DECISION-CERTIFIED RESIDUAL SAMPLING FOR AVERAGE-REWARD POLICY VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

An existing controller can perform well before the available observations justify accepting it. We study how historical data and additional samples can certify a policy in a finite average-reward Markov decision process. Two stationary-flow programs compare the best plausible gain with the worst plausible gain of the incumbent, giving an anytime-valid bound on policy loss. Separate dual potentials can improve this bound by an unbounded factor over the best common potential for the same incumbent. We then connect acquisition prices to actual sampling efficiency. With full-support dynamics, known rewards, and a unique optimal policy, normalized finite contractions consistently estimate the leading coefficients of the observed certificate excess above true policy loss, including its moving empirical centers. A price-tracking sampler realizes the optimal limiting allocation for this expansion and compares favorably with uniform sampling as the verification margin shrinks. A residual extension characterizes how heterogeneous historical counts reduce the additional samples. General validity and communicating-MDP completion require weaker conditions. Experiments distinguish certificate choice, allocation, policy quality, and computation, using controllers learned from behavior-policy logs and tests of the predicted stopping constants. An exact shared-model comparison also identifies uncertainty that the separate certificate can unnecessarily resolve.

1 INTRODUCTION

A maintenance controller balances service costs against future failure. A queueing controller trades service effort against congestion. An existing controller may already be good enough, yet its historical log may leave that judgment uncertain: common operating conditions are well represented, while other state–action pairs have little data. Hybrid reinforcement learning studies how such histories reduce subsequent sampling (Song et al., 2023; Wagenmaker & Pacchiano, 2023). We ask which further observations are needed to justify accepting a particular controller.

A uniformly accurate model is sufficient for policy evaluation, but can require observations with little effect on the decision. Conversely, a good policy may remain uncertified because the data permit a better competitor or a worse outcome for that policy. We retain the history and stop when a loss bound passes the requested tolerance.

Average reward makes local measures of relevance unreliable in two opposite ways. A rarely visited gateway can control a long excursion and therefore have a large effect on gain. A locally inferior action can also have a small average cost because the state is rarely visited. The second phenomenon persists in irreducible chains, not only at transient states. A useful verification rule must account for the long-run behavior of both a competitor and the policy being evaluated.

We compare these behaviors using stationary flows. One linear program finds the largest plausible gain; another finds the smallest plausible gain of the incumbent. Their difference certifies the policy, and changes in these programs assign prices to unresolved row uncertainty. This gives decision-certified residual sampling (DCR). A central question remains: do prices computed at the current data help after new observations move the estimates? Our main allocation result answers this question for a specified regular class, by accounting for the empirical centers that move with new 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.

Contributions. First, we give an anytime-valid verification procedure, including for multichain incumbents, and an unbounded separation from optimized common-potential verification. Second, we prove that finite-contraction prices identify the leading coefficients of the actual certificate excess. A tracking variant converts this relation into a stopping-time improvement, with a separate residual limit for growing historical coverage. Third, we test the full sequence from behavior-policy logs to a learned, frozen controller and certified supplementation. Matched comparisons separate certificate and acquisition effects, while an exact cancellation example delineates what this sufficient certificate can miss.

Prior-free average-reward stopping (Tuynman et al., 2024; Lee et al., 2025), structure-aware allocation (Zanette et al., 2019), and allocation tracking (Garivier & Kaufmann, 2016; Al Marjani & Proutiere, 2021) are established. Our allocation contribution is the connection from this policy-specific flow certificate to actual stopping cost. Its restricted-class efficiency theorem is distinct from the general safeguard in Section 5, which guarantees completion even when local scores are uninformative.

2 Problem and data

Let $M = (\mathcal{S}, \mathcal{A}, P, r)$ be a finite MDP, with S states, A actions per state, and rewards in $[0, 1]$. We index rows by $i = (s, a) \in \mathcal{I} = \mathcal{S} \times \mathcal{A}$ and write $d = SA$. For a stationary policy π, its gain from state s is

$$g_M^\pi(s) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_s^\pi\left[\sum_{t<T} R_t\right].$$

This limit exists in a finite MDP even when the policy is multichain. For $\varepsilon, \delta \in (0, 1)$, the goal is to return $\widehat{\pi}$ such that $\max_s[g_M^\star(s) - g_M^{\widehat{\pi}}(s)] \le \varepsilon$ with probability at least $1 - \delta$. In a communicating MDP, $g_M^\star(s) = g_M^\star$ is constant; its diameter D is the largest, over ordered state pairs, of the minimum expected hitting time. We first verify a retained incumbent; Section 5 adds policy proposals when it cannot meet the tolerance. The learner is not given D.

Historical and new observations. The historical data consist of prefixes of the reward/next-state observation stream of each row, with counts $n_{0,i}$ that can be zero or highly unequal. New observations may be requested directly, or obtained on a trajectory by choosing actions. In either case, the conditional law of the next observation at row i is the same stationary row law. This includes nonanticipatory logging and trajectory collection; it excludes outcome-filtered logs or unmodeled changes in dynamics. All observations, including those obtained while navigating, can update the row estimates.

Let $\mathcal{C}_t = \prod_i([r_{t,i}^-, r_{t,i}^+] \times \mathcal{P}_{t,i})$ be a nonempty rectangular confidence set, where $\mathcal{P}_{t,i} \subseteq \Delta_S$ is a polytope. We use a simultaneous event

$$\mathcal{E} = \{M \in \mathcal{C}_t \text{ at every evaluation time } t\}, \qquad \mathbb{P}(\mathcal{E}) \ge 1 - \delta. \tag{1}$$

The requirement concerns finite-dimensional row laws, not an unknown value function. Standard time-uniform row bounds suffice, so policy choice, allocation, and stopping can reuse the same data. Appendix A gives explicit constructions for unknown rewards, ℓ_1 transition balls, and coordinatewise binomial intervals. Known rewards or known structural zeros may be incorporated, but are not required by the theory.

