THE PRICE OF CROSS-ACTION SHARING: REGRET BOUNDS AND INFORMATION LIMITS IN RL

Anonymous authors
Paper under double-blind review

ABSTRACT

Observed environmental inputs can support cross-action reuse in reinforcement learning (RL), but an input sample alone does not reveal unexecuted outcomes when conditional responses are unknown. Which observations can then be pooled, and what does validating their reuse cost? We study finite episodic Markov decision processes with unknown input-sharing groups, conditional transition kernels, and reward means. A response-mediated transfer characterization distinguishes equal input laws from equal induced transitions and quantifies the bilinear interaction between input and response estimation errors. Certified Response-First Optimism (CRFO) separates group certification from estimation, retains individual estimates, and bounds conditional continuation values before averaging input laws. This ordering charges response uncertainty to executed transitions. Its separation-dependent, high-probability regret bound is at most twice the smaller of the complete individual-learning and certification-plus-pooling bounds, including identification, response, and reward costs. Targetwise confidence intersection can tighten jointly valid upper estimates without changing the proved regret order. A streaming temporal-difference extension reduces conditional-response storage through a distinct historical-target analysis. For two-policy selection with known reward formulas, matching information bounds reveal a sharper distinction. At a shared reference, let ε denote the cancellation scale, Δ be the common successor-mean gap, and κ the response contrast, so the policy gap is $\varepsilon\Delta$. Under uniform correctness over models admitting unequal input laws without positive separation, the pointwise optimal expected episode count is $\Theta(\max\{\Delta^{-2}, \kappa^2/(\varepsilon^2\Delta^2) + 1/(\varepsilon\Delta)\}\log(1/\delta))$, versus $\Theta(\Delta^{-2}\log(1/\delta))$ when input-law equality is known, where δ is the error probability. A product-aware stopping rule attains these orders with reference-tuned initialization and adapts to unknown resolution with an explicit iterated-logarithmic overhead. At a fixed successor gap, vanishing response contrast changes the small-cancellation cost from inverse-square to inverse-first-power because the sharing-induced alternative requires joint input–response perturbations. Thus, neither group count nor policy-gap size alone determines sharing cost.

1 INTRODUCTION

Reinforcement learning (RL) optimizes cumulative reward through interaction with an unknown dynamical system, so actions influence both immediate performance and future information (Sutton & Barto, 2018). This exploration–exploitation tradeoff distinguishes policy learning from prediction on a fixed dataset. Model-based methods estimate dynamics and plan optimistically over plausible continuations (Brafman & Tennenholtz, 2002; Jaksch et al., 2010; Azar et al., 2017). In contrast, model-free value methods update return estimates without storing a complete transition model, and can also achieve sample-efficient exploration (Jin et al., 2018; Zhang et al., 2020). In both approaches, statistical errors propagate through Bellman recursion, making their direction, dependence, and visitation counts consequential. Structural assumptions can reduce these costs, as in linear-function-approximation guarantees (Jin et al., 2020), but require specifying what observations justify transfer. Our question concerns a different source of reuse. Observed inputs provide evidence about category frequencies, whereas an unknown conditional-response kernel determines their consequences. Input evidence alone cannot replace missing response observations.

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.

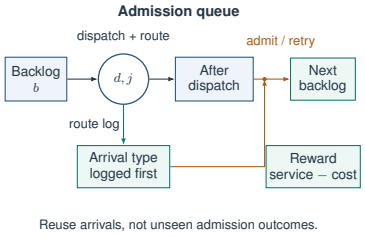


Figure 1: Input reuse and missing responses.

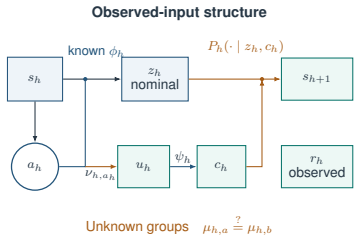


Figure 2: Observed-input transition structure.

A data-center admission queue makes this distinction concrete (Figure 1). Learning-based resource allocation and cluster scheduling motivate this domain (Mao et al., 2016; 2019). The state b is backlog, the action $a = (d, j)$ chooses dispatch capacity and an ingress route, and dispatch yields $z = (b - d)^+$, where $x^+ = \max\{x, 0\}$. The selected route reveals an arrival category $c = \psi(u)$ before admission, retries, and losses determine the next backlog. Reward balances served demand against backlog and dispatch costs. Routes with identical category laws can share frequency estimates, yet a category recorded at an empty queue does not reveal retry behavior at a congested queue. The open question is which arrivals can be pooled without imputing unobserved admission outcomes.

The abstraction applies when offered jobs are logged before buffer-dependent filtering and the remaining state–action dependence is captured by the nominal state (Section 2.1). Related examples include robotic sorting, buffered video delivery, and wireless scheduling (Kalashnikov et al., 2018; Mao et al., 2017; Mastronarde & van der Schaar, 2013). Appendix B.5 details more motivating examples, the required observations, and state dependence for each application.

Existing structural assumptions remove different sources of uncertainty. Post-decision states (PDS) distinguish modeled action effects from unresolved dynamics; additional independence or response structure can support virtual-experience updates (Sutton & Barto, 2018; Mastronarde & Van Der Schaar, 2011; Mastronarde & van der Schaar, 2013). Blind PDS learns a missing transition factor, while stochastic-sampling PDS reduces expectation-computation cost (Zhang et al., 2023; Corra et al., 2025). Observed exogenous-input methods exploit known response maps (Sinclair et al., 2023; Wan et al., 2026; Liang et al., 2026). Other structured-RL approaches use known dependency scopes, feature maps, value-function complexity, or value-equivalence relations (Tian et al., 2020; Jin et al., 2020; Jiang et al., 2017; Jin et al., 2021; Grimm et al., 2020). None of these premises turns an unverified equality into another action's observed transition.

Our focus is **unknown input-law sharing**. Compared with known-response exogenous models, our learner must still estimate each category's conditional consequences. Compared with a known partition, it must determine which observations admit pooling. State abstraction can describe equivalent predictions without providing evidence that certifies the equivalence (Ferns et al., 2004; Grimm et al., 2020). We therefore ask whether a learner can exploit valid pooling, pay to identify it, and retain the complete individual-learning comparison. Appendix A compares these objectives under their respective information assumptions.

First, *response-mediated transfer* characterizes the distinctions that matter for control. Our episodic MDP separates observed inputs from unknown responses, and Proposition 1 shows that unequal input laws can induce identical transitions when the response suppresses their discrepancy. Input equality is nevertheless necessary for zero transfer bias uniformly over unrestricted responses. An estimated cancellation also retains a bilinear error, even when its first-order terms vanish.

Second, **response-first optimism** converts valid input reuse into a regret guarantee. Certified Response-First Optimism (CRFO) separates group certification from estimation, keeps both individual and pooled input statistics, and resolves *input–response coverage mismatch* by constructing optimistic conditional-response values before averaging over input laws. Theorem 2 compares complete individual-learning and certification-plus-pooling guarantees within a factor of two, including unknown rewards and identification cost. Table 1 summarizes the induced special cases.

Third, we isolate the information price of unknown sharing in a fixed-policy task. Theorem 5 shows that *response-sensitive cancellation* can change the sample complexity order, distinguishing a first-

order response effect from a joint second-order perturbation. Without a positive separation margin and at a fixed successor gap, vanishing response contrast changes the cancellation-scale cost from inverse-square to inverse-first-power. The comparison separates reference-tuned optimality from scale-agnostic adaptation, rather than asserting a matching control-regret lower bound.

2 PROBLEM FORMULATION

We begin with a finite, undiscounted episodic Markov decision process (MDP). Write $M = (\mathcal{S}, \mathcal{A}, H, s_1, \{\bar{P}_h, R_h\}_{h=1}^H)$, where $\mathcal{S}, \mathcal{A}$ are nonempty state and action sets of sizes S, A, $H \geq 2$ is the horizon, $\bar{P}_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, and $R_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the reward-mean function. Here $\Delta(\mathcal{X})$ denotes probability laws on a finite set $\mathcal{X}$. Each of $K \geq 1$ episodes starts at the fixed state s_1. Write $[n] = \{1, \ldots, n\}$. In episode $k \in [K]$ at stage $h \in [H]$, the learner observes s_h^k, chooses $a_h^k \in \mathcal{A}$, and receives successor s_{h+1}^k and reward $r_h^k \in [0,1]$. All actions are available at every state. Kernels and means are stage-dependent but fixed across episodes.

Admissible decisions are determined by the observation history. Work on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ carrying environment observations and learner randomization. Use power-set sigma-algebras for finite alphabets and Borel sets for rewards. Let $\mathcal{F}_{k,0}$ be the episode-start history and $\mathcal{F}_{k,h}^-$ the history through the chosen action, before current feedback. Actions cannot use current feedback. Write $\mathcal{L}(\cdot \mid \mathcal{F})$ for conditional law and $\mathbb{E}$ for expectation.

2.1 OBSERVED-INPUT STRUCTURE AND UNKNOWN SHARING

The additional feedback separates an environmental input from its conditional consequence. Figure 2 shows its two branches: a known nominal action effect and an observed random input, which jointly index its unknown conditional law. A known *nominal-state map* $\phi_h : \mathcal{S} \times \mathcal{A} \to \mathcal{Z}$ gives $z_h = \phi_h(s_h, a_h)$, with $Z = |\mathcal{Z}|$. This label summarizes action effects, without requiring a physical afterstate. After acting, the learner also observes raw input $u_h \in \mathcal{U}$, with $U = |\mathcal{U}| < \infty$, and its *response-sufficient signature* $c_h = \psi_h(u_h) \in [C]$ under a known map $\psi_h : \mathcal{U} \to [C]$, where $C \geq 1$ is the category count. The unknown raw-input law $\nu_{h,a} \in \Delta(\mathcal{U})$ induces the signature law $\mu_{h,a}(c) = \sum_{u:\psi_h(u)=c} \nu_{h,a}(u)$. Identity ψ_h requires no compression. The unknown *response kernel* $P_h : \mathcal{Z} \times [C] \to \Delta(\mathcal{S})$ generates the successor from (z_h, c_h). Almost surely,

$$\mathcal{L}(u_h^k \mid \mathcal{F}_{k,h}^-) = \nu_{h,a_h^k}, \quad \mathcal{L}(s_{h+1}^k \mid \mathcal{F}_{k,h}^-, u_h^k) = P_h(\cdot \mid z_h^k, c_h^k), \quad \mathbb{E}[r_h^k \mid \mathcal{F}_{k,h}^-] = R_h(s_h^k, a_h^k). \quad (1)$$

The input law is state-independent conditional on action and stage. The response depends on state, action, and raw input only through (z_h, c_h). Reward may correlate with input and successor, while temporal dependence must respect (1). Appendix C.1 defines the nested filtrations.

There are two distinct sharing relations. The unknown *input-sharing partition* $\Pi_h^\star$ groups actions with identical $\mu_{h,a}$ and has $m_h^\star = |\Pi_h^\star|$ blocks. Separately, matching nominal labels share response rows (h, z, c). Neither implies equal raw-input laws, shared rewards, or cross-stage pooling. Alphabets and maps are known, but ν, μ, P and the partitions are unknown. The known *reward-information indicator* ζ_R is zero for known R_h and one for unknown means.

2.2 BELLMAN PREDICTION AND PERFORMANCE

Bellman prediction integrates response and input randomness separately. For $V : \mathcal{S} \to \mathbb{R}$, define $(P_h V)(z, c) = \sum_{s'} P_h(s' \mid z, c) V(s')$ and $(\bar{P}_h V)(s, a) = \sum_c \mu_{h,a}(c)(P_h V)(\phi_h(s, a), c)$. Accordingly, $\bar{P}_h(s' \mid s, a) = \sum_c \mu_{h,a}(c) P_h(s' \mid \phi_h(s, a), c)$ is the *marginal transition kernel*.

A deterministic Markov policy $\pi = (\pi_h)_{h=1}^H$ maps states to actions. Its value $V_h^\pi(s) = \mathbb{E}_{h,s}^\pi[\sum_{t=h}^H r_t]$ initializes the process at (h, s), including unvisited states. Define the *action-value operator* $\mathcal{B}_h V = R_h + \bar{P}_h V$, *policy-evaluation operator* $(\mathcal{T}_h^\pi V)(s) = (\mathcal{B}_h V)(s, \pi_h(s))$, and *optimality operator* $(\mathcal{T}_h^* V)(s) = \max_a (\mathcal{B}_h V)(s, a)$. Backward recursion from $V_{H+1}^\pi = V_{H+1}^* = 0$ yields policy values $V_h^\pi = \mathcal{T}_h^\pi V_{h+1}^\pi$, optimal values $V_h^* = \mathcal{T}_h^* V_{h+1}^*$, and action values $Q_h^\pi = \mathcal{B}_h V_{h+1}^\pi$ (choose the first action, then follow π). Also $Q_h^* = \mathcal{B}_h V_{h+1}^*$ and $0 \leq V_{h+1}^\pi \leq B_h := H - h$. The *occupancy probability* $\omega_{M,h}^\pi(s, a) = \mathbb{P}_M^\pi(s_h = s, a_h = a)$ measures exposure from the fixed initial state.

Table 1: Leading $\widetilde{O}$ regret with known exact structure.

Estimator and representation	Leading regret order
Input-factor estimator, m groups	$H^2\sqrt{mCK} + H^2\sqrt{SZCK}$
Collapsed groups, $y = (z, g)$	$H^2\sqrt{SZ_0K}$, $Z_0 \le \min\{SA, Zm\}$
Deterministic PDS, $y = f_h(s, a)$	$H^2\sqrt{SZ_0K}$
Ordinary MDP, $y = (s, a)$	$H^2S\sqrt{AK}$

For $\mathcal{F}_{k,0}$-measurable episode policies π^k, define *cumulative pseudo-regret*, including certification, by

$$\mathrm{Reg}(K) = \sum_{k=1}^{K} [V_1^*(s_1) - V_1^{\pi^k}(s_1)]. \tag{2}$$

The randomness is through the learned policies, not realized rewards. Section 5.3 defines the separate fixed-policy objective. We use total variation (TV), $\mathrm{TV}(p, q) = \frac{1}{2}\sum_x |p(x) - q(x)|$, vector span $\mathrm{span}(v) = \max_x v(x) - \min_x v(x)$, $x \vee y = \max\{x, y\}$, and $\widetilde{O}$ to suppress logarithms.

Motivating Example 1: In Figure 1, $s = b \in \{0, \dots, B_{\max}\}$ is backlog and $a = (d, j)$ selects dispatch and route. The known map $\phi_h(b, (d, j)) = (b - d)^+$ describes residual backlog, not downstream completion. Known ψ_h labels offered jobs before filtering. Route laws, equality groups, and admission/retry kernels are unknown. Reward is a $[0, 1]$-normalized service benefit minus backlog and dispatch costs, known in mean only when $\zeta_R = 0$. Pooling categories does not reveal responses at different (z, c). Accepted-job-only logs can violate the input premise (see Appendix B.5).

3 Warm-up: representations and valid input sharing

Representation and statistical information must be distinguished before specifying a pooling rule. Marginalizing the signature yields an ordinary MDP with kernel $\bar{P}_h$, preserving values but potentially losing information about sharing. Conversely, $C = 1, z = (s, a)$ recovers ordinary MDP feedback, and $C = 1, z = f_h(s, a)$ recovers a deterministic PDS with known afterstate map f_h. Retaining (z, c) as an intermediate state with no additional decision preserves the original feedback.

For an exact input partition Π_h with m_h blocks, let $G_h(a)$ and $g_h(a)$ denote the block containing a and its label. Then, $y_h = (\phi_h(s, a), g_h(a))$ is an integrated afterstate with at most $\min\{SA, Zm_h\}$ reachable labels. A single block makes z sufficient. Unequal laws can also permit collapse when their discrepancies cancel through the response (Proposition 7). Appendix B.2 gives stochastic-PDS reductions and their feedback and dimensional costs.

The statistical consequence is a source-to-target Bellman bias. Replacing the target law $\mu_{h,a}$ by a source law $\mu_{h,a'}$ at fixed z and continuation $V \in [0, B_h]^{\mathcal{S}}$ changes the expected continuation by at most $B_h\,\mathrm{TV}(\mu_{h,a'}, \mu_{h,a})$, by total-variation duality. This motivates certified within-block diameters

$$\max_{a,a' \in G} \mathrm{TV}(\mu_{h,a}, \mu_{h,a'}) \le \epsilon_{h,G} \quad (G \in \Pi_h), \qquad \epsilon_h = \max_{G \in \Pi_h} \epsilon_{h,G}. \tag{3}$$

The true equality partition $\Pi_h^\star$ has $m_h^\star$ blocks, distinct from the returned count m_h. Singletons and exact groups have zero diameter. The allowance $\epsilon_{h,G}$ enters the pooling radius and Theorem 2; it is distinct from the cancellation scale ε. Table 1 reports leading terms for known groups, with $m = \max_{h<H} m_h$ and Z_0 the maximum number of afterstate labels at a nonterminal stage. Add $\widetilde{O}(\zeta_R H\sqrt{SAK})$ for unrelated unknown rewards. These estimator-specific terms exclude certification and are not minimax rates or complete finite-time bounds (Corollary 8).

The factorized estimator reuses inputs across nominal states. The collapsed estimator instead learns integrated laws, so equivalent policy values need not entail equal statistical complexity. Moreover, a particular response can erase unequal inputs' effect on transitions. The following characterization connects this distinction to Theorem 5. At fixed h, z, write P_z for the $C \times S$ matrix with rows $P_h(\cdot \mid z, c)$ and $d = \mu_{h,b} - \mu_{h,a}$ for the signed source–target input discrepancy. Define the *transition-image discrepancy* $\mathsf{D}_z(a, b) = \frac{1}{2}\|d^\top P_z\|_1$ and the *Dobrushin response-contraction coefficient* $\chi(P_z) = \max_{c,c'} \mathrm{TV}(P_z(c, \cdot), P_z(c', \cdot))$. Here $\hat{P}_z$ estimates the response matrix, and $\chi(P_z)$ measures successor-law variation across categories.

Algorithm 1 Certified Response-First Optimism (CRFO)

Require: $H, K, \delta, \phi, \psi, \zeta_R$, known R if $\zeta_R = 0$, and a certification mode.
Obtain (τ, Π, ϵ) using Section 4.1; reset all estimation counters.
for $k = \tau + 1, \ldots, K$ **do**
  Set $\overline{V}^k_{H+1} = 0$.
  for $h = H, \ldots, 1$ **do**
    Compute $\bar{R}^k_h$, $W^{\mathrm{MB,k}}_h$ by (6), and $\overline{Q}^k_h$ by (7).
    Set $\overline{V}^k_h = \max_a \overline{Q}^k_h(\cdot, a)$ and select smallest-index maximizing actions.
  end for
  Freeze $\overline{V}^k, \pi^k$ and execute the episode, observing $(u^k_h, c^k_h, s^k_{h+1}, r^k_h)$.
  Update distinct input streams and only the executed response and unknown-reward rows.
end for

Proposition 1 (Response-mediated sharing and bilinear transfer)**.** *For every pair of input laws,*

$$\sup_{v \in [0,1]^S} |d^\top P_z v| = \mathsf{D}_z(a, b) \le \chi(P_z)\, \mathrm{TV}(\mu_{h,a}, \mu_{h,b}). \tag{4}$$

For $S \ge 2$, the supremum of $\mathsf{D}_z(a, b)$ over unrestricted response kernels equals $\mathrm{TV}(\mu_{h,a}, \mu_{h,b})$. Let probability estimates have total-variation errors at most e_a, e_b for the two input laws and e^P_c for response row c. Put $\hat{d} = \hat{\mu}_{h,b} - \hat{\mu}_{h,a}$ and $e_\mu = e_a + e_b$. Uniformly in bounded v,

$$|d^\top P_z v - \hat{d}^\top \hat{P}_z v| \le e_\mu \,\mathrm{span}(\hat{P}_z v) + \mathrm{span}(v) \Big[\sum\nolimits_c |\hat{d}_c| e^P_c + 2 e_\mu \max\nolimits_c e^P_c \Big]. \tag{5}$$

Estimated agreement does not establish exact cancellation. Even when $\hat{d} = 0$ and $\hat{P}_z v$ is constant, the interaction allowance remains necessary. Our algorithms therefore certify input-law equality rather than treating estimated cancellation as known structure.

Proof sketch. The main issue is that transfer errors involve two estimated factors. Total-variation duality gives (4). Expanding $d^\top P_z v - \hat{d}^\top \hat{P}_z v$ separates input error, response error, and their product, yielding (5) without assuming estimator independence. See Appendix B.4 for the complete proof, a sharp second-order example, and occupancy-weighted propagation. □

4 Algorithm Design

This section provides a new algorithm, (see Algorithm 1), Certified Response-First Optimism (CRFO), which separates structural certification from optimistic planning. Because reusable inputs do not reveal every conditional response, it first certifies groups and then estimates inputs, responses, and rewards separately. Its *response-first recursion* bounds conditional continuation values before averaging input frequencies, preserving actual response coverage. The model-based (MB) method stores successor histograms. Appendix D develops a lower-storage temporal-difference (TD) extension, which is a model-free RL algorithm for the conditional response.

4.1 Certification and input estimation

The first challenge is *selection-dependent pooling*. Grouping empirically similar actions and treating the same observations as a preselected shared sample can understate uncertainty. CRFO instead fixes the partition before collecting its estimation data. Conditional on the certification history, groups are fixed while subsequent actions may remain adaptive.

To make the confidence bounds simultaneous, for $\delta \in (0, 1)$ define $L = \log[(64H(2A + ZC + \zeta_R SA)(K + 1))/\delta]$ and $L_d = d \log 2 + L$. The *categorical radius* for a d-outcome law is $r_d(0) = 1$ and $r_d(n) = \min\{1, \sqrt{L_d/(2n)}\} \quad (n \ge 1)$. Here d is an outcome-alphabet size. The term L covers stages, streams, and counts up to K, while $d \log 2$ covers all outcome subsets. Distribution-level confidence therefore controls continuations fitted from the same history.

Exact pooling needs more than empirical agreement. Suppose a known $\gamma \in (0, 1]$ makes every pair of signature laws at $h < H$ equal or at least γ apart in total variation. Each round-robin cycle uses A episodes and samples every action once at each stage. After n cycles, let $\hat{\mu}^n_{h,a}$ be the certification histogram and $\hat{d}_h(a, a') = \mathrm{TV}(\hat{\mu}^n_{h,a}, \hat{\mu}^n_{h,a'})$ its pairwise discrepancy. With $r = r_C(n)$, certify equality when $\hat{d}_h + 2r < \gamma$ and difference when $\hat{d}_h - 2r > 0$. The two radii account for uncertainty in both laws. Stop when each pair at $h < H$ has one label and equality edges form disjoint cliques. Simultaneous confidence ensures resolution by $n_\gamma = \lceil 32 L_C/\gamma^2 \rceil$ cycles; otherwise return singletons.

Certification may be valid yet too costly. The *certification-cost gate* therefore runs this test only if $An_\gamma < K$ and $HAn_\gamma \le F^A_{\mathrm{MB}}$, the computable individual-learning bound defined in Section 5.1; otherwise it returns singletons immediately. This limits discovery expenditure without knowing the true group count. Known groups need no test. Without the separation premise, use singletons or valid diameter certificates with their explicit mismatch charge, not unverified exact pooling. These tests do not establish state independence or signature sufficiency.

At the certification stopping time τ, fix the returned partitions Π_h, group diameters $\epsilon_{h,G}$, and singleton groups at H, then reset all statistics. For target a, let $G_h(a)$ be its block and $I_h(a) = \{\{a\}, G_h(a)\}$ its distinct input streams. Define the input counter $N^k_{h,i,c} = \sum_{j=\tau+1}^{k-1} \mathbf{1}\{a^j_h \in i, c^j_h = c\}$, total $n^k_{h,i} = \sum_c N^k_{h,i,c}$, and histogram $\hat{\mu}^k_{h,i}(c) = N^k_{h,i,c}/n^k_{h,i}$, uniform at zero count. The *bias-aware radius* controls both sampling error and certified group mismatch $\lambda^k_{h,i} = \min\{1, r_C(n^k_{h,i}) + \epsilon_{h,i}\}$ and $\epsilon_{h,\{a\}} = 0$. Sampling error shrinks, but mismatch can persist. Each distinct stream is updated once, and confidence bounds allow the observations to overlap.

4.2 Response-first optimistic planning

The second challenge is *input–response coverage mismatch*. An input category updates both input streams, but its successor updates only the executed row (h, z, c). For example, a larger pool of arrival categories does not reveal retry behavior at an unvisited backlog. The planner must therefore use separate confidence terms for input frequencies, conditional responses, and rewards.

For unknown rewards, let $N^k_{h,s,a}$ count post-certification visits to (h, s, a) and let $\hat{R}^k_h(s, a)$ be their mean reward, zero when unvisited. Set $\rho_R(0) = 1$, $\rho_R(n) = \min\{1, \sqrt{L/(2n)}\}$ for $n \ge 1$, and $\bar{R}^k_h = \min\{1, \hat{R}^k_h + \rho_R(N^k_{h,s,a})\}$. For known means use $\bar{R}^k_h = R_h$. Input-law equality does not justify pooling these rewards. Appendix G gives the explicit reward counters and sums.

For each response row, $n^k_{h,z,c}$ counts actual post-certification visits and $\hat{P}^k_h$ is their successor histogram, uniform at zero count. With $\overline{V}^k_{H+1} = 0$, define

$$W^{\mathrm{MB,k}}_h(z, c) = \min\{B_h, (\hat{P}^k_h \overline{V}^k_{h+1})(z, c) + B_h r_S(n^k_{h,z,c})\}. \tag{6}$$

Here $W^{\mathrm{MB,k}}_h$ is the *upper conditional continuation*. Its bonus uses the actual response count, not a pooled input count. Since $\hat{P}^k_h$ and $\overline{V}^k_{h+1}$ depend on the same history, fixed-function concentration is insufficient. The uniform categorical event justifies this fitted continuation at adaptive counts.

Response-first planning uses $W^k_h = W^{\mathrm{MB,k}}_h$ to average each valid input stream, giving

$$\overline{Q}^k_h(s, a) = \min\Big\{B_h + 1, \bar{R}^k_h(s, a) + \min_{i \in I_h(a)} \Big[\sum\nolimits_c \hat{\mu}^k_{h,i}(c) W^k_h(\phi_h(s, a), c) + B_h \lambda^k_{h,i}\Big]\Big\}. \tag{7}$$

The *input-transfer allowance* $B_h \lambda^k_{h,i}$ controls each fitted response vector. Every candidate is an upper bound on the joint confidence event, so their minimum and return-range clipping preserve optimism. An individual estimate alone does not repair an invalid group certificate. Set $\overline{V}^k_h(s) = \max_a \overline{Q}^k_h(s, a)$, break ties by smallest action index, and freeze the policy and values for the episode.

Two extensions preserve this estimation logic. *Confidence-intersection planning* requires one target law to satisfy both input confidence sets before maximizing, eliminating mutually incompatible stream maximizers. Appendix C.9 defines its backup and numerical upper certificates, and Appendix C.10 retains the computed tightening in regret accounting. In contrast, TD reduces response storage from $O(HZCS)$ to $O(HZC)$ but reuses historical targets. Appendix D separates fixed-target noise from historical optimistic error. TD values need only dominate the optimal conditional continuation.

5 THEORETICAL GUARANTEES AND PROOF MECHANISMS

Theorem 2 bounds control regret including certification, conditional responses, and unknown rewards. Lemmas 3–4 connect input reuse to actual observations. Theorem 5 instead isolates fixed-policy identification. Its matching bounds do not characterize optimal control regret.

5.1 REGRET WITH CERTIFICATION AND RESPONSE-LEARNING COSTS

The first step converts local confidence radii into cumulative estimation costs. For J distribution rows with d outcomes each, the *categorical visit budget* for at most K total visits is $\mathcal{C}_d(J,K) = J + \sqrt{2L_dJK}$ (Lemma 12). The terms charge initialization and within-row summation followed by Cauchy–Schwarz across rows. Write $H_1 = \sum_{h<H} B_h = H(H-1)/2$ and $H_2 = \sum_{h<H} B_h^2 = H(H-1)(2H-1)/6$ for the sums of continuation ranges and their squares. Using the individual and pooled streams, define the cumulative signature-error bound

$$\mathfrak{S} = \min\Big\{ H_1\mathcal{C}_C(A,K),\ \sum\nolimits_{h<H} B_h[\mathcal{C}_C(m_h,K) + K\epsilon_h] \Big\}. \tag{8}$$

The branches count A individual laws or m_h pooled laws with persistent mismatch $K\epsilon_h$. Unknown reward means add the separate charge $\mathcal{G}_R = 2\zeta_R H[SA + \sqrt{2LSAK}]$. For a nonnegative input-error budget x, define the *complete post-certification regret envelope*

$$F_{\mathrm{MB}}(x) = 2x + 2H_1\mathcal{C}_S(ZC,K) + 4\sqrt{2KH_2L} + \mathcal{G}_R. \tag{9}$$

Its four terms account for input estimation, conditional-response estimation, trajectory fluctuations, and reward estimation, respectively. Signature equality reduces only the first. The individual-learning and true-group envelopes are $F^A_{\mathrm{MB}} = F_{\mathrm{MB}}(H_1\mathcal{C}_C(A,K))$ and $F^m_{\mathrm{MB}} = F_{\mathrm{MB}}(\sum_{h<H} B_h\mathcal{C}_C(m^\star_h,K))$. Here $m^\star_h$ counts true equality classes, whereas m_h in $\mathfrak{S}$ counts returned blocks. The gate uses only the computable F^A_{MB}, never $m^\star_h$.

Theorem 2 (Regret of certified signature distribution sharing)**.** *Under the model in Section 2, use CRFO with fresh post-certification estimates. Suppose a known $\gamma \in (0,1]$ makes each pair of signature laws at $h < H$ either equal or at least γ apart in total variation. With $n_\gamma = \lceil 32L_C/\gamma^2 \rceil$ and the certification gate of Section 4.1, with probability at least $1-\delta$,*

$$\mathrm{Reg}_{\mathrm{MB}}(K) \le \min\{HK,\ 2\min\{F^A_{\mathrm{MB}},\ HAn_\gamma + F^m_{\mathrm{MB}}\}\}. \tag{10}$$

More generally, any episode-end stopping time $0 \le \tau \le K$ returning history-measurable partitions and diameter bounds $\epsilon_{h,G} \in [0,1]$, valid with probability at least $1-\delta/4$, gives $\mathrm{Reg}_{\mathrm{MB}}(K) \le \min\{HK, H\tau + F_{\mathrm{MB}}(\mathfrak{S})\}$. Known exact groups take $\tau = \epsilon_h = 0$, where $\epsilon_h = \max_{G\in\Pi_h}\epsilon_{h,G}$.

The bound distinguishes two forms of reuse rather than replacing every action count by a group count. For exact groups with $m_h \le m$, its leading terms are $\widetilde{O}(H^2\sqrt{mCK} + H^2\sqrt{SZCK} + \zeta_R H\sqrt{SAK})$, plus certification. Here m measures input-law sharing, while ZC counts conditional-response rows shared through the nominal-state map. Consequently, input pooling may yield little benefit when response or reward estimation dominates. The factor-two comparison includes identification and compares theoretical bounds, not the realized regrets of independent learners.

Proof sketch. The difficulty is to account for selected groups without granting counterfactual responses. Lemmas 3–4 transfer confidence through the executed law and telescope Bellman errors. Counting actual visits gives $F_{\mathrm{MB}}(\mathfrak{S})$. If certification runs, its cost is at most $C_{\mathrm{cert}} = HAn_\gamma \le F^A_{\mathrm{MB}}$, and exact recovery bounds subsequent regret by both F^A_{MB} and F^m_{MB}. Thus total regret is at most $\min\{2F^A_{\mathrm{MB}}, C_{\mathrm{cert}} + F^m_{\mathrm{MB}}\}$. The gate handles skipped certification, while general certificates retain $H\tau$ and mismatch. See Appendices C and C.7 for the complete proof. □

5.2 PROOF MECHANISMS ON THE EXECUTED TRAJECTORY

To control all estimates used in planning, let $\mathcal{E}$ combine valid sharing certificates with simultaneous confidence bounds for signature laws, conditional transition kernels, and reward means, with failure probability below $\delta/2$ (Appendix C). Pooled histograms concentrate around the average conditional

signature law of the sampled actions. Certificates bound its target-law discrepancy, while uniform confidence covers adaptive counts and fitted continuations.

The first proof step must transfer confidence without transferring nonexistent response observations. The following inequality keeps the conditional-response error under the target action's true law, making that error accessible through the executed trajectory.

Lemma 3 (Response-first transfer and optimism)**.** *On $\mathcal{E}$, for every post-certification episode k, $h < H$, state–action pair (s, a), and $i \in I_h(a)$, with response vectors evaluated at $z = \phi_h(s, a)$,*

$$0 \le \hat{\mu}_{h,i}^{k\top} W_h^{\mathrm{MB,k}} + B_h \lambda_{h,i}^k - \mu_{h,a}^\top P_h \overline{V}_{h+1}^k \le 2B_h \lambda_{h,i}^k + \mu_{h,a}^\top (W_h^{\mathrm{MB,k}} - P_h \overline{V}_{h+1}^k). \tag{11}$$

For each signature c, $0 \le W_h^{\mathrm{MB,k}}(z, c) - (P_h \overline{V}_{h+1}^k)(z, c) \le 2B_h r_S(n_{h,z,c}^k)$. The planned values satisfy $\overline{Q}_h^k \ge Q_h^*$ and $\overline{V}_h^k \ge V_h^*$ at every stage.

Intuitively, the source stream estimates frequencies, but the executed target determines response visits. The remainder $\mu_{h,a}^\top (W_h^{\mathrm{MB,k}} - P_h \overline{V}_{h+1}^k)$ preserves this distinction. Accurate pooled frequencies do not eliminate uncertainty at an unvisited nominal state.

Proof sketch. Insert the true target-law average $\mu_{h,a}^\top W_h^{\mathrm{MB,k}}$. Uniform total-variation confidence controls the difference from each empirical average even though $W_h^{\mathrm{MB,k}}$ is fitted from the same history. The remaining error is nonnegative and uses the true target law. Backward induction gives optimism because upper reward estimates, the stream minimum, and clipping preserve valid upper bounds. See Appendix C.5 for the complete proof. □

The second step converts this conditional expectation into an actual-visit sum. A martingale accounts for the difference, but its concentration must hold before conditioning on future confidence or optimism. Write $\mathrm{Reg}_{\mathrm{post}} = \sum_{k=\tau+1}^K [V_1^*(s_1) - V_1^{\pi^k}(s_1)]$ for post-certification pseudo-regret.

Lemma 4 (Actual-visit regret accounting)**.** *Define the value-function error $d_h^k = \overline{V}_h^k - V_h^{\pi^k}$ and the reward-error bound $\beta_{h,R}^k = \zeta_R \min\{1, 2\rho_R(N_{h,s_h^k,a_h^k}^k)\}$. On $\mathcal{E}$,*

$$\begin{aligned}\mathrm{Reg}_{\mathrm{post}} \le\ & 2\sum_{k=\tau+1}^K \sum_{h<H} B_h \min_{i \in I_h(a_h^k)} \lambda_{h,i}^k + 2\sum_{k=\tau+1}^K \sum_{h<H} B_h r_S(n_{h,z_h^k,c_h^k}^k) \\ & + \sum_{k=\tau+1}^K \sum_{h=1}^H \beta_{h,R}^k + \mathcal{M}_K,\end{aligned} \tag{12}$$

where $\mathcal{M}_K$ is a martingale sum, with the unconditional tail $\mathbb{P}\{\mathcal{M}_K > 2\sqrt{2KH_2L}\} \le e^{-L}$.

The bound charges the tighter valid input stream, the response row reached, and the reward row executed, including stage H. Thus sharing can reduce one estimation cost without reducing the others. The martingale accounts for discrepancies between expected and observed response visits.

Proof sketch. Define $Y_{k,h} = 2B_h r_S(n_{h,z_h^k,c_h^k}^k) + d_{h+1}^k(s_{h+1}^k)$ for $k > \tau$ and zero otherwise. With the pre-observation sigma-algebra $\mathcal{F}_{k,h}^-$, set $\mathcal{M}_K = \sum_{k=1}^K \sum_{h<H} (\mathbb{E}[Y_{k,h} \mid \mathcal{F}_{k,h}^-] - Y_{k,h})$. Inserting these centered increments into the Bellman recursion yields (12), with terminal error $d_H^k(s_H^k) \le \beta_{H,R}^k$. The all-history range $Y_{k,h} \in [-B_h, 3B_h]$ gives the Hoeffding tail before intersection with $\mathcal{E}$, without input–successor independence. See Appendices C.5–C.6 for the complete proof. □

5.3 First-order and second-order costs of unknown sharing

The control bound does not determine whether discovering sharing is intrinsically expensive. We isolate this question through a fixed-policy selection task with known reward formulas. The key mechanism is *response-sensitive cancellation*: breaking an input equality affects the decision only through differences between conditional responses.

