conditional regret on the traffic light at $\rho = 0.05$. (c) Exact recovery of the three-stage Bellman quotient on the delayed task.

Table 4: Traffic light, $\rho = 0.05$: retention / stage-one code size K (mean over 100 replicates).

n	CBQ	Myopic	Identity	Predictive	Pred. + reward	Fixed 0.05	Fixed 0.5
10^4	0.00 / 1.0	0.00 / 1.0	1.00 / 9.0	0.00 / 3.0	0.00 / 3.0	1.00 / 6.2	0.62 / 2.0
10^5	0.46 / 2.0	0.46 / 2.0	1.00 / 9.0	0.00 / 3.0	0.00 / 3.0	1.00 / 3.9	0.46 / 2.0
3.2×10^5	1.00 / 3.0	1.00 / 3.0	1.00 / 9.0	0.00 / 3.0	0.00 / 3.0	1.00 / 3.1	0.51 / 2.0
10^6	1.00 / 3.0	1.00 / 3.0	1.00 / 9.0	0.00 / 3.0	0.00 / 3.0	1.00 / 3.0	0.41 / 2.0
3.2×10^6	1.00 / 3.0	1.00 / 3.0	1.00 / 9.0	0.00 / 3.0	0.00 / 3.0	1.00 / 3.0	0.46 / 2.0

B.4 RANDOM-POMDP BENCHMARK

Generator. Each instance has horizon $H \in \{2, 3, 4\}$, $|\mathcal{A}| \in \{2, 3\}$, two task states, $C \in \{2, 3\}$ cues, two nuisance factors with $m \in \{3, 4\}$ values each, and event rate $\rho \in \{0.02, 0.05, 0.1\}$; four instances per configuration give 288 instances with 54–128 atoms per stage. The task state moves by a random Dirichlet(1) kernel that depends on the full decision value and the action, so cues can act through delayed transitions. The next event and cue are exogenous, and each nuisance factor persists with probability 0.8 and is otherwise resampled. Base rewards are uniform on $[0.2, 0.6]$; working backward, the reward of the cue-appropriate action at every event atom is set so that its $Q^\star$-gap is exactly $\Delta = 0.25$, with a random permutation assigning actions to cues so that the cue flips the action. Instances whose planted rewards leave $[0, 1]$ are resampled. The behavior policy is uniform. Each instance uses its own child of `SeedSequence`$([42, \text{index}])$.

Factored candidate class. At every stage, Φ_t contains all pairs (partition of the decision values, subset of kept nuisance factors), with code (block, kept nuisance values). This class contains the identity and the Bellman quotient, and $\log|\Phi_t| = \log \mathrm{Bell}(k) + 2\log 2$, giving $\Gamma_{\mathrm{dev}} \in [79, 181]$. CBQ returns the certified candidate with the fewest supported codes. Baselines use the same class: fixed thresholds replace the radii by $\epsilon/2$ with $\epsilon \in \{0.02, 0.05, 0.1, 0.2, 0.5\}$, the myopic variant drops the transition test, and the predictive variants choose, at each stage, the candidate with at most the Bellman-quotient number of codes that minimizes next-atom log-loss (plus $20\times$ reward squared error) and plan on pooled rows. Every learner is evaluated on 15 sample sizes $n \in [10^4, 10^{11}]$; regret is the exact fresh-rollout regret from the initial law.

Certificate validity. For each run we computed the population charges $\mathfrak{S}_t$ (with $\mathfrak{A}_t = 0$), the count condition $nq_{D,s} \geq 8\Gamma_{\mathrm{cnt}}$, and the retention condition $2\mathfrak{S}_t < \Delta$ at every stage. In all 2,393 stage-cases where the condition held, CBQ retained the cue; in all 3,237 runs satisfying the count

Table 5: Delayed consequence ($H = 3$): retention / exact three-stage quotient recovery (mean over 100 replicates).

n	CBQ	Myopic	Identity	Predictive	Pred. + reward	Fixed 0.05	Fixed 0.5
10^4	0.14 / 0.14	0.00 / 0.00	1.00 / 0.00	0.00 / 0.00	0.00 / 0.00	1.00 / 0.30	0.01 / 0.01
3.2×10^4	1.00 / 1.00	0.00 / 0.00	1.00 / 0.00	0.00 / 0.00	0.00 / 0.00	1.00 / 0.91	0.00 / 0.00
10^5	1.00 / 1.00	0.00 / 0.00	1.00 / 0.00	0.00 / 0.00	0.00 / 0.00	1.00 / 1.00	0.00 / 0.00
10^6	1.00 / 1.00	0.00 / 0.00	1.00 / 0.00	0.00 / 0.00	0.00 / 0.00	1.00 / 1.00	0.00 / 0.00

condition, its exact regret was below $2\sum_s \mathfrak{B}_s$. Figure 8 shows where the certificate applies relative to observed retention.