Table 1 separates the validity of a stopping decision from the resources needed to reach it. Communication is a termination condition, not a prerequisite for evaluating an incumbent. Section 4 also extends the efficiency result to interior unknown reward means.

3 A certificate for the policy decision

For clarity, suppress t. Introduce a state–action occupancy u_i and transition flow y_{ij}. Define the stationary-flow polytope

$$\mathcal{F}(\mathcal{C}) = \left\{ (u, y) : \begin{array}{l} u \ge 0, \quad \sum_i u_i = 1, \quad \sum_j y_{ij} = u_i, \\ \sum_a u_{sa} = \sum_i y_{is} \quad (s \in \mathcal{S}), \\ y_i \in u_i \mathcal{P}_i \quad (i \in \mathcal{I}) \end{array} \right\}, \tag{2}$$

Table 1: Different tasks require different information. None of the learner's inputs includes a diameter, bias span, dynamical contraction bound, or relevance profile.

Task	Structural condition	What is obtained
Policy verification	Finite stationary MDP; simultaneous row confidence	Anytime-valid gain certificate for every initial state
Price-dependent efficiency	Full support; unique optimum; fixed deterministic incumbent; known rewards	Certificate-specific stopping constant
Guaranteed completion	True MDP communicating; unknown finite D	Residual ε^{-2} count envelope
Efficient trajectory collection	Same communication condition; a learned collector	Interaction bound including navigation learning
Route verification	Fixed proposed router; feasible hitting witness	A data-checkable expected hitting cost

where $0\mathcal{P}_i = \{0\}$. For a fixed policy, add $u_{sa} = \pi(a \mid s)\sum_b u_{sb}$ and denote the resulting polytope by $\mathcal{F}_\pi(\mathcal{C})$. For example, a probability box $\ell_{ij} \le p_j \le q_{ij}$ gives the linear constraints $\ell_{ij}u_i \le y_{ij} \le q_{ij}u_i$. An ℓ_1 ball similarly has an $O(S)$-size perspective formulation per row.

Starting from classical occupation-measure duality (Puterman, 1994), we obtain the following optimistic and robust-incumbent programs:

$$U(\mathcal{C}) = \max_{\mathcal{F}(\mathcal{C})} \sum_i r_i^+ u_i, \qquad L(\pi, \mathcal{C}) = \min_{\mathcal{F}_\pi(\mathcal{C})} \sum_i r_i^- u_i, \qquad C(\pi, \mathcal{C}) = U(\mathcal{C}) - L(\pi, \mathcal{C}). \tag{3}$$

The upper program asks whether some plausible recurrent behavior is better than the candidate. The lower program tests the candidate against every plausible recurrent class. In particular, a small transient reward does not automatically depress L.

Theorem 3.1 (Direct, anytime-valid decision certificate)**.** *On $\mathcal{E}$, simultaneously for every evaluation time, every data-selected stationary policy π, and every initial state s,*

$$g_M^\star(s) - g_M^\pi(s) \le C(\pi, \mathcal{C}_t). \tag{4}$$

Thus the probability of a finite data-dependent stop with $C(\widehat{\pi}, \mathcal{C}_t) \le \varepsilon$ and an incorrect returned policy is at most δ. Both programs have $O(S^2A)$ variables and constraints for row boxes or ℓ_1 balls. They require neither policy enumeration nor a supplied span or contraction bound.

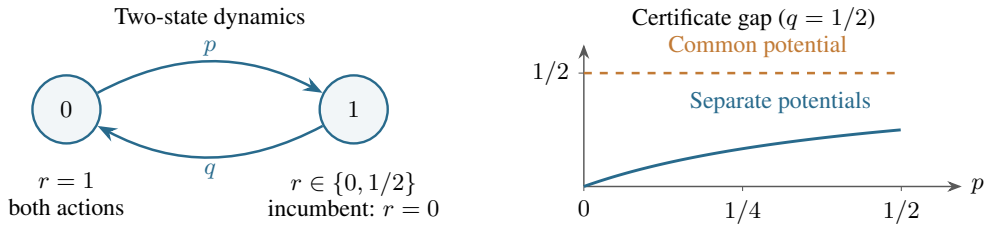
Proof idea. Every limiting expected state–action frequency in the true MDP induces a stationary flow in $\mathcal{F}(\mathcal{C}_t)$. Under π, it lies in $\mathcal{F}_\pi(\mathcal{C}_t)$. The two objectives therefore bound the corresponding gains. Conversely, divide each positive flow row by its occupancy to realize it with a plausible transition law. No division is needed for zero-mass rows. The common event (1) makes the argument valid after adaptive policy selection and stopping. Full details are in Appendix B. □

What is being optimized? The lower program fixes the candidate and varies plausible dynamics and long-run frequencies; the candidate may come from any controller or planner. The upper program is shared. At exact knowledge, $U = \max_s g_M^\star(s)$ and $L = \min_s g_M^\pi(s)$, giving worst-starting-state loss on a communicating model. Without communication, the gap can remain positive even for an optimal policy with different gains across starts. A common-potential bound requires the two dual witnesses to agree; shared-model regret instead evaluates both policies in the same plausible model. The former can exceed our separate bound (Proposition 3.2), while the latter can be smaller because uncertainty cancels (Appendix I).

Computation also supplies verification. For arbitrary finite potentials h, v, define

$$U_h = \max_{s,a}\{r_{sa}^+ + \max_{p \in \mathcal{P}_{sa}} p^\top h - h_s\}, \tag{5}$$

$$L_{\pi,v} = \min_s \sum_a \pi(a \mid s)\{r_{sa}^- + \min_{p \in \mathcal{P}_{sa}} p^\top v - v_s\}. \tag{6}$$



(a) Action-independent transitions.

(b) Separate gap: $p/(1+2p)$.