Consider $H = S = A = C = 2$, $Z = 1$, and raw input $u = c$, with states, actions, and categories relabeled by $\{0, 1\}$. First-stage action $a \in \{0, 1\}$ generates category $c \sim \mathrm{Bern}(p_a)$ and successor

$Y \mid c \sim \mathrm{Bern}(m_c)$, where $\mathrm{Bern}(p)$ denotes a Bernoulli law and $p_a, m_c \in (0,1)$ are unknown. Other terminal observations are parameter-independent. First-stage rewards are zero, and terminal rewards are $R_2(Y,0) = Y$ and $R_2(Y,1) = (1-\varepsilon)Y + \varepsilon/2$, where $0 < \varepsilon \le 1/2$ is the known *cancellation scale*. The target policies π and ρ choose actions zero and one, respectively, at both stages. With $q_a = (1-p_a)m_0 + p_a m_1$, their value contrast is

$$\theta = V_1^\pi - V_1^\rho = \varepsilon(q_0 - 1/2) - (1-\varepsilon)(p_1 - p_0)(m_1 - m_0). \tag{13}$$

Let $\mathcal{M}_\gamma$ be the model class with $\theta \neq 0$ and either $p_0 = p_1$ or $|p_1 - p_0| \ge \gamma$, and let $\mathcal{M}_=$ be its equality subclass. A procedure in $\mathcal{A}_\delta(\mathcal{M})$ stops almost surely and selects the better target with probability at least $1-\delta$ on every model in $\mathcal{M}$. Its stopping time T counts complete episodes revealing (c, Y). Thus correctness is uniform over the class, whereas expected cost below is evaluated at one reference instance. At the shared reference $M_{\Delta,\kappa}$, put $p_0 = p_1 = 1/2$ and $m_c = 1/2 + \Delta + (2c-1)\kappa/2$, where $\Delta = q_a - 1/2$ is the common successor-mean gap and $\kappa = m_1 - m_0$ is the response contrast. Define the *bilinear resolution* and *decision resolution* $r_* = \frac{\varepsilon\Delta}{\kappa + \sqrt{\varepsilon\Delta}}$ and $s_* = \min\{\Delta, \max\{\gamma, r_*\}\}$. The balance $\kappa r_* + r_*^2 \asymp \varepsilon\Delta$, where $\asymp$ denotes equality up to absolute constant factors, distinguishes linear sensitivity from joint perturbations.

Theorem 5 (Response-sensitive information price)**.** *For $0 < \delta \le 1/4$, $0 < \varepsilon \le 1/2$, $0 \le \gamma \le 1/7$, $0 < \Delta \le 1/20$, and $0 \le \kappa \le 7/10$,*

$$\inf_{\mathcal{A} \in \mathcal{A}_\delta(\mathcal{M}_=)} \mathbb{E}_{M_{\Delta,\kappa}} T = \Theta(\Delta^{-2}\log(1/\delta)), \inf_{\mathcal{A} \in \mathcal{A}_\delta(\mathcal{M}_\gamma)} \mathbb{E}_{M_{\Delta,\kappa}} T = \Theta(s_*^{-2}\log(1/\delta)). \tag{14}$$

The single rule initialized at one episode per action, without Δ or κ, satisfies $\mathbb{E}T \le 2^{28} s_*^{-2}[\log(32/\delta) + 2\log(2 + \log_2(s_*^{-2}))]$. For known equality, replace s_* by Δ.

The theorem shows that the same small policy gap can have different information costs in the same observation model. For $\gamma = 0$, the unknown-sharing price is $\Theta(\max\{\Delta^{-2}, \kappa^2/(\varepsilon^2\Delta^2) + 1/(\varepsilon\Delta)\}\log(1/\delta))$. At fixed Δ and $\kappa = 0$, the cost grows as ε^{-1} rather than ε^{-2} for fixed positive κ. Response sensitivity, not the policy gap alone, controls this information cost. The learner is not told that $\kappa = 0$ at a flat response; correctness still covers unequal responses.

Proof sketch of Theorem 5. One alternative shifts the common response by 2Δ. Another perturbs input and response by t, u, with $t \ge \gamma$ and $t(\kappa + u) \ge 2\varepsilon\Delta/(1-\varepsilon)$. Their complete-feedback Kullback–Leibler (KL) costs are $O(\Delta^2)$ and $O(\max\{\gamma, r_*\}^2)$. Conversely, Lemma 6 resolves at r_* or certifies equality before exploiting the common gap Δ. An unconditional stopping-time tail accounts for confidence failures in expected cost. See Appendix E.6 for the complete proof. □

Lemma 6 (Product-aware contrast certificate)**.** *Let $\hat{p}_a, \hat{m}_c \in [0,1]$ be empirical probability estimates with simultaneous error at most r, and define $\hat{d} = \hat{p}_1 - \hat{p}_0$, $\hat{\kappa} = \hat{m}_1 - \hat{m}_0$, and $\hat{q}_0 = (1-\hat{p}_0)\hat{m}_0 + \hat{p}_0\hat{m}_1$. Without independence of the estimates, $|\theta - \hat{\theta}| \le B$, where*

$$\hat{\theta} = \varepsilon(\hat{q}_0 - 1/2) - (1-\varepsilon)\hat{d}\hat{\kappa}, \qquad B = \varepsilon r + (2-\varepsilon)|\hat{\kappa}|r + 2(1-\varepsilon)|\hat{d}|r + (4-2\varepsilon)r^2. \tag{15}$$

The interaction term remains even when $\hat{d} = \hat{\kappa} = 0$. Thus, empirical cancellation cannot be treated as a known equality, and a purely first-order interval would miss the second-order regime.

Proof sketch. Expand both factors in (13). The terms in B bound common-mean error, the two sensitivities, and their product. Simultaneous mean confidence, not independence, controls this interaction. A sign-resolving interval or certified equality triggers stopping. See Lemma 37 and Appendix E.6 for the complete proof and stopping rule. □

6 Conclusion

Cross-action input sharing can reduce the cost of learning input frequencies, but it does not create conditional-response observations for unexecuted state–action configurations. Building on this distinction, CRFO accounts jointly for sharing identification and response coverage along the executed trajectory while retaining an individual-learning comparison. Our fixed-policy analysis further shows that the statistical cost of unknown sharing depends critically on conditional-response sensitivity, exhibiting a transition between first-order and second-order information regimes. Consequently, neither the number of sharing groups nor the policy gap alone characterizes the value or cost of exploiting shared experience. Numerical results illustrating these effects are provided in Appendix H.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and polishing for clarity, grammar, and readability, and to assist with the generation and editing of software code. All AI-assisted text was reviewed and revised by the authors, and all AI-assisted code was inspected, tested, and verified by the authors for correctness. The authors take full responsibility for the scientific content, correctness, and final presentation of this work.

ETHICS STATEMENT

This work is theoretical and simulation-based and does not involve human subjects, personal or sensitive data, or experiments on deployed systems. The application examples are used to motivate the mathematical model and should not be interpreted as validated deployment claims. All numerical results are generated from synthetic environments specified in the paper and supplementary materials. We have aimed to report assumptions, limitations, negative experimental findings, and the scope of the theoretical guarantees accurately and transparently.

REPRODUCIBILITY STATEMENT

The main text specifies the problem model, algorithms, assumptions, confidence parameters, and principal theoretical guarantees. The appendices provide complete proofs, supporting lemmas, boundary cases, implementation details, and the full experimental protocols and additional results. The supplementary materials include the simulation code, configurations, and saved numerical results used for the reported experiments. Random seeds and experimental settings are specified to support reproduction of the numerical findings.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pp. 12–38. PMLR, 2017.

Noga Alon, Nicolo Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Conference on Learning Theory*, pp. 23–35. PMLR, 2015.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International conference on machine learning*, pp. 263–272. PMLR, 2017.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. *Advances in Neural Information Processing Systems*, 30, 2017.

Andrew Corra, Nicholas Mastronarde, and Jacob Chakareski. Low-complexity physics-informed reinforcement learning using post-decision states with stochastic sampling. In *ICC 2025-IEEE International Conference on Communications*, pp. 4559–4564. IEEE, 2025.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pp. 2701–2709. PMLR, 2020.

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

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pp. 100–118, 1975.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pp. 998–1027. PMLR, 2016.

Stéphane Gaubert and Zheng Qu. Dobrushin's ergodicity coefficient for markov operators on cones. *Integral Equations and Operator Theory*, 81(1):127–150, 2015.

Avishek Ghosh, Sayak Ray Chowdhury, and Kannan Ramchandran. Model selection for generic reinforcement learning. *arXiv preprint arXiv:2107.05849*, 2021.

Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. *Advances in neural information processing systems*, 33: 5541–5552, 2020.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil' ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, volume 35, pp. 423–439. PMLR, 2014.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on learning theory*, pp. 2137–2143. PMLR, 2020.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, and Sergey Levine. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Proceedings of the 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pp. 651–673. PMLR, 2018.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

Hao Liang, Jiayu Cheng, Sean Sinclair, and Yali Du. Is pure exploitation sufficient in exogenous mdps with linear function approximation? In *International Conference on Learning Representations*, volume 2026, pp. 15304–15349, 2026.

Hongzi Mao, Mohammad Alizadeh, Ishai Menache, and Srikanth Kandula. Resource management with deep reinforcement learning. In *Proceedings of the 15th ACM workshop on hot topics in networks*, pp. 50–56, 2016.

Hongzi Mao, Ravi Netravali, and Mohammad Alizadeh. Neural adaptive video streaming with pensieve. In *Proceedings of the conference of the ACM special interest group on data communication*, pp. 197–210, 2017.

Hongzi Mao, Malte Schwarzkopf, Shaileshh Bojja Venkatakrishnan, Zili Meng, and Mohammad Alizadeh. Learning scheduling algorithms for data processing clusters. In *Proceedings of the ACM special interest group on data communication*, pp. 270–288. 2019.

Nicholas Mastronarde and Mihaela Van Der Schaar. Fast reinforcement learning for energy-efficient wireless communication. *IEEE Transactions on Signal Processing*, 59(12):6262–6266, 2011.

Nicholas Mastronarde and Mihaela van der Schaar. Joint physical-layer and system-level power management for delay-sensitive wireless communications. *IEEE transactions on mobile computing*, 12(4):694–709, 2013.

Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1392–1403, 2020.

Jungseul Ok, Alexandre Proutiere, and Damianos Tranos. Exploration in structured reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Matteo Papini, Andrea Tirinzoni, Aldo Pacchiano, Marcello Restelli, Alessandro Lazaric, and Matteo Pirotta. Reinforcement learning in linear mdps: Constant regret and representation selection. *Advances in Neural Information Processing Systems*, 34:16371–16383, 2021.

Clémence Réda, Andrea Tirinzoni, and Rémy Degenne. Dealing with misspecification in fixed-confidence linear top-m identification. *Advances in Neural Information Processing Systems*, 34: 25489–25501, 2021.

Geovani Rizk, Albert Thomas, Igor Colin, Rida Laraki, and Yann Chevaleyre. Best arm identification in graphical bilinear bandits. In *International Conference on Machine Learning*, volume 139, pp. 9010–9019. PMLR, 2021.

Amirhossein Roknilamouki, Arnob Ghosh, Ming Shi, Fatemeh Nourzad, Eylem Ekici, and Ness B Shroff. Provably efficient rl for linear mdps under instantaneous safety constraints in non-convex feature spaces. *arXiv preprint arXiv:2502.18655*, 2025.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

Ming Shi, Yingbin Liang, and Ness Shroff. A near-optimal algorithm for safe reinforcement learning under instantaneous hard constraints. In *International Conference on Machine Learning*, pp. 31243–31268. PMLR, 2023.

Sean R Sinclair, Felipe Vieira Frujeri, Ching-An Cheng, Luke Marshall, Hugo De Oliveira Barbalho, Jingling Li, Jennifer Neville, Ishai Menache, and Adith Swaminathan. Hindsight learning for mdps with exogenous inputs. In *International Conference on Machine Learning*, pp. 31877–31914. PMLR, 2023.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

Yi Tian, Jian Qian, and Suvrit Sra. Towards minimax optimal reinforcement learning in factored markov decision processes. *Advances in Neural Information Processing Systems*, 33:19896–19907, 2020.

Andrea Tirinzoni, Matteo Pirotta, and Alessandro Lazaric. A fully problem-dependent regret lower bound for finite-horizon mdps. *arXiv preprint arXiv:2106.13013*, 2021.

Jia Wan, Sean R Sinclair, Devavrat Shah, and Martin J Wainwright. Exploiting exogenous structure for sample-efficient reinforcement learning. *arXiv preprint arXiv:2409.14557*, 2026.

Yijun Wan. Ucrl with post-decision state: regret analysis. URL `https://www.normalesup.org/~wan/RLproject_Wan.pdf`.

Ming Yin and Yu-Xiang Wang. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3948–3958. PMLR, 2020.

Jingzhe Zhang, Xiaofan He, and Huaiyu Dai. Blind post-decision state-based reinforcement learning for intelligent iot. *IEEE Internet of Things Journal*, 10(12):10605–10620, 2023.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33: 15198–15207, 2020.

CONTENTS

1 Introduction **1**

2 Problem formulation **3**

2.1 Observed-input structure and unknown sharing 3

2.2 Bellman prediction and performance 3

3 Warm-up: representations and valid input sharing **4**

4 Algorithm design **5**

4.1 Certification and input estimation 5

4.2 Response-first optimistic planning 6

5 Theoretical guarantees and proof mechanisms **7**

5.1 Regret with certification and response-learning costs 7

5.2 Proof mechanisms on the executed trajectory 7

5.3 First-order and second-order costs of unknown sharing 8

6 Conclusion **9**

A Background and extended related work **18**

A.1 Optimistic exploration and conditional-response estimation 18

A.2 Post-decision states and observed exogenous inputs 19

A.3 Factored representations and model selection 19

A.4 Fixed-confidence identification and structural uncertainty 19

A.5 Function approximation and constrained exploration 20

B Model comparisons and motivating examples **20**

B.1 An exact same-feedback reduction 20

B.2 Representations, observations, and prior knowledge 21

B.2.1 Marginal MDPs and deterministic afterstates 21

B.2.2 When does a compact afterstate suffice? 21

B.2.3 General stochastic PDS and the observation requirement 22

B.3 Regret specialization with known components 23

B.3.1 Which earlier guarantees are recovered? 25

B.4 Response-mediated sharing and bilinear transfer 26

B.4.1 How sharing and approximation errors propagate through control 27

B.5 Application examples and measurement conditions 27

B.5.1 Data-center admission: shared inputs and unobserved responses 28

B.5.2 Robotic sorting: response-dependent transfer sensitivity 28

B.5.3 Buffered video delivery: observable capacity versus payload outcomes . . 29
B.5.4 Why the model itself is not an automatic complexity reduction 30

C Proof of Theorem 2 **30**
C.1 Notation, filtrations, and certification interface 30
C.2 Concentration and counting under adaptive sampling 31
C.3 Reward confidence and the terminal-stage contribution 33
C.4 Optimism for the model-based learner 33
C.5 Proofs of the main-text transfer and accounting lemmas 34
C.6 Post-certification regret on the executed trajectory 34
C.7 Certification and the certification-cost gate 35
C.8 Boundary cases and scope of the guarantee 36
C.9 Confidence-intersection optimism and certified optimization 36
C.9.1 Exact confidence geometry 36
C.9.2 Finite linear programs and upper certificates 38
C.9.3 A strict, observation-consistent example 39
C.9.4 Scope and computational tradeoff 39
C.9.5 When an intersection can add value 40
C.9.6 Representing statistical confidence before optimization 40
C.10 Confidence surplus and Bellman covariance 41
C.10.1 Computable residual allowances 41
C.10.2 The exact predictable variance 42
C.10.3 Joint rewards: why the variance terms cannot be separated arbitrarily . . . 43
C.11 Self-normalization and response complexity 44
C.11.1 A uniform vector certificate for data-dependent continuations 44
C.11.2 An overlap correction that independent-stream reasoning misses 45
C.11.3 What covering and eluder dimensions measure here 45

D Streaming TD extension and common-policy comparison **46**
D.1 Complete response-dependent bounds 46
D.2 Streaming conditional-response estimator 46
D.3 Weight identities and optimism 47
D.4 Complete regret bound and its proof 48
D.5 Statistical and storage implications 50

E Proof of Theorem 5 **50**
E.1 Information lower bounds 51
E.2 Correctness of the sequential rule 52
E.3 Expected sample cost at the shared instance 52

E.4 Completion of the fixed-contrast specialization . . . 53
E.5 Known responses and the role of the successor gap . . . 53
E.6 Response-sensitive information geometry . . . 54
E.6.1 Reference family, effective scale, and proof dependencies . . . 54
E.6.2 Product-aware sequential certification . . . 55
E.6.3 Expected time, pointwise optimization, and adaptation . . . 56

F Additional limits of experience sharing **57**
F.1 Factored information on observed histories . . . 58
F.2 Proof of Theorem 39 . . . 59
F.3 Proof of the certification bound in Proposition 40 . . . 59
F.4 Proof of the response lower bound in Proposition 40 . . . 59

G Implementation details and computational requirements **60**
G.1 Online inputs and categorical estimators . . . 60
G.2 Reward observations and their confidence bound . . . 61
G.3 Certification modes and stopping rules . . . 61
G.4 Computation and estimator storage . . . 62

H Experimental protocols and additional results **62**
H.1 Finite-buffer environment . . . 62
H.2 A population pooling counterexample . . . 63
H.3 Matched control comparisons . . . 63
H.3.1 Matched queue runs and instrumentation . . . 63
H.3.2 Margin and state/horizon sensitivity . . . 64
H.4 Policy-selection reliability . . . 65
H.4.1 Predeclared experiment and data . . . 66
H.4.2 Exact sampling at large stopping budgets . . . 66
H.4.3 Results and uncertainty . . . 66
H.4.4 Explicit lower bounds and stopping branches . . . 67
H.4.5 Sensitivity within the binary two-policy family . . . 67
H.4.6 Affordability is not positive net benefit . . . 68
H.4.7 Unknown-reward implementation and a boundary of the sharing benefit . . . 69
H.5 A same-information ablation of confidence intersection . . . 70

NOTATION REFERENCE

Table 2 indexes the main-text symbols. Appendix C.1 specifies filtrations and locally reused letters. Appendix G gives the complete estimator and certification definitions.

Table 2: Notation.

Symbol	Meaning
$\mathcal{S}, \mathcal{A}, \mathcal{Z}, \mathcal{U}$	
S, A, Z, U, C	State, action, nominal-state, and raw-input sets, their sizes, and category count C.
k, h, K, H	Episode/stage indices, episode budget, and horizon.
$s_h, a_h, z_h, u_h, c_h, r_h$	State, action, nominal state, raw input, signature, and observed reward.
ϕ_h, ψ_h	Known nominal-state and response-sufficient signature maps.
$\nu_{h,a}, \mu_{h,a}, P_h, \bar{P}_h$	Raw-input law, induced category law, conditional-response kernel, and marginal transition kernel.
$R_h, \zeta_R, \hat{R}_h^k, \bar{R}_h^k$	Reward mean, its known/unknown indicator $(0/1)$, empirical mean, and upper estimate.
$\mathcal{F}_{k,0}, \mathcal{F}_{k,h}^-$	Episode-start and pre-feedback histories; the latter includes the current action.
$\mathbb{P}, \mathbb{E}, \mathcal{L}(\cdot \mid \mathcal{F})$	Probability, expectation, and conditional distribution.
$V_h^\pi, Q_h^\pi, V_h^*, Q_h^*$	Policy state/action values and their optimal counterparts.
$\mathcal{B}_h, \mathcal{T}_h^\pi, \mathcal{T}_h^*$	Action-value, policy-evaluation, and optimality Bellman operators.
$\omega_{M,h}^\pi, \mathrm{Reg}(K)$	State–action occupancy and cumulative pseudo-regret, including certification.
B_h, H_1, H_2	Continuation range $H-h$, sum $\sum_{h<H} B_h$, and squared sum $\sum_{h<H} B_h^2$.
$\Pi_h^\star, m_h^\star; \Pi_h, m_h$	True equality partition/count and returned partition/count.
$G_h(a), g_h(a), I_h(a)$	Target block, group label, and distinct individual/group streams.
$\epsilon_{h,G}, \epsilon_h$	Certified block diameter and its stagewise maximum.
$\delta, \gamma, n_\gamma, \tau$	Failure probability, nonzero-law separation, certification cycle cap, and certification episode count.
$N_{h,i,c}^k, n_{h,i}^k, \hat{\mu}_{h,i}^k$	Input outcome count, stream total, and category histogram.
$N_{h,z,c,s'}^k, n_{h,z,c}^k, \hat{P}_h^k$	Successor count, actual response-row total, and response histogram.
$N_{h,s,a}^k, T_{h,s,a}^k$	Actual reward-row count and accumulated reward.
$L, L_d, r_d, \lambda_{h,i}^k$	Confidence logarithms, categorical radius, and capped input sampling-plus-mismatch radius.
$\rho_R, \mathcal{G}_R$	Reward-mean radius and cumulative reward-estimation allowance.
$W_h^{\mathrm{MB,k}}, \overline{V}_h^k, \overline{Q}_h^k, \pi^k$	Upper conditional continuation, optimistic state/action values, and frozen greedy episode policy.
$\mathcal{C}_d(J,K), \mathfrak{S}$	Visit bound for J rows with d outcomes, and cumulative input-error budget.
$F_{\mathrm{MB}}, F_{\mathrm{MB}}^A, F_{\mathrm{MB}}^m$	Complete regret envelope and its individual/true-group evaluations (Section 5.1).
$\mathcal{E}, d_h^k, \beta_{h,R}^k, \mathcal{M}_K$	Joint confidence event, executed-policy value error, reward allowance, and trajectory martingale sum.
$P_z, d, \hat{d}, \mathsf{D}_z, \chi(P_z)$	Response matrix, input discrepancy and estimate, induced transition discrepancy, and contraction coefficient.
e_a, e_b, e_c^P, e_μ	Input/response TV errors and $e_\mu = e_a + e_b$ in Proposition 1.
m, Z_0	Maximum stagewise group and nonterminal afterstate counts (Table 1).
p_a, m_c, q_a, θ	Binary input, conditional/marginal successor probabilities, and policy contrast (Section 5.3).
$\varepsilon, \Delta, \kappa, r_*, s_*$	Cancellation scale, common successor gap, response contrast, and bilinear/decision resolutions.
$\hat{d}, \hat{\kappa}, \hat{q}_0, \hat{\theta}, B$	Estimated input/response contrasts, successor mean, policy contrast, and interval half-width in Lemma 6.
$\mathcal{M}_=, \mathcal{M}_\gamma, \mathcal{A}_\delta(\mathcal{M})$	Equality/separation model classes and almost-surely terminating, uniformly δ-correct procedures.
T, n_0	Policy-selection episode cost and initial per-action sample count (Appendix E.6).

A Background and extended related work

This appendix expands the positioning in Section 1. The comparison distinguishes a representation of the dynamics from the observations and structural information supplied to the learner. Appendix B provides the explicit reductions and regret specializations.

A.1 Optimistic exploration and conditional-response estimation

Optimistic exploration provides the statistical foundation, not the new sharing premise. R-MAX uses optimistic initialization; upper-confidence reinforcement learning (UCRL) plans over transition confidence sets; upper-confidence-bound value iteration (UCBVI) uses value-dependent bonuses and variance analysis (Brafman & Tennenholtz, 2002; Jaksch et al., 2010; Azar et al., 2017). Optimistic Q-learning and reference-advantage decomposition establish streaming alternatives (Jin et al., 2018; Zhang et al., 2020). CRFO adds accounting for a selected input partition while response samples remain restricted to actual rows. Its TD extension requires historical-target propagation (Appendix D). Uniform probably approximately correct (uniform-PAC) guarantees are a separate criterion (Dann et al., 2017); neither main theorem automatically establishes them or last-iterate accuracy.

Section 1 uses a common episode notation but not a common information model. In Azar et al. (2017, Theorem 2), write $\ell_{\mathrm{UCB}} = \log(5HSAT_{\mathrm{step}}/\delta)$ and $T_{\mathrm{step}} = HK$. The known-reward, stage-independent-kernel bound is

$$\mathrm{Reg}(K) \le 30H\ell_{\mathrm{UCB}}\sqrt{SAK} + 2500H^2S^2A\ell_{\mathrm{UCB}}^2 + 4H^{3/2}\sqrt{K\ell_{\mathrm{UCB}}}.$$

The three-term expression retains finite-budget terms beyond its asymptotic leading order. When $K \ge H^2S^3A$ and $SA \ge H$, it simplifies to $\widetilde{O}(H\sqrt{SAK})$. Our stage-dependent model and categorical response bonus are different, so the ordinary-MDP embedding does not import this sharper bound.

For the factored EULER algorithm (F-EULER), let d_P count conditionally independent successor components and X_ℓ count assignments to the known state–action parent scope of component ℓ. These are not our numbers of input groups or signature categories. With known rewards, Tian et al. (2020, Corollary 3) give the leading bound

$$\widetilde{O}\left(\sum_{\ell=1}^{d_P}\sqrt{HX_\ell T_{\mathrm{step}}}\right) = \widetilde{O}\left(H\sum_{\ell=1}^{d_P}\sqrt{X_\ell K}\right).$$

Their stated large-sample regime makes additional polynomial-in-structure, logarithmic-in-time terms lower order. This leading term applies in that regime, not after dropping the additional terms for every budget. For unknown factored rewards their result additionally has $\widetilde{O}(\sum_{j=1}^{d_R}\sqrt{Y_jT_{\mathrm{step}}})$, where d_R counts reward factors and Y_j their scope cardinalities. Our transition decomposition is serial, and its unknown sharing partition is not a known conditional-independence scope.

For PDS convergence, $\widetilde{Q}_n$ denotes the corresponding learned afterstate action-value array and $\widetilde{Q}^\star$ its optimal discounted value. The conventional-PDS convergence statement and the blind-PDS statement are, respectively, Lemma 1 and Proposition 2 in Zhang et al. (2023), under their stated visitation and stepsize conditions. They concern convergence as the update index n tends to infinity, not an episodic regret bound. Additional virtual-experience updates need the independence and response structure used in Mastronarde & van der Schaar (2013). This distinction does not imply that every PDS method lacks finite-time regret guarantees; Appendix B also discusses the continuing UCRL-PDS result.

For the observed exogenous Markov decision process (Exo-MDP) result in the cited version of Wan et al. (2026), the response and reward maps are known, and a common input is independent and identically distributed (i.i.d.) across stages. The effective dimension r_{exo} is determined by those maps. The reported $\widetilde{\Theta}(H\sqrt{r_{\mathrm{exo}}K})$ is a minimax order in that information setting. Unknown conditional responses and stage-dependent action laws preclude interpreting our result as an improvement of that rate.

For comparison, our complete bound in Equation (9) retains input, response, reward, and trajectory terms. A returned exact-group count m_h equals the true count $m_h^\star$ only after exact recovery; a singleton fallback has $m_h = A$. The certification cost HAn_γ remains inside Theorem 2's strategy

comparison. Theorem 5 instead concerns pointwise expected episodes under uniform fixed-policy correctness, not general control regret.

A.2 Post-decision states and observed exogenous inputs

PDS and exogenous-input methods exploit different supplied structure. Conventional afterstates separate modeled action effects from unresolved dynamics; virtual experience requires additional invariance and response structure (Sutton & Barto, 2018; Mastronarde & Van Der Schaar, 2011; Mastronarde & van der Schaar, 2013). Blind PDS estimates a missing factor, whereas stochastic-sampling PDS reduces expectation-computation cost (Zhang et al., 2023; Corra et al., 2025). Hindsight and observed-input methods exploit known response maps (Sinclair et al., 2023; Wan et al., 2026); linear exogenous-MDP work studies another structured specialization (Liang et al., 2026). Action dependence and a two-factor decomposition are therefore not novelty claims here. The distinction is discovering input pooling while the conditional consequences remain unknown. An input equality licenses frequency reuse, not an unobserved successor.

A.3 Factored representations and model selection

Factored MDPs exploit known parent scopes (Tian et al., 2020); our factorization instead separates an observed input from a conditional response. Appendix B shows how marginalization, afterstate construction, and state augmentation alter feedback or dimensions. Value equivalence and state-similarity metrics distinguish full-model equality from equality of decision-relevant predictions (Grimm et al., 2020; Ferns et al., 2004). Proposition 1 specializes that distinction to the observed factors and retains their bilinear estimation error; CRFO is not assumed to discover every weaker value-equivalent relation.

Ghosh et al. (2021) study adaptation to smaller transition families under explicit structural assumptions. Our equality certificate instead assumes a lower bound on nonzero discrepancies between signature laws. It does not require separation under every bounded continuation function, which would fail for a constant test. The restart conditions estimation on the selected partition using fresh data, avoiding an uncharged search over all partitions. Its cost is explicit because certification observations are discarded and all actions are investigated. Without separation, the alternative procedure returns a quantitative diameter rather than certifying equality.

Model-based and value-based optimism share important principles (Neu & Pike-Burke, 2020). Optimistic representation selection, including local choices, already has precedent (Papini et al., 2021); taking a minimum of upper estimates is not by itself new. The hybrid retained in Appendix D is justified by simultaneous inequalities on the same executed history. Black-box corralling instead manages the feedback given to distinct learners (Agarwal et al., 2017). Its results do not justify taking the minimum of arbitrary independently trained algorithms' regret bounds. Similarly, a known feedback graph grants actual side observations (Alon et al., 2015); a conjectured equality between two input laws does not grant counterfactual response observations.

A.4 Fixed-confidence identification and structural uncertainty

Change-of-measure arguments characterize the information needed to distinguish decisions under alternative models (Kaufmann et al., 2016; Garivier & Kaufmann, 2016). Réda et al. (2021) analyze unknown misspecification in fixed-confidence identification. Their misspecification radius is an upper bound on deviation from a structural model. In contrast, γ here lower-bounds an admitted nonzero difference between observed input laws, separating the exact-equality component from unequal alternatives.

Theorem 5 characterizes a reference family with varying common-response gap and response contrast, with correctness uniform over the original model class. Its fixed-contrast specialization gives the earlier cancellation–separation law. The new second-order regime requires perturbing both input and response factors and retaining their interaction in the stopping interval. Time-uniform confidence sequences provide more refined boundaries (Howard et al., 2021); the proof uses an explicit repeated-look boundary and an unconditional expected-time tail. Reference-tuned pointwise optimality is distinguished from scale-agnostic adaptation (Jamieson et al., 2014).

Bilinear bandit identification also studies structured targets, but, for example, Rizk et al. (2021) associate bilinear rewards with pairs of node actions on a graph. Here the learner observes the input category and its conditional response separately, and the bilinear term arises in a policy contrast with unknown distribution-sharing restrictions. We therefore do not claim that bilinearity or structured testing is itself new. The contribution is the closed-form first-order/second-order transition for this complete observation law.

Policy evaluation studies information about returns rather than complete kernels (Kallus & Uehara, 2020; Yin & Wang, 2020; Duan et al., 2020). Our known-equality estimator is not claimed to improve on an equally informed equality-constrained estimator; the separation concerns correctness under a larger model class. Likewise, structured exploration analyzes information requirements for general control (Combes et al., 2017; Ok et al., 2018; Tirinzoni et al., 2021). The uniform certification gate does not solve their decision-directed allocation problem.

A.5 Function approximation and constrained exploration

Linear MDPs, low Bellman rank, and value-targeted regression impose different conditions for efficient generalization (Jiang et al., 2017; Jin et al., 2020; Ayoub et al., 2020; Jin et al., 2021). Covering and eluder complexities control statistical uniformity and sequential distinguishability, respectively (Russo & Van Roy, 2013; Abbasi-Yadkori et al., 2011). The present main theorems are finite-space results. They neither require these function-class quantities nor establish an arbitrary neural-network guarantee.

The present model has no safe-exploration constraint. Related linear-MDP work (Shi et al., 2023; Roknilamouki et al., 2025) concerns admissible actions, rather than certificates for pooling observations; its safety guarantees are not assumed here.

B Model comparisons and motivating examples

The reduction below proves the same-feedback comparison in Section 3. It also gives the exact-cancellation construction behind Section 5.3, while Appendix E supplies the matching class-dependent selection theorem.

B.1 An exact same-feedback reduction

To justify the same-feedback comparison in Section 3, first take known deterministic rewards. Augment stage h by a pre-decision phase with state s and an intermediate phase with state (z, c). At the first phase choose a, receive $R_h(s, a)$, and draw (z, c) by (16); at the intermediate phase take the unique dummy action, receive zero reward, and draw s' from $P_h(\cdot \mid z, c)$. The last response can be retained with zero terminal continuation. This is a finite, stage-indexed MDP of $2H$ phases and at most $S + ZC$ phase-tagged states. Its only freely chosen actions are the original actions, so deterministic Markov policies correspond exactly after dummy actions are suppressed. A chronological induction shows equality of the joint laws of $(s_h, a_h, z_h, c_h, s_{h+1})$ and of accumulated rewards.

To retain raw feedback when ψ is many-to-one, draw u according to $\nu_{h,a}$ and attach u to the first-phase observation, with $c = \psi_h(u)$. Both learners observe this mark. No information has been added or suppressed. The online algorithms use its signature, whereas the lower-bound likelihood uses the whole raw observation. A signature-only augmented MDP is therefore sufficient for the algorithmic comparison, but raw-law KL is still the correct information measure for the actual feedback model.

The structured baseline is specified without importing an unrelated theorem. It estimates HA individual signature rows and HZC conditional response rows, with the radii in Sec. 4.1. At an intermediate state its upper value is exactly (6); at the preceding phase its optimistic expectation is (7) with only the individual stream. Thus it is precisely the certified learner of Section 4 with singleton groups and MB responses. The proof of Lemma 16 gives F^A_{MB} directly. If known equality constraints are used, the group stream yields F^m_{MB}. With the same known equality constraints, the proposed MB learner is not being claimed to improve on an optimally implemented equality-constrained baseline. Its additional guarantee concerns discovering those ties with a cost gate. The augmented representation is also compatible with other factored/optimistic designs, whose sharper guarantees require their own bonus and structural analysis (Tian et al., 2020; Papini et al., 2021).

For unknown bounded rewards, the structured baseline uses the same (h, s, a) reward counts and upper estimates as the certified learner of Section 4. The Bellman operator depends on the mean reward R_h, so the identical optimistic-planning comparison and the $\mathcal{G}_R$ charge remain valid. If a reward and its transition are correlated, their original joint feedback is kept as an observed mark; for a stationary joint reward-transition law, a full-feedback Markov representation can retain the predecessor (s, a) at the dummy phase. Under the more general conditional-mean reward premise, only the Bellman comparison and retention of the original marks are asserted. No conditional reward-independence claim is required to justify the reward-confidence proof.

B.2 Representations, observations, and prior knowledge

The alternating-phase kernels in Section 3 are

$$T_{2h-1}(z, c \mid s, a) = \mathbf{1}\{z = \phi_h(s, a)\}\mu_{h,a}(c), \qquad T_{2h}(s' \mid z, c) = P_h(s' \mid z, c). \tag{16}$$

Three different comparisons must be distinguished. A representation preserves policy values. A statistical reduction must also preserve the information available to the learner, without using unknown parameters. A complexity comparison must additionally account for changes in the number of states, signatures, and free probability rows. PDS is an intermediate-state representation of an MDP, not a class of dynamics disjoint from MDPs. The known-versus-unknown sharing classes in Table 3 vary prior knowledge within one fixed interaction model.

B.2.1 Marginal MDPs and deterministic afterstates

Let $\mathcal{M}$ have the factorization (1). Integrating u, c gives $\bar{P}_h$ in Section 2.2. For every fixed Markov policy on the original state, backward induction in Section 2.2 gives the same expected return in the marginal model. An ordinary learner can ignore the extra observations. This direction is always available under the stipulated reward-mean condition, even when reward noise is correlated with the transition.

The reverse direction is not generally an equivalence of learning experiments. Consider two actions, one nominal state, two signatures, and $P_h(1 \mid z, c) = 1/2$ for every c. In one model both signature laws are Bernoulli$(1/4)$, whereas in another they are Bernoulli$(1/4)$ and Bernoulli$(3/4)$. Give both models identical rewards. Their marginal state–action transition and reward laws agree for every action, but their observed signature laws and equality partitions differ. Any parameter-independent transformation of the unmarked history has the same law in the two models, so cannot produce the two different signature histories. The example establishes loss of information under marginalization, not a regret separation on this reward-neutral family.