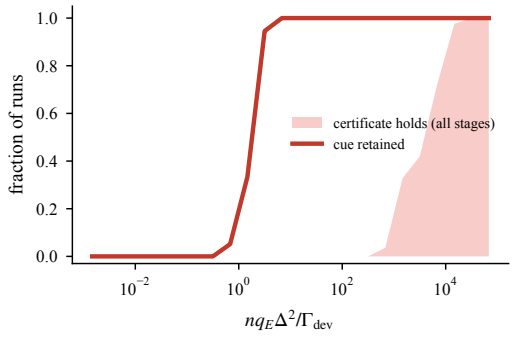

Figure 8: Random-POMDP benchmark: fraction of runs in which the retention certificate holds at every stage (shaded) and fraction in which CBQ retains the cue (line), against normalized coverage.

Table 6: Random-POMDP benchmark (288 instances): retention / exact recovery of the Bellman quotient at all stages.

Learner	$n = 10^6$	10^8	10^9	10^{10}	10^{11}
CBQ	0.00 / 0.00	0.98 / 0.28	1.00 / 0.65	1.00 / 0.88	1.00 / 0.95
Myopic	0.00 / 0.00	0.95 / 0.15	1.00 / 0.53	1.00 / 0.84	1.00 / 0.94
Identity	0.95 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00
Fixed $\epsilon = 0.02$	0.95 / 0.00	1.00 / 0.62	1.00 / 0.82	1.00 / 0.79	1.00 / 0.78
Fixed $\epsilon = 0.05$	0.95 / 0.00	1.00 / 0.60	1.00 / 0.57	1.00 / 0.56	1.00 / 0.55
Fixed $\epsilon = 0.1$	0.95 / 0.08	1.00 / 0.30	1.00 / 0.27	1.00 / 0.26	1.00 / 0.26
Fixed $\epsilon = 0.2$	0.93 / 0.07	0.91 / 0.02	0.93 / 0.02	0.90 / 0.02	0.91 / 0.02
Fixed $\epsilon = 0.5$	0.03 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00
Predictive	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00
Predictive + reward	0.05 / 0.00	0.05 / 0.00	0.05 / 0.00	0.05 / 0.00	0.05 / 0.00

At $n = 10^{11}$ the mean ratio of learned to Bellman-quotient code size is 1.00 for CBQ, 12.5 for identity, 0.96 for $\epsilon = 0.05$, and 0.60 for reward-free predictive compression; exact fresh-rollout regret is 0 for CBQ, 0.070 for predictive, and 0.177 for $\epsilon = 0.5$. Below the coverage threshold, CBQ keeps atoms that it cannot yet certify apart in shared blocks, so its code never contains an uncertified split; identity, which never compresses, acts per atom throughout.

B.5 Bias cannot be bought back with data

We take the one-stage traffic light ($\Delta = 0.4$, $\rho = 0.1$, $m = 3$) and let every watermark value $z > 0$ pull each cue's rewards toward the opposite cue by ϵ, so the reference coarsening that drops the watermark has reward tolerance ϵ and the true gap at watermarked atoms is $|\Delta - 2\epsilon|$. CBQ runs with $\tau = \epsilon$, 100 replicates per point. Figure 9 shows event-conditional regret. For $\epsilon \leq \Delta/2$ more data drives regret to zero; beyond $\epsilon = \Delta/2$ regret plateaus at a positive level from $n = 10^6$ to 10^9 (and for some ϵ rises slightly with n as the biased merge becomes certain), as the effective-gap condition predicts. The shaded region is where Corollary 2 certifies retention ($\Delta_{\text{eff}} > 0$).