Figure 1: A local action gap can overstate a long-run policy gap. The same fixed incumbent is evaluated in both panels; self-loops are omitted on the left. For $p > 0$, the transitions in Proposition 3.2 have full support; $p = 0$ is shown only as a limit. As state 1 becomes rare, its inferior action has vanishing average cost, but the best common-potential certificate remains $1/2$. Separate potentials recover the exact loss.

Stationarity cancels the potential terms, giving $U \le U_h$ and $L \ge L_{\pi,v}$. Linear-program duality gives $U = \inf_h U_h$ and $L = \sup_v L_{\pi,v}$. The potentials need not agree. Row support evaluations check returned witnesses. Rigorous numerical validity requires outward error allowances; experiments instead use double-precision checks and a reserved 10^{-8} margin, not verified arithmetic (Appendix C).

Proposition 3.2 (A separation in full-support dynamics)**.** *For every fixed π, $C(\pi, \mathcal{C}) \le \inf_h[U_h - L_{\pi,h}]$. Consider two states with action-independent transitions $(1-p, p)$ from state 0 and $(q, 1-q)$ from state 1, where $0 < p, q < 1$. Both actions at 0 earn reward one; the actions at 1 earn zero and $1/2$. For the incumbent choosing the zero-reward action,*

$$C(\pi, \{M\}) = \frac{p}{2(p+q)}, \qquad \inf_h[U_h - L_{\pi,h}] = \frac{1}{2}.$$

Every policy induces the same irreducible, aperiodic chain. The ratio of the two certificates is unbounded as $p \downarrow 0$ with q fixed.

The incumbent sacrifices $1/2$ only on a $p/(p+q)$ fraction of visits. The flow programs recover this average loss. A common potential, however, must simultaneously justify the competitor's larger local reward and the incumbent's smaller one; their difference survives every choice of potential. Figure 1 illustrates the distinction, and Appendix D proves it by direct substitution. The comparison holds the incumbent fixed: an identification algorithm can instead propose the other action at state 1.

3.1 When can an existing policy be certified?

A valid certificate is useful only if it becomes small with enough relevant data. For an incumbent, this question depends on the two witnesses needed to verify that policy, rather than on the worst time needed to navigate the MDP.

Theorem 3.3 (Incumbent verification without supplied spans)**.** *Let the true MDP be communicating, and fix an incumbent π with loss $\Delta_\pi = g^\star_M - \min_s g^\pi_M(s) < \varepsilon$. Let $h^\star, v^\pi$ be exact-model upper and lower optimal dual witnesses, with spans $H_\star, H_\pi$. For the empirical ℓ_1 confidence sets, write $e_i(w, N) = 2b_R(N_i) + b_P(N_i)\,\mathrm{span}(w)$. On $\mathcal{E}$,*

$$C(\pi, \mathcal{C}_N) \le \Delta_\pi + \max_i e_i(h^\star, N) + \max_s \sum_a \pi(a \mid s) e_{sa}(v^\pi, N). \tag{7}$$

Let n_π be the smallest positive integer such that $4b_R(n_\pi) + (H_\star + H_\pi)b_P(n_\pi) \le \varepsilon - \Delta_\pi$. If π is re-evaluated after each design–safeguard pair, alternating arbitrary design requests with least-filled sampling certifies π, or stops earlier on another certified policy, within $2\sum_i [n_\pi - n_{0,i}]_+ + 2$ additional observations. In particular,

$$n_\pi = \widetilde{O}\left(\frac{(1 + H_\star + H_\pi)^2[S + \log(1/\delta)]}{(\varepsilon - \Delta_\pi)^2}\right).$$

The witnesses and their spans are used only in the bound, not supplied to the learner.

The proof tests the two uncertain programs with their respective true-model witnesses. A row error changes a potential expectation by at most its ℓ_1 error times half the potential span; subtracting the

resulting upper and lower bounds gives (7). Appendix D supplies each step. In the two-state family, $H_\star + H_\pi \le 3$ when $q = 1/2$. For $p \le \varepsilon/2$, verification therefore needs $\widetilde{O}(\varepsilon^{-2})$ observations per row, independently of the diverging diameter $1/p$. A common-potential test still has gap at least $1/2$. Appendix D.4 extends the distinction to finite-state action-independent chains. Section 5 treats the complementary case in which a suitable policy must first be found.

Retaining earlier evidence. We retain the smallest checked upper bound and the largest checked policy lower bound, together with that policy. Their gap is nonincreasing even for nonnested confidence sets. New proposals share one upper LP and require one lower LP each. Theorem 3.3 requires re-evaluating its incumbent, not merely retaining its old lower bound (Appendix C).

4 From an unresolved decision to residual acquisition

The certificate is also a source of acquisition information. Its upper and lower optimizers describe the two long-run explanations currently furthest apart. A row that has substantial mass in either explanation, and whose uncertainty constraint is binding, is a natural sampling target. The following formula makes that statement precise.

For a probability-box confidence set, fix its empirical center $\widehat{p}_i$ and contract row i by a scalar θ_i:

$$\ell_i(\theta_i) = \widehat{p}_i + \theta_i(\ell_i - \widehat{p}_i), \qquad q_i(\theta_i) = \widehat{p}_i + \theta_i(q_i - \widehat{p}_i).$$

Let u^+, u^- be optimal occupancies for U and $-L$, respectively, and let $\lambda^{+,q}, \lambda^{+,\ell}, \lambda^{-,q}, \lambda^{-,\ell} \ge 0$ be their dual multipliers for the upper/lower flow-row inequalities.