Conversely, take any finite episodic MDP with kernel $T_h(s' \mid s, a)$ and the bounded reward feedback allowed in Section 2. Set

$$\mathcal{U} = [C] = \{1\}, \qquad \phi_h(s, a) = (s, a), \qquad P_h(s' \mid (s, a), 1) = T_h(s' \mid s, a). \tag{17}$$

The added input is constant and deterministically reconstructible. Hence the original interaction, policy values, and regret are preserved, with $Z = SA$. All input laws are equal, but all transition uncertainty remains in the response. A small number of input groups alone therefore does not imply an easy control problem.

A deterministic PDS kernel $T_h(s' \mid s, a) = K_h^{\mathrm{post}}(s' \mid f_h(s, a))$ has the smaller embedding $\phi_h = f_h$, $C = 1$, and $P_h = K_h^{\mathrm{post}}$. Its afterstate is computable from the observed state and action, so no additional feedback is assumed. If a post-PDS kernel still depends on a, use the afterstate $(f_h(s, a), a)$ and count its reachable labels instead. This augmentation may eliminate a proposed action-sharing saving. The singleton input law in either embedding is known from its alphabet, not estimated from observations. These are representational reductions, not claims of new dynamics beyond finite MDPs.

B.2.2 When does a compact afterstate suffice?

The next proposition formalizes the nominal-state sufficiency condition used in Section 3. Define the set of reachable labels of a deterministic map as its image over all state–action pairs, whether or not a particular learning policy has visited them.

Proposition 7 (Controlled afterstate sufficiency)**.** *At a fixed stage, $z = \phi_h(s,a)$ alone admits a postdecision kernel $K_h^{\mathrm{post}}(\cdot \mid z)$ if and only if, for every pair with $\phi_h(s,a) = \phi_h(t,b) = z$,*

$$\sum\nolimits_c [\mu_{h,a}(c) - \mu_{h,b}(c)] P_h(\cdot \mid z, c) = 0. \tag{18}$$

In particular, if a supplied partition has equal signature laws within each group g, then $y = (\phi_h(s,a), g_h(a))$ is a valid deterministic afterstate with

$$\begin{aligned} K_h^{\Pi}(s' \mid z, g) &= \sum\nolimits_c \mu_{h,g}(c) P_h(s' \mid z, c), \\ Z_{0,h} &= \big|\{(\phi_h(s,a), g_h(a)) : (s,a) \in \mathcal{S} \times \mathcal{A}\}\big| \le \min\{SA, Zm_h\}. \end{aligned} \tag{19}$$

Here $\mu_{h,g}$ denotes the common signature law, which need not be known numerically.

Proof. If $K_h^{\mathrm{post}}(\cdot \mid z)$ exists, both mixtures in Section 2.2 for the indicated pairs equal it. Subtraction gives (18). Conversely, that identity makes the mixture independent of the representative pair chosen from each nonempty fiber of ϕ_h. Define K_h^{post} by that common mixture and arbitrarily on labels outside the image. It is a probability kernel and yields the original transition law. For $y = (z, g)$, any two pairs in the same fiber have the same signature law and conditional-response kernel, so the criterion holds. The number of labels is bounded both by the number SA of pairs mapped to them and by the product Zm_h. □

Equality of input laws is sufficient but not necessary for this collapse. If the response is independent of c, every input-law difference vanishes in (18). Conversely, different action laws can make z alone insufficient. With one nominal state, $\mu_0(1) = 1/4$, $\mu_1(1) = 3/4$, and deterministic response $s' = c$, the two marginal successor laws differ. A single postdecision law at that z cannot represent both actions. Appending their group or action labels repairs the representation at the corresponding dimensional cost.

The collapsed learner estimates K_h^{Π} directly from its observed (y, s') pairs and ignores c for transition estimation. It retains fewer rows when $Z_{0,h}$ is small, but no longer exploits equality of the original $P_h(\cdot \mid z, c)$ across different groups. The factored and collapsed learners therefore exploit different parameter sharing. Neither representation's generic confidence bound dominates for every m, C, S, Z.

B.2.3 General stochastic PDS and the observation requirement

To clarify the feedback requirements in Section 3, consider the stochastic PDS model reviewed by Zhang et al. (2023),

$$T_h(s' \mid s, a) = \sum\nolimits_y K_h^{\mathrm{pre}}(y \mid s, a) K_h^{\mathrm{post}}(s' \mid y, a). \tag{20}$$

The pre-PDS kernel is known a priori in conventional PDS and estimated in blind PDS. The post-PDS kernel can still depend on the action, and the intermediate y is observed in the cited algorithms. Their reward decomposition can include an unknown component. None of these facts follows merely from one common disturbance group in our model.

Verification of the feedback-preserving recoding. Step 1. Construct a common input law. When the finite pre-PDS kernel is known, fix an ordering of M possible y values. At stage h, combine all cumulative-probability endpoints of $K_h^{\mathrm{pre}}(\cdot \mid s, a)$, for all (s, a), with $0, 1$, and sort the distinct endpoints as $0 = e_0 < e_1 < \cdots < e_{C_h} = 1$. Each interval has length $w_{h,c} = e_c - e_{c-1} > 0$, and $C_h \le SA(M-1) + 1$. Every (s, a) has a known map $g_{h,s,a}(c)$ giving the inverse cumulative distribution function (CDF) outcome on that entire interval. Thus

$$K_h^{\mathrm{pre}}(y \mid s, a) = \sum_{c : g_{h,s,a}(c) = y} w_{h,c}.$$

Step 2. Recode the observed intermediate state. Given the actually observed y, the learner can use its own independent randomization to draw

$$\mathbb{P}(c \mid s, a, y) = \frac{w_{h,c} \mathbf{1}\{g_{h,s,a}(c) = y\}}{K_h^{\mathrm{pre}}(y \mid s, a)}. \tag{21}$$

The denominator is positive whenever that y can occur; define an arbitrary distribution on impossible observations. The resulting c has the known common law w_h, independent of the state and action. With $z = (s, a)$ set

$$P_h(s' \mid z, c) = K_h^{\mathrm{post}}(s' \mid g_{h,s,a}(c), a).$$

Step 3. Verify the joint feedback law. Multiplying (21) by $K_h^{\mathrm{pre}}(y \mid s, a)K_h^{\mathrm{post}}(s' \mid y, a)$ gives the required joint law $w_{h,c}P_h(s' \mid z, c)$. The reverse transformation recovers $y = g_{h,s,a}(c)$. This preserves the PDS feedback up to parameter-independent internal randomization, rather than granting an unobserved environmental seed. Variable C_h can be padded by zero-probability categories to a common finite C. Original reward observations are retained. □

This reduction does not preserve the compact PDS parameterization: $Z = SA$ and C can grow with SAM. The displayed learner also does not exploit every equality in the induced response rows. More importantly, the construction uses the known numerical K_h^{pre}. It is not an available transformation for an arbitrary blind-PDS learner whose pre-kernel is unknown. Directly declaring $c = y$ fails the state-independent-law premise whenever $K_h^{\mathrm{pre}}(y \mid s, a)$ varies with s for a fixed a. A blind-PDS process can always be represented through its marginal MDP, but that reduction discards intermediate observations unless additional marks or state augmentation are retained. Extra conditional-independence or observed-input structure can make a compact reduction possible. The distinction is not lack of MDP expressivity, but preservation of feedback, supplied knowledge, and dimension.

An unknown pre-kernel does not preclude a feedback-preserving embedding with a larger state. Let v denote the preceding observed PDS label, with a separate initial symbol $\perp$, and use the augmented state $\hat{s} = (s, v)$. Set $C = 1$, $\phi_h(\hat{s}, a) = (s, a)$, and

$$P_h((s', y) \mid (s, a), 1) = K_h^{\mathrm{pre}}(y \mid s, a)K_h^{\mathrm{post}}(s' \mid y, a). \tag{22}$$

The observed successor now retains precisely the new physical state and PDS observation. The augmentation is computed from observed history without knowing either kernel. It uses at most $S(M + 1)$ states and SA nominal labels. Original Markov policies lift by ignoring v. Conversely, a Markov policy on $\hat{s}$ is implementable in the original process with memory of the preceding PDS observation. Since both the next physical-state law and the mean reward depend only on (s, a), backward induction gives the same unrestricted optimal value, independent of v. Exact correspondence with the original Markov-policy class uses the restricted lifted policies. Thus general blind PDS is representable with its feedback retained, but neither compact dimensions nor its particular algorithm's convergence theorem follow from this embedding. Its two unknown factors are combined rather than separately learned.

B.3 Regret specialization with known components

The reductions above acquire statistical meaning only after the corresponding known quantities are used by the estimator. A one-point probability simplex has zero uncertainty, whereas blindly substituting $C = 1$ into a generic radius leaves an unnecessary positive term. The following corollary formalizes the exact-component convention stated in Section 4. It does not modify Theorem 2's unknown-component algorithm or its proof.

For a positive integer row count J, a logarithmic confidence allowance ℓ, and the original S, A, H, K, define

$$\begin{aligned}
\mathcal{G}_R(\ell) &= 2\zeta_R H[SA + \sqrt{2\ell SAK}], \\
B_{\mathrm{MB}}(J; \ell) &= 2H_1\left[J + \sqrt{2(S\log 2 + \ell)JK}\right] + 4\sqrt{2KH_2\ell} + \mathcal{G}_R(\ell), \\
B_{\mathrm{TD}}(J; \ell) &= e\left[H_1\{J + 12\sqrt{(H+1)\ell JK}\} + 4\sqrt{2KH_2\ell} + \mathcal{G}_R(\ell)\right], \\
B_{\mathrm{HYB}}(J; \ell) &= \min\{B_{\mathrm{MB}}(J; \ell), B_{\mathrm{TD}}(J; \ell)\}.
\end{aligned} \tag{23}$$

Fix $b \in \{\mathrm{MB}, \mathrm{TD}, \mathrm{HYB}\}$ before sampling. These are response and reward costs, with the input-estimation term removed. Here B_b is distinct from the stagewise continuation bound B_h.

Corollary 8 (Exact-component and afterstate specializations)**.** *Under the finite episodic model and reward assumptions of Theorem 2, the following statements hold, each with probability at least* $1 - \delta$ *for the specified procedure.*

Known input laws. *If all $\mu_{h,a}$ are known a priori, use $\mu_{h,a}^\top W_h$ instead of the inner stream minimum in* (7)*, with no input bonus or certification. For the original L and $J = ZC$,*

$$\mathrm{Reg}_b(K) \le \min\{HK, B_b(ZC; L)\}. \tag{24}$$

In particular, this applies to a known singleton signature.

Deterministic afterstates and collapsed exact groups. *Suppose a deterministic afterstate map is known a priori and has at most $Z_0 \le SA$ reachable labels at every stage, with conditional successor law depending only on that label. Use its $C = 1$ representation and the exact singleton input. Put*

$$\ell_{\mathrm{col}} = \log \frac{128H(2A + SA + \zeta_R SA)(K+1)}{\delta}.$$

Then $\mathrm{Reg}_b(K) \le \min\{HK, B_b(Z_0; \ell_{\mathrm{col}})\}$*. If instead an episode-end procedure returns an exact partition with failure probability at most $\delta/4$, discard its τ episodes, form the map in* (19)*, and run this learner with confidence parameter $\delta/2$. For $Z_0 = \max_{h<H} Z_{0,h} \le \min\{SA, Z \max_h m_h\}$,*

$$\mathrm{Reg}_b(K) \le \min\{HK, H\tau + B_b(Z_0; \ell_{\mathrm{col}})\}. \tag{25}$$

Terminal continuation is zero, so terminal afterstate labels need not enlarge Z_0.

Known response. *If P_h is known a priori, use $W_h = P_h \overline{V}_{h+1}^k$ and retain the certified input streams and reward estimator. Then*

$$\mathrm{Reg}(K) \le \min\{HK, H\tau + 2\mathfrak{S} + \sqrt{2KH_2L} + \mathcal{G}_R(L)\}. \tag{26}$$

If μ, P, and the reward means are all known, exact backward planning incurs zero pseudo-regret.

Proof. Step 1. Known input laws. With known μ, its average of either upper response is at least $\mu_{h,a}^\top P_h V_{h+1}^*$. Thus the optimism proof in Appendix D holds without a categorical input event. In the MB one-step inequality (38), the input error is identically zero. The actual response-row count and reward terms remain, and the unconditional trajectory bound uses the same all-history ranges. Their summation is exactly the MB expression in (23). In the TD proof, raw unrolling and chronological optimism still hold; the input term in (67) is zero, while initialization, the historical multiplier, the weighted noise, and the reward term are unchanged. This gives B_{TD}. When both response estimators are evaluated on the same history, their pointwise minimum remains optimistic and is no larger than either candidate. Both derivations therefore apply to that hybrid execution. The original L is conservative because fewer events are needed.

Step 2. Known afterstates. For a known deterministic afterstate, $c = 1$ is known and the actual observed successor at row y has the fixed conditional law $K_h^{\mathrm{post}}(\cdot \mid y)$. At most Z_0 nonterminal rows are counted per stage. Relabeling each stage's image, using that exact input, and applying the preceding argument with $J = Z_0$ proves the known-map assertion. The reward rows remain (h, s, a), so no reward sharing is inferred. The logarithm ℓ_{col} pays for every possible row since $Z_0 \le SA$ and is at least the prescribed core logarithm with confidence $\delta/2$.

Step 3. Data-selected exact groups. For data-selected groups, condition on a completed certification history whose returned groups are exact. Proposition 7 then gives a fixed deterministic map and a fixed successor law at each collapsed row. Fresh future samples obey those conditional laws even though their visits are adaptive. The preceding proof fails with conditional probability at most $\delta/2$, uniformly over every such history. Integrating over valid histories and adding certification failure $\delta/4$ gives total failure at most $3\delta/4 < \delta$. Each certification episode costs at most H, and K upper-bounds the remaining horizon. This proves (25). No claim of exactness is made for groups with a positive diameter certificate.

Step 4. Known responses and exact planning. With known P, the MB response error vanishes exactly. The one-step inequality leaves only reward error, $2B_h \min_i \lambda_{h,i}^k$, and $\bar{P}_h d_{h+1}^k$. Replacing the last conditional expectation by its actual successor value yields one trajectory martingale, with unconditional range lengths $2B_h$. Range-form Azuma bounds its sum by $\sqrt{2KH_2L}$. Counting input and reward radii gives (26); the same certificate and reward event budgets suffice. Finally, when both factors and the mean rewards are known a priori, the computed Bellman values are exact. Selecting their maximizers gives $V_1^{\pi^k} = V_1^*$ in every episode, regardless of realized reward noise. □

Table 1 summarizes these specializations in the main text. Its rows refer to the indicated factorized or collapsed estimator, not a universal ranking of representations. Table 1 follows by substituting $Z_0 = SA$ for an ordinary MDP, $Z_0 = \max_h |\operatorname{im} f_h|$ for a deterministic PDS, or $Z_0 \le \min\{SA, Zm\}$ for known collapsed groups in (23). For the full factored representation, Theorem 2 with known groups has $\tau = 0$ and disturbance term $\sum_h B_h \mathcal{C}_C(m_h, K)$. Known input-law equality removes identification, not estimation of an unknown common distribution. Conversely, the collapsed estimator learns the integrated law directly and therefore does not estimate that distribution separately.

A known partition permits a choice between factored and collapsed estimators. Their response costs scale with ZC and $Z_0 \le Zm$, respectively, while only the former pays the explicit input-estimation cost and shares the same response rows across groups. This is a comparison of proved bounds for specified learners. Neither Theorem 2 nor the numerical results assert that the unmodified factored learner automatically attains the collapsed bound, or the minimum realized regret of separately trained representations.

B.3.1 Which earlier guarantees are recovered?

For $C = 1$ and $Z_0 = SA$, the MB bound has leading order $\widetilde{O}(H^2 S\sqrt{AK})$ and TD has $\widetilde{O}(H^{5/2}\sqrt{SAK})$, with unrelated unknown-reward cost $\widetilde{O}(H\sqrt{SAK})$. The TD order agrees with the conservative Hoeffding-bonus Q-learning order after converting $T = HK$ (Jin et al., 2018). These specializations do not recover sharper Bernstein or reference-advantage guarantees. For example, Zhang et al. (2020) obtain leading $\widetilde{O}(H^{3/2}\sqrt{SAK})$ under their episodic assumptions. The stationary-kernel UCBVI bound in Azar et al. (2017) has a different horizon normalization and additional lower-order terms. A valid representation-level embedding does not transfer those algorithms' sharper bonuses into ours.

PDS also has finite-time regret precedents. The supplied technical note of Wan considers a communicating, average-reward MDP with deterministic afterstate map and reports a high-probability bound of the form

$$25 D_{\mathrm{diam}} \sqrt{S|\mathcal{Y}|T \log(2|\mathcal{Y}|T/\delta)} + 15\sqrt{SAT \log(SAT/\delta)}.$$

This is a reported result of that note, not a theorem newly verified here. It exhibits the same response-row reduction from SA to $|\mathcal{Y}|$, but D_{diam} is a communicating diameter and T counts continuing interactions. Our fixed-start episodic regret and horizon H do not result from replacing D_{diam} by H. Likewise, the cited conventional and blind PDS convergence results use discounted stochastic approximation and their own visitation assumptions (Zhang et al., 2023). The present episodic pseudo-regret theorem neither proves convergence of those iterates nor implies almost-sure convergence of our last iterate. Its sublinear bound instead controls the average of the executed policies' value gaps.

With one supplied sharing group and known P, (26) becomes $2H_1\mathcal{C}_C(1, K) + \sqrt{2KH_2L} + \mathcal{G}_R(L)$. For known reward means, its conservative leading rate is $\widetilde{O}(H^2\sqrt{CK})$. It is not the minimax effective-rank rate $\widetilde{O}(H\sqrt{r_{\mathrm{exo}}K})$ with r_{exo} the effective dimension, for the common time-homogeneous observed-input model with known response and reward maps in Wan et al. (2026). Our bound uses stagewise categorical confidence and does not exploit their rank reduction or common-law estimation across stages. Known reward maps depending on the disturbance can also permit reward sharing that our arbitrary state–action reward estimator does not exploit. These differences prevent a dominance claim.

Virtual experience requires a further distinction. If $s' = g_h(z, c)$ is a known deterministic map, an observed valid c can produce counterfactual successors $g_h(z', c)$ at other nominal states. A known stochastic P_h instead permits exact conditional expectations without claiming observed successor samples. With unknown P_h, equality of input laws alone provides neither operation. Thus the familiar benefit of known afterstate or virtual-experience structure is retained when its knowledge is actually supplied, while certification and response learning are additional costs only in the corresponding unknown-information settings.

Finally, Theorem 5 does not assign one of its three rates to "MDP", another to "PDS", and the third to our model. All three experiments admit an MDP representation. When equality is known a priori, the first-stage transition can also be collapsed into a common afterstate. When alternatives may break equality, assuming that same collapse without evidence can invalidate the policy decision.

The admissible alternatives and observations, not the name of the representation, determine the information lower bound.

B.4 Response-mediated sharing and bilinear transfer

The range-only transfer bound used in Section 3 is

$$|(\mu_{h,a'} - \mu_{h,a})^\top (P_h V)(z,\cdot)| \le B_h \,\mathrm{TV}(\mu_{h,a'}, \mu_{h,a}). \tag{27}$$

This subsection proves Proposition 1. It distinguishes equality of input laws, equality of their induced transition laws, and equality only for a particular continuation. The contraction inequality is the finite-kernel Dobrushin principle (Gaubert & Qu, 2015); the empirical remainder below makes both unknown factors explicit. Neither is presented as a new general concentration inequality.

Proof of Proposition 1. Step 1: the exact transition image. Suppress h, z. The signed measure $d^\top P_z$ has total mass zero. Total-variation duality therefore gives $\sup_{v\in[0,1]^S} |d^\top P_z v| = \frac{1}{2}\|d^\top P_z\|_1 = \mathsf{D}_z(a,b)$. If $t = \mathrm{TV}(\mu_a,\mu_b) > 0$, write $d = t(p^+ - p^-)$ with p^+, p^- probability laws, obtained by normalizing its positive and negative parts. Then

$$\mathrm{TV}((p^+)^\top P_z, (p^-)^\top P_z) \le \sum_{c,c'} p^+_c p^-_{c'} \,\mathrm{TV}(P_z(c,\cdot), P_z(c',\cdot)) \le \chi(P_z).$$

Multiplication by t proves the contraction bound; $t = 0$ is immediate. For $S \ge 2$, map categories with $d_c > 0$ deterministically to one successor and the other categories to a second successor. This kernel attains $\mathsf{D}_z(a,b) = t$. Thus zero bias for *every unrestricted response kernel* is equivalent to input-law equality, although a particular response can annihilate a nonzero discrepancy. No cross-action independence is used.

Step 2: retain the bilinear remainder. For any v, decompose exactly

$$d^\top P_z v - \hat{d}^\top \hat{P}_z v = (d-\hat{d})^\top \hat{P}_z v + \hat{d}^\top (P_z - \hat{P}_z) v + (d-\hat{d})^\top (P_z - \hat{P}_z) v. \tag{28}$$

The vector $d - \hat{d}$ has zero sum and $\|d - \hat{d}\|_1 \le 2e_\mu$, where $e_\mu = e_a + e_b$. Its first term is therefore bounded by $e_\mu \,\mathrm{span}(\hat{P}_z v)$. Each row discrepancy has zero sum and obeys $|(P_z(c,\cdot) - \hat{P}_z(c,\cdot))v| \le e^P_c \,\mathrm{span}(v)$. Applying this once with weights $|\hat{d}_c|$ and once with $|d_c - \hat{d}_c|$ proves (5). The inequalities are deterministic on the stated simultaneous confidence event, so they apply to a continuation selected from the same data. They are not fixed-test concentration after conditioning on that continuation. □

The transfer remainder gives a computable certificate for a proposed continuation: $|d^\top P_z v| \le |\hat{d}^\top \hat{P}_z v|$ plus the right side of (5). For the actual post-certification histograms, take $e_a = r_C(n^k_{h,\{a\}})$, $e_b = r_C(n^k_{h,\{b\}})$, and $e^P_c = r_S(n^k_{h,z,c})$. Lemma 11 provides joint coverage, including zero counts. A value-specific zero transfer does not imply equality of transition laws, and the current CRFO certification rule still targets distributional groups, not a learned value-specific partition.

A second-order witness. Let $C = S = 2$, $v = (0,1)$, $0 < t < 1/2$, and both estimated input laws be $\mathrm{Bern}(1/2)$. Let both estimated response rows be $\mathrm{Bern}(1/2)$. Take true input probabilities $p_a = 1/2 - t$, $p_b = 1/2 + t$ and true response probabilities $m_0 = 1/2 - t$, $m_1 = 1/2 + t$. Then $\hat{d} = 0$ and $\hat{P}_z v$ is constant, so both first-order terms vanish. Nevertheless $d^\top P_z v = 4t^2$, exactly the interaction allowance $2e_\mu \max_c e^P_c$ with $e_\mu = 2t$. Deleting the product term would certify zero transfer incorrectly. This is the multi-category Bellman counterpart of the product obstruction in Theorem 5.

What the response can erase. Write $r_z = \mathrm{rank}(P_z)$. Because $P_z \mathbf{1} = \mathbf{1}$, every element of $\ker(P_z^\top)$ has zero sum. Hence the space of control-null input directions has dimension $C - r_z$, and the image of the input simplex has affine dimension $r_z - 1$. To verify the latter, restrict $P_z^\top$ to the $(C-1)$-dimensional zero-sum tangent space and subtract the nullity $C - r_z$. At an interior input law, sufficiently small perturbations in every such direction are feasible. A rank-one response erases every input discrepancy after marginalization, whereas observed signatures still distinguish their input laws. These are representation dimensions, not a regret bound in an unknown low-dimensional feature space. In particular, learning the response image itself is not free.

B.4.1 How sharing and approximation errors propagate through control

This corollary translates the transfer errors in Proposition 1 into a policy-value bound for a frozen surrogate model. Let M be the true model and $\widetilde{M}$ any frozen surrogate MDP on the same states and actions, with rewards in $[0,1]$. Define

$$\rho_h(s,a) = |R_h(s,a) - \widetilde{R}_h(s,a)|, \quad \eta_h(s,a) = \mathrm{TV}(\bar{P}_h(\cdot \mid s,a), \widetilde{P}_h(\cdot \mid s,a)), \quad c_h^{\mathrm{err}} = \rho_h + B_h \eta_h.$$

Let $\omega^{\pi}_{M,h}(s,a)$ be the stage-h occupancy probability under M from s_1. The following specialization of Bellman perturbation makes the relevant occupancy explicit, rather than adding unrelated error terms.

Corollary 9 (Occupancy-weighted pooling and planning errors)**.** *For every deterministic Markov policy π,*

$$V^{\pi}_{M,1} - V^{\pi}_{\widetilde{M},1} = \sum_{h=1}^{H} \mathbb{E}^{\pi}_{M}\Big[(R_h - \widetilde{R}_h)(s_h,a_h) + (\bar{P}_h - \widetilde{P}_h)V^{\pi}_{\widetilde{M},h+1}(s_h,a_h)\Big],$$
$$|V^{\pi}_{M,1} - V^{\pi}_{\widetilde{M},1}| \le \sum_{h,s,a} \omega^{\pi}_{M,h}(s,a) c_h^{\mathrm{err}}(s,a). \tag{29}$$

If $\widetilde{\pi}$ is ϵ_{opt}-optimal in $\widetilde{M}$ at s_1, then, for a true optimal policy π^*,

$$V^{*}_{M,1} - V^{\widetilde{\pi}}_{M,1} \le \epsilon_{\mathrm{opt}} + \sum_{h,s,a} [\omega^{\pi^*}_{M,h}(s,a) + \omega^{\widetilde{\pi}}_{M,h}(s,a)] c_h^{\mathrm{err}}(s,a). \tag{30}$$

For pure input reassignment $a \mapsto b_h(s,a)$ with the true response, $\eta_h(s,a) = \mathsf{D}_{h,\phi_h(s,a)}(a, b_h(s,a))$. If the reassigned input and response are also estimated, then

$$\eta_h(s,a) \le \min\left\{1, \mathsf{D}_{h,z}(a,b) + \chi(P_{h,z}) e_b + \sum_c \hat{\mu}_{h,b}(c) e_c^P \right\}, \quad z = \phi_h(s,a), \quad b = b_h(s,a). \tag{31}$$

Proof. Subtract the two policy Bellman equations and add $\bar{P}_h V^{\pi}_{\widetilde{M},h+1}$. Backward iteration of the remaining true-kernel difference yields the first identity. Its terminal continuation is zero; total-variation duality and $\mathrm{span}(V^{\pi}_{\widetilde{M},h+1}) \le B_h$ give the occupancy bound. Insert surrogate values for π^* and $\widetilde{\pi}$ between $V^{*}_{M,1}$ and $V^{\widetilde{\pi}}_{M,1}$. Their surrogate gap is at most ϵ_{opt}, and applying the preceding bound to each policy proves (30). For (31), insert $\mu_b^\top P_z$ and $\hat{\mu}_b^\top P_z$ between $\mu_a^\top P_z$ and $\hat{\mu}_b^\top \hat{P}_z$, then use the triangle inequality, response contraction, and convexity of total variation. All statements hold pathwise for a data-selected surrogate on its confidence event; no independence of reward and transition estimation is assumed. □

Equation (30) is a surrogate-policy guarantee, not the online regret theorem. It generally contains an optimal-policy occupancy that may not have been sampled. Response-first optimism is precisely what lets Lemma 4 instead charge actual visits of the executed learner. The solver tolerance ϵ_{opt}, statistical estimation errors, and sharing mismatch have different roles. Confidence-intersection CRFO uses upper certificates rather than a possibly downward-biased approximate maximizer, with its numerical error handled in Appendix C.9. No new low-rank, mixing, or minimum-coverage assumption is needed for these identities.

B.5 Application examples and measurement conditions

These examples make the observation model operational. The data-center queue develops Figure 1; the sorting and video examples isolate additional consequences of response-mediated transfer. The cited application papers motivate the domains, not the exact factorization assumed here. All displayed probability values are constructed examples, not measurements. Binary alphabets are relabeled $\{0,1\}$ for convenience.

B.5.1 DATA-CENTER ADMISSION: SHARED INPUTS AND UNOBSERVED RESPONSES

The queue in Figure 1 isolates workload reuse motivated by resource allocation and cluster scheduling (Mao et al., 2016; 2019). Our example abstracts a finite admission queue, not the full scheduler in either system. Let $B_{\max} = 4$ be its capacity, $b \in \{0, \dots, 4\}$ its backlog, $d \in \{0, 1, 2\}$ a dispatch command, $j \in \{1, 2, 3\}$ an ingress-route index, and $c \in \{0, 1\}$ an offered-work indicator. All commands are defined even when $d > b$, with known nominal backlog $z = (b - d)^+$. The bounded reward mean $R_h(b, (d, j)) = 1 - [b/4 + d/2]/2$ balances queue occupancy against dispatch effort. The category is logged before capacity truncation and has probabilities

$$\mu_{h,(d,1)}(1) = \mu_{h,(d,2)}(1) = 1/4, \qquad \mu_{h,(d,3)}(1) = 3/4.$$

Conditional on (z, c), an unobserved retry indicator ξ is Bernoulli with unknown parameter $\vartheta_h(z, c) \in [0, 1]$, and $b' = \min\{4, z + c + \xi\}$. The symbol ξ is distinct from the observed reward r_h. This defines the stochastic response row exactly:

$$P_h(b' \mid z, c) = (1 - \vartheta_h(z, c))\mathbf{1}\{b' = \min(4, z + c)\} + \vartheta_h(z, c)\mathbf{1}\{b' = \min(4, z + c + 1)\}.$$

One possible true instance has $\vartheta_h(z, c) = 0.05 + 0.10z + 0.05c$. The learner is not given this formula. Modes 1 and 2 permit exact pooling; mode 3 does not. Distinct telemetry tags can be included as $u = (c, j)$, with known $\psi(u) = c$. Raw distributions then differ even for modes 1 and 2, but their response-sufficient signature laws are equal.

This specialization explains three aspects of the learner. The signature removes known irrelevant information, rather than requiring equality of all raw telemetry. A pooled category sample updates a disturbance count, not every response-row count. Finally, equality of the input laws does not resolve the retry probability at an unvisited load: even infinitely many category observations cannot identify that row without actual visits. If categories were observed only after overflow censoring, or their conditional law depended on backlog, the stated model would not apply without an explicitly analyzed extension.

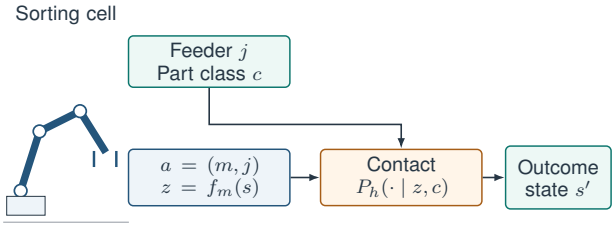
B.5.2 ROBOTIC SORTING: RESPONSE-DEPENDENT TRANSFER SENSITIVITY

This instance of Proposition 1 separates observed object information from robotic contact outcomes (Kalashnikov et al., 2018). The following finite feeder-and-contact abstraction isolates that distinction; it is not a reduction of general vision-based robot control. Use a finite state set of ready, successful, and recovery configurations. Each nominal motion m has a known map f_m on that set; include its contact/configuration attributes in $z = f_m(s)$. A feeder j supplies $c \in \{0, 1\}$ independently of the current configuration conditional on j. Action $a = (m, j)$ applies the map and selects the feeder. A successful state receives known reward one on the next stage, with any specified command cost incorporated in a bounded current reward. Figure 3 gives an explicit unknown response: success probabilities are $(0.9, 0.8)$ at z_1 and $(0.9, 0.1)$ at z_2, ordered by category $c = 0, 1$. Failure transitions to a designated recovery state, and success/recovery rows are absorbing. These are entries of the true unknown response, not a response table supplied to the learner.

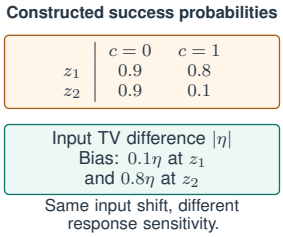
The nominal label z retains motion and configuration attributes, and c is observed before contact. For the success-indicator continuation, the same signed input shift η produces eight times as much bias at z_2 as at z_1. These probabilities define a constructed example, not robot measurements.

With a common class law $(1/2, 1/2)$, the expected success probabilities are 0.85 and 0.50. Observing a contact class at z_1 therefore does not determine success at z_2. More subtly, replace a target class law $(1/2, 1/2)$ by a source law $(1/2 + \eta, 1/2 - \eta)$, where η is an illustrative signed distribution shift with $|\eta| \le 1/2$. For the success-indicator continuation, the induced bias is 0.1η at z_1 but 0.8η at z_2. The TV mismatch is the same, $|\eta|$, yet its decision effect differs by a factor of eight. This is why signature-law equality is a robust sufficient certificate, whereas finer decision-sensitive reuse must also account for response uncertainty. The robust geometry in Proposition 1 explains that distinction.

The example assumes that every command is defined at every state (including absorbing success/recovery states), that the complete influence of motion and configuration on the response is retained in z, and that feeder classes have no additional state dependence. A state-dependent scene sensor, unknown contact classification, continuous motion model, or safety constraint is not automatically covered by this finite abstraction. The numerical experiments in Appendix H remain the separately specified synthetic queue experiments; no robot performance measurement is inferred from this example.

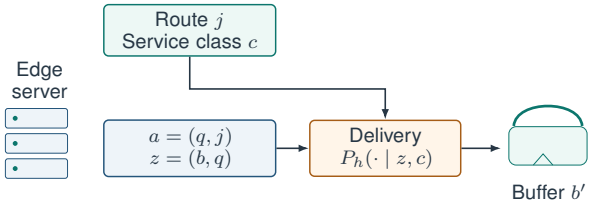


(a) Observed feeder class and unknown contact response.

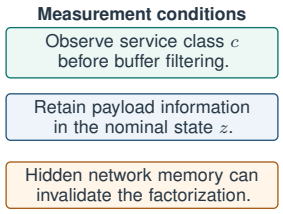


(b) Equal input mismatch, different value bias.

Figure 3: Robotic sorting and response-dependent transfer.



(a) Input telemetry and payload-dependent response.



(b) Measurement conditions.

Figure 4: Buffered video delivery and observable service opportunities.

B.5.3 Buffered video delivery: observable capacity versus payload outcomes

This video example illustrates the measurement conditions of Section 2.1 (Mao et al., 2017). Consider a fixed-slot delivery abstraction with receiver buffer $b \in \{0, \dots, B_{\max}\}$, payload-batch command $q \in \{1, 2\}$, and route j selected from a finite menu. The action is $a = (q, j)$ and the known nominal label is $z = (b, q)$. A monitor records an offered service category $c \in \{1, 2, 3\}$ before receiver-buffer filtering. Its stage-specific law depends on j, but not on b. The unknown on-time batch-success indicator $Y \mid (z, c)$ has probability $\vartheta_h(z, c)$, and the next buffer is $b' = \min\{B_{\max}, (b-1)^+ + qY\}$. Together these rules define a finite response kernel, while a bounded utility can combine delivered quality and playback stalls.

Equal route-category laws allow frequency pooling across payload commands, but do not identify how payload size changes delivery success. Including q in z prevents precisely that erroneous response pooling. Figure 4 distinguishes the capacity observation from the payload-dependent outcome. As in Proposition 1, the effect of a route-law discrepancy depends on the differences between its conditional delivery responses, not just its total variation.

Here q denotes a payload command, not the successor probability q_a in Theorem 5. The headset depicts a client, not head motion, rendering, or perception. Reuse requires the stated state-independence and response-sufficiency conditions.

This abstraction deliberately separates the theorem from unrestricted adaptive streaming. Throughput inferred only from completed downloads can be action-dependent or censored. Persistent link conditions must be included in the state, and closed-loop congestion may violate state-independent input laws even after that augmentation. Likewise, a headset client introduces no guarantee for motion prediction or real-time rendering. Wireless packet scheduling offers an established related PDS/virtual-experience interpretation (Mastronarde & van der Schaar, 2013), but its Markov channel dependence must be retained rather than assumed away. Theorems 2–5 apply only when the specific finite abstraction satisfies their premises.

B.5.4 WHY THE MODEL ITSELF IS NOT AN AUTOMATIC COMPLEXITY REDUCTION

The boundary cases of Section 3 constrain the application interpretations. For $C = 1$ the model becomes ordinary known-afterstate aggregation. Taking the injective map $\phi_h(s, a) = (s, a)$ additionally recovers unrestricted tabular dynamics. Thus a factorization alone does not guarantee a statistical gain. Conversely, supplying the full conditional response removes response-estimation cost but leaves identification of valid input-law sharing to be handled. The contribution is the joint information and error accounting between these limits, not the claim that a new notation enlarges the class of finite MDPs.