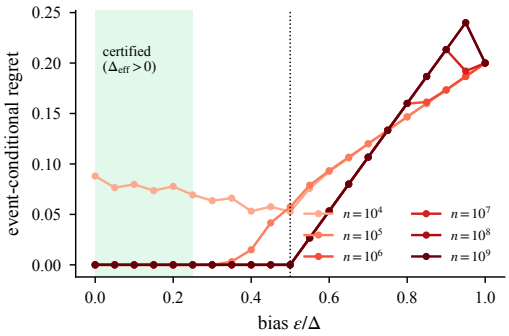

Figure 9: Approximation bias: event-conditional regret against ϵ/Δ for six sample sizes.

B.6 Pixel study

Scene. Frames are 64×64 RGB: sky, road with lane dashes, a 24×24 textured watermark in the top-left corner (8 fixed textures generated from seed 42), and a pole. With probability ρ a traffic-light housing appears with a lit red (top) or green (bottom) lamp. Every frame receives a global brightness jitter of $\pm 2.5\%$ and i.i.d. Gaussian pixel noise ($\sigma = 0.03$). Rewards follow the traffic-light table (Appendix B); the next frame keeps the watermark with probability 0.9 and draws a fresh event and cue. Actions are uniform. Evaluation uses 6,000 test frames with balanced cue states.

Tokenizer and CBQ. Each frame is average-pooled by 4 and cut into 16 patches of $4 \times 4 \times 3$ values. Per patch position we run k-means ($k = 32$, k-means++ initialization, 15 Lloyd iterations) on an unlabeled pool of 20,000 frames and merge centroids closer than 0.3 in ℓ_2, the tokenizer's resolution. In every run this yields exactly 24 master atoms, one per (watermark, light state), and red and green never share an atom. The candidate class keeps a subset of the 16 patch factors ($|\Phi| = 2^{16}$, $\Gamma_{\text{dev}} = 21.0$); the fixed atom order ranks atoms by pool frequency, which is independent of the training sample. Training tuples are tokenized in streams of 10^5 frames, so CBQ scales to 3×10^6 episodes without storing frames.

Deep baselines. All use a four-layer convolutional encoder (32–64–128–128 channels, stride 2) and a 16-dimensional latent, AdamW, batch 256, and 4,000 training steps, on the same offline episodes (up to 10^6, stored as 8-bit frames on the GPU). The autoencoder reconstructs the current frame and the world model the next frame, both with a summed Gaussian log-likelihood; the world model + reward adds a per-sample reward regression head on (z, a); the self-predictive latent predicts the normalized EMA-target embedding of the next frame from (z, a) (target momentum 0.99). Frozen latents are probed with a two-layer Q-head trained for 2,000 steps on reward regression. End-to-end Q trains encoder and head jointly on reward regression. The watermark probe is a linear classifier trained on half of the test codes and evaluated on the other half.

Table 7: Pixel study: fraction of three seeds with full cue retention (deep baselines run up to $n = 10^6$).

ρ	n	CBQ	Autoencoder	World model	WM + reward	Self-predictive	End-to-end Q
0.01	3×10^4	0.00	0.00	0.00	0.00	0.00	0.67
0.01	10^5	0.00	0.00	0.00	0.00	0.00	1.00
0.01	3×10^5	0.00	0.00	0.00	0.00	0.00	0.67
0.01	10^6	0.33	0.33	0.00	0.00	0.00	1.00
0.01	3×10^6	1.00	—	—	—	—	—
0.03	3×10^4	0.00	0.00	0.00	0.00	0.00	1.00
0.03	10^5	0.00	0.33	0.00	0.00	0.00	1.00
0.03	3×10^5	0.00	0.67	0.00	0.00	0.00	1.00
0.03	10^6	1.00	0.33	0.00	0.00	0.00	1.00
0.03	3×10^6	1.00	—	—	—	—	—
0.1	3×10^4	0.00	1.00	0.00	0.33	0.00	0.67
0.1	10^5	0.00	1.00	0.00	0.00	0.00	1.00
0.1	3×10^5	1.00	1.00	0.00	0.00	0.00	1.00
0.1	10^6	1.00	0.67	0.00	0.00	0.00	1.00
0.1	3×10^6	1.00	—	—	—	—	—