Theorem 4.1 (Decision uncertainty prices)**.** *Fix the policy, centers, and rewards. Suppose each LP has a differentiable primal–dual optimal selection near $\theta = \mathbf{1}$ (for example, a nonsingular optimal basis that stays optimal). Then the derivative of C with respect to θ_i, evaluated at $\theta = \mathbf{1}$, equals*

$$d_i = \sum_{\sigma\in\{+,-\}} u_i^\sigma \sum_j \left[\lambda_{ij}^{\sigma,q}(q_{ij} - \widehat{p}_{ij}) + \lambda_{ij}^{\sigma,\ell}(\widehat{p}_{ij} - \ell_{ij})\right] \ge 0. \tag{8}$$

For an ℓ_1 ball with radius ρ_i, the corresponding derivative with respect to ρ_i is $u_i^+\lambda_i^+ + u_i^-\lambda_i^-$. An unknown reward interval contributes the analogous occupancy-weighted endpoint derivatives.

The formula differentiates the primal–dual Lagrangian. Occupancy alone omits the value consequence of moving probability within a row. At degenerate optima a solver's multiplier remains a local score; a finite contraction avoids choosing among ambiguous derivatives. The next result connects those finite prices to the actual observed certificate under additional conditions.

A simple uncertainty price. In one state, with two reward intervals and a candidate that takes action one,

$$C = \max\{r_1^+, r_2^+\} - r_1^-.$$

When $r_2^+ > r_1^+$, only the competitor's upper endpoint and the incumbent's lower endpoint obstruct certification. Contracting toward the empirical means gives prices $d_1 = \widehat{r}_1 - r_1^-$ and $d_2 = r_2^+ - \widehat{r}_2$. In a transition problem, occupancy measures how often a row matters and the multiplier measures the consequence of tightening it.

For a batch budget B, approximate a row's remaining contribution by $d_i\sqrt{(N_i \vee 1)/(N_i + m_i)}$. This yields the residual subproblem

$$\min_{m\ge 0,\ \sum_i m_i = B} \sum_i \frac{\alpha_i}{\sqrt{N_i + m_i}}, \qquad \alpha_i = d_i\sqrt{N_i}, \quad N_i + m_i = \max\{N_i, \zeta\alpha_i^{2/3}\} \tag{9}$$

for positive counts. The water level ζ sets the budget. A largest discrete-marginal rule gives the exact integer optimizer; zero-count rows are handled by the exploration safeguard. This is a local design rule; it is not an instance-optimal sample-complexity characterization.

From a local price to a batch. For fixed empirical centers, replacing N_i by $N_i + m_i$ suggests a contraction $\theta_i = \sqrt{N_i/(N_i + m_i)}$. The first-order certificate expansion is a constant plus $\sum_i d_i\theta_i$, yielding (9). Appendix F gives the derivation and rounding bound. After collection the algorithm checks the observed certificate again, since new data also move its centers.

4.1 FROM PRICES TO VERIFICATION TIME

A safeguard can ensure completion even with uninformative design scores. To explain the benefit of prices, we study a restricted class: full-support transitions and a unique optimal policy make the active decision locally stable. Here prices determine the leading term of the *observed* certificate, including the effect of moving empirical centers.

Assume known rewards and unknown transitions with $P_i(j) > 0$. Let ψ be the unique optimal deterministic policy and fix a deterministic incumbent π, possibly selected from historical data and then retained. For $\nu \in \{\psi, \pi\}$, let u^ν be its true stationary state–action occupancy and h^ν an average-reward bias. Define

$$w_i = \tfrac{1}{2}\big[u_i^\psi \operatorname{span}(h^\psi) + u_i^\pi \operatorname{span}(h^\pi)\big], \qquad q_i^\star = \frac{w_i^{2/3}}{\sum_j w_j^{2/3}}, \tag{10}$$

and suppose some $w_i > 0$. Per unit ℓ_1 radius, the largest first-order gain change is half the bias span times the row occupancy: probability moves from a lowest-bias to a highest-bias successor. The two policies give the two terms in w_i.

At each evaluation, DCR-TRACK contracts row i's transition ball by a fixed factor $1 - \kappa$, obtaining $\mathcal{C}_N^{i,\kappa}$, and estimates

$$\widehat{w}_i = \frac{C(\pi, \mathcal{C}_N) - C(\pi, \mathcal{C}_N^{i,\kappa})}{\kappa b_P(N_i)}, \qquad 0 < \kappa < 1.$$

It tracks cumulative proportions proportional to $\widehat{w}_i^{2/3}$. Square-numbered requests cycle through all rows, giving infinite coverage with a vanishing sampling fraction. Checks have vanishing relative spacing, and verify the fixed incumbent with the current certificate and positive buffer $C_t + (1 + \sum_i N_i(t))^{-1} \le \varepsilon$.

Theorem 4.2 (Price-dependent stopping for a fixed incumbent)**.** *Under the preceding assumptions, use the empirical ℓ_1 balls, fixed δ, and finite historical counts. Let τ count added direct observations, including forced requests, and put $\Delta_\pi = g_M^\star - g_M^\pi$, $m = \varepsilon - \Delta_\pi > 0$, and $T_m = 8m^{-2}\log(1/m)$. On an event of probability at least $1 - \delta$,*

$$\lim_{m \downarrow 0} \frac{\tau_{\mathrm{Track}}(\Delta_\pi + m)}{T_m} = \left(\sum_i w_i^{2/3}\right)^3, \qquad \lim_{m \downarrow 0} \frac{\tau_{\mathrm{uniform}}(\Delta_\pi + m)}{T_m} = d\left(\sum_i w_i\right)^2. \tag{11}$$

The uniform comparator uses the same incumbent, confidence sets, checks, and buffer. Their stopping-time ratio therefore tends to $(\sum_i w_i^{2/3})^3/[d(\sum_i w_i)^2] \le 1$, strictly below one unless all d weights are equal.

Why this concerns observations. Appendix G proves, for arbitrary diverging row counts, $C_t - \Delta_\pi = (1 + o(1))\sum_i w_i b_P(N_i(t))$ almost surely. Smooth gain perturbation gives the coefficients; the unique optimal policy eventually remains optimal throughout the confidence set. Empirical-center errors are $O(\sqrt{\log\log n/n})$, smaller than the chosen $b_P(n) \sim 2\sqrt{\log n/n}$. Finite contraction prices consistently estimate w_i, and tracking yields $N_i(t)/t \to q_i^\star$. The buffer forces first passing times to diverge as m decreases to zero. Appendix H gives an explicit stable neighborhood and finite error bounds.