C PROOF OF THEOREM 2

The proof follows the order of the algorithm. First, predictable categorical concentration justifies the input and response confidence bounds. Reward confidence and backward induction establish optimism. A response-first Bellman decomposition then yields the post-certification bound. Finally, a pairwise test and the certification-cost gate charge the cost of recovering the groups. All estimates use only post-certification observations.

C.1 NOTATION, FILTRATIONS, AND CERTIFICATION INTERFACE

The notation and observation order below complete Section 2 and Algorithm 1 and its appendix TD extension. Appendix G then gives the corresponding operational parameters.

The divergences are $\mathrm{TV}(p, q) = \frac{1}{2}\sum_x |p_x - q_x|$ and $\mathrm{KL}(p\|q) = \sum_x p_x \log(p_x/q_x)$. The Bernoulli specialization of KL is denoted by kl. The nominal state $z_h = \phi_h(s_h, a_h)$ records the known action effect, while $c_h = \psi_h(u_h)$ is the observed response-sufficient disturbance category. We write $\Delta(\mathcal{X})$ for probability laws on a finite set and $\Delta_d = \{w \in \mathbb{R}^d_+ : \sum_i w_i = 1\}$. The convention $0\log(0/q) = 0$ is used in KL, with positive mass against zero giving $+\infty$. Also $H_1 = H(H-1)/2$ and $H_2 = H(H-1)(2H-1)/6$. All statements use rewards in $[0, 1]$ *per stage*; an episode reward is therefore at most H. The disturbance and response kernels may differ across stages. The initial state is fixed, so (2) is not interchanged with a regret conditional on randomly realized initial states. A random-start extension is not used here.

The main-text symbols are collected in the opening notation reference (Table 2). The conventions below specify their probabilistic and algorithmic meaning.

Work on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ supporting the environment kernels and learner randomization. Finite state/input alphabets and bounded real rewards are standard Borel spaces, so the displayed conditional kernels and expectations are well-defined. The order of observation determines the correct conditioning for each estimator. Let $\mathcal{F}_{k,0}$ be all information before episode k, including the selected partition, all past transitions and rewards, and algorithmic randomization already used. Let $\mathcal{F}^-_{k,h}$ include the within-episode observations through the chosen a^k_h, but not u^k_h. Let $\mathcal{F}^c_{k,h}$ additionally include u^k_h and c^k_h but not s^k_{h+1}, and let $\mathcal{F}^+_{k,h}$ also include s^k_{h+1}. Reveal r^k_h last and let $\mathcal{F}^r_{k,h} = \sigma(\mathcal{F}^+_{k,h}, r^k_h)$ be the completed-step history. The reward mean conditional on $\mathcal{F}^-_{k,h}$ is $R_h(s^k_h, a^k_h)$, even if conditioning on $\mathcal{F}^c_{k,h}$ or $\mathcal{F}^+_{k,h}$ changes it. Reward concentration uses the coarse pre-step-to-completed-step filtration, not a mean-zero claim after the same-step successor is revealed. Then

$$\mathbb{P}(c^k_h = c \mid \mathcal{F}^-_{k,h}) = \mu_{h,a^k_h}(c), \qquad \mathbb{P}(s^k_{h+1} = s' \mid \mathcal{F}^c_{k,h}) = P_h(s' \mid z^k_h, c^k_h).$$

Every $\overline{V}^k_h$ and π^k_h is $\mathcal{F}_{k,0}$-measurable. These distinctions are needed because observing c changes the conditional response law, whereas selecting a does not reveal c.

The implementation is defined on all histories, including zero-count and tied-value cases. An empirical law with zero count is uniform on its declared finite support and is always a probability law. At a positive count it is the normalized histogram. For a response row, the TD visit counter is needed even when the next-state histogram is not stored. Duplicate self/block candidates have the same empirical law. They may reference one stored stream or two separately maintained identical copies, but no counter is incremented twice for the same observation. All argmax ties use the smallest canonical action index. The TD raw array is updated after each actual transition, but the values used

as targets and the policy for the current episode are not recomputed. At $h = H$, all conditional response predictions are zero and (7) returns $\bar{R}_H^k$, which equals the true immediate reward mean only when it is known a priori. Every source observation enters one individual stream and one block stream; this is reuse of a single observation, not two independent samples. Concentration is applied to each stream separately and union bounded, without assuming independence between streams.

For the general certificate result, $\tau \in \{0, \dots, K\}$ is a stopping time for completed episodes. The returned partitions and certificates are measurable at τ. Each certification episode, like each learning episode, uses a policy fixed before that episode. Every history returns genuine partitions and certificates in $[0, 1]$, with singleton radius zero. Replacing a proposed radius x by $\min\{1, \max\{0, x\}\}$ preserves a valid certificate because total variation lies in $[0, 1]$. These conventions ensure bounded predictions even on certificate-failure histories.

A test cycle uses actual actions rather than additional data access. A cycle consists of A episodes. In episode k of the test and stage h, use action $1 + ((k + h - 2) \bmod A)$ in one-based indexing. Hence each stage–action pair receives exactly one signature observation per cycle. Pair tests and the consistency check are performed only at the end of a cycle. If the cap is reached without a consistent complete classification, return singleton blocks. A test interrupted by the learning horizon has incurred at most HK regret. The gated construction never intentionally begins a test whose cap is as long as the horizon.

In Appendix C.11, D is a vector-feature dimension, Λ_t its regularized Gram matrix, $S_t(f)$ a vector martingale indexed by a test function, N_α a sup-norm covering number, and $\dim_E$ an eluder dimension. The scalar regularization λ there is distinct from the indexed pooling radius $\lambda_{h,i}^k$. In the product-aware rule, n_j and r_j are its cumulative per-action count and simultaneous mean radius, as defined in Appendix E.6.

The context of locally reused symbols is fixed by their indices and domain. In the queue, d is dispatch capacity. In categorical confidence, d is an alphabet size. In Proposition 1, d is a signed input-law vector, whereas d_h^k is an executed-policy value error. The scalar m_c in Section 5.3 is a conditional Bernoulli mean, not a group count m_h. The simplex $\Delta(\mathcal{X})$ is distinct from the scalar successor gap Δ, and $B_h = H - h$ is distinct from the product interval's half-width B. Hats always denote estimates; a bar on V or Q denotes a planned upper value, while $\bar{P}_h$ denotes the true marginal kernel.

C.2 Concentration and counting under adaptive sampling

The confidence events below justify the radii in Section 4.1 and supply the statistical event used in the proof of Theorem 2. The filtration is that of Appendix C.1.

Write $J_P = ZC$ for the number of conditional-response rows per stage. The learner chooses where to sample using past observations. The proof must therefore control visit-indexed streams without conditioning on the final sample counts. A single categorical argument covers individual laws, pooled laws, and response rows.

Lemma 10 (Adaptive categorical-stream confidence)**.** *Let $X_i \in [d]$ be a categorical stream with predictable conditional law p_i, not necessarily constant, let e_x be the xth coordinate vector, and let $\hat{p}_n = n^{-1} \sum_{i \le n} e_{X_i}$ and $\bar{p}_n = n^{-1} \sum_{i \le n} p_i$. For a fixed $n \ge 1$,*

$$\mathbb{P}\{\mathrm{TV}(\hat{p}_n, \bar{p}_n) > \sqrt{(d \log 2 + x)/(2n)}\} \le e^{-x}.$$

The same bound applies to visit-indexed streams selected before the categorical observation, and a union bound over $n \le K$ is valid.

Proof. For any subset $D \subset [d]$, $\mathbf{1}\{X_i \in D\} - p_i(D)$ has conditional mean zero and conditional range of length one. Hoeffding's conditional moment bound gives

$$\mathbb{E}[\exp\{t(\mathbf{1}\{X_i \in D\} - p_i(D))\} \mid \text{past}] \le e^{t^2/8}.$$

Iteration and Chernoff optimization give probability at most e^{-2nr^2} that its average exceeds r. Since two probability vectors have equal mass, $\mathrm{TV}(\hat{p}_n, \bar{p}_n) = \max_{D \subset [d]} (\hat{p}_n(D) - \bar{p}_n(D))$. Union bounding over at most 2^d subsets proves the display. For a predictable visit selection, apply the same exponential-supermartingale argument along the successive selected observations. Unfinished

streams can be extended with valid independent observations on an enlarged probability space; this is a proof device, not extra data available to the algorithm. Equivalently, optional stopping of the nonnegative exponential supermartingale at the nth visit gives the bound on the event that this visit occurs. Union bounding over the finitely many possible prefix lengths handles adaptive counts; conditioning on a realized count is neither needed nor asserted. □

Lemma 11 (Disturbance and response events)**.** *With L, L_d in Sec.4.1, there is a sampling event $\mathcal{E}_{\mathrm{samp}}$ of probability at least $1-\delta/8$. On its intersection with the valid-certificate event, simultaneously at every episode and row,*

$$\mathrm{TV}(\hat{\mu}^k_{h,i}, \mu_{h,a}) \le \lambda^k_{h,i} \quad (i \in I_h(a)),$$
$$\mathrm{TV}(\hat{P}^k_h(\cdot \mid z,c), P_h(\cdot \mid z,c)) \le r_S(n^k_{h,z,c}), \tag{32}$$

The event $\mathcal{E}_{\mathrm{samp}}$ controls each pooled stream relative to its average predictable source law even if its proposed certificate is false. A test using the individual signature radii is itself simultaneously correct as a confidence procedure with failure probability at most $\delta/8$.

Proof. Condition on the complete test history, but not on future data. Each individual stream has fixed source law. In a pooled stream, the ith source action is predictable at observation time and belongs to that fixed block. Lemma 10 controls deviation from its average source law, and (3) gives

$$\mathrm{TV}(\bar{\mu}_{G,n}, \mu_{h,a}) \le \frac{1}{n}\sum_{i\le n} \mathrm{TV}(\mu_{h,a_i}, \mu_{h,a}) \le \epsilon_{h,G}$$

for every target $a \in G$. The triangle inequality and clipping at one yield the first line. There are at most $2HA$ streams and K nonempty prefixes per stream. Each actual response row has conditional next-state law $P_h(\cdot \mid z,c)$, so the second line follows from the same lemma over at most HJ_PK prefixes. For $n=0$, the total-variation bound one is deterministic. The combined failure probability is at most $H(2A+J_P)Ke^{-L} < \delta/64$, which is smaller than the stated budget. This bound is uniform in the conditioned test history; integrating removes the conditioning without any union over partitions or stopping times. Test streams add at most $HAKe^{-L} < \delta/64$, also within their stated budget. Arbitrary certificates from another procedure instead use their separately assumed $\delta/4$ failure budget. □

Lemma 12 (Counting actual denominators)**.** *Suppose K' visits are distributed over J cells, and a cell's counter is incremented exactly when visited. Then, for $K' \le K$,*

$$\sum_{\text{visits}} r_d(n_{\text{before}}) \le J + \sqrt{2L_dJK} = \mathcal{C}_d(J,K). \tag{33}$$

Consequently,

$$\sum_{k>\tau, h<H} B_h \min_{i\in I_h(a^k_h)} \lambda^k_{h,i} \le \mathfrak{S}. \tag{34}$$

Proof. A cell with N visits contributes at most $1+\sqrt{L_d/2}\sum_{i=1}^{N-1} i^{-1/2} \le 1+\sqrt{2L_dN}$. Summing and using $\sum_j \sqrt{N_j} \le \sqrt{JK'}$ proves (33). For the self branch, every executed action increments its own counter, giving $H_1\mathcal{C}_C(A,K)$. For the block branch, every executed action increments the counter of its one block, giving $\sum_h B_h\mathcal{C}_C(m_h,K)$. Its bias terms sum to at most $K\sum_h B_h\epsilon_h$. The sum of pointwise minima is at most each of these two complete sums, proving (34). No count is incremented for an unobserved counterfactual response. □

Lemma 13 (Range-form martingale concentration)**.** *For a finite martingale-difference sequence X_t with deterministic conditional range lengths c_t, and $\eta \in (0,1)$,*

$$\mathbb{P}\left\{\sum_t X_t > \sqrt{\tfrac{1}{2}\sum_t c_t^2 \log(1/\eta)}\right\} \le \eta. \tag{35}$$

Proof. Conditional Hoeffding gives $\mathbb{E}[e^{uX_t} \mid \mathcal{F}_{t-1}] \le e^{u^2c_t^2/8}$. Iterating and applying Markov's inequality gives $\mathbb{P}(\sum_t X_t > x) \le \exp(-ux + u^2\sum_t c_t^2/8)$. Optimize over $u>0$. If all c_t vanish, every increment is zero and the result is immediate. The ranges must hold on the entire probability space, not only on a future optimism event. □

C.3 Reward confidence and the terminal-stage contribution

This step supplies the unknown-reward charge $\mathcal{G}_R$ in Theorem 2, including the terminal stage where transition uncertainty vanishes. Write $d_h^k = \overline{V}_h^k - V_h^{\pi^k}$ for the optimistic value error relative to the executed policy, and $D_h = \sum_{k>\tau} d_h^k(s_h^k)$ for its accumulated value at stage h.

Lemma 14 (Actual-row reward confidence)**.** *There is an event $\mathcal{E}_R$, with failure probability at most $2\zeta_R HSAKe^{-L} < \delta/32$, on which every upper reward estimate satisfies*

$$0 \le \bar{R}_h^k(s,a) - R_h(s,a) \le \zeta_R \min\{1, 2\rho_R(N_{h,s,a}^k)\}.$$

If U_h^R sums this upper bound over the actual post-certification visits at stage h, then $\sum_{h=1}^H U_h^R \le \mathcal{G}_R$. In particular, the terminal value error is at most U_H^R.

Proof. Step 1. Predictable reward sampling. The original known-reward specialization has $\zeta_R = 0$. For $\zeta_R = 1$, define $\rho_R(0) = 1$ and $\rho_R(n) = \min\{1, \sqrt{L/(2n)}\}$ for $n > 0$, as in the certified learner of Section 4. The statistic $\hat{R}_h^k(s,a)$ uses only rewards collected at that actual row after testing. Given the full test history, visit selection is predictable and each reward has conditional mean $R_h(s,a)$. For each actual row visit, conditional Hoeffding gives $\mathbb{E}[\exp\{t(r_h^k - R_h(s_h^k, a_h^k))\} \mid \mathcal{F}_{k,h}^-] \le \exp(t^2/8)$. Row selection is measurable before this joint observation. The visit-indexed exponential argument of Lemma 10 therefore applies to the reward even if it is correlated with the signature and successor. Define $\mathcal{E}_R$ as simultaneous coverage of all reward means by their radii. A union bound over at most $HSAK$ nonempty prefixes gives

$$\mathbb{P}(\mathcal{E}_R^c) \le 2HSAKe^{-L} < \delta/32.$$

Step 2. Simultaneous upper estimates. No claim that $r_h^k - R_h$ has zero mean conditional on the same-step successor is used. At count zero the radius one covers the mean deterministically. For $\zeta_R = 0$ set $\mathcal{E}_R$ to the whole space. On $\mathcal{E}_R$, the clipped upper estimate satisfies

$$0 \le \bar{R}_h^k(s,a) - R_h(s,a) \le \zeta_R \min\{1, 2\rho_R(N_{h,s,a}^k)\}.$$

Step 3. Actual-row summation and the terminal stage. For known reward means the difference is identically zero. Let U_h^R sum the right side over actual post-test visits at stage h. Counting SA possible reward rows per stage gives

$$\sum_{h=1}^H U_h^R \le 2\zeta_R H[SA + \sqrt{2LSAK}] = \mathcal{G}_R. \tag{36}$$

In particular the terminal error obeys $D_H \le U_H^R$, rather than $D_H = 0$ when reward means are unknown. The reward charge applies also at the terminal stage. The propagation of these terms is treated in the model-based proof below and the TD proof in Appendix D. □

C.4 Optimism for the model-based learner

The following induction justifies the optimistic planning step in Algorithm 1 before regret is accumulated.

Lemma 15 (Response-first optimism)**.** *On the valid-certificate event, $\mathcal{E}_{\mathrm{samp}}$, and $\mathcal{E}_R$, the model-based learner satisfies $\overline{Q}_h^k \ge Q_h^*$ and $\overline{V}_h^k \ge V_h^*$ for every post-certification episode and state. Its response prediction satisfies $W_h^{\mathrm{MB,k}} \ge P_h \overline{V}_{h+1}^k$.*

Proof. Step 1. Conditional response. For $0 \le V \le B_h$, total-variation duality and Lemma 11 imply $|(\hat{P}_h^k - P_h)V| \le B_h r_S(n_{h,z,c}^k)$ simultaneously for every V, including the data-dependent $\overline{V}_{h+1}^k$. Since $P_h \overline{V}_{h+1}^k \le B_h$, clipping the upper estimate at B_h preserves this inequality.

Step 2. Input average and backward induction. Initialize $\overline{V}_{H+1}^k = V_{H+1}^* = 0$. If $\overline{V}_{h+1}^k \ge V_{h+1}^*$, then, for each candidate stream i,

$$\hat{\mu}_{h,i}^{k\top} W_h^{\mathrm{MB,k}} + B_h \lambda_{h,i}^k \ge \mu_{h,a}^\top W_h^{\mathrm{MB,k}} \ge \mu_{h,a}^\top P_h V_{h+1}^*.$$

The first inequality applies to a bounded vector and does not require independence between its entries and the empirical law. Adding $\bar{R}_h^k \geq R_h$, minimizing over valid streams, and clipping at $B_h + 1 \geq Q_h^*$ preserve the lower bound on Q_h^*. Maximizing over actions completes the induction. At $h = H$, the continuation is zero and reward confidence supplies the base case even when rewards are unknown. □

C.5 Proofs of the main-text transfer and accounting lemmas

This subsection proves Lemmas 3–4, connecting valid input reuse to the actual response rows reached by CRFO.

Proof of Lemma 3. On the certificate and sampling event, $\mathrm{TV}(\hat{\mu}_{h,i}^k, \mu_{h,a}) \leq \lambda_{h,i}^k$. Hence $|(\hat{\mu}_{h,i}^k - \mu_{h,a})^\top W_h^{\mathrm{MB,k}}| \leq B_h \lambda_{h,i}^k$ for the actual, possibly data-dependent bounded response vector. Response confidence gives $0 \leq W_h^{\mathrm{MB,k}} - P_h \overline{V}_{h+1}^k \leq 2B_h r_S(n_{h,z,\cdot}^k)$. Adding and subtracting $\mu_{h,a}^\top W_h^{\mathrm{MB,k}}$ proves both sides of (11). Lemma 15 proves the remaining backward-optimism claim, including the terminal stage. These are simultaneous statements, not concentration for a random test function conditional on its training data. □

Proof of Lemma 4. For $k > \tau$ and $h < H$, Lemma 3, the greedy choice, and the Bellman equation imply

$$d_h^k(s_h^k) \leq \beta_{h,R}^k + 2B_h \min_{i \in I_h(a_h^k)} \lambda_{h,i}^k + \mathbb{E}[Y_{k,h} \mid \mathcal{F}_{k,h}^-].$$

Subtract and add $Y_{k,h}$, sum over $h < H$ and then over post-certification episodes. The $d_{h+1}^k(s_{h+1}^k)$ terms telescope. The terminal inequality $d_H^k(s_H^k) \leq \beta_{H,R}^k$ follows from reward confidence and the greedy terminal action. On optimism, $V_1^* - V_1^{\pi^k} \leq d_1^k$, yielding (12).

For the tail argument, the values and counters at episode start are frozen and measurable before each current observation. The conditional mean of $Y_{k,h}$ is the sum of the response-radius average under μ_{h,a_h^k} and $(\bar{P}_h d_{h+1}^k)(s_h^k, a_h^k)$. The model gives these conditional expectations irrespective of whether $\mathcal{E}$ occurs. Regard $Y_{k,h}$ as revealed with the completed-step history, which also contains the reward; no independence of the reward and successor is needed. Since $r_S \in [0,1]$ and $d_{h+1}^k \in [-B_h, B_h]$ on all histories, $Y_{k,h}$ has conditional range length at most $4B_h$. Set $Y_{k,h} = 0$ throughout certification. This is predictable at an episode start and defines the martingale on all histories without evaluating a post-certification estimate before it exists. Lemma 13 with $\sum_{k,h<H}(4B_h)^2 = 16KH_2$ gives the stated bound $\sqrt{8KH_2L} = 2\sqrt{2KH_2L}$ at failure probability e^{-L}. This bound was not conditioned on certificate validity or future optimism. If $\tau = K$, the sum is zero and both conclusions are immediate. □

C.6 Post-certification regret on the executed trajectory

The next argument sums the local bounds of Lemma 4 and establishes the envelope $F_{\mathrm{MB}}(\mathfrak{S})$ used in Theorem 2.

Lemma 16 (Complete model-based bound)**.** *Under the certificate interface in Appendix C.1, with certificate failure at most $\delta/4$, the model-based learner has, with probability at least $1 - \delta$,*

$$\mathrm{Reg}_{\mathrm{MB}}(K) \leq \min\{HK, H\tau + F_{\mathrm{MB}}(\mathfrak{S})\}. \tag{37}$$

The same pathwise model-based argument applies to an optimistic hybrid whose action-value candidate is no larger than the model-based candidate evaluated on that hybrid's own frozen values.

Proof. Step 1. Bellman recursion. If $\tau = K$, the deterministic cap suffices. Otherwise condition the sampling calculation on the completed certification history, but define trajectory increments on every history using the predictable indicator $\mathbf{1}\{\tau < k\}$. Lemmas 3–4, proved completely above, give the executed-policy recursion

$$d_h^k(s_h^k) \leq (\bar{R}_h^k - R_h)(s_h^k, a_h^k) + 2B_h \min_i \lambda_{h,i}^k + 2B_h \sum_c \mu_{h,a_h^k}(c) r_S(n_{h,z_h^k,c}^k) + (\bar{P}_h d_{h+1}^k)(s_h^k, a_h^k). \tag{38}$$

Step 2. Concentration and visit counting. Their complete-step martingale tail does not condition on certificate validity or future optimism. Count the actual input and response rows by Lemma 12, and the terminal as well as nonterminal reward rows by Lemma 14. The telescoped bound is

$$\mathrm{Reg}_{\mathrm{post}} \le 2\mathfrak{S} + 2H_1\mathcal{C}_S(J_P, K) + 2\sqrt{2KH_2L} + \mathcal{G}_R \le F_{\mathrm{MB}}(\mathfrak{S}).$$

No nonterminal response is imputed to an unexecuted row. If the response is known exactly, its error vanishes and the only trajectory increment is $\bar{P}_h d^k_{h+1} - d^k_{h+1}(s^k_{h+1})$, of all-history range $2B_h$. The corresponding bound is $2\mathfrak{S} + \sqrt{2KH_2L} + \mathcal{G}_R$.

Step 3. Joint probability and certification cost. The certificate, sampling, reward, and trajectory failures have total probability at most

$$\frac{\delta}{4} + H(2A + ZC)Ke^{-L} + 2\zeta_R HSAKe^{-L} + 2e^{-L} < \delta.$$

The last budget even covers two trajectory tails, though the joint-increment proof needs only one. The sampling bound conditional on the certification history is uniform whether or not its output is valid; diameter implications are applied only on the intersection of events. Add at most $H\tau$ for certification and intersect with HK. For an optimistic hybrid no larger than the MB candidate on its own frozen continuation, the same upper recursion and counts hold on that hybrid's history, so the conclusion is unchanged. □

C.7 Certification and the certification-cost gate

This step completes Theorem 2 by bounding group-identification cost and treating both branches of the gate in Section 4.1.

Assumption 17 (Equality or known separation)**.** A known $\gamma \in (0, 1]$ satisfies, for every $h < H$ and pair a, b, either $\mu_{h,a} = \mu_{h,b}$ or $\mathrm{TV}(\mu_{h,a}, \mu_{h,b}) \ge \gamma$.

Lemma 18 (Exact recovery and approximate diameter certificates)**.** *For the cycle-by-cycle rule in Section 4.1, under Assumption 17, every pair is correctly classified by*

$$n_\gamma = \lceil 32L_C/\gamma^2 \rceil \quad \textit{cycles}, \qquad \tau \le An_\gamma, \tag{39}$$

on the simultaneous input-confidence event. Without separation, a fixed $n_a \ge 1$ cycles and the complete-link rule in Appendix G give valid certificates

$$\epsilon_{h,G} = \min\{1, \max_{a,b\in G} \hat{d}_h(a, b) + 2r_C(n_a)\}, \tag{40}$$

with singleton radius zero. Both validity statements have failure probability smaller than $\delta/8$, provided the required samples fit within K episodes.

Proof. Step 1. Uniform pair confidence. The triangle inequality gives, simultaneously over the cycle prefixes,

$$|\hat{d}_h(a, b) - \mathrm{TV}(\mu_{h,a}, \mu_{h,b})| \le 2r_C(n). \tag{41}$$

Thus $\hat{d} + 2r < \gamma$ forces equality under Assumption 17, and $\hat{d} - 2r > 0$ forces inequality. The tests cannot conflict on this event.

Step 2. Termination of exact certification. At $n = n_\gamma$, $r_C(n) \le \gamma/8$. For an equal pair, $\hat{d} + 2r \le 4r \le \gamma/2 < \gamma$. For an unequal pair, $\hat{d} - 2r \ge \gamma - 4r \ge \gamma/2 > 0$. Therefore every pair is resolved and the equality graph is exactly the union of the true class cliques. The all-history clique check and singleton fallback define valid partitions even off this event. A gate-authorized run has $An_\gamma < K$, so the confidence prefixes used in the proof are available within the horizon.

Step 3. Approximate certification. At a fixed n_a, apply (41) to every pair in each data-selected block. Taking the maximum and clipping at one proves (40) for all blocks simultaneously. No union bound over candidate partitions is needed because the pair event already holds for all actions. The pre-certification part of Lemma 11 bounds failure by $HAKe^{-L} < \delta/64$. □

Corollary 19 (Gated model-based comparison)**.** *The exact-certification gate of Section 4.1 and the model-based learner satisfy the bound in Theorem 2.*

Proof of Theorem 2 and Corollary 19. The general certificate statement follows from Lemma 16. For the exact gate, write $C_{\mathrm{cert}} = HAn_\gamma$, $B_A = F^A_{\mathrm{MB}}$, and $B_m = F^m_{\mathrm{MB}}$. The first two are known before sampling; the unknown B_m is used only in the comparison.

Case 1. Certification exceeds the individual bound. If $C_{\mathrm{cert}} > B_A$, the learner starts with singletons. Its regret is at most B_A, and $C_{\mathrm{cert}} + B_m > B_A$, so the desired comparison holds without a factor of two.

Case 2. Certification does not fit within the horizon. If $An_\gamma \geq K$, then $C_{\mathrm{cert}} \geq HK$. The individual bound and the deterministic cap both hold, and $HK \leq C_{\mathrm{cert}} + B_m$. This again proves the required comparison.

Case 3. The gate permits certification. Now $C_{\mathrm{cert}} \leq B_A$. On the exact-recovery event, the returned counts equal the true equality-class counts. The individual stream bounds subsequent regret by B_A, while the group stream bounds it by B_m. Both are valid for this one execution, giving

$$\mathrm{Reg}_{\mathrm{MB}}(K) \leq \min\{2B_A, C_{\mathrm{cert}} + B_m\} \leq 2\min\{B_A, C_{\mathrm{cert}} + B_m\}. \tag{42}$$

Intersect with HK. For this case use the pre-certification event from Lemma 18 as the certificate event in Lemma 16; its probability budget covers exact recovery as well as certificate validity. No second uncharged failure event is introduced. The leading rate follows from the displayed definitions of $L_C, L_S, \mathcal{C}_d, H_1, H_2$, and $\mathcal{G}_R$. □

For returned exact groups with $m_h \leq m$, the leading model-based consequence stated in Section 1 is

$$H\tau + \widetilde{O}\Big(H^2\sqrt{mCK} + H^2\sqrt{SZCK} + \zeta_R H\sqrt{SAK}\Big). \tag{43}$$

Here $\widetilde{O}$ suppresses logarithms and the initialization terms retained in (9).

A supplied lower bound on nonzero discrepancies is essential for the equality implication. Failure to reject a difference is not itself evidence of equality. Approximate certification instead retains the resulting $K\epsilon_h$ contribution. On the shared family of Theorem 39, a cycle costs $(A-1)\Delta$ actual regret and the choice $\gamma = 2\Delta$ yields $O(AL_C/\Delta + A\Delta)$ certification regret. This is a family-specific interpretation, not a matching general-horizon minimax result.

C.8 Boundary cases and scope of the guarantee

These boundary cases specify when the preceding proof reduces to individual learning, known-group learning, or a horizon-limited guarantee. The preceding proof allows zero probabilities, deterministic responses, unvisited rows, and S, A, C, or Z equal to one. Zero-count confidence is deterministic. If $A = 1$, pair classification is vacuous. Small K or very small positive γ are handled by the skip rule. Known groups use $\tau = 0$; approximate groups retain their diameter charge. A false supplied separation, an insufficient signature, or additional state dependence of the input law is outside the stated model, not an event certified by this learner.

C.9 Confidence-intersection optimism and certified optimization

This subsection defines and proves the confidence-intersection refinement introduced in Section 4.2. It changes the optimistic planning operation, not the observation model, certification procedure, post-certification restart, or empirical counters. The original scalar algorithm and all reported historical experiments remain as specified earlier. Optimization over transition confidence sets is an established form of model-based optimism (Jaksch et al., 2010; Neu & Pike-Burke, 2020). The question here is how to combine overlapping individual and pooled confidence constraints while preserving executed-target accounting.

C.9.1 Exact confidence geometry

The scalar backup optimizes each input allowance separately. To require simultaneous target-law feasibility, define

$$\mathcal{C}^{\mu,k}_{h,a} = \bigcap\nolimits_{i \in I_h(a)} \{q \in \Delta([C]) : \mathrm{TV}(q, \hat{\mu}^k_{h,i}) \leq \lambda^k_{h,i}\}. \tag{44}$$

With response set $\mathcal{C}^{P,k}_{h,z,c} = \{p \in \Delta(\mathcal{S}) : \mathrm{TV}(p, \hat{P}^k_h(\cdot \mid z, c)) \le r_S(n^k_{h,z,c})\}$, compute the conditional support and then the input-mixture support,

$$w^k_h(z,c) = \max_{p \in \mathcal{C}^{P,k}_{h,z,c}} p^\top \overline{V}^k_{h+1}, \qquad U^k_h(s,a) = \max_{q \in \mathcal{C}^{\mu,k}_{h,a}} q^\top w^k_h(\phi_h(s,a), \cdot). \tag{45}$$

Use $\overline{Q}^k_h(s,a) = \min\{B_h + 1, \bar{R}^k_h(s,a) + U^k_h(s,a)\}$. The construction shares information targetwise, not through one globally tied optimistic MDP. The following proof handles empty intersections and numerically certified upper values as well as exact optimization.

Fix an episode-start history and a stage $h < H$. Suppress (h,k) when no ambiguity results. For a target a at nominal state z, let $\mathcal{C}^{\mu,k}_{h,a}$ be the input intersection in (44) and let $\mathcal{C}^{P,k}_{h,z,c} = \{p \in \Delta(\mathcal{S}) : \mathrm{TV}(p, \hat{P}^k_h(\cdot \mid z, c)) \le r_S(n^k_{h,z,c})\}$ be the response ball. The response sets are nonempty compact polytopes because they contain their empirical histograms. The input intersection can be empty on a confidence-failure history. In that case the algorithm uses the original scalar backup (7), with (6), at that target. The fallback is specified independently of the good event. Known component laws are represented by singleton sets, and terminal continuations are zero.

The two-stream input intersection is defined in (44). Its radii already contain any certified within-group mismatch, so its simultaneous coverage uses the original statistical event.

Proposition 20 (Confidence-intersection refinement)**.** *Run CRFO with its original certification rule and counters, replacing scalar planning by* (45) *whenever the input intersection is nonempty. On the original joint confidence event $\mathcal{E}$, write $w = w^k_h(z,\cdot)$, $v = \overline{V}^k_{h+1}$, and*

$$\ell^k_{h,a} = \min\{1, 2 \min_{i \in I_h(a)} \lambda^k_{h,i}\}.$$

Then the intersection contains $\mu_{h,a}$, and

$$0 \le w(c) - (P_h v)(z,c) \le 2B_h r_S(n^k_{h,z,c}), \tag{46}$$

$$0 \le U^k_h(s,a) - \mu^\top_{h,a} P_h v \le \mathrm{span}(w)\ell^k_{h,a} + \mu^\top_{h,a}(w - P_h v). \tag{47}$$

The geometry-aware candidate is no larger than the original scalar CRFO candidate evaluated on the same history and continuation*. It is optimistic, and its executed policy satisfies the complete guarantee of Theorem 2, with the same L, certification charge, reward charge, and failure probability. The same conclusions hold with w and U replaced by the computed upper responses and continuations in the capped upper-certificate implementation below, without a prescribed numerical optimization accuracy.*

Proof. Step 1. Simultaneous containment, including approximate groups. The original sampling and certificate event gives $\mathrm{TV}(\mu_{h,a}, \hat{\mu}^k_{h,i}) \le \lambda^k_{h,i}$ for every stream containing the target and $\mathrm{TV}(P_h(\cdot \mid z,c), \hat{P}^k_h(\cdot \mid z,c)) \le r_S(n^k_{h,z,c})$. Thus each true law belongs to the relevant polytope. Intersection needs no independence between streams and no additional union bound. The existing prefix-uniform event applies at the adaptive visit counts. The diameter allowance in λ is still required when the pool is only approximate.

Step 2. Conditional response. The maximization defining $w(c)$ ranges over a set containing the true response, so $w(c) \ge (P_h v)(z,c)$. Every feasible p also satisfies

$$p^\top v \le (\hat{P}^k_h v)(z,c) + \mathrm{span}(v) r_S(n^k_{h,z,c}) \le (\hat{P}^k_h v)(z,c) + B_h r_S(n^k_{h,z,c}),$$

and $p^\top v \le \max v \le B_h$. Consequently $w(c) \le W^{\mathrm{MB,k}}_h(z,c)$. Comparing both feasible laws with their common center proves (46). These are deterministic inequalities on the simultaneous event, so the test vector v can depend on the same data.

Step 3. Target-law disagreement and the span charge. Let q^+ maximize $q^\top w$ over the input intersection. Compactness gives a maximizer. For every stream i,

$$\mathrm{TV}(q^+, \mu_{h,a}) \le \mathrm{TV}(q^+, \hat{\mu}^k_{h,i}) + \mathrm{TV}(\hat{\mu}^k_{h,i}, \mu_{h,a}) \le 2\lambda^k_{h,i}.$$

Probability vectors have total-variation distance at most one. Hence $(q^+ - \mu_{h,a})^\top w \le \mathrm{span}(w)\ell^k_{h,a}$. The lower bound follows because both $\mu_{h,a}$ and the true response are feasible. Adding the nonnegative

response error proves (47). The factor $\mathrm{span}(w)$ expresses *value relevance*: redistributing probability among equally valued optimistic responses cannot alter their expectation. It does not certify that the true response is constant, and does not remove its estimation cost.

Step 4. Comparison with scalar candidates. For every stream and every feasible q,

$$q^\top w \le \hat{\mu}_{h,i}^{k\top} w + \lambda_{h,i}^k \,\mathrm{span}(w) \le \hat{\mu}_{h,i}^{k\top} W_h^{\mathrm{MB,k}} + B_h \lambda_{h,i}^k.$$

Taking the maximum on the left and the minimum over streams on the right proves pointwise tightening. Adding the same reward upper bound and clipping preserve the comparison. Backward induction establishes $\overline{Q}_h^k \ge Q_h^*$ and $\overline{V}_h^k \ge V_h^*$, starting with reward confidence at the terminal stage. The scalar comparator in this argument uses the refined learner's continuation, not the continuation of an independently trained scalar learner.

Step 5. Regret on the actually executed history. Equation (47) is bounded above by $2B_h \min_i \lambda_{h,i}^k + 2B_h \sum_c \mu_{h,a}(c) r_S(n_{h,z,c}^k)$. It therefore gives the same one-step inequality as Lemma 4, with the refined learner's d_h^k and executed actions. The martingale is defined on all histories, where clipping gives $d_{h+1}^k \in [-B_h, B_h]$ and the same range bound for $Y_{k,h}$. The empty-intersection fallback is bounded and measurable and is never activated on $\mathcal{E}$. Actual-row counting, the unknown-reward charge, and all certification-gate cases are unchanged. This proves Theorem 2 for the refined execution. It does not compare its realized regret with that of a separately trained scalar learner.