Tracking and convex allocation are established tools (Garivier & Kaufmann, 2016; Al Marjani & Proutiere, 2021; Boyd & Vandenberghe, 2004); the additional step is relating them to this observed certificate. This is a fixed-δ, small-margin limit, not finite-sample dominance or an expected-time bound. It applies to DCR-TRACK, whose vanishing safeguard differs from the half-least-filled rule. Joint prices extend the result to unknown interior reward means (Corollary G.5). Direct empirical-model weights attain the same limit (Corollary G.6); experiments examine their different behavior before the confidence set stabilizes.

Historical coverage and the remaining cost. Fixed histories vanish from the leading asymptotic term. To retain their contribution, let $n_{0,i}/T_m \to \lambda_i$ while the historical total diverges. The variant

DCR-FROZEN freezes prices after a negligible pilot. Its added cost divided by T_m converges to $B^*(\lambda)$ (Theorem G.4), where x_i denotes total row counts on the same scale:

$$B^*(\lambda) = \min_{x \geq \lambda} \left\{ \sum_i (x_i - \lambda_i) : \sum_{i:w_i>0} \frac{w_i}{\sqrt{x_i}} \leq 1 \right\}.$$

If the history already meets the constraint, $B^*(\lambda) = 0$ and $x^* = \lambda$. Otherwise the optimal target credits each row's existing coverage: $x_i^* = \max\{\lambda_i, cw_i^{2/3}\}$, with c chosen to meet the constraint. The algorithm uses observed counts and prices, not m or λ. Appendix K distinguishes this certificate-specific limit from information-theoretic benchmarks.

5 COMPLETION BEYOND THE REGULAR CLASS

Full support and unique optimality explain the price-tracking rate; they are not needed to validate the certificate. A broader communicating MDP may require a different safeguard and new policy proposals. The general DCR variant (Algorithm 1) combines the local design with equal numbers of least-filled requests. An optimistic-witness proposal uses the anchored iteration of Lee et al. (2025); Appendix E specifies and proves its residual test without supplying a diameter or bias span.

Let b_R, b_P be the simultaneous radii in Appendix A, and let $n_\star$ be the smallest integer satisfying $2b_R(n_\star) + 2Db_P(n_\star) \leq 3\varepsilon/4$. These are analysis quantities.

Theorem 5.1 (General residual completion)**.** *On $\mathcal{E}$, alternating arbitrary design and least-filled requests, with the optimistic-witness proposal checked after each pair, returns an ε-optimal policy after at most*

$$2\sum_i [n_\star - n_{0,i}]_+ + 2$$

additional direct observations, where

$$n_\star = \widetilde{O}\big((D+1)^2[S + \log(1/\delta)]/\varepsilon^2\big)\,. \tag{12}$$

The rule does not use D or $n_\star$ and can stop immediately on sufficient historical evidence.

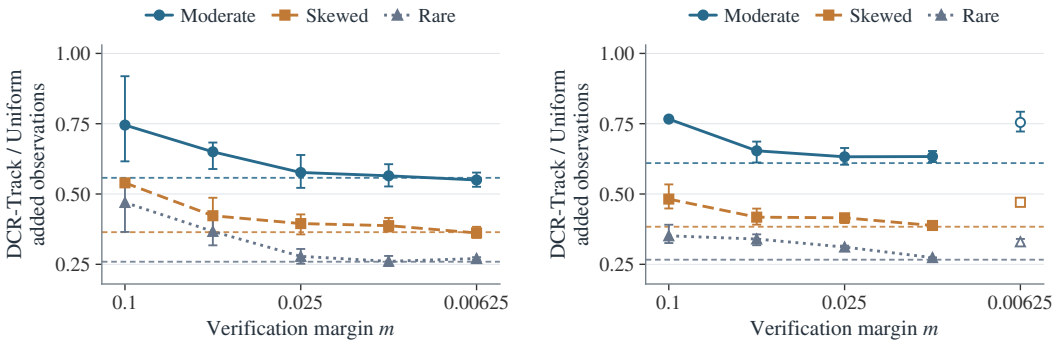
Key argument. Anchoring supplies a potential with residual at most $\varepsilon/4$ and span at most $2D$. Its lower-greedy policy has Bellman gap at most $\varepsilon/4 + \max_i[2b_R(N_i) + 2Db_P(N_i)]$, and the separate flow bound is no larger. Before all rows reach $n_\star$, each safeguard request reduces the missing-count sum by one. Appendix E gives the complete construction. □

This guarantee holds for arbitrary scores. It is intentionally distinct from the price-dependent asymptotic theorem: its constant exploration fraction buys general completion, while DCR-Track uses a vanishing fraction to attain its allocation limit.

Trajectory collection. Direct row requests do not include travel costs. GOSPRL (Tarbouriech et al., 2021) learns to satisfy a fixed demand vector of total size B in $\widetilde{O}(DB + D^{3/2}S^2A)$ interactions. Appendix J applies this existing collector in frozen-demand phases and gives a separate hitting-witness test. The empirical router evaluated below is a lighter procedure and does not inherit the collector's bound.

6 EXPERIMENTS: PREDICTED RATES AND LOGGED CONTROLLERS

We test stopping costs and controllers learned from behavior-policy logs. The rate study uses DCR-TRACK: ℓ_1 balls, repeated finite contractions, sparse forcing, and current-certificate checks. The logged-controller study uses DCR: probability boxes, dual prices, half least-filled sampling, and retained complete gaps. DCR-FROZEN instead freezes prices after a pilot for the growing-history limit; its diagnostic is separate. Protocols and raw records accompany the paper.



(a) Known rewards.

(b) Unknown rewards.

Figure 2: Actual sample costs approach the certificate-specific predictions. Curves with markers show ratios of mean added observations over eight paired seeds for each model; horizontal dashed lines are theoretical limits. Error bars are 95% paired seed-bootstrap intervals within a model. Hollow points use capped Uniform costs in the denominator. Smaller margins require more precise verification.

Does the predicted stopping constant appear in data? Three full-support, three-state MDPs have progressively more unequal stationary occupancies. On eight paired seeds per model, the learner estimates every transition row and verifies a fixed suboptimal incumbent. Figure 2 compares the observed stopping ratios with Theorem 4.2. With known rewards and margin $m = 0.00625$, the ratios are 0.550, 0.362, and 0.270, against predicted limits 0.558, 0.364, and 0.259. The thresholds $\Delta_\pi + m$ and true coefficients are used only for this post-hoc mechanism diagnostic; acquisition receives neither.

With unknown Bernoulli rewards and joint uncertainty prices, the smallest margin at which both methods always stop is 0.0125. Ratios are 0.633, 0.388, and 0.274, versus the corresponding limits 0.610, 0.384, and 0.266. At $m = 0.00625$, DCR-TRACK stops in all 24 cases, while Uniform reaches the 2^{24} cap without stopping; those capped comparisons remain visible. These controlled tests support the asymptotic mechanism, not broad task generalization. Appendix M gives all margins, counts, numerical checks, and computation costs.

A simpler allocation estimate. Plug-in empirical occupancies and biases attain the same first passing checkpoints as finite prices in every known-reward case at $m = 0.00625$. With unknown rewards at $m = 0.0125$, however, plug-in passes in only 4 of 24 cases, versus 24 for finite prices; its mean capped costs are 5.13–9.86 times larger. The empirical optimum is correct at the terminal budget, yet a row assigned zero plug-in weight still supports a plausible competitor. Plug-in is cheaper to compute: about 3.66 ms versus 22 ms per check on these small problems. Appendix M.1 gives every comparison and its scope.

From behavior-policy logs to a frozen controller. We use maintenance, queueing, RiverSwim, and a cyclic-patrol family, each at five and twelve states with two parameter settings: 16 tasks in total. A continuing trajectory of 8,192 or 65,536 steps follows a myopic reward policy with 0.1 or 0.5 random-action exploration. An empirical planner learns a controller solely from that log, then freezes it. Eight seeds yield 512 incumbents. All remain in the analysis, including the 63 short-history and 50 long-history policies whose true loss exceeds 0.1. Rewards are known in this study; all transition rows and their successor support are unknown.

We then supplement each history through a generative row oracle. DCR, the Initial-price baseline, and Uniform share the same incumbent, historical observations, and row streams. At every checkpoint, both separate and globally optimized common-potential certificates use that exact dataset and retain minimum complete gaps. This separates acquisition effects from certificate effects without changing the policy or retention convention (Table 2). Appendix O details the 39,936 checks and independent replay.

Table 2: Controllers from behavior-policy logs at $\varepsilon = 0.1$. Each historical length $H = \sum_i n_{0,i}$ has 256 incumbent cases, all included. Costs are mean capped added observations, assigning 2^{21} to unpassed cases. Acquisition methods share the historical trajectory prefix and supplemental row streams. On each acquisition path, both checkers use the same frozen policy and data and retain the smallest complete gap.

		Separate potentials		Optimized common potential	
H	Acquisition	Passes	Capped cost	Passes	Capped cost
8,192	Uniform	192/256	701,380	179/256	942,034
	Initial price	192/256	694,972	168/256	1,031,059
	DCR	192/256	626,236	153/256	1,044,430
65,536	Uniform	206/256	556,875	206/256	796,887
	Initial price	206/256	538,070	196/256	819,073
	DCR	206/256	490,895	189/256	916,436

Benefits depend on the certificate being improved. For the separate certificate, DCR lowers mean capped additional counts by 10.7% and 11.8% relative to Uniform at the two historical sizes, and by 9.9% and 8.8% relative to the initial prices. All three methods certify the same number of policies by the cap: 192 and 206 of 256. No observed stop accepts a policy with loss above tolerance. Along those same DCR sampling paths, however, the common-potential checker has *higher* capped costs than under Uniform, by 10.9% and 15.0%, and certifies fewer policies. Prices help their specified flow objective; they do not universally improve another verifier. Longer histories also change the learned controller, so their comparison is not a coverage-only causal effect.

Additional evidence and remaining costs. A historical-scaling diagnostic exposes delays from inaccurate frozen prices and finite-radius corrections (Appendix N); it does not establish convergence of the residual costs. Earlier Bellman and oracle-incumbent diagnostics remain in Appendices S.4 and S.5. Navigation is also unresolved: at 65,536 RiverSwim trajectory steps, the empirical router has mean policy loss 0.2278, versus 0.1161 for random actions, despite a tighter certificate (Appendix R.5).

7 Related work and implications

Hybrid RL and offline-data-dependent exploration (Song et al., 2023; Wagenmaker & Pacchiano, 2023) motivate retaining historical coverage. Offline best-arm identification already incorporates counts into information constraints and tracking (Agrawal et al., 2023); BESPOKE and KLB-TS exploit nonuniform policy-identification difficulty (Zanette et al., 2019; Al Marjani & Proutiere, 2021). Our result concerns fixed-incumbent average-reward verification and the actual flow certificate, with a shrinking verification margin at fixed confidence rather than best-policy identification as $\delta \downarrow 0$.

Prior-free average-reward stopping (Tuynman et al., 2024; Lee et al., 2025) and sharp discovery bounds (Zurek & Chen, 2024; 2025; Wang et al., 2024) assume generative sampling, with different span, diameter, or mixing conditions. Interval evaluation (Tewari & Bartlett, 2007), robust average-reward evaluation (Wang et al., 2023), and maximum regret in reward-uncertain discounted MDPs (Regan & Boutilier, 2009) also precede this paper. Robust evaluation, convex allocation, and tracking are prior tools. Our contribution connects them through the observed-certificate expansion, finite-price consistency, and actual stopping and residual-history constants. The fixed-incumbent common-potential separation explains why certificate choice matters. These results do not rank complete policy-identification algorithms.