Step 6. Certified numerical upper solutions. The construction below produces response and input values between their exact support values and valid scalar upper bounds. These two inequalities are the only properties used above, except that the input-distance argument is replaced by the capped span inequality proved below. Thus all conclusions remain valid without assuming that a floating-point optimizer returns an exact optimum. □

C.9.2 Finite linear programs and upper certificates

This implementation of (45) must return an upper certificate, not merely a feasible primal value, to preserve Proposition 20. For any nonnegative value vector $v \in \mathbb{R}^d$, probability centers $\hat{p}_i \in \Delta([d])$, and radii $r_i \in [0,1]$, consider

$$\begin{aligned} \max_{q,t_1,\dots,t_J} \quad & v^\top q \qquad \text{subject to} \quad \sum_j q_j = 1, \quad 0 \le q_j \le 1, \\ & q_j - \hat{p}_{i,j} \le t_{i,j}, \quad \hat{p}_{i,j} - q_j \le t_{i,j}, \quad 0 \le t_{i,j} \le 1, \quad \sum_j t_{i,j} \le 2r_i. \end{aligned} \tag{48}$$

The auxiliary variables represent absolute deviations. Their upper bounds lose no feasible probability vector. The response problem uses $d = S, J = 1$, and each input problem uses $d = C, J \le 2$. There are $d(J+1)$ variables, $J(2d+1)$ non-bound inequalities, and one equality. Thus the refinement requires finite linear programs (LPs), not enumeration of all action partitions. For known laws, the corresponding maximization is their exact expectation.

Writing this program as $\max c^\top x$ subject to $Ax \le b$, $Ex = d_0$, and $0 \le x \le 1$ gives a convenient certificate. Here $c = (v, 0, \dots, 0)$ and $d_0 = 1$; d_0 is an equality right side, not the alphabet size.

Lemma 21 (Box-corrected dual upper certificate)**.** *For any nonnegative multipliers $y \ge 0$ and unrestricted multiplier z, every feasible x satisfies*

$$c^\top x \le \mathcal{D}(y,z) := b^\top y + d_0 z + \sum_j [c - A^\top y - E^\top z]_j^+. \tag{49}$$

The formula does not require the remaining dual residual to vanish or to have a prescribed sign.

Proof. Decompose $c = A^\top y + E^\top z + r$. Feasibility and $y \ge 0$ give $c^\top x \le b^\top y + d_0 z + r^\top x$. Since $x_j \in [0,1]$, $r^\top x \le \sum_j \max\{r_j, 0\}$. This proves the claim. Strong duality for a nonempty compact linear program shows that minimizing this box-form dual recovers the exact support value. Exact optimality is not required for an upper certificate. □

Define the analytic support bound

$$B_{\mathrm{an}}(v) = \min\{\max v,\ \min_i[\hat{p}_i^\top v + r_i \operatorname{span}(v)]\}.$$

It upper-bounds the exact support whenever the intersection is nonempty. At any optimization iteration, use

$$\widehat{U}(v) = \min\{B_{\mathrm{an}}(v), \max\{\min v, \mathcal{D}(y,z)\}\}.$$

Both arguments of the outer minimum are valid upper bounds, so this cap never becomes overconfident on a nonempty feasible set. In the response program, it yields a vector $\widetilde{w}$ with $P_h v \le w \le \widetilde{w} \le W_h^{\mathrm{MB,k}}$ on $\mathcal{E}$. For the input program, use $v = \widetilde{w}$. Because $\mu_{h,a}$ lies in every ball,

$$0 \le \widehat{U}(\widetilde{w}) - \mu_{h,a}^\top \widetilde{w} \le \operatorname{span}(\widetilde{w}) \min\{1, 2 \min_i \lambda_{h,i}^k\}.$$

The $2\lambda_i$ inequality follows by comparing the analytic bound with the true-law expectation through center i. The factor one follows from the range $[\min \widetilde{w}, \max \widetilde{w}]$. This proves the capped-solver part of Proposition 20. A solver may stop early, or use B_{an} alone if no useful certificate is available. A feasible *primal* value is a lower bound on the maximum and must not be substituted for this optimistic upper value.

Floating-point solver output is not by itself a mathematical certificate. The accompanying reference routine recomputes (49) in exact rational arithmetic for the supplied finite-precision coefficients and rounds the returned upper value outward. This certifies that represented polytope. Appendix C.9.6 specifies conservative confidence inputs, the implemented solve modes, and the deterministic allowance for radius enlargement. Exact count ratios avoid center-rounding error. No guarantee for an unverified numerically underestimated radius is asserted. If a numerical feasibility decision is uncertain, using the original scalar backup preserves its confidence proof.

C.9.3 A STRICT, OBSERVATION-CONSISTENT EXAMPLE

This construction shows that the refinement in Proposition 20 can be strict even when the individual and pooled streams overlap. Take three signatures and a known response-value vector $w = (0, 1/2, 1)$. Let an individual stream have center $(3/5, 2/5, 0)$ and radius $1/5$, and its pooled stream have center $(3/5, 1/10, 3/10)$ and radius $1/10$. Their intersection is the singleton $q = (3/5, 1/5, 1/5)$. Indeed, the centers have TV distance $3/10$, the sum of the radii. Equality in the triangle inequality fixes the first coordinate at $3/5$ and requires moving exactly $1/5$ from the second coordinate of the individual center to the third.

The two separate exact support values are $2/5$ and $9/20$. Taking their minimum gives $2/5$, whereas their intersection has value $3/10$. Thus

$$\max_{q \in \cap_i \mathcal{C}_i} q^\top w = \frac{3}{10} < \min_i \max_{q \in \mathcal{C}_i} q^\top w = \frac{2}{5}.$$

The gain is due to eliminating mutually incompatible optimistic laws, not additional observations. The empirical centers are consistent with overlapping streams: the pool can contain 1000 individual observations with counts $(600, 400, 0)$ and 1000 observations from each of three other actions, each with counts $(600, 0, 400)$. Under the true common law $(3/5, 1/5, 1/5)$, every individual center is within its radius $1/5$, and this finite history has positive probability. There are four actions in this example. The radii have the required inverse-square-root ratio for sample counts 1000 and 4000. Taking $K = 4001$ and setting $L_C = 80$ through an admissible $\delta = 64H(2A + ZC)(K + 1)\exp(C \log 2 - 80) \in (0, 1)$ makes the radii exactly $1/5$ and $1/10$. A two-stage instance with terminal reward equal to a binary successor and $P(Y = 1 \mid c) = (0, 1/2, 1)$ realizes the value vector. This example proves strict improvement in a Bellman upper bound, not a regret separation between complete learning trajectories.

C.9.4 SCOPE AND COMPUTATIONAL TRADEOFF

The confidence-intersection mode spends computation to reduce a conservative optimistic relaxation. It does not remove the known-separation premise for exact grouping, the sampling cost of certification, response-sufficiency assumptions, or the need to visit unknown response rows. It leaves the worst-case rate in Theorem 2 unchanged. The sharper span charge can be informative even when the input laws remain uncertain, but cannot replace response learning. The historical TD experiments use the scalar construction and do not validate this new refinement empirically.

C.9.5 When an intersection can add value

A comparison with the scalar bonus combines several changes: simplex constraints, value-span dependence, response supports, and intersection. The appropriate comparator for the last change holds the response vector fixed and optimizes each stream separately.

Lemma 22 (Binary and nested-set boundary)**.** *For $C = 2$, any nonempty intersection of simplex TV balls $\mathcal{C}_i$ and any common vector w satisfy*

$$\max_{q \in \cap_i \mathcal{C}_i} q^\top w = \min_i \max_{q \in \mathcal{C}_i} q^\top w.$$

The same identity holds in any dimension if one $\mathcal{C}_i$ is contained in every other ball.

Proof. Write $q = (1 - t, t)$ and $\mathcal{C}_i = [\ell_i, u_i]$ in the coordinate t. The objective is $w_0 + (w_1 - w_0)t$. For nonnegative slope, its intersection maximum is at $\min_i u_i$; for negative slope it is at $\max_i \ell_i$. These equal the smallest individual maxima in the respective cases. A constant objective is immediate. Under nesting the intersection is simply the smallest set. □

Thus binary-input experiments cannot isolate an extra intersection gain over separate exact supports, although they can reveal improvements over scalar bonuses. With $C \geq 3$, a strict intersection improvement also certifies that neither set contains the other. Its absence does not prove nesting: non-nested sets can have the same support in a particular direction.

The construction is targetwise, rather than a global optimistic model with shared parameters. For example, a common binary probability $p \in [0.4, 0.6]$ and known response vectors $(0, 1)$ and $(1, 0)$ yield pointwise upper values 0.6 at two targets. They require $p = 0.6$ and $p = 0.4$, respectively, so no single common law attains both. This does not harm pointwise optimism or the executed-history proof.

A strict useful decision, conditional on a realizable history. Extend the three-category witness above to $H = 2$, $A = 4$, and $Z = 2$. Action zero has known continuation vector $(0, 1/2, 1)$; the other three actions have constant known continuation $7/20$, implemented by a binary terminal reward with success probability $7/20$. First rewards are zero. All four signature laws are $(3/5, 1/5, 1/5)$, with their equality supplied. Use the same overlapping counts, $K = 4001$, and the stated formula for δ with these dimensions to obtain $L_C = 80$. Separate exact supports choose action zero, since $2/5 > 7/20$. Intersection chooses a constant-response action, since $3/10 < 7/20$, improving its true value by $1/20$.

This proves a decision consequence beyond a smaller confidence number. It remains a conditional existence result: the prescribed fixed-collection history has positive probability, but is exceptionally rare and is not asserted to arise along CRFO's greedy trajectory. The small confidence level is also not representative. Appendix H.5 separately reports every run of an ordinary-confidence on-policy diagnostic, including any zero-activation cases.

C.9.6 Representing statistical confidence before optimization

The numerical and statistical certificates have different responsibilities. The support routine certifies the exact rational polytope supplied to it. The `confidence_inputs.py` interface first constructs conservative inputs to that routine. For integer category counts it returns exact ratios $N_c / \sum_j N_j$ (uniform for zero count). For rational δ, it brackets the argument of the logarithm by directed decimal division, encloses each correctly rounded logarithm and square root by adjacent decimals, and directs the remaining arithmetic outward. It returns rational endpoints $[r_d^-, r_d^+]$ containing the ideal radius. Bias diameters, if used, must likewise be exact rationals or outward upper bounds. Clipping at one is monotone. No unverified center normalization or inward-rounded radius is hidden in this interface.

Supply r_d^+ and exact centers to the support primitive, and retain its `exact_upper` rational value for nested supports and action comparisons. The floating `upper` field is an outward display approximation. The solver may approximate the rational LP while searching; the box-corrected rational calculation certifies its candidate multipliers against the represented coefficients, independently of that approximation.

Theorem 2 states the ideal-radius bound. To account explicitly for representation enlargement, suppose every used signature/response radius exceeds its ideal value by at most η; center ratios and all other quantities are exact. The same one-step proof adds at most $2B_h\eta$ to each of the input and response charges, hence at most $4H_1K\eta$ to the complete regret bound. This statement concerns numerical enlargement, not statistical failure. An implementation can retain the original gate and append this deterministic allowance, or use correspondingly enlarged complete bounds in the gate. The new diagnostic uses known responses, so only the input allowance is needed; its reported endpoint widths bound the enlargement. Finite boundary checks test this construction but do not replace the confidence proof.

The supplied optimizer has two modes: a completed HiGHS solve followed by rational upper verification, or the analytic no-solve upper bound. The earlier 300 analytic-path checks are labeled accordingly. Lemma 21 also applies to arbitrary intermediate multipliers, but the package does not expose or benchmark an interrupted-optimizer iterate. A failed or inconclusive solve returns the analytic bound; no primal feasible objective is used as an optimistic certificate.

C.10 Confidence surplus and covariance-resolved Bellman propagation

This subsection proves Proposition 25. Unlike a comparison of two independently trained policies, the result evaluates every confidence width and every tightening on the single history actually generated by CRFO. It also separates deterministic estimation allowances from the conditional variance of the trajectory martingale. No minimum visitation probability or independence between the input and response estimators is imposed.

C.10.1 Computable residual allowances

These residuals retain the tightening achieved by the planner in Proposition 25, rather than replacing every optimized value by its analytic cap. For a nonterminal target, suppress (k,h,s,a) and set $v=\overline{V}^k_{h+1}$, $z=\phi_h(s,a)$, $r_c=r_S(n^k_{h,z,c})$ and $\lambda_*=\min_{i\in I_h(a)}\lambda^k_{h,i}$. Let w_c be the exact response support or its capped upper certificate, and let U be the corresponding exact or certified input upper value from Appendix C.9. Thus

$$\max_{p\in\mathcal{C}^P_c} p^\top v \le w_c \le T_c, \qquad \max_{q\in\mathcal{C}^\mu} q^\top w \le U \le A(w),$$

where the analytic caps are

$$T_c=\min\{\max v, \hat{P}_c^\top v + r_c\,\mathrm{span}(v)\}, \qquad A(w)=\min\{\max w, \min_i[\hat{\mu}_i^\top w+\lambda_i\,\mathrm{span}(w)]\}. \tag{50}$$

The *confidence surpluses* $g^P_c=T_c-w_c$ and $g^\mu=A(w)-U$ quantify the tightening already achieved by planning, not unknown statistical errors. Define

$$a^P(c)=\mathrm{span}(v)\min\{1,2r_c\}-g^P_c, \qquad a^\mu=\mathrm{span}(w)\min\{1,2\lambda_*\}-g^\mu. \tag{51}$$

A no-solve certificate has zero surplus at the relevant maximization; a better certified solve can increase it. An exact solve is not assumed. On an empty-intersection history use the original scalar backup and set $a^P(c)=2B_hr_c$, $a^\mu=2B_h\lambda_*$, and both surpluses to zero. This all-history convention is essential for the concentration argument.

Lemma 23 (Surplus-preserving transfer)**.** *Whenever the input intersection is nonempty, the quantities in* (51) *are measurable and satisfy* $0\le a^P(c),a^\mu\le B_h$ *on every history. On* $\mathcal{E}$*,*

$$0\le w_c-P_cv\le a^P(c), \qquad 0\le U-\mu_a^\top Pv\le a^\mu+\mu_a^\top a^P. \tag{52}$$

The same upper inequalities hold for the scalar fallback with its stated allowances.

Proof. Step 1: bounds that do not condition on confidence. The empirical law belongs to its response ball, so $w_c\ge\hat{P}_c^\top v$ and $0\le g^P_c\le r_c\,\mathrm{span}(v)\le\mathrm{span}(v)\min\{1,2r_c\}$. Also $a^P(c)\le\mathrm{span}(v)\le B_h$. For any q in a nonempty input intersection, $U\ge q^\top w\ge\hat{\mu}_i^\top w-\lambda_i\,\mathrm{span}(w)$ for every i. Consequently $0\le g^\mu\le 2\lambda_i\,\mathrm{span}(w)$ for every i, and $g^\mu\le\mathrm{span}(w)$ because both caps lie in the range of w. These give $0\le a^\mu\le B_h$. Centers and radii are history-measurable, including data-selected diameter certificates. Each bounded polytope has finitely many candidate

bases. On a nonzero-determinant basis its candidate vertex is a Borel function of the coefficients; feasibility is a Borel condition. Maximizing over the feasible candidates gives a measurable support on the nonempty-set event. This argument does not require the radii to take finitely many values. Deterministic tie-breaking and the rational upper-certificate operations are measurable as well.

Step 2: subtract surplus before discarding statistical geometry. On $\mathcal{E}$, P_c belongs to its response ball. Hence $w_c \geq P_c v$ and $T_c - P_c v \leq \mathrm{span}(v)\min\{1, 2r_c\}$ by comparison through the empirical center and by the range cap. Subtracting g_c^P gives the first bound. Similarly μ_a belongs to every input ball, so $U \geq \mu_a^\top w$ and $A(w) - \mu_a^\top w \leq \mathrm{span}(w)\min\{1, 2\lambda_*\}$. Subtract g^μ, and insert $\mu_a^\top w$ between U and $\mu_a^\top P v$. This proves (52). Scalar fallback follows from Lemma 3. The two surplus subtractions are justified separately; no product of estimated laws is interpreted as a physical response sample. □

C.10.2 The exact predictable variance

The following decomposition refines the trajectory martingale of Lemma 4 while retaining input–successor dependence. Let $d_h^k = \overline{V}_h^k - V_h^{\pi^k}$ and, for $k > \tau, h < H$, define

$$Y_{k,h} = a_{k,h}^P(c_h^k) + d_{h+1}^k(s_{h+1}^k), \quad \xi_{k,h} = \mathbb{E}[Y_{k,h} \mid \mathcal{F}_{k,h}^-] - Y_{k,h}, \quad v_{k,h} = \mathrm{Var}(Y_{k,h} \mid \mathcal{F}_{k,h}^-).$$

Use zero for $k \leq \tau$, and write $\mathcal{V}_K = \sum_{k,h<H} v_{k,h}$. Conditional on pre-step information, put $f_c = a_{k,h}^P(c) + (P_h d_{h+1}^k)(z_h^k, c)$. Response sufficiency gives the exact identity

$$v_{k,h} = \mathrm{Var}_{c\sim\mu_{h,a_h^k}}(f_c) + \sum_c \mu_{h,a_h^k}(c)\,\mathrm{Var}_{s'\sim P_h(\cdot|z_h^k,c)}(d_{h+1}^k(s')). \tag{53}$$

Equivalently the first term equals $\mathrm{Var}_c(a^P(c)) + \mathrm{Var}_c(P_h d) + 2\,\mathrm{Cov}_c(a^P(c), P_h d)$. This covariance need not be positive and must not be dropped. Splitting ξ at the intermediate signature sigma-algebra gives two orthogonal increments, not two independent random variables. The second has conditional mean zero given the input; tower conditioning makes their cross expectation zero. Across stages, the complete-step martingale increments are likewise orthogonal in expectation even though the states, actions, counts, and value estimates are dependent. The quantity $\mathcal{V}_K$ is an analysis-only predictable quadratic variation, not an extra input to the planner.

Lemma 24 (A finite-horizon Bernstein–Freedman envelope)**.** *Let $N = K(H-1)$, $b = 4H$, $J = \lceil \log_2(N \vee 1)\rceil$, and $\ell = \log(16(J+1)/\delta)$. For the above all-history increments, with probability at least $1 - \delta/8$,*

$$\sum_{k,h<H} \xi_{k,h} \leq \mathfrak{B}_K(\mathcal{V}_K) := \min\left\{2\sqrt{2KH_2\log(16/\delta)},\ 2\sqrt{(\mathcal{V}_K + b^2)\ell} + \frac{2b\ell}{3}\right\}. \tag{54}$$

Proof. Always $d_{h+1}^k \in [-B_h, B_h]$ and $a^P \in [0, 2B_h]$, including the fallback. Thus $Y_{k,h} \in [-B_h, 3B_h]$, $|\xi_{k,h}| \leq 4B_h \leq b$, and its conditional range has length $4B_h$. Conditional Hoeffding iteration yields the first bound with failure $\delta/16$.

For completeness, conditional Taylor expansion and $|\xi| \leq b$ give, for $0 < t < 3/b$,

$$\mathbb{E}[e^{t\xi} \mid \mathcal{F}^-] \leq \exp\{t^2\mathbb{E}[\xi^2 \mid \mathcal{F}^-]/[2(1 - tb/3)]\}.$$

Indeed, $\mathbb{E}\xi = 0$, $|\mathbb{E}\xi^j| \leq b^{j-2}\mathbb{E}\xi^2$ for $j \geq 2$, and $j! \geq 2\,3^{j-2}$. Multiplying the conditional bounds produces a nonnegative supermartingale. Exponential Markov inequality, or stopping before its variance exceeds a deterministic v, and optimization in t give

$$\mathbb{P}\{\textstyle\sum \xi \geq \sqrt{2vx} + 2bx/3,\ \mathcal{V}_K \leq v\} \leq e^{-x}.$$

One may verify the displayed threshold by solving $u^2/[2(v + bu/3)] \geq x$; it upper-bounds the positive quadratic root. This is the scalar Bernstein–Freedman argument (Freedman, 1975), included to specify the constants and conditioning.

Since $\mathcal{V}_K \leq Nb^2$, partition its range into $[0, b^2]$ and $(2^{j-1}b^2, 2^j b^2]$, $j = 1, \ldots, J$. Apply the preceding bound with upper endpoint v in each bin and $x = \ell$. Every endpoint is at most $2(\mathcal{V}_K + b^2)$ on its bin. The union failure is $(J+1)e^{-\ell} = \delta/16$. Intersecting this event with the Hoeffding event proves (54). No confidence or optimism event has been conditioned on. □

Proposition 25 (Covariance-resolved regret with geometric surplus)**.** *For confidence-intersection CRFO with the original certification gate,*

$$\mathrm{Reg}(K) \le \min\left\{ HK,\ H\tau + \sum_{k>\tau,h} \beta^k_{h,R} + \sum_{k>\tau,h<H} [a^\mu_{k,h} + a^P_{k,h}(c^k_h)] + \mathfrak{B}_K(\mathcal{V}_K) \right\} \tag{55}$$

holds simultaneously with the regret bound (10) *at probability at least* $1-\delta$*. The explicit Bernstein–Freedman/Azuma minimum* $\mathfrak{B}_K$ *is given in* (54)*. Input and response surpluses enter with negative signs; the variance retains conditional covariance. No independence beyond the conditional laws in* (1) *is added. The complete proof follows.*

Complete proof of Proposition 25. On $\mathcal{E}$, subtract the executed-policy Bellman equation from the optimistic update and use Lemma 23:

$$d^k_h(s^k_h) \le \beta^k_{h,R} + a^\mu_{k,h} + \mathbb{E}[a^P_{k,h}(c^k_h) + d^k_{h+1}(s^k_{h+1}) \mid \mathcal{F}^-_{k,h}].$$

Insert $Y_{k,h} + \xi_{k,h}$. Sum over h and $k > \tau$; the value errors telescope, and at H retain $d^k_H \le \beta^k_{H,R}$. Optimism at the initial state gives

$$\mathrm{Reg}(K) \le H\tau + \sum_{k>\tau,h} \beta^k_{h,R} + \sum_{k>\tau,h<H} (a^\mu_{k,h} + a^P_{k,h}(c^k_h)) + \sum_{k,h<H} \xi_{k,h}.$$

Apply Lemma 24 and the deterministic cap HK. Substituting (51) exposes both cumulative gain terms with negative signs. This does not compare trajectories of separately trained algorithms.

The certificate, sampling, and reward events already have total failure below $\delta/2$. The new two-tail event costs at most $\delta/8$. Intersect also with the original trajectory events (at most $2e^{-L}$). The explicit L leaves the sum below δ, so the new bound and Theorem 2 hold simultaneously at the stated level. No extra modeling assumption or additional observations have been introduced. Numerically enlarged radii require the deterministic enlargement allowance of Appendix C.9.6; optimization error itself is retained through the computable surplus, since any unextracted gain remains in a^P, a^μ. □

C.10.3 Joint rewards: why the variance terms cannot be separated arbitrarily

The preceding trajectory variance concerns Bellman *estimation error*, not the physical return variance. Their relationship must not be asserted without additional arguments. The following identity records the exact physical decomposition permitted by our reward premise.

Lemma 26 (Joint-reward Bellman variance budget)**.** *Fix a deterministic Markov policy* π *selected measurably at episode start. Put* $Z_h = r_h + V^\pi_{h+1}(s_{h+1})$ *and* $\eta_h = Z_h - V^\pi_h(s_h)$*. Then* η_h *is a complete-step martingale difference and*

$$\begin{aligned} \mathrm{Var}(Z_h \mid \mathcal{F}^-_h) &= \mathrm{Var}(r_h \mid \mathcal{F}^-_h) + \mathrm{Var}_{c\sim\mu}(P_h V^\pi_{h+1}) \\ &\quad + \mathbb{E}_{c\sim\mu}\,\mathrm{Var}_{P_h(\cdot|z,c)}(V^\pi_{h+1}) + 2\,\mathrm{Cov}(r_h, V^\pi_{h+1}(s_{h+1}) \mid \mathcal{F}^-_h), \\ \mathbb{E}\left[\sum_{h=1}^H \mathrm{Var}(Z_h \mid \mathcal{F}^-_h) \mid \mathcal{F}_{k,0}\right] &= \mathrm{Var}\left(\sum_{h=1}^H r_h \mid \mathcal{F}_{k,0}\right) \le H^2/4. \end{aligned} \tag{56}$$

Proof. The conditional reward mean in (1) and the Bellman equation give $\mathbb{E}[\eta_h \mid \mathcal{F}^-_h] = 0$. Expanding the variance of $r_h + V^\pi_{h+1}(s_{h+1})$ retains its covariance, and conditioning the continuation on c gives the first display. The sum of η_h telescopes to $\sum_h r_h - V^\pi_1(s_1)$. Distinct increments have zero conditional cross expectation by the tower property, even when rewards at different times have a dependent joint law. Squaring the sum, conditioning on $\mathcal{F}_{k,0}$, and using $\sum_h r_h \in [0, H]$ prove the second display. This is a covariance-complete specialization of Bellman total-variance reasoning (Azar et al., 2017), not a claim that our confidence radii achieve the minimax Bernstein regret rate. □

For example, let $H = 2$, one action, $Y \sim \mathrm{Bern}(1/2)$, $r_1 = 1 - Y$, and $r_2 = Y$. The episode return is exactly one, but the first-step reward and continuation each have variance $1/4$ and covariance $-1/4$. Dropping the covariance predicts positive noise where the actual return has none. Reward concentration still uses the pre-step conditional mean and remains valid.

C.11 SELF-NORMALIZATION, OVERLAPPING STREAMS, AND SEQUENTIAL COMPLEXITY

Theorem 2 uses categorical confidence, not a linear-MDP assumption. This subsection makes the corresponding vector-martingale and function-class structure explicit. It explains why adaptive continuation values and overlapping streams cannot be treated as independent samples, and why a small number of input groups does not imply a small response complexity. The concentration and eluder tools below are established tools (Abbasi-Yadkori et al., 2011; Russo & Van Roy, 2013; Jin et al., 2020); the overlap calculation and observation-specific specializations identify their role in our model. They do not assert a regret theorem for arbitrary neural function approximation.

C.11.1 A UNIFORM VECTOR CERTIFICATE FOR DATA-DEPENDENT CONTINUATIONS

This certificate explains the uniformity required when Section 4.2 evaluates a continuation fitted from overlapping adaptive observations. Let $(\mathcal{G}_t)_{t\ge 0}$ be a filtration, $x_t \in \mathbb{R}^D$ be $\mathcal{G}_{t-1}$-measurable with $\|x_t\|_2 \le 1$, and $Y_t \mid \mathcal{G}_{t-1}$ have kernel Q_t. Let $\mathcal{F} \subset [0,B]^{\mathcal{Y}}$ be a deterministic, sup-norm-separable class of measurable functions admitting a finite sup-norm α-net of size N_α, for fixed $\alpha > 0$. Define

$$\Lambda_t = \lambda I + \sum_{i=1}^{t} x_i x_i^\top, \qquad S_t(f) = \sum_{i=1}^{t} x_i [f(Y_i) - Q_i f], \quad \lambda > 0.$$

Here $Q_i f$ is the integral of a *fixed test function* against the predictable kernel. After obtaining a simultaneous event, one may substitute a data-selected f; $Q_i f$ must not be replaced by an unjustified conditional expectation of a future-selected random function.

Lemma 27 (Uniform self-normalized continuation certificate)**.** *With probability at least $1-\delta$, simultaneously for every $t \ge 0$ and $f \in \mathcal{F}$,*

$$\|S_t(f)\|_{\Lambda_t^{-1}} \le \frac{B}{2}\sqrt{\log\frac{\det \Lambda_t}{\det(\lambda I)} + 2\log(N_\alpha/\delta)} + 2\alpha\sqrt{t}. \tag{57}$$

Sup-norm separability makes the supremum event measurable. The certificate remains valid at stopping times and for functions chosen from the class after observing the data.

Proof. Step 1: fixed functions and predictable features. For deterministic f, the noise $\eta_i = f(Y_i) - Q_i f$ is conditionally mean zero and $B/2$-sub-Gaussian by the conditional Hoeffding lemma. Suppose $B > 0$ and normalize it to unit sub-Gaussian noise. For fixed $v \in \mathbb{R}^D$,

$$M_t(v) = \exp\left\{ v^\top \frac{2S_t(f)}{B} - \frac{1}{2} v^\top (\Lambda_t - \lambda I) v \right\}$$

is a nonnegative supermartingale. Integrate v against the independent Gaussian measure $\mathcal{N}(0,(\lambda I)^{-1})$. Tonelli's theorem applies because the integrand is nonnegative. Completing the square gives the mixture

$$\overline{M}_t = \frac{\det(\lambda I)^{1/2}}{\det(\Lambda_t)^{1/2}} \exp\left\{ \frac{2}{B^2} \|S_t(f)\|^2_{\Lambda_t^{-1}} \right\}.$$

Its initial mean is one. Stopping at its first crossing of $1/\delta$, applying the bounded-stopping supermartingale inequality, and then monotone convergence give $\mathbb{P}(\sup_t \overline{M}_t \ge 1/\delta) \le \delta$. This proves the fixed-function self-normalized bound of Abbasi-Yadkori et al. (2011). If $B = 0$, the process is identically zero.

Step 2: uniformity rather than a false fixed-function argument. Apply the first step with error probability δ/N_α to each deterministic net element, then take a union bound. For $\|f - f_\alpha\|_\infty \le \alpha$, write $e_i = (f - f_\alpha)(Y_i) - Q_i(f - f_\alpha)$, so $|e_i| \le 2\alpha$ on every history. If X has columns $x_1, \dots, x_t$, then

$$X^\top (\lambda I + XX^\top)^{-1} X \preceq I, \quad \|Xe\|_{\Lambda_t^{-1}} \le \|e\|_2 \le 2\alpha\sqrt{t}.$$

The matrix inequality follows from a singular-value decomposition, whose nonzero eigenvalues are $\sigma_j^2/(\lambda + \sigma_j^2) \le 1$. The triangle inequality proves (57). This contraction keeps the discretization remainder at $\sqrt{t}$ rather than replacing it by a triangle bound proportional to $t/\sqrt{\lambda}$. The covering

argument itself is the value-aware uniform-concentration device of Jin et al. (2020), not a new linear-RL result.

Step 3: measurability and the paper's feedback order. A countable dense subset determines the supremum, since each finite-time expression is sup-norm continuous in f. On a finite alphabet, $\mathcal{F}$ may be taken to be the compact cube $[0, B]^{\mathcal{Y}}$; all involved maps are continuous in its finitely many coordinates. The singleton identity class also covers bounded real-valued rewards. For response rows choose x_t to be their one-hot label, after c_t but before s_{t+1}; response sufficiency makes $Q_t = P_h(\cdot \mid z_t, c_t)$. For reward estimation use the state–action label *before* the joint observation, because its centered reward need not have mean zero after conditioning on c_t. The corresponding row-visit stopped processes are predictable. Conditioning on the completed certification history is permissible; conditioning on future optimism is not. □

For $|\mathcal{Y}| = S$, a coordinate grid gives $N_\alpha \le (1 + \lceil B/\alpha \rceil)^S$. Thus this result applies to the actual learned continuations without an extra realizability assumption. It does not remove state dependence from their complexity. In the tabular response design, $D = ZC$ and $\log\det(\Lambda_t/\lambda I) = \sum_{z,c} \log(1 + n_{z,c}/\lambda)$: only observed response rows contribute information. Separate stages use separate predictable kernels and a union bound, not an independence assertion across stages.

C.11.2 An overlap correction that independent-stream reasoning misses

The next calculation quantifies the dependence between the individual and pooled input streams maintained in Section 4.1. Fix a true exact block G after certification, one stage, and a fixed bounded category function f. Let $n_a > 0$, $n_G = \sum_{b \in G} n_b$, and let r_G be the number of visited actions in the block. Use one-hot action features in Lemma 27; its common conditional mean is $\mu_G^\top f$. Write β_t for the right side of (57) with δ replaced by $\delta/(HA)$, which covers at most HA returned stage–block pairs. For visited actions set $u_b = \mathbf{1}\{b = a\}/n_a - 1/n_G$, and set $u_b = 0$ at unvisited actions. Then

$$|\hat{\mu}_a^\top f - \hat{\mu}_G^\top f| \le \beta_t \sqrt{\frac{1}{n_a} - \frac{1}{n_G} + \lambda \left[\left(\frac{1}{n_a} - \frac{1}{n_G} \right)^2 + \frac{r_G - 1}{n_G^2} \right]}. \tag{58}$$

Proof. The common mean cancels because $\sum_b n_b u_b = 0$, leaving $u^\top S_t(f)$. Weighted Cauchy–Schwarz gives $|u^\top S_t(f)| \le \|u\|_{\Lambda_t} \|S_t(f)\|_{\Lambda_t^{-1}}$. Since Λ_t is diagonal, direct expansion gives $\sum_b n_b u_b^2 = 1/n_a - 1/n_G$ and $\sum_b u_b^2 = (1/n_a - 1/n_G)^2 + (r_G - 1)/n_G^2$. All coefficients may depend on the observed counts because the vector inequality already holds simultaneously for every direction. Uniformity over f follows from Lemma 27, not from conditioning on the selected f. □

When $n_a = n_G$, the two estimates are identical and the right side is exactly zero. For deterministic allocation and independent draws with variance σ_f^2, the exact variance is $\mathrm{Var}(\hat{\mu}_a^\top f - \hat{\mu}_G^\top f) = \sigma_f^2(1/n_a - 1/n_G)$, since their covariance is σ_f^2/n_G. This last variance formula is not asserted conditional on arbitrary adaptive counts. Equation (58) is the corresponding adaptive certificate. Approximate groups additionally retain their mean-mismatch term; equality cannot be inferred from this formula without the declared class restriction.

A related feedback-order failure is easy to miss. With $c \sim \mathrm{Bern}(1/2)$ and $r = c$, the reward mean is $1/2$ before the input. However, $\mathbb{E}[c(r - 1/2)] = 1/4$, so using the already-observed category as a reward-regression feature while retaining this unconditional centering produces a biased score. This is why the response and reward sigma-algebras in Appendix C.1 are not interchangeable.

C.11.3 What covering and eluder dimensions measure here

This comparison explains why reducing the number of input groups in Theorem 2 does not reduce unrestricted response complexity. For a scalar function class $\mathcal{F}$ on a query domain, a query x is η-independent of $x_1, \ldots, x_n$ if two functions have squared discrepancy at previous queries at most η^2 but discrepancy at x larger than η. The η-eluder dimension is the largest such sequence at a common scale at least η (Russo & Van Roy, 2013). It measures extrapolation from previous queries, whereas a covering number controls uniform selection among functions. These are different proof obligations.

Proposition 28 (Observation-specific response complexity)**.** *Let $J = ZC$ at a fixed stage and $S \geq 2$. For arbitrary response rows, their coordinate-probability class $\mathcal{F}_P = \{(j, s') \mapsto P_j(s') : P_j \in \Delta([S])\}$ satisfies, for $0 < \eta < 1$ and $\alpha > 0$,*

$$J(S-1) \leq \dim_E(\mathcal{F}_P, \eta) \leq JS, \qquad \log N_\alpha(\mathcal{F}_P, \|\cdot\|_\infty) \leq J(S-1)\log(1+\lceil (S-1)/\alpha \rceil). \tag{59}$$

For a fixed nonconstant continuation v, rescale its range to $[0, 1]$. The conditional-value class $j \mapsto P_j v$ has eluder dimension exactly J for $0 < \eta < 1$. With one known exact input group, these response complexities remain unchanged. Unconstrained unknown reward means retain their SA-coordinate complexity.

Proof. For the coordinate lower bound, query the first $S-1$ coordinates of each row. At a new coordinate, two distributions concentrated at that coordinate and at the unqueried last coordinate agree on all preceding coordinates and differ by one at the current query. Other rows can agree identically. This constructs $J(S-1)$ independent queries. A repeated query is dependent on itself, so a domain of size JS gives the upper bound. For the cover, round the first $S-1$ probabilities down on a grid of spacing $\alpha/(S-1)$ and assign the remaining mass to the last coordinate. The rounded vector remains a probability law and has sup-norm error at most α; counting grid choices proves the displayed bound.

For fixed v, mix independently within each row between states attaining its minimum and maximum. Every vector in $[0, 1]^J$ is realized, so J distinct row queries are independent and no longer sequence is possible. A common input law imposes no constraints on these conditional rows. Likewise, the model imposes no sharing relation on its arbitrary reward means, whose scalar query class is $[0, 1]^{SA}$. □

These dimensions concern the stated coordinate or continuation query classes, not a claim that the online learner may query any row on demand. Reachability and actual counts are handled separately by Lemma 4. The same distinction prevents an unjustified reduction to known-feature linear RL: $\tilde{P} = \mu P$ is bilinear in two unknown factors, and observing one factor's realization does not supply the other factor's unknown features. The one-hot representation is always available, but replacing its dimension by a smaller one requires a separate representation premise. No covering-number or eluder term is inserted into Theorem 2 merely by analogy with a different observation model.

D Streaming TD extension and common-policy comparison

The main model-based learner stores a successor histogram at each response row. This extension replaces that histogram by a scalar optimistic continuation estimate. Input certification and reward estimation are unchanged, but the scalar estimate uses historical targets, so its optimism and regret require a different argument. The extension is model-free for the conditional response, not for every factor. We use HYB to denote the common-policy hybrid that retains both MB and TD estimates.

D.1 Complete response-dependent bounds

The following envelopes make the storage–estimation comparison with Theorem 2 explicit before the TD update is introduced. Keep $\mathcal{C}_d, \mathfrak{S}, L, H_1, H_2$, and $\mathcal{G}_R$ from the main text and define

$$\begin{aligned} F_{\mathrm{TD}}(x) &= e\left[2x + H_1\{ZC + 12\sqrt{(H+1)LZCK}\} + 4\sqrt{2KH_2L} + \mathcal{G}_R\right], \\ F_{\mathrm{HYB}}(x) &= \min\{F_{\mathrm{MB}}(x), F_{\mathrm{TD}}(x)\}, \qquad e = \exp(1). \end{aligned} \tag{60}$$

For $b \in \{\mathrm{MB}, \mathrm{TD}, \mathrm{HYB}\}$, put $F_b^A = F_b(H_1\mathcal{C}_C(A, K))$ and $F_b^m = F_b(\sum_{h<H} B_h\mathcal{C}_C(m_h^\star, K))$. The backend b is fixed before sampling. The certification gate uses its corresponding F_b^A; no true group count is an input.

D.2 Streaming conditional-response estimator

Storing a full successor distribution is unnecessary when only its continuation expectation is queried. Following optimistic Q-learning (Jin et al., 2018), the TD implementation replaces each successor

Algorithm 2 Streaming TD variant of CRFO

Require: The same inputs as Algorithm 1, using the TD certification bound F^A_{TD}.
1: Obtain (τ, Π, ϵ) by Section 4.1. Reset signature and reward statistics.
2: Initialize every response-row counter to zero and $\widetilde{W}_h(z,c) = B_h$.
3: **for** $k = \tau + 1, \dots, K$ **do**
4: Form clipped $W^{\mathrm{TD,k}}$ and plan backward using (7). Freeze $\overline{V}^k, \pi^k$.
5: Execute the episode and update the observed signature streams and reward row if unknown.
6: For each visited response row, increment its count to i and apply (61) with the frozen target.
7: **end for**

histogram by a scalar $\widetilde{W}_h(z,c)$ initialized at B_h. At the row's ith actual visit in episode k, use

$$\alpha_i = \frac{H+1}{H+i}, \qquad b_{h,i} = 2B_h\sqrt{(H+1)L/i},$$
$$\widetilde{W}_h(z,c) \leftarrow (1-\alpha_i)\widetilde{W}_h(z,c) + \alpha_i[\overline{V}^k_{h+1}(s^k_{h+1}) + b_{h,i}]. \tag{61}$$

The target is frozen before its successor is observed. The bonus $b_{h,i}$ dominates weighted noise about the fixed optimal continuation, while α_i limits how often historical optimistic errors propagate. Clipping only the prediction $W^{\mathrm{TD,k}}_h = \min\{B_h, \widetilde{W}_h\}$ preserves the raw recursion's exact weight expansion. TD certifies the *optimal*, not necessarily the current, conditional continuation. The proof below resolves that difference without relabeling old targets.

TD is model-free *for the conditional response*, but still estimates signature frequencies. MB stores $O(HZCS)$ response statistics, versus $O(HZC)$ for TD. Both retain $O(HAC)$ signature statistics and $O(HSA)$ values and reward statistics (Appendix G.4). TD's smaller response memory incurs the historical-target propagation cost in the bounds in (69).

A hybrid maintains both responses, uses $W^{\mathrm{HYB}}_h = \min\{W^{\mathrm{MB}}_h, W^{\mathrm{TD}}_h\}$ in (7), and updates both on the same frozen continuation values and executed trajectory. It retains model-based storage. The minimum is optimistic on the simultaneous confidence event; the proof below does not compare the realized regret of separately trained policies.

D.3 Weight identities and optimism

This subsection proves the historical-weight and optimism identities for Algorithm 2, used by the TD regret proof in Appendix D. It uses the categorical and martingale events of Appendix C.2.

A temporal-difference target is reused through later updates. The relevant question is not just its current estimation error, but how often that error reappears downstream. The following identities quantify this reuse before the optimism argument.

Lemma 29 (TD weight identities)**.** *For $\alpha_t = (H+1)/(H+t)$, define $\alpha^0_t = \prod_{j=1}^t(1-\alpha_j)$ and $\alpha^i_t = \alpha_i\prod_{j=i+1}^t(1-\alpha_j)$ for $1 \le i \le t$. For $t \ge 1$,*

$$\alpha^0_t = 0, \quad \sum_{i=1}^t \alpha^i_t = 1, \quad \sum_{i=1}^t (\alpha^i_t)^2 \le \frac{H+1}{t},$$
$$\sum_{i=1}^t \frac{\alpha^i_t}{\sqrt{i}} \le \frac{2}{\sqrt{t}}, \qquad \sum_{t=i}^\infty \alpha^i_t = 1 + \frac{1}{H}. \tag{62}$$

For $t = 0$ set $\alpha^0_0 = 1$.

Proof. Because $\alpha_1 = 1$, the initial weight vanishes for every positive count. The weights telescope to one. Direct multiplication gives

$$\alpha^i_t = (H+1)\frac{(t-1)!(H+i-1)!}{(i-1)!(H+t)!}.$$

These weights increase with i, so their maximum is $\alpha_t^t = (H+1)/(H+t) \le (H+1)/t$. A probability vector's squared norm is at most its largest coordinate, proving the third assertion.

Let $s_t = \sum_i \alpha_t^i/\sqrt{i}$. The recursion is $s_t = (1-\alpha_t)s_{t-1} + \alpha_t/\sqrt{t}$. Starting with $s_1 = 1 \le 2$, induction reduces $s_t \le 2/\sqrt{t}$ to

$$2\sqrt{t(t-1)} + H + 1 \le 2(H+t),$$

which holds since $\sqrt{t(t-1)} \le t$ and $H \ge 1$.

For fixed i, write $w_t = \alpha_t^i$ and $u_t = (H+t)w_t/H$. Since $w_{t+1} = tw_t/(H+t+1)$, $u_t - u_{t+1} = w_t$. Also $u_t \to 0$ by the factorial formula, while $u_i = (H+i)\alpha_i/H = (H+1)/H$. Summing the telescoping relation proves the final identity. These are the familiar optimistic Q-learning weights (Jin et al., 2018); their downstream accounting in this model is given below. □

Lemma 30 (TD noise and upper-response optimism)**.** *There is an event of probability at least* $1-\delta/8$ *on which, for every response row and every positive prefix* $t \le K$*,*

$$\left| \sum_{i=1}^{t} \alpha_t^i \{V_{h+1}^*(s_i') - (P_h V_{h+1}^*)(z,c)\} \right| \le B_h \sqrt{(H+1)L/t}. \tag{63}$$

On this event, the reward-confidence event of Lemma 14, and the certified disturbance and response events, all backends satisfy $\overline{V}_h^k \ge V_h^*$*, including the hybrid using a common policy.*

Proof. For a fixed row, the summands before multiplying by deterministic weights are visit-indexed martingale differences with conditional range length B_h. Lemma 29 and the two-sided form of (35) give failure at most $2e^{-2L}$ for each fixed t. Union bounding over at most HJ_PK prefixes is below $\delta/8$. Importantly, the test is the fixed V^*, not a data-dependent relabeling of past targets.

Proceed chronologically through episodes and backwards through the planning stages of the current episode. At the beginning, TD raw values equal $B_h \ge P_h V_{h+1}^*$. At any later positive count, unroll (61). By chronological induction, each historical target $\overline{V}_{h+1}^j(s_i')$ is at least $V_{h+1}^*(s_i')$. The weighted bonus is at least $2B_h\sqrt{(H+1)L/t}$ because $i \le t$ and the weights sum to one. It dominates the noise in (63). Thus $W^{\mathrm{raw}} \ge P_h V_{h+1}^*$, and clipping at B_h preserves that inequality. The MB branch satisfies $W^{\mathrm{MB}} \ge P_h \overline{V}_{h+1}^k \ge P_h V_{h+1}^*$ by the response confidence event and backward induction. Their minimum is therefore also an upper bound on $P_h V_{h+1}^*$.

Every W is in $[0, B_h]$ even outside the good event. The disturbance confidence event implies

$$\hat{\mu}_i^\top W + B_h \lambda_i \ge \mu_a^\top W \ge \mu_a^\top P_h V_{h+1}^*$$

for both streams. Minimizing over streams, adding the upper reward $\bar{R}_h^k \ge R_h$, and clipping at the valid upper bound $B_h + 1$ preserves $\overline{Q}_h^k \ge Q_h^*$. Maximization over actions completes both inductions. This argument uses no claim that a TD upper response dominates the current $P_h \overline{V}_{h+1}^k$. □

D.4 Complete regret bound and its proof

This theorem establishes the TD and common-policy hybrid extensions announced in Section 4, using the historical-error identities proved above.

Theorem 31 (One-policy TD and hybrid guarantees)**.** *Under the model, reward, and general certificate conditions of Theorem 2, fix* $b \in \{\mathrm{MB}, \mathrm{TD}, \mathrm{HYB}\}$ *before interaction and restart its estimators after certification. Then, with probability at least* $1-\delta$*,*

$$\mathrm{Reg}_b(K) \le \min\{HK, H\tau + F_b(\mathfrak{S})\}. \tag{64}$$

For the hybrid, both defining bounds hold on the same history. With known exact groups there is no certification cost. With unknown exact groups and Assumption 17, the same gate using F_b^A *gives*

$$\mathrm{Reg}_b(K) \le \min\{HK, 2\min\{F_b^A, HAn_\gamma + F_b^m\}\}.$$

This is a per-execution statement, not a simultaneous assertion about separately trained backends.

Proof. Step 1. Common confidence and optimism. Use Lemmas 11, 14, and 30. The latter controls the TD noise at fixed V^* and proves optimism by chronological induction. If $\tau = K$, only the deterministic cap is needed. Otherwise, all sums below are over $k > \tau$, padded with $\mathbf{1}\{\tau < k\}$ for martingale concentration. Write

$$d_h^k = \overline{V}_h^k - V_h^{\pi^k}, \quad \varphi_h^k = \overline{V}_h^k - V_h^*, \quad g_h^k = V_h^* - V_h^{\pi^k}, \quad D_h = \sum_{k>\tau} d_h^k(s_h^k).$$

On optimism, $d_h^k = \varphi_h^k + g_h^k$ with both latter terms nonnegative. On all histories their ranges remain bounded. Let $\Lambda_h = \sum_{k>\tau} B_h \min_i \lambda_{h,i}^k$ and let U_h^R be as in Lemma 14.

Step 2. Model-based comparison on a hybrid history. Pointwise $W^{\mathrm{HYB}} \le W^{\mathrm{MB}}$, so nonnegative empirical probabilities imply that the hybrid action-value candidate is no larger than its model-based candidate on the hybrid's own continuation. The proof of Lemma 16 therefore gives $F_{\mathrm{MB}}(\mathfrak{S})$ on this history as well. No separately run model-based trajectory is invoked.

Step 3. Unroll historical TD targets. Let

$$e_h^k(z,c) = W_h^{\mathrm{TD,k}}(z,c) - (P_h V_{h+1}^*)(z,c).$$

This is in $[-B_h, B_h]$ always and is nonnegative on the optimism event. Suppose the row has t previous visits, in episodes $k_1, \dots, k_t$, with actual successor states $s_1', \dots, s_t'$. Unrolling the *unclipped raw* recursion, using clipping only as an upper bound, and then using (63) and Lemma 29 yield

$$e_h^k(z,c) \le B_h \mathbf{1}\{t=0\} + \sum_{i=1}^{t} \alpha_t^i \varphi_{h+1}^{k_i}(s_i') + 6B_h\sqrt{(H+1)L/(t \vee 1)}\,\mathbf{1}\{t>0\}. \tag{65}$$

Indeed, the weighted bonus is at most $4B_h\sqrt{(H+1)L/t}$ and the noise contributes at most one further copy; the coefficient six is conservative.

Step 4. Count historical reuse. Sum (65) over actual current visits at stage h. There are at most J_P first visits. Each historical nonnegative $\varphi_{h+1}^{k_i}(s_i')$ is charged at most $\sum_{t=i}^{\infty} \alpha_t^i = 1 + 1/H$. Summing inverse square roots within each row and then using Cauchy–Schwarz gives

$$\sum_k e_h^k(z_h^k, c_h^k) \le B_h J_P + \left(1 + \frac{1}{H}\right) \sum_k \varphi_{h+1}^k(s_{h+1}^k) + 12B_h\sqrt{(H+1)LJ_PK}. \tag{66}$$

The inequality remains true for unvisited rows and for $h = H$, where all terms vanish.

Step 5. Propagate errors along the executed policy. For the TD execution, and also for the hybrid because its action value is no larger than its TD candidate, the bounded-stream inequality gives

$$d_h^k(s_h^k) \le (\bar{R}_h^k - R_h)(s_h^k, a_h^k) + 2B_h \min_i \lambda_{h,i}^k + \mathbb{E}[e_h^k(z_h^k, c_h^k) + g_{h+1}^k(s_{h+1}^k) \mid \mathcal{F}_{k,h}^-]. \tag{67}$$

Denote the conditional mean minus the actual quantity in this equation by ζ_h^k. On the entire probability space, the actual quantity lies in $[-B_h, 2B_h]$, so its conditional range length is at most $3B_h$. Combining (66) with (67) and using $\varphi_{h+1}^k \le d_{h+1}^k$ gives, on the good event,

$$D_h \le U_h^R + 2\Lambda_h + B_h J_P + 12B_h\sqrt{(H+1)LJ_PK} + \left(1 + \frac{1}{H}\right) D_{h+1} + \sum_k \zeta_h^k. \tag{68}$$

Apply (68) for $h < H$ and terminate at $D_H \le U_H^R$. Thus the terminal reward is charged with weight $(1+1/H)^{H-1}$, not discarded. Let $\omega_h = (1+1/H)^{h-1} \le e$. Unroll the recursion; the weighted martingale $\sum_{k,h} \omega_h \zeta_h^k$ has range lengths at most $3eB_h$. Equation (35) bounds it by $3e\sqrt{KH_2L/2}$ except on probability e^{-L}. Since $\Lambda_h \ge 0$, Lemma 12 and $\omega_h \le e$ imply

$$\mathrm{Reg}_{\mathrm{post}} \le D_1 \le e\left[2\mathfrak{S} + H_1\{J_P + 12\sqrt{(H+1)LJ_PK}\} + 4\sqrt{2KH_2L} + \mathcal{G}_R\right] = F_{\mathrm{TD}}(\mathfrak{S}).$$

A global martingale bound was applied after assigning the deterministic weights ω_h; this avoids summing stagewise square-root bounds and losing another factor of $\sqrt{H}$.

Step 6. Joint event and the gate. The event budget is

$$\delta/4 + H(2A + ZC)Ke^{-L} + 2\zeta_R HSAKe^{-L} + 2HZCKe^{-2L} + 3e^{-L} < \delta.$$

It includes the two model-based trajectory tails and the weighted TD tail, all proved without conditioning on future optimism. Hence their minimum is valid for one hybrid execution. Add $H\tau$ and the cap HK. Under separation, exact recovery is included in the certificate event. The three scalar gate cases in the proof of Corollary 19 then apply verbatim with $B_A = F_b^A$ and $B_m = F_b^m$, because F_b is nondecreasing in its input-error argument. This proves the second display without an additional probability charge. □

D.5 Statistical and storage implications

These specializations interpret the preceding theorem and the response-storage comparison in Section 4.2. For exact groups with $m_h \le m$, the leading terms are

$$\begin{aligned}\mathrm{Reg}_{\mathrm{MB}} &= \widetilde{O}\Big(H\tau + H^2\sqrt{mCK} + H^2\sqrt{SZCK} + \zeta_R H\sqrt{SAK}\Big),\\ \mathrm{Reg}_{\mathrm{TD}} &= \widetilde{O}\Big(H\tau + H^2\sqrt{mCK} + H^{5/2}\sqrt{ZCK} + \zeta_R H\sqrt{SAK}\Big).\end{aligned} \tag{69}$$

The TD branch removes the successor-alphabet factor from its response estimate but pays for historical-target propagation. It stores $O(HZC)$ response scalars and counters rather than $O(HZCS)$ successor counts. The hybrid retains both data structures, so its statistical comparison does not imply a memory saving. A smaller bound for one branch is not a promise of smaller realized regret than a separately trained algorithm. Appendix H.3 retains the experiments in which the hybrid coincides with MB.

E Proof of Theorem 5

This appendix first proves the fixed-contrast specialization (70), including the original attaining rule used in the numerical studies. Appendix E.6 then proves the response-sensitive family in Theorem 5 and the new product certificate of Lemma 6. This comparison uses known reward formulas; it is a realizable special case of the more general control model. The class $\mathcal{M}_\gamma$ fixes the rewards, maps, supports, and ε in Section 5.3; $p_0, p_1, m_0, m_1 \in (0,1)$ are unknown. Its only structural restriction is $p_0 = p_1$ or $|p_0 - p_1| \ge \gamma$, with zero contrasts excluded. All first actions are available. Set the raw first-stage input to $u = c$ and the initial state to zero. One complete episode supplies one (c, Y) pair, the deterministic zero first reward, and the deterministic terminal reward specified in Section 5.3. For definiteness, the terminal disturbance is an independent $\mathrm{Bern}(1/2)$ variable and the last successor is deterministically zero, identically in every compared model. Thus rewards and terminal observations reveal no additional parameter information conditional on the recorded (c, Y) and actions. This fixes a full two-stage observation law, rather than only its reward means. The task is fixed-policy selection, the cost is complete episodes, and correctness is uniform over the indicated class. The expectation in (70) is evaluated at M_0 only. It is not a worst-instance bound over arbitrarily small nonzero contrasts.

Table 3 summarizes the three information premises at the fixed reference. Expected cost is evaluated at the same M_0, with correctness uniform over the class in each row, as specified in Theorem 5.

Table 3: Information price at the shared fixed-contrast reference.

Sharing premise	Admissible input laws	Optimal order of $\mathbb{E}_{M_0}[T]$
Known input-law equality	$p_0 = p_1$	$\Theta(\log(1/\delta))$
Known separation	$p_0 = p_1$ or $\lvert p_0 - p_1\rvert \ge \gamma > 0$	$\Theta\big(\max\{\varepsilon,\gamma\}^{-2}\log(1/\delta)\big)$
No separation	Arbitrary $p_0, p_1 \in (0,1)$	$\Theta\big(\varepsilon^{-2}\log(1/\delta)\big)$

At this reference, Theorem 5 specializes to

$$\inf_{\mathcal{A}\in\mathcal{A}_\delta(\mathcal{M}_=)} \mathbb{E}_{M_0} T = \Theta(\log(1/\delta)), \qquad \inf_{\mathcal{A}\in\mathcal{A}_\delta(\mathcal{M}_\gamma)} \mathbb{E}_{M_0} T = \Theta(\max\{\varepsilon,\gamma\}^{-2}\log(1/\delta)). \tag{70}$$

E.1 INFORMATION LOWER BOUNDS

This lemma gives the fixed-contrast lower bound underlying Theorem 5 and the reference-instance experiments.

Lemma 32 (Pointwise lower bounds with uniform correctness)**.** *At the fixed-contrast instance M_0, let $\sigma_0 = \max\{\varepsilon, \gamma\}$ and $t_* = \max\{\gamma, \varepsilon/[7(1-\varepsilon)]\}$. Every uniformly δ-correct procedure for $\mathcal{M}_\gamma$ satisfies*

$$\mathbb{E}_{M_0} N_1(T) \geq \mathrm{kl}(1-\delta, \delta)/[-\tfrac{1}{2}\log(1-4t_*^2)], \tag{71}$$

where $N_1(T)$ counts first-action-one episodes, and therefore $\mathbb{E}_{M_0} T \geq c\sigma_0^{-2}\log(1/\delta)$ for an absolute $c > 0$. Every uniformly correct procedure for $\mathcal{M}_=$ satisfies $\mathbb{E}_{M_0} T \geq c\log(1/\delta)$.

Proof. Step 1. Construct admissible decision-changing alternatives. Taking expectations of the specified terminal rewards yields (13). Under M_0, $q_0 = q_1 = 11/20$ and $\theta = \varepsilon/20$. For

$$t_\varepsilon = \frac{\varepsilon}{7(1-\varepsilon)}, \qquad t_* = \max\{\gamma, t_\varepsilon\},$$

we have $0 < t_* \leq 1/7$ when $\gamma \leq 1/7$ and $\varepsilon \leq 1/2$. The alternative M_* changes only p_1 to $1/2 + t_*$. It belongs to $\mathcal{M}_\gamma$, and

$$\theta_* = \frac{\varepsilon}{20} - \frac{7}{10}(1-\varepsilon)t_* \leq -\frac{\varepsilon}{20} < 0.$$

Unlike the particular t_ε alternative alone, this construction remains admissible for a fixed positive γ as ε shrinks. The response kernel is unchanged. With $\sigma_0 = \max\{\varepsilon, \gamma\}$,

$$\sigma_0/7 \leq t_* \leq \sigma_0, \qquad 2t_*^2 \leq -\tfrac{1}{2}\log(1-4t_*^2) \leq \frac{2t_*^2}{1-4/49} = \frac{98}{45}t_*^2. \tag{72}$$

Here $t_\varepsilon \in [\varepsilon/7, 2\varepsilon/7]$, and the logarithm bounds follow from $x \leq -\log(1-x) \leq x/(1-x)$.

Step 2. Use the likelihood of the complete stopped history. For a deterministic episode cap n, the chain rule gives

$$\mathrm{KL}(\mathbb{P}_{M_0}^{T\wedge n}, \mathbb{P}_{M_*}^{T\wedge n}) = \mathbb{E}_{M_0} N_1(T \wedge n)\ \mathrm{kl}(1/2, 1/2 + t_*).$$

The conditional successor, reward, and extra terminal likelihoods cancel, as do the identical action-selection and stopping kernels. Let E_n be the event that the procedure has stopped by n and recommends π. Binary data processing bounds this finite-history KL below by $\mathrm{kl}(\mathbb{P}_{M_0}(E_n), \mathbb{P}_{M_*}(E_n))$. Almost-sure termination under both models gives convergence of these event probabilities to those of the terminal recommendation. Monotone convergence of $N_1(T \wedge n)$ and lower semicontinuity of binary KL yield the stopped lower bound, even if $\mathbb{E}_{M_0} T = \infty$ but $\mathbb{E}_{M_0} N_1(T) < \infty$. Infinite total expected time alone would not justify a bound on this particular action count. The finite-expected-time likelihood equality can alternatively be proved by Fubini; only action-one increments are nonzero, so finite $\mathbb{E}N_1(T)$ suffices for their absolute integrability. This is the complete-observation change-of-measure argument of Kaufmann et al. (2016).

Let E be the event that the procedure recommends π. Uniform δ-correctness implies $\mathbb{P}_{M_0}(E) \geq 1-\delta$ and $\mathbb{P}_{M_*}(E) \leq \delta$. Binary data processing and monotonicity give (71). For $0 < \delta \leq 1/4$, $\mathrm{kl}(1-\delta, \delta) = (1-2\delta)\log((1-\delta)/\delta) \geq \frac{1}{4}\log(1/\delta)$, because $1-2\delta \geq 1/2$ and $1-\delta \geq \sqrt{\delta}$. Since $N_1(T) \leq T$, (72) therefore gives $\mathbb{E}_{M_0} T \geq (45/392)\sigma_0^{-2}\log(1/\delta)$.

For the known-equality lower bound, keep $p_0 = p_1 = 1/2$ and change the response means to $m_0' = 1/10, m_1' = 4/5$. Their common successor mean is $9/20$, so the contrast reverses to $-\varepsilon/20$. Every episode, under either first action, has information

$$d_= \tfrac{1}{2}\mathrm{kl}(1/5, 1/10) + \tfrac{1}{2}\mathrm{kl}(9/10, 4/5) \in (0, \infty).$$

Consequently $\mathbb{E}_{M_0} T \geq \mathrm{kl}(1-\delta, \delta)/d_=$. This proves the $\Omega(\log(1/\delta))$ lower bound even when the equality is known; this particular alternative charges response learning. It does not establish that unknown responses are necessary for the order law; Corollary 35 removes that uncertainty.

□

E.2 Correctness of the sequential rule

Our *two-route stopping* rule resolves the decision directly or certifies equality first, as appropriate for the classes in Table 3. Let $\sigma_0 = \max\{\varepsilon, \gamma\}$, $b_0 = \lceil \sigma_0^{-2} \rceil$, and $b_j = b_0 2^j$. At look j, reach *cumulative* count b_j per first action by collecting only $b_j - b_{j-1}$ new episodes per action, with $b_{-1} = 0$. Define signature means $\hat{p}_a$, successor means $\hat{q}_a$, pooled mean $\bar{Y} = (\hat{q}_0 + \hat{q}_1)/2$, and

$$r_j = \sqrt{\frac{\log[16(j+1)^2/\delta]}{2b_j}}, \qquad I_= = [\varepsilon(\bar{Y} - 1/2) \ \pm \varepsilon r_j],$$
$$I_{\mathrm{dir}} = [\hat{q}_0 - (1-\varepsilon)\hat{q}_1 - \varepsilon/2 \ \pm (2-\varepsilon) r_j]. \tag{73}$$

Return the sign of I_{dir} when it excludes zero. Otherwise stop with the sign of $I_=$ only if $|\hat{p}_0 - \hat{p}_1| + 2r_j < \gamma$ and $I_=$ excludes zero. With known input-law equality, start at $b_0 = 1$ and use $I_=$ alone. The direct route can finish without identifying a group, while the tied route cancels common error. The $(j+1)^2$ allocation controls repeated looks. Fast stopping on that event alone does not bound $\mathbb{E}T$; Appendix E.3 controls the unconditional tail.

Lemma 33 (Uniform correctness and almost-sure termination)**.** *The two-branch rule in this subsection, with cumulative per-action counts b_j, is δ-correct over $\mathcal{M}_\gamma$ and stops almost surely at every model in that class. With known input-law equality, its tied-interval version has the same properties over $\mathcal{M}_=$.*

Proof. At deterministic look j, each first action has been collected b_j times cumulatively. Only $b_j - b_{j-1}$ additional samples are drawn per action, so the stopping cost is $2b_j$, not $2\sum_{i \le j} b_i$. The four bounded means $\hat{p}_0, \hat{p}_1, \hat{q}_0, \hat{q}_1$ are individually averages of independent and identically distributed variables. For this upper construction, fix an alternating schedule of the first actions. Define the infinite observation streams even after the procedure stops, solely for the proof. Every deterministic look therefore has well-defined cumulative means. The signature and successor within an episode need not be independent. Hoeffding's inequality and a union bound imply

$$\mathbb{P}(\exists j, a : \ |\hat{p}_a - p_a| > r_j \text{ or } |\hat{q}_a - q_a| > r_j) \le \sum_{j \ge 0} 8e^{-2b_j r_j^2} = \frac{\delta}{2} \sum_{j \ge 0} (j+1)^{-2} < \delta.$$

On the complementary simultaneous event, I_{dir} always covers (13). If $|\hat{p}_0 - \hat{p}_1| + 2r_j < \gamma$, then $|p_0 - p_1| < \gamma$. The declared class forces equality. Under that equality, $q_0 = q_1 = q$ and $|\bar{Y} - q| \le r_j$, so $I_=$ covers θ as well. Hence either stopping branch has the correct sign. If both apply, the direct branch has priority on every history, including failure histories on which the intervals could disagree. At $\gamma = 0$ the strict equality test can never hold. Known input-law equality allows $I_=$ without the empirical test. All statements hold simultaneously over looks; no unaccounted post-selection or optional-stopping step remains.

Since $r_j \to 0$ and the empirical successor means converge almost surely, I_{dir} eventually excludes zero at every $\theta \neq 0$. The known-tie rule also stops almost surely because its class excludes $q = 1/2$. These properties prove uniform δ-correctness over both classes. The rule does not require the user to supply a sign gap. The expected-time calculation below is at M_0, whose gap is fixed in terms of ε by the construction.

□

E.3 Expected sample cost at the shared instance

The following tail argument controls expected sample cost without conditioning away confidence failures, completing the upper-bound component of the fixed-contrast comparison.

Lemma 34 (Unconditional stopping-time bound)**.** *For the rules in Lemma 33, at M_0 and uniformly over the parameter ranges in Theorem 5,*

$$\mathbb{E}_{M_0} T \le 2{,}000{,}000 \max\{\varepsilon, \gamma\}^{-2} \log(1/\delta)$$

for the two-branch rule. The same bound without the inverse-square factor holds for the known-equality rule. These are conservative absolute constants, not optimized stopping thresholds.

Proof. Let $\sigma_0 = \max\{\varepsilon, \gamma\}$, $b_0 = \lceil \sigma_0^{-2} \rceil$, $b_j = b_0 2^j$, and define

$$J_\delta = \min\{j \geq 0 : \ 2^j \geq 12800 \log[16(j+1)^2/\delta]\}. \tag{74}$$

For every $j \geq J_\delta$, $r_j \leq \sigma_0/160$. The ratio of the logarithm to 2^j decreases from this index onward. Suppose also that all four empirical means at that look are within $\sigma_0/160$ of their true means.

If $\sigma_0 = \varepsilon \geq \gamma$, the direct estimation error is at most $(2-\varepsilon)\sigma_0/160 \leq \varepsilon/80$, and its radius is at most $\varepsilon/80$. Their sum is at most $\varepsilon/40 < \theta = \varepsilon/20$, so the direct branch stops. If $\sigma_0 = \gamma \geq \varepsilon$, the empirical signature difference plus $2r_j$ is at most $\gamma/40 < \gamma$. The equality is certified. The pooled successor error plus r_j is at most $\gamma/80 \leq 1/560 < 1/20 = q - 1/2$, so the tied interval excludes zero. Either way the procedure has stopped by this look. Therefore, at M_0,

$$\mathbb{P}(T > 2b_j) \leq 8 \exp(-b_j \sigma_0^2/12800), \qquad j \geq J_\delta. \tag{75}$$

For dyadic b_j, the intervals $[b_j/2, b_j]$ have disjoint interiors and $b_j e^{-a b_j} \leq 2 \int_{b_j/2}^{b_j} e^{-ax}\, dx$. Summing the tail of the discrete stopping time gives the explicit bound

$$\begin{aligned} \mathbb{E}_{M_0} T &= 2b_0 + \sum_{j \geq 0} 2 b_j \mathbb{P}(T > 2b_j) \\ &\leq 2b_0 2^{J_\delta} + 16 \sum_{j \geq J_\delta} b_j e^{-b_j \sigma_0^2/12800} \leq 2b_0 2^{J_\delta} + 409600 \sigma_0^{-2}. \end{aligned} \tag{76}$$

The absolute constant can be made explicit. Put $\ell = \log(16/\delta) \geq \log 64 > 4$ and $j_\delta = \lceil \log_2(2^{16}\ell) \rceil$. Then $j_\delta + 1 \leq 18 + \log_2 \ell \leq 18 + \ell < e^\ell$, so $\log[16(j_\delta+1)^2/\delta] \leq 3\ell$. Consequently $2^{j_\delta} \geq 2^{16}\ell > 12800(3\ell)$ and $J_\delta \leq j_\delta$. Thus $2^{J_\delta} < 2^{17}\ell \leq 393216 \log(1/\delta)$. Since $b_0 \leq 2\sigma_0^{-2}$, (76) gives the explicit, deliberately conservative bound

$$\mathbb{E}_{M_0} T \leq 2{,}000{,}000\, \sigma_0^{-2} \log(1/\delta).$$

This proves the required upper order uniformly over all stated parameter ranges. The class-dependent initial look is what avoids an extra $\log\log(1/\sigma_0)$ from starting at one sample for every margin. The constants in this proof are conservative and are not used to set the practical stopping threshold.

For known input-law equality use $b_0 = 1$. Once $r_j \leq 1/160$ and both successor means are within $1/160$ of $11/20$, the tied interval excludes zero. The identical dyadic tail argument, with $\sigma_0 = 1$, yields $\mathbb{E}_{M_0} T = O(\log(1/\delta))$. The same upper constant 2,000,000 suffices with $\sigma_0 = 1$. This proves the stated expected-episode bound. The constants are analytical bounds, not optimized constants or measured stopping costs.

□

E.4 Completion of the fixed-contrast specialization

Combining the preceding three lemmas yields the specialization of Theorem 5 used by the numerical studies.

Proof of the specialization (70). Lemma 32 applies to every uniformly correct procedure, including adaptive collection rules. Lemma 33 supplies a member of each required correctness class, and Lemma 34 bounds its expected episode count at the same M_0. Taking the infimum over each class proves both matching orders. The expectation is pointwise at M_0, whereas correctness is uniform over the stated class; no worst-instance or unrestricted-control conclusion is used. □

E.5 Known responses and the role of the successor gap

This specialization separates uncertain sharing from response estimation by making the response probabilities known while retaining the same fixed-policy decision task.

Corollary 35 (The equality price persists with known responses)**.** *Let $\mathcal{M}_\gamma^P$ and $\mathcal{M}_=^P$ restrict the classes of Theorem 5 to $m_0 = 1/5, m_1 = 9/10$ and supply these two probabilities to the learner. Only p_0, p_1 are unknown. At the same M_0, uniformly over the theorem's $(\varepsilon, \gamma, \delta)$ ranges,*

$$\inf_{\mathcal{A} \in \mathcal{A}_\delta(\mathcal{M}_=^P)} \mathbb{E}_{M_0} T = \Theta(\log(1/\delta)), \qquad \inf_{\mathcal{A} \in \mathcal{A}_\delta(\mathcal{M}_\gamma^P)} \mathbb{E}_{M_0} T = \Theta(\max\{\varepsilon, \gamma\}^{-2} \log(1/\delta)).$$

Proof. The upper bounds follow by ignoring the known response parameters and using the already proved rules. For $\mathcal{M}_{\gamma}^{P}$, the sign-changing alternative $p_1 = 1/2 + t_*$ retains exactly these response probabilities, so (71) holds unchanged. For $\mathcal{M}_{=}^{P}$, change both disturbance probabilities to $p_0' = p_1' = 5/14$. Then $q' = 1/5 + (7/10)(5/14) = 9/20$, and the contrast changes from $\varepsilon/20$ to $-\varepsilon/20$. Every episode contributes $\mathrm{kl}(1/2, 5/14) \in (0, \infty)$ information, independent of the first action, because the conditional response cancels from the likelihood ratio. Stopped change of measure gives $\mathbb{E}_{M_0} T \geq \mathrm{kl}(1-\delta, \delta)/\,\mathrm{kl}(1/2, 5/14)$. This establishes the remaining lower bound. □

This specialization isolates the structural difficulty: one must rule out a disturbance discrepancy that destroys a cancellation. In the control model, unknown response rows still incur the explicit costs in Theorem 2 and the response-only obstruction in Appendix F. These are distinct statements.

The response-sensitive reference family below quantifies a varying common successor gap. Its pointwise constants are uniform only over the stated interior reference ranges, not every off-reference model. The fixed-contrast rule above remains the rule used by the historical experiments.

E.6 Response-sensitive information geometry and the second-order regime

This subsection proves the reference-family extension in Theorem 5. The ambient correctness classes $\mathcal{M}_{=}$ and $\mathcal{M}_{\gamma}$ are unchanged: their response probabilities are unknown and unrestricted in $(0, 1)$. Only the instance at which expected cost is evaluated varies. In particular, a small or zero response contrast is not an additional prior promise supplied to the learner.