8 Conclusion

Two flow programs turn uncertainty about a controller into a checkable decision and sampling prices. In a regular class, these prices yield explicit stopping and residual-history constants. A different safeguard ensures general completion. Experiments show targeted sample savings and their limits: gains depend on the certificate, and tighter certificates need not produce better controllers.

AI USE STATEMENT

Generative AI tools were used extensively in checking proofs, implementing and debugging code, polishing the manuscript. The submitting authors are responsible for the final text, mathematical claims, code, and interpretation.

ETHICS STATEMENT

Our experiments use synthetic Markov decision processes and simulated behavior-policy logs; they involve no human participants or personal data. The proposed certificates assess policy performance under explicit statistical and modeling assumptions. A certificate of near-optimal average reward does not guarantee safety, fairness, or satisfaction of constraints that are absent from the model. In particular, distribution shift, model misspecification, and selective logging can invalidate the stated guarantees. Applications with consequential real-world effects therefore require domain-specific validation and safeguards beyond these certificates.

REPRODUCIBILITY STATEMENT

Complete proofs, assumptions, algorithms, and experimental task definitions are included in the appendix. A separate supplementary package contains the implementation and the raw observations summarized in the figures and tables; the paper source compiles independently of that package.

REFERENCES

Shubhada Agrawal, Sandeep Juneja, Karthikeyan Shanmugam, and Arun Sai Suggala. Optimal best-arm identification in bandits with access to offline data, 2023. URL https://arxiv.org/abs/2306.09048.

Aymen Al Marjani and Alexandre Proutiere. Adaptive sampling for best policy identification in Markov decision processes. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7459–7468. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/marjani21a.html.

J. Frédéric Bonnans and Alexander Shapiro. *Perturbation Analysis of Optimization Problems*. Springer, 2000. doi: 10.1007/978-1-4612-1394-9. URL https://doi.org/10.1007/978-1-4612-1394-9.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511804441. URL https://web.stanford.edu/~boyd/cvxbook/.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 998–1027. PMLR, 2016. URL https://proceedings.mlr.press/v49/garivier16a.html.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010. URL https://www.jmlr.org/papers/v11/jaksch10a.html.

Jongmin Lee, Mario Bravo, and Roberto Cominetti. Near-optimal sample complexity for MDPs via anchoring. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 32907–32929. PMLR, 2025. URL https://proceedings.mlr.press/v267/lee25c.html.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994. doi: 10.1002/9780470316887.

Kevin Regan and Craig Boutilier. Regret-based reward elicitation for Markov decision processes. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pp. 444–451, 2009. URL https://www.cs.toronto.edu/~cebly/Papers/Regan_Boutilier_uai09.pdf.

Kevin Regan and Craig Boutilier. Robust policy computation in reward-uncertain MDPs using nondominated policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pp. 1127–1133, 2010. doi: 10.1609/aaai.v24i1.7740. URL https://ojs.aaai.org/index.php/AAAI/article/view/7740.

Yuda Song, Yifei Zhou, Ayush Sekhari, Drew Bagnell, Akshay Krishnamurthy, and Wen Sun. Hybrid RL: using both offline and online data can make RL efficient. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=yyBis80iUuU.

Alexander L. Strehl and Michael L. Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008. doi: 10.1016/j.jcss.2007.08.009. URL https://doi.org/10.1016/j.jcss.2007.08.009.

Jean Tarbouriech, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. A provably efficient sample collection strategy for reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 7611–7624, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/3e98410c45ea98addec555019bbae8eb-Abstract.html.

Ambuj Tewari and Peter L. Bartlett. Bounded parameter Markov decision processes with average reward criterion. In *Learning Theory*, volume 4539 of *Lecture Notes in Computer Science*, pp. 263–277. Springer, 2007. doi: 10.1007/978-3-540-72927-3_20. URL https://dept.stat.lsa.umich.edu/~tewaria/research/tewari07bounded.pdf.

Adrienne Tuynman, Rémy Degenne, and Emilie Kaufmann. Finding good policies in average-reward Markov decision processes without prior knowledge. In *Advances in Neural Information Processing Systems*, volume 37, pp. 109948–109979, 2024. doi: 10.52202/079017-3489. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c6b2921f24f82dc05cda53bbe50bab90-Abstract-Conference.html.

Andrew Wagenmaker and Aldo Pacchiano. Leveraging offline data in online reinforcement learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35300–35338. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/wagenmaker23a.html.

Shengbo Wang, Jose Blanchet, and Peter Glynn. Optimal sample complexity for average reward Markov decision processes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jOm5p3q7c7.

Yue Wang, Alvaro Velasquez, George Atia, Ashley Prater-Bennette, and Shaofeng Zou. Robust average-reward Markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 15215–15223, 2023. doi: 10.1609/aaai.v37i12.26775. URL https://ojs.aaai.org/index.php/AAAI/article/view/26775.

Andrea Zanette, Mykel J. Kochenderfer, and Emma Brunskill. Almost horizon-free structure-aware best policy identification with a generative model. In *Advances in Neural Information Processing Systems*, volume 32, pp. 5625–5634, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/a724b9124acc7b5058ed75a31a9c2919-Abstract.html.

Matthew Zurek and Yudong Chen. Span-based optimal sample complexity for weakly communicating and general average reward mdps. In *Advances in Neural Information Processing Systems*, volume 37, pp. 33455–33504, 2024. doi: 10.52202/079017-1055. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3acbe9dc3a1e8d48a57b16e9aef91879-Abstract-Conference.html.