E.6.1 Reference family, effective scale, and proof dependencies

This argument proves the lower bounds in Theorem 5 over the response-sensitive reference family, with explicit constants. For $0 < \Delta \leq 1/20$ and $0 \leq \kappa \leq 7/10$, define $M_{\Delta,\kappa}$ by

$$p_0 = p_1 = 1/2, \quad m_0 = 1/2 + \Delta - \kappa/2, \quad m_1 = 1/2 + \Delta + \kappa/2.$$

The common successor mean is $1/2 + \Delta$, the conditional-response contrast is κ, and the policy gap is $a = \varepsilon\Delta$. Set

$$r_* = \frac{a}{\kappa + \sqrt{a}}, \qquad s_* = \min\{\Delta, \max\{\gamma, r_*\}\}, \qquad \mathcal{L}(s, \delta) = \log(32/\delta) + 2\log(2 + \log_2(s^{-2})). \tag{77}$$

All scales are strictly positive except the allowed $\gamma = 0$ and $\kappa = 0$. The resolution r_* solves the first-order/second-order balance up to absolute constants: $\kappa r_* + r_*^2 \leq a$, while $\kappa(6r_*) + (6r_*)^2 \geq 4a$. We will prove

$$\begin{aligned} 2^{-12}\Delta^{-2}\log(1/\delta) &\leq \inf_{\mathcal{A} \in \mathcal{A}_\delta(\mathcal{M}_=)} \mathbb{E}_{M_{\Delta,\kappa}} T \leq 2^{32}\Delta^{-2}\log(1/\delta), \\ 2^{-12}s_*^{-2}\log(1/\delta) &\leq \inf_{\mathcal{A} \in \mathcal{A}_\delta(\mathcal{M}_\gamma)} \mathbb{E}_{M_{\Delta,\kappa}} T \leq 2^{32}s_*^{-2}\log(1/\delta). \end{aligned} \tag{78}$$

The constants are deliberately conservative, not algorithm tuning constants. Lemma 36 proves the lower bounds, Lemma 37 gives a uniformly correct rule, and Lemma 38 proves its expected-time bounds, including failure histories.

Lemma 36 (Response and bilinear decision-changing alternatives)**.** *At $M_{\Delta,\kappa}$, let $a = \varepsilon\Delta$ and*

$$t_0 = \frac{\sqrt{\kappa^2 + 8a/(1-\varepsilon)} - \kappa}{2}, \quad t = \max\{\gamma, t_0\}, \quad u = \max\{0, 2a/[(1-\varepsilon)t] - \kappa\}.$$

Every uniformly δ-correct rule over $\mathcal{M}_{\gamma}$ obeys

$$\mathbb{E}_{M_{\Delta,\kappa}} T \geq \frac{\mathrm{kl}(1-\delta, \delta)}{\min\{45\Delta^2,\ 648\max\{\gamma, r_*\}^2\}}. \tag{79}$$

Over $\mathcal{M}_{=}$ the same claim holds with denominator $45\Delta^2$.

Proof. Step 1: an alternative inside the equality submodel. Keep both input probabilities equal to $1/2$ and lower both response probabilities by 2Δ. They remain in $[1/10, 9/10]$ and change the policy gap

from a to $-a$. For Bernoulli parameters, $\mathrm{kl}(x, y) \le (x-y)^2/[y(1-y)]$, obtained from $\log z \le z-1$. Thus every episode has KL at most $4\Delta^2/(9/100) < 45\Delta^2$. This alternative is admissible in both correctness classes, including when $\gamma > 0$.

Step 2: a margin-respecting bilinear alternative. Change p_1 to $1/2 + t$ and change the response means to $m_0' = m_0 - u/2$, $m_1' = m_1 + u/2$; retain $p_0 = 1/2$. Their common midpoint is unchanged, while the response contrast becomes $\kappa + u$. Because $t_0(\kappa + t_0) = 2a/(1-\varepsilon)$, $0 \le u \le t_0 \le t$, and

$$\theta' = a - (1-\varepsilon)t(\kappa + u) \le -a.$$

This alternative is in $\mathcal{M}_\gamma$. To check support without asymptotics, $t_0^2 \le 2a/(1-\varepsilon) \le 1/10$, so $t \le \sqrt{1/10} < 1/2$. Moreover $\kappa + u \le \kappa + t_0 \le (7/10 + \sqrt{89/100})/2 < 17/20$. The altered response probabilities therefore lie in $[3/40, 39/40]$. No support singularity is introduced.

Step 3: information in the full feedback. Under either first action the reference signature is Bernoulli$(1/2)$. The per-episode joint KL splits into the input KL and the reference-signature average of the two response KLs. Action zero has no input term, while for action one

$$\mathrm{kl}(1/2, 1/2 + t) = -\tfrac{1}{2}\log(1 - 4t^2) \le 7t^2.$$

Since $m_c'(1 - m_c') \ge 39/1600$, the response contribution is at most $\frac{1}{2}\sum_c (u/2)^2/[m_c'(1-m_c')] \le 11u^2$. Thus every action yields at most $18t^2$ information per episode. With $x = \kappa/\sqrt{a}$, the equation for t_0 gives $r_* \le t_0 \le 6r_*$: substitution of $r_*/\sqrt{a} = 1/(x+1)$ gives a quadratic value at most 1, whereas substitution of $6/(x+1)$ gives at least 4; the true right side $2/(1-\varepsilon)$ is in $[2, 4]$. Consequently $18t^2 \le 648 \max\{\gamma, r_*\}^2$.

Step 4: stopped information and the two obstructions. Apply the complete-feedback stopped likelihood argument in Appendix E.1 to each alternative separately. Action and stopping kernels cancel; the conditional observation KL is bounded for every chosen action. For finite expected stopping time, bounded log-likelihood increments justify Fubini; infinite expectation already satisfies the result. Binary data processing for the recommendation yields both $\mathbb{E}T \ge \mathrm{kl}(1-\delta, \delta)/(45\Delta^2)$ and $\mathbb{E}T \ge \mathrm{kl}(1-\delta, \delta)/(648 \max\{\gamma, r_*\}^2)$. Their maximum is (79). Using $\mathrm{kl}(1-\delta, \delta) \ge \frac{1}{4}\log(1/\delta)$ proves the lower half of (78). □

The new alternative changes *both* factors. At $\kappa = 0$, changing only an input law leaves the policy gap unchanged, so the original same-response construction cannot establish the correct local rate. Balancing two perturbations of order $\sqrt{\varepsilon\Delta}$ costs order $\varepsilon\Delta$ in KL, rather than order $(\varepsilon\Delta)^2$.

E.6.2 Product-aware sequential certification

This construction implements Lemma 6 and supplies the uniformly correct procedure used in Theorem 5. Fix any integer initial resolution $n_0 \ge 1$ before observing data. Alternate the first-stage actions, and at look j reach cumulative count $n_j = 2^j n_0$ per action. Let $N_{c,j}$ be the number of occurrences of category c across these $2n_j$ episodes, and let $\hat{m}_c$ be the corresponding successor mean, initialized to $1/2$ if its count is zero. Write $\hat{p}_a$ for the action-specific category means, $\bar{Y}$ for the mean of all successors, and

$$\ell_j = \log(32(j+1)^2/\delta), \qquad r_{p,j} = \min\{1, \sqrt{\ell_j/(2n_j)}\},$$
$$r_{m,j} = \max_c \min\{1, \sqrt{\ell_j/(2(N_{c,j} \vee 1))}\},$$

where the radius for an empty response row is explicitly one. Let $r_j = \max\{r_{p,j}, r_{m,j}\}$, $\hat{d} = \hat{p}_1 - \hat{p}_0$, $\hat{\kappa} = \hat{m}_1 - \hat{m}_0$, $\hat{q}_0 = (1-\hat{p}_0)\hat{m}_0 + \hat{p}_0\hat{m}_1$, and

$$\hat{\theta}_j = \varepsilon(\hat{q}_0 - 1/2) - (1-\varepsilon)\hat{d}\hat{\kappa},$$
$$B_j = \varepsilon r_j + (2-\varepsilon)|\hat{\kappa}| r_j + 2(1-\varepsilon)|\hat{d}| r_j + (4-2\varepsilon) r_j^2. \tag{80}$$

Stop with the sign of $[\hat{\theta}_j \pm B_j]$ if it excludes zero. Otherwise, if $|\hat{d}| + 2r_{p,j} < \gamma$, stop with the sign of $[\varepsilon(\bar{Y} - 1/2) \pm \varepsilon r_{p,j}]$ when that interval excludes zero. If equality is known a priori, use only this latter interval without the empirical equality test. The priority rule defines a unique output even on confidence-failure histories.

Lemma 37 (Uniform product certificate)**.** *For every fixed n_0, this rule is uniformly δ-correct and almost surely terminating on $\mathcal{M}_\gamma$. The known-equality version has the same properties on $\mathcal{M}_=$. It needs $\varepsilon, \gamma, \delta, n_0$, but not Δ, κ, or a promised policy gap.*

Proof. Step 1: simultaneous means with random response counts. Use infinite alternating-action streams as a proof construction, even after stopping. At each deterministic look, Hoeffding bounds the two category means and the pooled successor mean around their respective expectations. Conditional on the entire finite category sequence, the observations in a response row are independent Bernoulli(m_c): the joint feedback law factors into the action-dependent category law and the common conditional-response law. This conditional argument is valid for this fixed collection schedule; it is not an assertion that conditioning on arbitrary adaptive counts preserves independence in the control algorithm. It gives the two response-mean bounds at their random $N_{c,j}$, with deterministic coverage at zero count. The union failure over five means and all looks is at most $\sum_{j\geq 0} 10e^{-\ell_j} = (10\delta/32)\sum_{j\geq 0}(j+1)^{-2} < \delta$.

Step 2: keep the cross term. On this event, $|d - \hat{d}| \leq 2r_j$, $|\kappa - \hat{\kappa}| \leq 2r_j$, and

$$|q_0 - \hat{q}_0| \leq r_j + |\hat{\kappa}|r_j + 2r_j^2, \quad |d\kappa - \hat{d}\hat{\kappa}| \leq 2|\hat{\kappa}|r_j + 2|\hat{d}|r_j + 4r_j^2.$$

The first inequality uses the convex combination with empirical $\hat{p}_0 \in [0,1]$; the second follows by expanding both factor errors. Combining them gives $|\theta - \hat{\theta}_j| \leq B_j$. No independence of the factor estimates is used. In particular, the r_j^2 term cannot be deleted when both first derivatives vanish. If equality is certified, the declared class forces $d = 0$, and the pooled successor interval covers its common mean. Either stopping branch therefore gives the correct sign.

Step 3: termination. Because $p_0, p_1 \in (0,1)$, both response rows are visited infinitely often under the alternating schedule. All radii tend to zero almost surely, and all empirical means converge. The product interval therefore eventually excludes zero whenever $\theta \neq 0$. Under known equality the pooled interval does so whenever $q \neq 1/2$. Every threshold and statistic is history-measurable, so the stopping rule is a stopping time. □

E.6.3 Expected time, pointwise optimization, and adaptation

This final step proves the initialization-dependent sample-cost statements in Theorem 5, while keeping uniform correctness distinct from pointwise evaluation.

Lemma 38 (Unconditional local sample cost)**.** *At $M_{\Delta,\kappa}$, the product-aware rule with $n_0 = \lceil s_*^{-2} \rceil$ satisfies $\mathbb{E}T \leq 2^{27} s_*^{-2} \log(32/\delta)$. The single rule with $n_0 = 1$ instead satisfies*

$$\mathbb{E}T \leq 2^{28} s_*^{-2} \mathcal{L}(s_*, \delta). \tag{81}$$

For known equality replace s_* by Δ in both displays.

Proof. Set $s = s_*$ and $e = s/64$. At a deterministic look with n episodes per action, suppose that (i) each category count is at least $n/2$, (ii) the five empirical means differ from their reference expectations by at most e, and (iii) $\sqrt{\ell_j/n} \leq e$. Then $r_j \leq e$, $|\hat{d}| \leq 2e$, and $|\hat{\kappa}| \leq \kappa + 2e$.

If $r_* > \gamma$, we have $s \leq \Delta$, $\kappa s \leq a$, and $s^2 \leq a$. Thus both the computed radius and the actual plug-in error are at most $\varepsilon e + 2\kappa e + 12e^2 < a/16$. The product interval is strictly positive. If $\gamma \geq r_*$, then $s \leq \gamma, \Delta$, so $|\hat{d}| + 2r_{p,j} \leq 4e < \gamma$ and the tied interval is strictly positive. The known-equality case uses the same argument with $s = \Delta$.

At the reference, the total category count is binomial with $2n$ trials and probability $1/2$. Hoeffding gives failure probability at most $2e^{-n/4}$ for (i). Use unconditional Hoeffding for the two input means and the pooled successor mean. Conditional on each complete category sequence satisfying (i), the two response means have total failure at most $4e^{-ne^2}$. Therefore $\mathbb{P}(\text{(ii) fails}, \text{(i) holds}) \leq 4e^{-2ne^2} + 2e^{-4ne^2} + 4e^{-ne^2} \leq 10e^{-ne^2}$. This does not condition the input-mean tail on a selected count event. Hence, whenever (iii) holds,

$$\mathbb{P}_{M_{\Delta,\kappa}}(T > 2n_j) \leq 12\exp(-n_j s^2/4096). \tag{82}$$

This is unconditional; it includes earlier confidence failures.

To make the expected-time constants explicit, let $L_0 = \log(32/\delta)$. For $n_0 = \lceil s^{-2} \rceil$, choose the first dyadic count $n_{j_*} \geq 2^{20} n_0 L_0$. Then $n_{j_*} < 2^{21} n_0 L_0$ and $j_* + 1 \leq 22 + \log_2 L_0$. For $L_0 \geq \log 128$, elementary differentiation gives $L_0 + 2\log(22 + \log_2 L_0) \leq 3L_0$. Thus (iii) holds at j_* and at every later look, since $\ell_j / 2^j$ decreases. For $n_0 = 1$, set $M = \log_2(s^{-2})$ and $A = L_0 + 2\log(2 + M)$. Choose the first dyadic count at least $2^{20} s^{-2} A$. Its index satisfies $j_* + 1 \leq 22 + M + \log_2 A$, and

$$\ell_{j_*} \leq L_0 + 2\log[(M+2)(16+A)] = A + 2\log(16 + A) \leq 4A.$$

Here $\log_2 A \leq A$ and $A \geq \log 128$ justify the first and last elementary bounds. Again (iii) holds thereafter.

Finally, writing $n_* = n_{j_*}$, the dyadic tail sum gives

$$\mathbb{E}T \leq 2n_* + 24 \sum_{j \geq 0} n_* 2^j e^{-(n_* s^2/4096) 2^j} \leq 26 n_*.$$

For the last step the leading exponent is at least two and $\sum_{j \geq 0} 2^j e^{-a 2^j} \leq \sum_{j \geq 0} e^{-2^j} \leq 1$ for $a \geq 2$. Substitution of n_* and $n_0 \leq 2s^{-2}$ proves the stated conservative constants. □

Completion of Theorem 5 for the reference family. The lower bounds follow from Lemma 36. Every fixed initial resolution n_0 defines a rule that is correct on the *entire* indicated class by Lemma 37. Therefore the pointwise infimum may select $n_0 = \lceil s_*^{-2} \rceil$ at the reference; this is algorithm selection in a pointwise optimum, not prior knowledge required for correctness. Lemma 38 and $\log(32/\delta) \leq (7/2)\log(1/\delta)$ give (78). A single scale-agnostic rule is obtained by $n_0 = 1$, with (81); we do not erase its iterated logarithm. The distinction between a pointwise infimum and simultaneous adaptation is familiar in sequential identification (Jamieson et al., 2014), but the effective resolution here comes from the two-factor observation model.

At $\Delta = 1/20$, $\kappa = 7/10$, $r_* \asymp \varepsilon$ and Δ is fixed, so $s_* \asymp \max\{\varepsilon, \gamma\}$ throughout the original parameter range (with absolute constants). This recovers (70). The simpler original rule in Appendices E.2–E.3 still attains that specialization without knowing a varying reference scale. □

For $\gamma = 0$, the inverse effective scale satisfies

$$s_*^{-2} = \max\left\{ \Delta^{-2}, \frac{\kappa^2 + 2\kappa\sqrt{\varepsilon\Delta} + \varepsilon\Delta}{\varepsilon^2 \Delta^2} \right\} \asymp \max\{\Delta^{-2}, \kappa^2/(\varepsilon^2\Delta^2) + 1/(\varepsilon\Delta)\}.$$

The cross term is retained in the exact expression and bounded by the other two terms for the order statement. At $\kappa = 0$ and fixed Δ, the small-ε price is ε^{-1}, not ε^{-2}. At fixed positive κ the first-order price returns. This is a property of the same ambient model and full feedback, not of changing the observations or declaring the response contrast known. The historical selection experiments concern only M_0 and the original two-route rule; they do not constitute an empirical evaluation of this new product-aware rule.

F Additional limits of experience sharing

These constructions explain why the certification and conditional-response costs in Theorem 2 cannot generally be discarded. The matching fixed-menu result has its own proof in Appendix E.

A sharing bound is meaningful only if it retains uncertainty that pooling cannot remove. The constructions below separate easy disturbance learning from hard response learning and identify the corresponding information costs. The following construction separates knowing that sharing is valid from having to remain reliable when it is not. It already has a known response and $H = 2$; it is not caused by difficult long-horizon planning.

Theorem 39 (Regret on a fully shared instance)**.** *Let $A \geq 2$, $0 < \Delta \leq 1/8$. There are instances $M_0, M_1, \ldots, M_{A-1}$ of* (1) *with known response and $C = 2$ such that M_0 has a single common disturbance law and a sharing-aware oracle has zero regret. For any algorithm, put $\mathcal{R} = \max_{1 \leq j < A} \mathbb{E}_j \operatorname{Reg}(K) > 0$. Then*

$$\mathbb{E}_0 \operatorname{Reg}(K) \geq \min\left\{ \frac{\Delta K}{4},\ \frac{A-1}{16\Delta} \left[\frac{1}{2} \log \frac{\Delta K}{2\mathcal{R}} - \log 2 \right]_+ \right\}. \tag{83}$$

Under M_0, all success probabilities are $1/2$, action 0 has first-stage reward 1, and every other action has reward $1-\Delta$. Under M_j, only action j changes, to success probability $1/2+2\Delta$. A terminal success earns one. Observations from other actions cannot distinguish these instances. When $\mathcal{R}=O(K^\alpha)$ with $\alpha<1$ and fixed A,Δ, (83) gives an $\Omega((A-1)\log K/\Delta)$ easy-instance cost. This lower bound does not require the algorithm to output a partition.

Proposition 40 (Certification and response obstructions)**.** *For Bernoulli disturbance laws $1/2$ versus $1/2+\gamma$, $0<\gamma\le 1/4$, any sequential procedure that identifies the full sharing partition with error at most $\delta<1/2$ on the all-equal instance and every single-action alternative must satisfy*

$$\mathbb{E}_0 N_j \ge \frac{\mathrm{kl}(1-\delta,\delta)}{4\gamma^2} \quad (1\le j<A). \tag{84}$$

Consequently, uniform exact certification over arbitrarily small nonzero separations cannot have finite expected sampling cost at equality. Separately, for $K\ge A\ge 2$, even with one deterministic signature law shared by all actions, an unknown response can force

$$\mathbb{E}\,\mathrm{Reg}(K) \ge \frac{H-1}{32}\sqrt{AK}. \tag{85}$$

The first statement concerns certification procedures; it is not an impossibility theorem for all useful approximate sharing. The second is a bandit embedded in the conditional response. These two statements rule out deleting identification or response costs solely because the number of true disturbance groups is small. Proofs use the usual adaptive likelihood-ratio chain rule (Kaufmann et al., 2016), with the complete constructions given in Appendix F.

F.1 Factored information on observed histories

The input and response counts in Lemma 4 also have an information-theoretic meaning. The following likelihood calculation applies to finite-feedback subfamilies of our model; it does not require access to counterfactual samples. We state explicitly when the other observed marks can be ignored.

Lemma 41 (Complete-feedback factor accounting)**.** *Let two models have the same known maps and common conditional laws for additional observed marks, including rewards, given (h,s,a,c,s'). For simplicity take raw input $u=c$; a shared conditional raw-input law gives the same conclusion. Put $D^{\mu}_{h,a}=\mathrm{KL}(\mu_{h,a},\mu'_{h,a})$ and $D^{P}_{h,z,c}=\mathrm{KL}(P_h(\cdot\mid z,c),P'_h(\cdot\mid z,c))$. For any adaptive learner and deterministic episode budget n,*

$$\mathrm{KL}(\mathbb{P}^{\mathrm{hist}}_M,\mathbb{P}^{\mathrm{hist}}_{M'})=\sum_{h,a}\mathbb{E}_M N^{\mu}_{h,a}(n)D^{\mu}_{h,a}+\sum_{h,z,c}\mathbb{E}_M N^{P}_{h,z,c}(n)D^{P}_{h,z,c}, \tag{86}$$

where the two counts record actual action inputs and actual conditional-response visits. If a procedure stops at an episode end almost surely at T in both models and its recommendation is correct with error at most $\delta<1/2$ for opposite answers, then the right side of (86) *at T is at least* $\mathrm{kl}(1-\delta,\delta)$. *This conclusion does not require $\mathbb{E}_M T<\infty$.*

Proof. Factor each full step law as $\mu_{h,a}(c)P_h(s'\mid\phi_h(s,a),c)$ times the common conditional mark law. For each realized history, the algorithm's action and stopping probabilities are the same functions in both environments and cancel from the likelihood ratio, as do the common marks. Conditional expectation of the input log ratio gives $D^{\mu}_{h,a}$; conditioning next on c gives $D^{P}_{h,z,c}$. Summing proves (86). This is the adaptive likelihood chain rule underlying bandit change of measure (Kaufmann et al., 2016).

For an unbounded stopping time, apply the finite calculation to $T\wedge n$ and the event E_n that the procedure has stopped by n and recommends the answer correct under M. Data processing lower-bounds the capped divergence by $\mathrm{kl}(\mathbb{P}_M(E_n),\mathbb{P}_{M'}(E_n))$. Almost-sure stopping gives limits at least $1-\delta$ and at most δ. Monotone convergence of nonnegative accumulated conditional divergences and lower semicontinuity of binary KL prove the claim, including infinite information. If only one action j differs, this reduces to

$$\mathrm{KL}(\mathbb{P}^{\mathrm{hist}}_0,\mathbb{P}^{\mathrm{hist}}_1)=\mathbb{E}_0 N_j\,\mathrm{KL}(p_0,p_1) \tag{87}$$

at finite horizons, with the same stopped-decision lower-bound consequence. Natural logarithms and continuous boundary conventions are used throughout. □

A category observation contributes information about its action's input law at any state; response information is acquired only at the visited (h, z, c). This is the information counterpart of actual-visit accounting, not a matching characterization of controlled exploration. Marginalizing c can strictly decrease KL, even while preserving policy values. If raw inputs or rewards have *different* conditional mark laws, their nonnegative conditional KL contributions must be added. Omitting them cannot upper-bound the information available to a learner. Our lower-bound constructions fix these marks precisely to exclude such extra channels.

F.2 Proof of Theorem 39

Proof. Take raw input $u = c$, deterministic rewards, and parameter-independent additional terminal observations. The first-stage state is fixed. Taking action a gives a known first reward $r_0 = 1$ if $a = 0$ and $r_a = 1 - \Delta$ otherwise. A binary signature is drawn with success probability p_a and becomes the second-stage state deterministically. All second-stage actions give reward equal to that binary state and reveal no additional informative randomness. This is (1) with a known identity response. All first-stage actions remain available and second-stage actions are interchangeable.

Under M_0, set $p_a = 1/2$ for every a. Under M_j, change only p_j to $1/2 + 2\Delta$. The null's optimal action is zero and its nonzero actions have gap Δ. Under M_j, action j is optimal, action zero has gap Δ, and every other nonzero action has gap 2Δ. Let N_j count first-stage choices of j over K episodes. Thus

$$\mathbb{E}_0 \operatorname{Reg}(K) = \Delta \sum_{j=1}^{A-1} \mathbb{E}_0 N_j, \qquad \mathbb{E}_j \operatorname{Reg}(K) \geq \Delta \mathbb{E}_j(K - N_j).$$

If the left-hand side exceeds $\Delta K/4$, the theorem follows. Otherwise each $\mathbb{E}_0 N_j \leq K/4$. For $E_j = \{N_j < K/2\}$, Markov's inequality gives

$$\mathbb{P}_0(E_j) \geq \tfrac{1}{2}, \qquad \mathbb{P}_j(E_j) \leq \frac{2\mathcal{R}}{\Delta K}.$$

If the positive part in (83) is zero, there is nothing further to prove. Otherwise the latter upper bound is below one. Using $\operatorname{kl}(p, q) \geq p \log(1/q) - \log 2$ and (87),

$$\mathbb{E}_0 N_j \ \operatorname{kl}(1/2, 1/2 + 2\Delta) \geq \tfrac{1}{2} \log \frac{\Delta K}{2\mathcal{R}} - \log 2.$$

Since $\Delta \leq 1/8$,

$$\operatorname{kl}(1/2, 1/2 + 2\Delta) = -\tfrac{1}{2} \log(1 - 16\Delta^2) \leq 16\Delta^2.$$

Multiply by Δ and sum over $j = 1, \ldots, A - 1$. Together with the first case this proves (83). On M_0, an oracle informed that every action has the same signature law chooses action zero solely by comparing the known first-stage rewards; it need not know that common success probability. Hence its regret is exactly zero. □

F.3 Proof of the certification bound in Proposition 40

Proof. Let the null have A identical Bernoulli$(1/2)$ laws. Alternative j changes only law j to Bernoulli$(1/2 + \gamma)$. For the event E that the procedure outputs the one-block partition, correctness requires $\mathbb{P}_0(E) \geq 1 - \delta$ and $\mathbb{P}_j(E) \leq \delta$. Since $\delta < 1/2$, binary divergence is minimized over these two constraints at $(1 - \delta, \delta)$. Therefore (87) implies

$$\mathbb{E}_0 N_j \ \operatorname{kl}(1/2, 1/2 + \gamma) \geq \operatorname{kl}(1 - \delta, \delta).$$

For $\gamma \leq 1/4$, $-\frac{1}{2} \log(1 - 4\gamma^2) \leq 4\gamma^2$, proving (84). If a single procedure were uniformly correct against all positive γ while having finite $\mathbb{E}_0 N_j$, letting $\gamma \downarrow 0$ would contradict the bound. This proves the asserted obstruction without claiming that approximate pooling is impossible. □

F.4 Proof of the response lower bound in Proposition 40

Proof. Take $C = 1$ and a deterministic disturbance under every action and stage. The state set contains the initial state and two absorbing states g, b. Set $R_1 = 0$, and for $h \geq 2$ set $R_h(g, a) = 1$ and $R_h(b, a) = 0$ for every a. At the initial state, $\phi(s_1, a)$ assigns a separate nominal state z_a to

each action. At g, b, it assigns separate absorbing nominal states. Thus $Z = A + 2$ suffices. The unknown first-stage response at z_a reaches g with probability p_a; subsequent responses stay in the corresponding absorbing state. No later observation gives additional information about a first-stage transition.

Let M_0 have $p_a = 1/2$ for all a. For $j \in [A]$, let M_j change only p_j to $1/2 + \eta$, where $\eta = \frac{1}{8}\sqrt{A/K} \le 1/8$. Under M_j, episode regret is $(H-1)\eta$ unless action j is selected first. For the first-stage count N_j, (87) and Pinsker's inequality imply

$$\mathbb{E}_j N_j \le \mathbb{E}_0 N_j + K\sqrt{\tfrac{1}{2}\mathbb{E}_0 N_j \,\mathrm{kl}(1/2, 1/2+\eta)} \le \mathbb{E}_0 N_j + K\eta\sqrt{2\mathbb{E}_0 N_j}.$$

Averaging, using $\sum_j \mathbb{E}_0 N_j = K$ and Cauchy–Schwarz, gives

$$\frac{1}{A}\sum_j \mathbb{E}_j N_j \le \frac{K}{A} + K\eta\sqrt{2K/A} \le \left(\frac{1}{2} + \frac{\sqrt{2}}{8}\right) K.$$

The average expected regret is therefore at least

$$(H-1)\eta K\left(\frac{1}{2} - \frac{\sqrt{2}}{8}\right) \ge \frac{H-1}{32}\sqrt{AK}.$$

At least one alternative attains the average lower bound, proving the claim. The observed disturbance and even its entire law are uninformative in this construction. □

G Implementation details and computational requirements

This appendix specifies the operational statistics used by the main model-based learner and by the TD extension. It does not introduce an additional sampling oracle. The following definitions complete Algorithm 1, separating known inputs, empirical statistics, and unknown analysis quantities. The filtration is specified in Appendix C.1. All counters reset after certification; true groups, responses, and gaps are not inputs unless explicitly known in the stated specialization.

G.1 Online inputs and categorical estimators

These explicit counters and initializations complete the estimator definitions in Section 4. The inputs are the finite sets, $H, K, s_1, \phi, \psi, \zeta_R$, known R when $\zeta_R = 0$, a fixed response estimator $b \in \{\mathrm{MB}, \mathrm{TD}, \mathrm{HYB}\}$, and $\delta \in (0,1)$. Cardinalities and confidence constants are

$$B_h = H - h, \quad J_P = ZC, \quad L = \log\frac{64H(2A + J_P + \zeta_R SA)(K+1)}{\delta}, \quad L_d = d\log 2 + L.$$

The radius r_d is in Sec. 4.1; the queue studies use $\delta = 0.05$.

For stream $i \subseteq \mathcal{A}$ and a possible signature c, define the outcome counter and empirical law by

$$N^k_{h,i,c} = \sum_{j=\tau+1}^{k-1} \mathbf{1}\{a^j_h \in i,\ c^j_h = c\}, \quad n^k_{h,i} = \sum_{c=1}^{C} N^k_{h,i,c}, \tag{88}$$

$$\hat{\mu}^k_{h,i}(c) = \begin{cases} N^k_{h,i,c}/n^k_{h,i}, & n^k_{h,i} > 0, \\ 1/C, & n^k_{h,i} = 0. \end{cases}$$

For action a, the streams are $\{a\}$ and $G_h(a)$. Coincident streams can share storage, but one counter must not be incremented twice. The actual-response counters are

$$N^k_{h,z,c,s'} = \sum_{j=\tau+1}^{k-1} \mathbf{1}\{z^j_h = z, c^j_h = c, s^j_{h+1} = s'\}, \quad n^k_{h,z,c} = \sum_{s' \in \mathcal{S}} N^k_{h,z,c,s'}, \tag{89}$$

$$\hat{P}^k_h(s' \mid z, c) = \begin{cases} N^k_{h,z,c,s'}/n^k_{h,z,c}, & n^k_{h,z,c} > 0, \\ 1/S, & n^k_{h,z,c} = 0. \end{cases}$$

On a new observation, a histogram with previous count n and empirical vector $\hat{p}$ updates to $(n\hat{p} + e_x)/(n+1)$, where e_x is the unit vector of the observed outcome. The uniform zero-count initializer is multiplied by zero. TD stores the row count but not the successor histogram.

The stream penalty is $\lambda^k_{h,i} = \min\{1, r_C(n^k_{h,i}) + \epsilon_{h,i}\}$, with zero individual bias and certified group diameter. MB uses (6). TD uses learning weight $\alpha_i = (H+1)/(H+i)$ and bonus $b_{h,i} = 2B_h\sqrt{(H+1)L/i}$, where $i \geq 1$ counts updated row visits, not episodes. Raw scalars start at B_h; clipping only predictions to $[0, B_h]$ preserves the recursion analyzed in Appendix D. Terminal response predictions are zero.

G.2 Reward observations and their confidence bound

These statistics implement the separate reward component of Algorithm 1 and the reward-confidence lemma. For unknown means, keep actual-row counts and sums, including the terminal stage:

$$N^k_{h,s,a} = \sum_{j=\tau+1}^{k-1} \mathbf{1}\{s^j_h = s, a^j_h = a\}, \qquad T^k_{h,s,a} = \sum_{j=\tau+1}^{k-1} \mathbf{1}\{s^j_h = s, a^j_h = a\} r^j_h.$$

Use $\hat{R}^k_h(s,a) = T^k_{h,s,a}/N^k_{h,s,a}$ when the count is positive and zero when it is empty. Increment this count by one and its sum by the observed reward at the actual (h,s,a) only. Before planning, form $\bar{R}^k_h = \min\{1, \hat{R}^k_h + \rho_R(N^k_{h,s,a})\}$ with $\rho_R(0) = 1$ and $\rho_R(n) = \min\{1, \sqrt{L/(2n)}\}$ otherwise. This scalar reward radius uses L, without the categorical alphabet term. For supplied means set $\bar{R}^k_h = R_h$ and omit these counters. Appendix C.3 proves the additive charge, including its terminal contribution. Reward observations are not pooled on the basis of equal disturbance laws.

G.3 Certification modes and stopping rules

Four modes define the certificate interface of the certified learner of Section 4 without assuming an oracle for unknown sharing groups. For approximate certification, the *canonical complete-link rule*

Table 4: Certification inputs and outputs. The test cost remains part of regret.

Mode	Explicit prescription
Individual	Set $\tau = 0$, $\Pi_h = \{\{a\} : a \in \mathcal{A}\}$, and all $\epsilon = 0$.
Known groups	Supply valid groups, set $\tau = 0$, and use their known diameter bounds (zero for exact groups).
Unknown, separated	Supply the equality-or-separation premise with $\gamma \in (0,1]$. Use the pair tests, cap $n_\gamma = \lceil 32 L_C/\gamma^2 \rceil$ cycles, and consistency check in Appendix C.7. The gate compares HAn_γ to the fully explicit singleton bound F^A_b.
Unknown, approximate	Supply test cycles $n_a \geq 1$ with $An_a < K$ and empirical complete-link threshold $\eta_{\mathrm{pool}} \in [0,1]$. Use (40) and the canonical complete-link rule; return the computed nonzero biases.

starts from singleton blocks in increasing action-index order. Merge the lexicographically first pair of blocks whose union has empirical pairwise diameter at most η_{pool}, then repeat until no admissible pair remains. Order blocks by their smallest action index after each merge. This finite, deterministic rule is measurable from the certification sample. Its output need not minimize the number of blocks. Validity follows because (41) holds simultaneously for every action pair, irrespective of which blocks are selected. Assign each nonsingleton its diameter bound (40) and each singleton zero. The historical experiments use exact separated certification, not this approximate mode.

The pairwise thresholds are those in Section 4.1. The stopping time τ is the first consistent completed cycle or the capped fallback time, not an unknown tuning parameter. Replacing the test by another certificate requires accounting for its failure probability.

One entirely specified no-separation default is $n_a = \lfloor K^{2/3}/A \rfloor$ and $\eta_{\mathrm{pool}} = \min\{1, 2r_C(n_a)\}$. Use singleton mode when $n_a < 1$ or $An_a \geq K$. This is an admissible finite-budget choice, not an optimality claim: its statistical guarantee retains the resulting $K\epsilon_h$ term. The experiments reported in Appendix H use the separated test and their listed separation values, not this no-separation prescription.

G.4 Computation and estimator storage

This accounting supports the MB–TD memory comparison in Section 4, using the implemented counters of Appendix G. It distinguishes response storage from the planning and signature statistics common to both methods.

In a tabular implementation that explicitly stores all planned action values, all response estimators require $O(HSA)$ planned-value storage, $O(HS)$ policy and state-value storage, and $O(HAC)$ individual and block signature statistics. The latter includes at most HA blocks across stages and does not assume block observations are independent of the individual streams. MB adds $O(HZCS)$ next-state counts and $O(HZC)$ row counts. TD response adds $O(HZC)$ raw scalar estimates and visit counters instead of next-state histograms. The hybrid retains both and therefore has MB-order memory. Unknown rewards add $O(HSA)$ counts and reward sums and $O(1)$ work per observed reward. Assuming unit-cost lookup of reward upper estimates and evaluation of the known action map, a tabular MB planning pass costs $O(HZCS + HSAC)$ arithmetic operations: first form all conditional response expectations, then integrate each candidate against its two signature streams. A TD response pass costs $O(HZC + HSAC)$ because its conditional predictions are already scalar arrays. An individual observed transition increments constant-many histogram entries or scalar estimates, apart from the cost of computing the known maps. Both methods still pay for action enumeration; independence of a statistical term from A is not independence of runtime from A.

At each test-cycle boundary, all empirical pair distances require $O(HA^2C)$ operations. The deliberately simple transitivity check in the original experimental implementation costs $O(HA^3)$; a connected-component construction followed by a pairwise clique check can reduce that check, without changing any statistical claim. Pair flags need $O(HA^2)$ storage.

The separate response-learning term limits the improvement suggested by the input term alone. For $m \le A$, where m bounds the exact input-group count, compare the leading individual and pooled expressions using the elementary inequalities

$$\frac{\sqrt{A}+x}{\sqrt{m}+x} \le \frac{\sqrt{A}}{\sqrt{m}}, \qquad \frac{\sqrt{A}+x}{\sqrt{m}+x} \le \frac{\sqrt{A}+x}{x}, \quad x>0.$$

Substituting $x = \sqrt{SZ}$ gives the MB expression, and $x = \sqrt{HZ}$ gives the leading-order TD expression. This is purely an interpretation of the displayed rate terms; constants, logarithms, and test costs remain in the full theorem.

H Experimental protocols and additional results

This appendix documents the scalar-control, policy-selection, and confidence-intersection experiments. These protocols support the control and fixed-policy results. Appendix H.3 gives matched response estimator and grouping comparisons, and Appendix H.4 studies fixed-menu reliability.

H.1 Finite-buffer environment

This environment instantiates the queue in Figure 1 and the control comparisons. The state set is $\{0,1,2,3\}$ and the initial state is 2. There are eight actions, two signature categories, and four nominal states. Horizons and counts are $H = 4$, $K = 20{,}000$; all methods use confidence input $\delta = 0.05$. For $a \in \{0,\dots,7\}$ define service $d(a) = a \bmod 4$ and mode $j(a) = \lfloor a/4 \rfloor$. The known nominal map is $z = (s - d(a))^+$ and the known reward is $R(s,a) = 1 - 0.2s/3 - 0.15d(a)/3 - 0.08j(a)$. The hidden response at every stage is

$$P(s' \mid z, c) = 0.85\mathbf{1}\{s' = \min(3, z+2c)\} + 0.15\mathbf{1}\{s' = \max(0, \min(3, z+2c) - 1)\}.$$

When the two designated successors coincide, the probabilities add to one at that state. The signature is Bernoulli with success parameter

$$p_a = \begin{cases} 0.5, & \text{one-law regime,} \\ 0.85 \text{ if } j(a) = 0, \quad 0.15 \text{ otherwise,} & \text{two-law regime,} \\ 0.1 + 0.8a/7, & \text{eight-law regime.} \end{cases}$$

These stage laws happen to be time homogeneous, but the learner keeps stage-specific estimators as required by the more general theorem. A raw signature interpretation is $u = (c, a)$ with known $\psi(c, a) = c$; the action tag is deterministic and does not require simulation of another random variable. The law of c has no state dependence, matching the formulation.

The compared algorithms use the following information and confidence inputs. The six methods are action-specific MB; MB with the true groups supplied; TD response with the true groups supplied; the two-response hybrid with the true groups supplied; the gated hybrid with sequentially learned groups; and forced full pooling. Known groups are a strictly stronger-information comparator, not information given to the adaptive algorithm. The adaptive test uses $\gamma = 0.7$ for the first two regimes and $\gamma = 0.8/7$ for the third. The equality-or-separation promise holds in all three cases. The core algorithms use the categorical radii in Sec. 4.1, the input confidence in Sec. 4.1, and the TD learning rate and bonus in (61), with no tuned multiplicative constants. Forced full pooling assumes a single zero-bias block regardless of the regime. Its certificate is invalid in the two- and eight-law regimes, so Theorem 31 is not asserted for it.

Evaluation uses the true policy value rather than a noisy realized return. Each setting uses seeds $0, \ldots, 11$. Two uniform random numbers per stage select the categorical signature and successor, with seeds shared across methods to couple exogenous random numbers. Each method still induces its own trajectory and data. Every episode's deterministic Markov policy is evaluated exactly by backward dynamic programming under the true MDP; the resulting difference from the true optimal value is the episode's pseudo-regret. These true-model evaluations are for measurement only and are not inputs to the learner. Saved curves contain 251 checkpoints per run, including episode one and the final episode. Bands represent one standard error (SE) across seeds, not a confidence band uniformly over time.

The learned-group test lengths averaged 1157.3 episodes in the one-law setting (range 1008–1360) and 1038.7 in the two-law setting (range 928–1224). In the eight-law setting, the gate skipped the test on every seed. All observed test decisions matched the true partition when a test was completed. The minimum computed optimism gap was nonnegative for the correctly specified methods in these runs. These finite checks cannot establish a uniform high-probability theorem. The zero standard error for one-law TD response occurs because its conservative predictions induce the same evaluated policy sequence across those seeds at this horizon, not because the environmental observations are deterministic. The hybrid and known-group MB curves coincide here; the experiments do not establish a universal practical advantage from keeping the TD branch.

H.2 A POPULATION POOLING COUNTEREXAMPLE

This calculation illustrates the transfer-bias obstruction in Section 3 when an input-sharing premise is false. A separate two-stage instance has two actions, known first rewards $R_1(0) = 1$, $R_1(1) = 0.9$, and an identity response $s_2 = c$ with terminal reward c. The true success probabilities are $\mu_0(1) = 0.2$ and $\mu_1(1) = 0.9$. True action values are therefore 1.2 and 1.8, respectively. Replacing both laws by *any* common law with success probability p produces estimated values $1 + p$ and $0.9 + p$. Its exact optimal policy always chooses action zero, which has true per-episode regret 0.6. Thus even an exactly known, incorrectly pooled model has regret $0.6K$. This is a population calculation, not a fitted simulation curve or a claim about the queue benchmark above.

H.3 MATCHED CONTROL COMPARISONS

These diagnostics compare the estimator and information choices in Algorithm 1 and Appendix D. They separate three comparisons: individual versus known sharing at a fixed response estimator; known versus learned sharing at a fixed response estimator; and MB/TD/hybrid with the same supplied sharing information. The scripts record every specified setting, including unfavorable outcomes. True model values are used for evaluation and analytical certificates only.

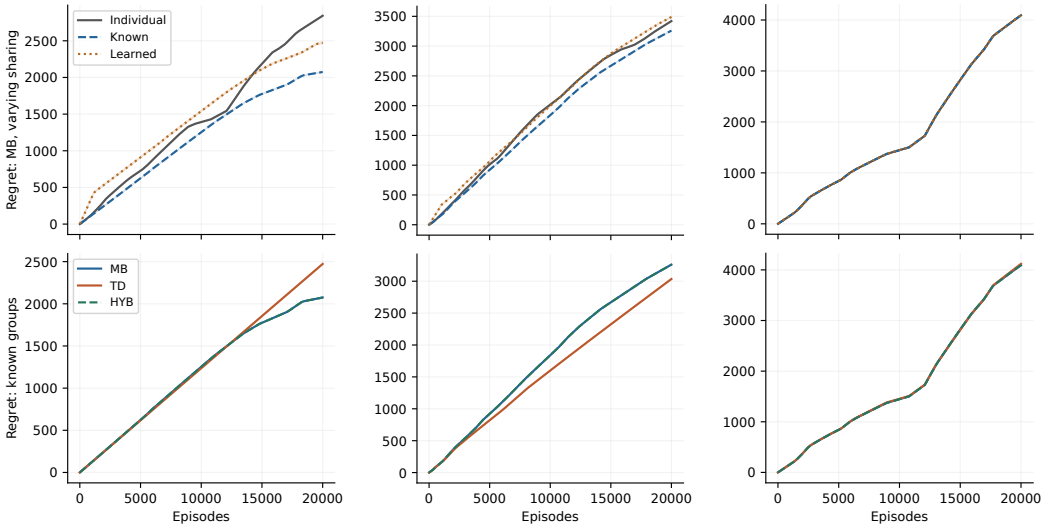
H.3.1 MATCHED QUEUE RUNS AND INSTRUMENTATION

The following protocol isolates the effects of grouping information and conditional-response estimation in Algorithm 1 and its TD extension. We execute 576 runs: 324 matched runs (three regimes, three response estimators, three information modes, twelve seeds), 36 margin runs (two regimes,

three valid margins, six seeds), and 216 size runs (three horizons, four state sizes, three response estimators, six seeds). All runs use 20,000 episodes and $\delta = 0.05$. Seeds match across compared methods, which share random-number streams but generate their own trajectories. Reported standard errors are across seeds.

The instrumented runner copies the original executable update rules, records branch and clipping counts, and exposes a supplied margin argument. Three 1000-episode checks verify bitwise equality of its learning curves and test diagnostics with the original MB, TD, and hybrid runner. The exact test additionally rejects conflicting equal/different marks on exceptional histories, matching the certified learner of Section 4. At a visited nonterminal row, a strict MB selection means $W^{\mathrm{MB}} < W^{\mathrm{TD}} - 10^{-12}$; a strict TD selection is defined symmetrically, and remaining comparisons are ties. These are pointwise response selections, not the branch choices of the retrospective regret certificate. A clipped visit has $\overline{V}_h(s_h)$ at its remaining-reward cap. Test visits are excluded from these denominators.

The top row fixes MB and varies sharing information; the bottom row fixes known groups and varies the response estimator. Bands show one standard error over twelve seeds. Coincident MB/hybrid curves are retained.



(a) One input law. (b) Two input laws. (c) Eight input laws.

Figure 5: Queue learning with matched information and estimators.

The regenerated overlapping settings agree with the saved historical curves to numerical precision. Individual TD has final regret 2954.6 ± 1.4, 3548.4 ± 1.0, and 4119.7 ± 2.1 in the one-, two-, and eight-law regimes. Individual MB and hybrid coincide. The remaining matched results are in Table 5. No test in the tested runs returns an incorrect partition.

The mean actual test regrets are 432.1 and 342.8 in the first two regimes, compared with average durations 1157.3 and 1038.7 episodes. The difference between known and learned final regret need not equal test regret, since restarting changes the later sample counts and policy history. For known-group hybrid, strict MB fractions are 46.3%, 40.8%, and 11.7%; strict TD fractions are zero. The corresponding fractions of visited optimistic values at their reward cap are 21.1%, 21.6%, and 69.7%. Heavy clipping helps explain why response-estimator distinctions can fail to change the selected policy. The one-law standalone TD's zero measured standard error reflects a common evaluated policy sequence despite random observations.

H.3.2 Margin and state/horizon sensitivity

This sweep tests the gate and the separate input- and response-learning terms in Theorem 2. At supplied $\gamma = 0.7$, the one- and two-law test gate runs. At smaller but still valid margins 0.35 and 0.175, the theoretical cap exceeds the episode budget and the gate skips on every seed. The learned-

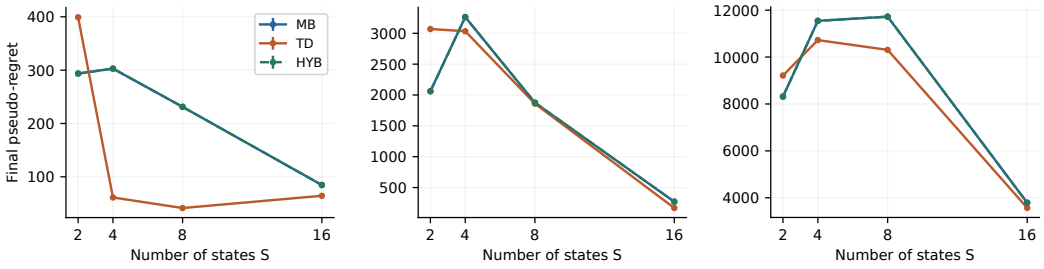
Table 5: Matched queue pseudo-regret.

Method	One law	Two laws	Eight laws
MB, individual	2841.7 ± 2.4	3420.3 ± 0.7	4091.9 ± 4.0
MB, known groups	2074.8 ± 2.0	3260.0 ± 1.7	4091.9 ± 4.0
TD, known groups	2472.6 ± 0.0	3034.5 ± 0.3	4119.7 ± 2.1
Hybrid, known groups	2074.8 ± 2.0	3260.0 ± 1.7	4091.9 ± 4.0
MB/hybrid, learned groups	2472.4 ± 8.9	3488.7 ± 5.7	4091.9 ± 4.0
TD, learned groups	2761.7 ± 6.5	3229.7 ± 4.3	4119.7 ± 2.1

group algorithm then coincides with its individual-stream response estimator. For the eight-law regime the valid $\gamma = 0.8/7$ also causes a skip. Consequently, a smaller valid margin does not imply more finite-budget sampling when the certification-cost gate is used: it can disable certification entirely.

The size sweep uses $H \in \{2, 4, 8\}$ and $S = Z \in \{2, 4, 8, 16\}$, keeping $A = 8$, $C = 2$ and two disturbance groups. The same formulas are used with state set $\{0, \dots, S-1\}$, initial state $\min\{2, S-1\}$, $\phi = (s - a \bmod 4)^+$, reward $1 - 0.2s/(S-1) - 0.15(a \bmod 4)/3 - 0.08\lfloor a/4 \rfloor$, and response 0.85 at $\min(S-1, z+2c)$ plus 0.15 one state below. This varies an environment family; changing S also changes reward normalization and the reachable buffer geometry, so it is not a pure scaling experiment for a minimax exponent.

The sweep uses six seeds and $20{,}000$ episodes for every setting. Error bars show one standard error. MB and hybrid coincide, while standalone TD can cross MB without contradicting a same-history bound.



(a) Horizon $H = 2$. (b) Horizon $H = 4$. (c) Horizon $H = 8$.

Figure 6: State and horizon sensitivity with known groups.

The sweep probes values on both sides of $S = H$, but all tested dimensions remain far below the regime in which the TD global confidence constants can offset the MB categorical factor. Comparing only the leading powers $S^{1/2}$ and $H^{1/2}$ would miss that issue. At $H = 2, S = 4$, for example, MB/hybrid regret is 302.9 and TD regret is 61.3; at $H = 4, S = 2$, MB/hybrid regret is 2061.1 and TD regret is 3068.9. These are different policy histories, not evaluations of two bounds on the hybrid history. A broadly useful practical hybrid advantage remains unestablished by this suite.

Entries are mean pseudo-regret and one standard error over twelve seeds. Known groups provide additional information; MB and hybrid coincide in these runs.

The reported inactivity of the TD branch is retained. These runs evaluate the implementations with their stated confidence bonuses; they do not establish superiority of the common-policy hybrid over either separately trained algorithm.

H.4 Policy-selection reliability under valid and invalid sharing

These experiments test the fixed-contrast specialization of Theorem 5, not the full response-sensitive reference family.

H.4.1 PREDECLARED EXPERIMENT AND DATA

The family, class, reward, and target policies are those of Theorem 5. We fix $\delta = 0.05$, $\varepsilon \in \{0.5, 0.2, 0.1, 0.05\}$, and $\gamma \in \{0, 0.01, 0.05, 0.1\}$. For every pair we test both the shared instance M_0 and the admitted alternative $p_1 = 1/2 + \max\{\gamma, t_\varepsilon\}$, with all other parameters unchanged. Each of the 32 cases has 2000 independent replications. All selected settings are reported.

The methods observe identical prefixes of the same complete-episode categorical stream in each replication. We compare the supplied-tie rule, the direct-successor rule, and the margin-certificate rule in (73). The direct rule uses $b_0 = \lceil \varepsilon^{-2} \rceil$; the margin rule uses $b_0 = \lceil \max\{\varepsilon, \gamma\}^{-2} \rceil$; known input-law equality uses $b_0 = 1$. Each uses its own look index in the stated simultaneous boundary. At $\gamma = 0$, direct and margin rules coincide exactly. The supplied-tie rule is labeled misspecified on every broken-tie case and carries no correctness guarantee there.

At fixed total budgets 3200, 12800, and 51200, we also compare the tied empirical-factor estimator (identically the direct tied estimator), the separate-disturbance estimator with pooled conditional-response estimates, and the direct-successor contrast. Their sign errors and squared errors use the same observations. The untied factor estimate is

$$\hat{\theta}_{\text{untied}} = \hat{m}_0 + \hat{p}_0(\hat{m}_1 - \hat{m}_0) - (1 - \varepsilon)[\hat{m}_0 + \hat{p}_1(\hat{m}_1 - \hat{m}_0)] - \varepsilon/2.$$

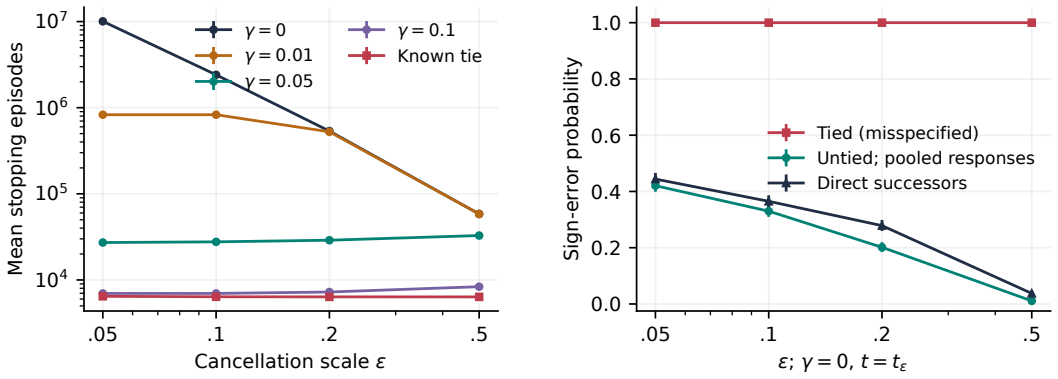
Both response estimates pool the two first actions, and empty response rows take value $1/2$. The tied estimator exploits an extra equality; the untied one remains meaningful when that equality fails. It is not an oracle competitor deprived of response information.

H.4.2 EXACT SAMPLING AT LARGE STOPPING BUDGETS

This sampling construction reproduces the complete-feedback law of the fixed-policy experiment without replacing it by expected counts. For each first action, the four complete-trajectory categories $(c, Y) = (0,0), (0,1), (1,0), (1,1)$ have probabilities $((1-p_a)(1-m_0), (1-p_a)m_0, p_a(1-m_1), p_a m_1)$. Independent multinomial increments between the union of all method looks generate exactly the sufficient-statistic law of full episodes. This avoids explicitly materializing millions of binary trajectories per replication; it is not a Gaussian or asymptotic approximation. The same count arrays drive every method. All checkpoint counts, per-replication stopping decisions, fixed-budget estimates, and summaries are saved. No run hit the administrative look cap, so no censored outcomes are omitted. The counts replay every reported decision.

H.4.3 RESULTS AND UNCERTAINTY

Figure 7 and Table 6 summarize the stopping-cost and error measurements for the preceding protocol. The shared-instance panel reports mean stopping episodes at M_0 with one-standard-error bars. The unequal-law panel reports sign errors at 3200 episodes with 95% Wilson intervals. The tied estimator assumes a false equality there; the untied comparator still pools conditional-response observations.



(a) Same shared instance.

(b) Broken equality, 3200 episodes.

Figure 7: Fixed-policy cost and reliability under uncertain sharing.

Table 6 reports the shared-instance stopping results for every setting. The direct rule column uses the paired $\gamma = 0$ case; its minor Monte Carlo variation across other margins is retained in the machine-readable summaries. Larger margins reduce the admissible alternative class and can make the equality certificate useful. These are conservative confidence rules, not an evaluation of optimal leading constants. For $\gamma = 0$, reducing ε from 0.5 to 0.05 raises mean stopping episodes from 58,368 to 10,073,702, while the supplied-tie rule stays near 6400. At $\gamma = 0.1$, the margin rule stops near 7000 at the two smallest ε values.

Table 6 reports mean stopping episodes and one standard error over 2000 replications per cell. All listed methods have zero observed sign errors at these shared instances.

Table 6: Stopping episodes at the shared fixed-contrast instance.

ε	Known tie	$\gamma = 0$	$\gamma = .01$	$\gamma = .05$	$\gamma = .1$
0.5	$6,357 \pm 51$	$58,368 \pm 477$	$58,425 \pm 468$	$32,784 \pm 87$	$8,352 \pm 25$
0.2	$6,362 \pm 52$	$535,501 \pm 4,662$	$525,158 \pm 4,538$	$28,928 \pm 193$	$7,245 \pm 48$
0.1	$6,376 \pm 51$	$2,409,267 \pm 20,006$	$829,030 \pm 1,995$	$27,699 \pm 157$	$6,973 \pm 41$
0.05	$6,458 \pm 50$	$10,073,702 \pm 79,211$	$829,030 \pm 1,995$	$27,200 \pm 139$	$6,973 \pm 41$

Across all 32 cases, the direct and margin-certificate procedures have zero observed errors. Each valid per-case result has a 95% Wilson binomial interval $[0, 0.00192]$ from 2000 replications; this is sampling uncertainty in the observed error rate, not a replacement for the theorem's $\delta = 0.05$ guarantee and not a simultaneous band over cases. The supplied-tie rule is correct in every shared replication and incorrect in all 2000 replications of every broken-tie case, with per-case interval $[0.99808, 1]$ for those error rates. The deliberately misspecified rule confidently uses the wrong identity; the declared alternatives are therefore essential.

At $\varepsilon = 0.1$, $\gamma = 0$, and 3200 episodes under the broken tie, the sign-error rates are 1.000 for the tied plug-in, 0.330 for the response-sharing untied plug-in, and 0.365 for direct successors. Their Wilson intervals are $[0.9981, 1]$, $[0.3097, 0.3509]$, and $[0.3442, 0.3863]$, respectively. The sequential direct rule eventually stops with no observed error, at a much larger budget; comparing these outcomes requires keeping fixed-budget error and stopping-time sample cost distinct. The same-data untied estimator's smaller fixed-budget error illustrates the benefit of retaining the known shared response structure even without tying disturbances. The elementary direct certificate does not claim an optimal constant.

H.4.4 Explicit lower bounds and stopping branches

The lower bound in (71) can be computed directly without new simulation. At $\delta = 0.05$, its numerator is $0.9 \log 19$ and its denominator is $-\frac{1}{2}\log(1 - 4t_*^2)$. It is a bound on $\mathbb{E}_{M_0} N_1(T)$ and therefore on $\mathbb{E}_{M_0} T$ for any uniformly correct procedure. Table 7 compares it with measured total stopping episodes for the margin-certificate rule. The ratio is to this particular lower bound, not to the unknown optimal expected sample size; the single alternative may give a loose constant. These large ratios preclude presenting this implementation as practically sample efficient merely because its asymptotic parameter orders match.

Costs and branch fractions use the saved 2000-replication data. The explicit lower bound is analytical, and the mean/bound ratio is not an approximation factor to the unknown optimum.

Direct stopping is checked first, so simultaneous direct and equality certificates are recorded as direct. The remaining fraction in each row is the direct branch. At $\varepsilon = 0.05$, the no-margin rule always stops directly; all positive margins tested use the certified-tie branch in every replication. At larger ε, both branch regimes occur across settings. These frequencies test the mechanism conditional on the particular family and chosen boundary, not an optimal allocation claim.

H.4.5 Sensitivity within the binary two-policy family

To separate replication count from environmental variation, an additional sweep changes (p_0, m_0, m_1) to the four tuples in Table 8. All 16 cases are retained. There are 16,000 independent replicated datasets and 48,000 paired method outcomes. The old 32-case dataset is unchanged. We use the same exact categorical-increment sampling and stopping rules, with independent case-specific streams.

Table 7: Analytical lower bounds and observed stopping branches.

ε	γ	Mean episodes	Explicit lower bound	Mean / bound	Certified tie (%)
0.05	0	10,073,702	23,435.2	429.9	0.0
0.05	0.01	829,030	13,247.3	62.6	100.0
0.05	0.05	27,200	527.3	51.6	100.0
0.05	0.1	6,973	129.8	53.7	100.0
0.1	0	2,409,267	5,256.3	458.4	0.0
0.1	0.01	829,030	5,256.3	157.7	96.7
0.1	0.05	27,699	527.3	52.5	100.0
0.1	0.1	6,973	129.8	53.7	100.0
0.2	0	535,501	1,036.1	516.8	0.0
0.2	0.01	525,158	1,036.1	506.8	0.0
0.2	0.05	28,928	527.3	54.9	100.0
0.2	0.1	7,245	129.8	55.8	100.0
0.5	0	58,368	62.2	937.8	0.0
0.5	0.01	58,425	62.2	938.8	0.0
0.5	0.05	32,784	62.2	526.8	71.6
0.5	0.1	8,352	62.2	134.2	99.9

For a shared mean $q_0 > 1/2$ and $m_1 > m_0$, the broken tie sets

$$t = \max\left\{\gamma, \frac{2\varepsilon(q_0 - 1/2)}{(1-\varepsilon)(m_1 - m_0)}\right\}, \qquad p_1 = p_0 + t.$$

Every chosen p_1 is interior, belongs to the declared class, and gives contrast at most $-\varepsilon(q_0 - 1/2)$. This generalizes the diagnostic alternative, not the uniform-in-parameters theorem at M_0.

Entries are mean total episodes over 1000 replications. Known-equality and direct columns use $\gamma = 0$; the last two use the margin certificate at $\gamma = 0.1$.

Table 8: Shared-instance sensitivity at $\varepsilon = 0.1$.

p_0	(m_0, m_1)	d	Known tie	Direct	Margin 0.1	Certified (%)
0.5	$(0.2, 0.85)$	0.025	26,624	9,666,560	22,835	100.0
0.5	$(0.25, 0.95)$	0.100	1,528	584,294	6,790	100.0
0.5	$(0.45, 0.65)$	0.050	6,479	2,383,872	6,944	100.0
0.7	$(0.2, 0.7)$	0.050	6,386	2,392,883	6,842	100.0

Changing d from 0.025 to 0.1 changes the known-tie mean from 26,624 to 1528 episodes, illustrating the residual sign difficulty even after equality is known a priori. At $d = 0.05$, changing the response separation or base signature frequency retains the qualitative structural comparison in the tested settings. The margin rule's cost at $d = 0.025$ also shows that certification does not remove the successor-gap cost. Both uniformly valid rules make zero observed errors in each of these 16 cases (per-case 95% Wilson interval $[0, 0.00383]$); the supplied-tie rule is wrong in every broken-tie replication. Exact count checkpoints replay all 48,000 stopping outcomes.

This expands the parameter diagnostic within one small realizable family. It does not test larger policy libraries, deeper decision problems, arbitrary response kernels, or misspecification outside the declared class. Together with the queue results, it supports structural sensitivity and reliability under the tested premises, without establishing broad deployment robustness or small stopping constants.

H.4.6 AFFORDABILITY IS NOT POSITIVE NET BENEFIT

The gate in Theorem 2 compares the maximum test charge with the individual bound. It does not use the unknown number of true groups. For known reward means ($\zeta_R = 0$), at $H = S = Z = C = 2$, $A = 100$, $K = 10^6$, $\delta = 0.05$, $\gamma = 1$, and $m = 1$, its full constants give

$$n_\gamma = 908, \quad HAn_\gamma = 181600, \quad F^A_{\mathrm{MB}} = 210366.68, \quad F^m_{\mathrm{MB}} = 74587.48.$$

The gate runs, although the sum of the test charge and the grouped bound is $256187.48 > F^A_{\mathrm{MB}}$. This is compatible with the factor-two theorem. The fully shared family can satisfy this declared class

even though a fully shared instance by itself cannot justify choosing a particular numerical separation for alternatives.

The existing queue data also reveal both signs of net effect. Table 9 decomposes the learned-MB history itself into actual test and subsequent pseudo-regret; their sum is its total. Its paired difference from individual MB uses the same twelve seeds. It is a comparison of complete learning procedures, not a causal isolation of the test: their post-test policies and restart histories differ. Entries report means and standard errors over twelve seeds, with paired standard errors for the final column. The

Table 9: Certification and post-certification regret.

Laws	Test	After test	Total learned	Learned − individual
One	432.1 ± 9.7	2040.3 ± 2.6	2472.4 ± 8.9	-369.3 ± 9.1
Two	342.8 ± 7.3	3145.9 ± 2.3	3488.7 ± 5.7	68.4 ± 6.0
Eight	0.0 ± 0.0	4091.9 ± 4.0	4091.9 ± 4.0	0.0 ± 0.0

original margin and size sweeps remain in Appendix H.3. Reducing a valid margin from 0.7 to 0.35 or 0.175 causes the queue gate to skip at $K = 20000$. The method cannot select a favorable margin from apparent equality in those data.

H.4.7 Unknown-reward implementation and a boundary of the sharing benefit

The unknown-reward term in Theorem 2 requires a separate implementation check: equality of disturbance laws does not imply equality of reward means, and uncertainty in the immediate reward changes exploration even when the transition model is unchanged. We therefore retain the queue mean rewards and transition laws and replace supplied means by observed Bernoulli rewards with those means. The planner receives only their actual (h, s, a) counts and sums. The true means remain available only to the environment and exact pseudo-regret evaluator. This is an implementation diagnostic of the extra term in Theorem 2, not an empirical claim that reward estimation is a new method.

The fixed configuration uses $K = 20000$, $H = S = Z = 4$, $A = 8$, $C = 2$, two law regimes, all three response estimators, three grouping modes, and seeds 903–906. Each setting runs once with supplied means and once with unknown means: 144 runs in 72 pairs. A separate reward random stream preserves the transition random numbers in the supplied-mean specialization. Policies may differ, so paired random numbers do not imply identical observations after different actions. No reward observation from the testing phase enters the restarted learner. Table 10 reports every independent-reward mean comparison. HYB agrees exactly with MB in these runs.

Each cell gives mean pseudo-regret and one standard error over four seeds. MB and HYB coincide. Known groups provide additional information but do not guarantee lower realized regret at every finite budget.

Table 10: Unknown-reward control diagnostic.

Laws / response	Grouping	Supplied means	Unknown means
1 / MB	Individual	2837.9 ± 2.5	3152.3 ± 18.8
1 / MB	Known	2063.3 ± 3.3	3633.3 ± 41.8
1 / MB	Learned	2448.6 ± 12.5	3848.6 ± 30.4
1 / TD	Individual	2954.9 ± 0.7	3168.1 ± 33.9
1 / TD	Known	2472.6 ± 0.0	3598.2 ± 34.8
1 / TD	Learned	2750.9 ± 7.8	3831.6 ± 29.7
2 / MB	Individual	3420.4 ± 1.7	4843.8 ± 37.1
2 / MB	Known	3257.9 ± 2.7	4487.9 ± 35.7
2 / MB	Learned	3478.9 ± 4.8	4666.7 ± 29.8
2 / TD	Individual	3547.1 ± 0.6	4795.2 ± 37.2
2 / TD	Known	3033.2 ± 0.7	4315.4 ± 30.1
2 / TD	Learned	3221.4 ± 3.5	4505.0 ± 23.5

The one-law MB result changes from 2837.9 (individual) and 2063.3 (known groups) with supplied means to 3152.3 and 3633.3 with unknown means. Learned sharing similarly changes from 2448.6 to 3848.6. Thus the favorable known-reward comparison does not transfer uniformly to this

unknown-reward execution. With two laws, known and learned sharing have smaller mean regret than individual learning in both reward modes. These are four-seed synthetic diagnostics, not broad performance guarantees. In particular, actionwise reward uncertainty can dominate the reduction in input estimation: the two uncertainty sources must remain distinct.

To test the permitted within-step dependence, an additional 12 runs use two law regimes, all three response estimators, learned groups, and seeds $903, 904$. Here the same uniform variate generates the successor and the Bernoulli reward, preserving the reward mean conditional on the pre-transition history while correlating the two observations. MB/HYB mean regrets are 3813.3 ± 27.3 and 4555.9 ± 20.3 for one and two laws; TD gives 3786.4 ± 72.9 and 4353.4 ± 24.8. These two-seed checks exercise the observation model and are not a precision performance comparison.

Across all 156 runs, the recorded reward-confidence failures, partition-consistency errors, and optimism violations are zero. The largest realized sum of reward-radius charges is 23881.35, below its analytical bound $\mathcal{G}_R = 43159.68$. Every unknown-reward run records exactly $H(K - \tau)$ post-test rewards, including terminal observations. HYB selects no strictly better TD response at visited rows. These checks support implementation consistency on the executed histories. They neither prove a high-probability theorem nor estimate its failure probability precisely.

H.5 A SAME-INFORMATION ABLATION OF CONFIDENCE INTERSECTION

The historical queue uses two signatures, so Lemma 22 rules out attributing a gain over separate exact supports to intersection alone. This diagnostic isolates that mechanism with a common response vector at each target. It uses known responses and supplied equality; it therefore does not test response learning or group discovery. Those questions remain in the historical matched experiments above.

Strict targets and action changes use 732 target and 366 decision checkpoints per row. The regret difference is intersection minus separate supports, reported as a mean and paired standard error.

Table 11: Same-information confidence-intersection diagnostic.

C	Strict targets	Action changes	Δ regret
2	0/732	0/366	0.000 ± 0.000
3	56/732	2/366	-0.033 ± 0.037
8	158/732	2/366	$+0.020 \pm 0.119$

Environment and declared protocol. Use $H = S = A = Z = 2$, $K = 1500$, and $\delta = 1/20$. The initial state is fixed, $z = a$, and the common signature is uniform on $\{0, \dots, C - 1\}$. First rewards are $a/50$. The terminal reward is the binary successor Y, whose known probabilities are

$$w_0(c) = \frac{1}{20} + \frac{9c}{10(C-1)}, \qquad w_1(c) = w_0(C - 1 - c).$$

Terminal actions have no effect; terminal inputs can be fixed. Thus true first-action values are $1/2$ and $13/25$, and action one is uniquely optimal. All learners receive the response probabilities and one equality group, but must learn its common input law. This is a realizable specialization of the paper's model. Known responses give singleton response confidence sets, so the identical w_a isolates input geometry.

We fixed $C \in \{2, 3, 8\}$, seeds 1901–1906, and the three methods before running this diagnostic. The scalar method uses $\min_i(\hat{\mu}_i^\top w_a + r_C(n_i))$, capped at one. The separate method minimizes the two exact single-ball supports; the intersection method uses the certified support of both balls. Each adds the same first reward. All use individual and pooled counts, smallest-index ties, and episodewise greedy planning. Known groups imply $\tau = 0$. Each seed supplies the same sequence of uniform signatures and successor uniforms across methods; their actions and observations can differ. True-model values are used for regret evaluation only. No configuration or seed was excluded after observing results.

Computation and measurement. Integer histograms and the interface in Appendix C.9.6 define the input balls. Single-ball TV support is computed by exact rational mass transfer from low-valued to high-valued coordinates. For binary inputs the intersection identity avoids a linear program (LP).

In other dimensions, the sufficient nesting test $\mathrm{TV}(\hat{\mu}_i, \hat{\mu}_j) + r_i \le r_j$ also avoids an LP. Otherwise the completed HiGHS solve supplies multipliers for the rational upper certificate. The policy uses this exact upper value, capped by separate exact supports; no approximate primal objective selects actions. An infeasible intersection uses the scalar fallback. The largest radius bracket width in these runs is 10^{-79}, and no true-law confidence violation was observed; neither fact proves a uniform coverage theorem.

Every 25 episodes from episode one, and at episode 1500, all three backups are evaluated on the same current history. This gives 61 decision checkpoints per run, 366 per method/category, and 732 target comparisons. "Strict" means separate minus intersection support exceeds 10^{-10}. A strict difference certifies non-nesting; failure of the sufficient nesting test alone does not. "Both active" means both TV constraints of the approximate primal optimizer are within 10^{-7} of their radius, a diagnostic unrelated to validity of the upper certificate. Action changes compare separate and intersection on that same history. Their beneficial/harmful classification uses true action values. These local diagnostics must not be confused with the regret of independently trained learners.

Regret and planning seconds are means over six paired seeds (54 runs). Policy linear-program calls exclude counterfactual checkpoint probes, which are counted separately. Timing includes probes, Python overhead, exact arithmetic, and input preparation; it is an implementation measurement, not a complexity guarantee.

Table 12: On-policy confidence-intersection comparison.

C	Method	Regret	Seconds	Policy LPs	Probe LPs
2	Scalar	14.410	1.004	0	0
	Separate	14.430	1.175	0	0
	Intersection	14.430	1.144	0	0
3	Scalar	9.493	1.009	0	43
	Separate	9.517	1.317	0	69
	Intersection	9.483	2.023	1548	0
8	Scalar	7.837	1.374	0	220
	Separate	7.880	1.862	0	192
	Intersection	7.900	5.456	4799	0

Counts are totals over six seeds. Target columns have denominator 732; the final column gives beneficial/harmful changes in 366 decisions, using separate exact supports as the local comparator.

Table 13: Confidence-intersection activation on recorded histories.

C	History generated by	Strict	Both active	Nesting certified	Changes $+/-$
2	Each method	0	0	—	0/0
3	Scalar	32	32	689	0/0
	Separate	58	58	663	1/1
	Intersection	56	56	667	0/2
8	Scalar	148	149	512	0/1
	Separate	157	158	540	1/2
	Intersection	158	159	544	1/1

Interpretation and a replayable useful decision. On the intersection learner's histories, strict tightening occurs at 7.65% of three-category targets and 21.58% of eight-category targets, versus zero in the binary boundary case. The paired regret differences in Table 11 are small relative to their standard errors and change sign across dimensions. Thus, there is no demonstrated uniform control advantage. The extra solves and rational verification increase computation, and binary inputs need no intersection LP.

One ordinary-confidence history gives a useful three-category decision beyond the rare construction in Appendix C.9.5. For the separate learner with seed 1901, immediately before episode 1451 the action-specific counts are $(397, 373, 403)$ and $(78, 94, 105)$. Their sums are 1173 and 277, and the pool is their sum. The separate action scores are approximately $(0.584125, 0.583642)$; intersection scores are $(0.583368, 0.583642)$. Thus the former chooses action zero and the latter action one, improving true value by 0.02 at this decision. This is a recorded greedy separate-learner history with

supplied groups, not a claim about the intersection learner's own history or typical improvement. Table 13 includes the harmful switches as well.