Matthew Zurek and Yudong Chen. The plug-in approach for average-reward and discounted MDPs: Optimal sample complexity analysis. In Gautam Kamath and Po-Ling Loh (eds.), *Proceedings of The 36th International Conference on Algorithmic Learning Theory*, volume 272 of *Proceedings of Machine Learning Research*, pp. 1386–1387. PMLR, 24–27 Feb 2025. URL https://proceedings.mlr.press/v272/zurek25a.html.

CONTENTS

A Statistical model and simultaneous confidence **17**
A.1 Notation, gain, and the sampling filtration 17
A.2 Explicit confidence sets for unknown rewards and transitions 17
A.3 Probability boxes for practical DCR experiments 18
A.4 Data reuse, nesting, and retained evidence 19

B Stationary-flow geometry and policy guarantees **19**
B.1 Perspective constraints, zero occupancies, and realizability 19
B.2 Limiting occupancy for a fixed true policy 20
B.3 Proof of the direct certificate theorem 20
B.4 What the certificate measures 20
B.5 Exact meaning of the two uncertain extrema 21

C Dual witnesses and strict incumbent-wise tightening **21**
C.1 Weak duality by cancellation 21
C.2 Strong duality and finite formulations 22
C.3 Checked approximate solutions 22
C.4 A transient limiting example 23
C.5 A finite candidate portfolio 23
C.6 Retaining bounds across nonnested confidence sets 23

D How much data does an incumbent need? **23**
D.1 Exact-model witnesses and row perturbations 24
D.2 Proof of the incumbent verification bound 24
D.3 Rare recurrent decisions: a separation with full-support transitions 25
D.4 A finite-state class with an exact verification gap 26

E Planning and completion with unknown diameter **27**
E.1 Existence of a bounded optimistic potential 27
E.2 A finite residual solver with an explicit bound 27
E.3 From an observed residual to a policy certificate 28
E.4 Proof of the residual sample-count theorem 29
E.5 Almost-sure termination of the unrestricted count implementation 30
E.6 Batches, checking frequency, and the implemented planner 30

F Decision prices and residual design **30**
F.1 Perturbing a row confidence set 30
F.2 Degeneracy, inactive rows, and boundary effects 31
F.3 A finite alternative at degenerate optima 32
F.4 Continuous residual allocation 32
F.5 Exact integer allocation and the implemented rounding rule 33
F.6 Positive floors and arbitrary offline coverage 33
F.7 The general safeguarded variant 34

G From uncertainty prices to actual verification time **34**
G.1 Model, events, and limiting coefficients 34
G.2 An expansion for the actual empirical certificate 35
G.3 Finite contraction prices consistently estimate the coefficients 36
G.4 Cumulative tracking with sparse forced observations 36
G.5 Proof of the stopping-time theorem 37
G.6 Historical data on the critical scale 38
G.7 Unknown rewards with interior means 39
G.8 An empirical-model alternative 40
G.9 Boundaries of the result 40

H A quantitative neighborhood for the certificate expansion **41**
H.1 Finite perturbation of a fixed-policy gain 41
H.2 Policy stability and the observed certificate 42
H.3 A finite error bound for contraction prices 42
H.4 A sufficient time to enter the neighborhood 43
H.5 A finite tail bound for relative center errors 43

I What the separate certificate can miss **44**

J Unknown navigation and trajectory realization **45**
J.1 A sample-collection reduction with learned dynamics 45
J.2 Doubling floors with frozen design demands 46
J.3 A data-checkable hitting-time witness 47
J.4 The implemented router and its qualitative safeguard 47

K An adaptive information constraint after a warm start **48**
K.1 Transcript divergence with historical observations 48
K.2 Alternatives that require different decisions 49
K.3 A two-action example with an explicit residual law 49
K.4 Why navigation is not multiplied into this lower bound 50

L Transient effects, exact examples, and interpretation **50**
L.1 A rare gateway with constant decision relevance 50
L.2 Transient mistakes that do not change the gain 50
L.3 A limitation from decoupling the plausible models 51

M Testing the price-dependent stopping theorem **51**
M.1 A simpler empirical-model allocation baseline 52

N A bounded test of residual historical coverage **54**

O Controllers learned from behavior-policy logs and then frozen **56**

P Finite implementations and numerical objectives **59**
P.1 Compact programs without policy enumeration 59
P.2 Policy proposals and the role of smoothing 59
P.3 How dual prices are extracted 59
P.4 Numerical termination and fallback behavior 60
P.5 A finite LP for the hitting witness 60

Q Experimental tasks and evaluation protocol **60**
Q.1 Three continuing-control families 60
Q.2 Historical counts and generative completion 61
Q.3 Allocation methods and fair policy comparison 61
Q.4 True policy loss and multichain evaluation 62
Q.5 A continuous-trajectory comparison 62
Q.6 Evaluation metrics and uncertainty 62

R Additional empirical results **63**
R.1 Stopping counts by task and accuracy 63
R.2 Certificate contraction and fixed-budget policy loss 64
R.3 Proposal-pool ablation 65
R.4 Trajectory results and censoring 65
R.5 Numerical tie handling and trajectory sensitivity 65
R.6 Finite computation and routing witnesses 67
R.7 Reading the empirical evidence 67

S Additional verification and allocation studies **68**
S.1 A finer batching and checking schedule 68
S.2 Optimizing the common-potential baseline 68
S.3 Finite-data verification in the full-support example 69
S.4 Operational stopping and an isolated price-refresh ablation 69
S.5 Fixed incumbents and historical coverage in the control tasks 71

T Limitations and research directions **73**

NOTATION AT A GLANCE

Symbol	Meaning
$\mathcal{I} = \mathcal{S} \times \mathcal{A}$, $d = SA$	State–action rows and their number
$n_{0,i}$, N_i, m_i	Historical, total, and requested additional row counts
$\mathcal{C}$, $\mathcal{P}_i$	Rectangular model set and transition set for row i
r_i^-, r_i^+; ℓ_i, q_i	Reward interval and coordinatewise probability bounds
u_i, y_{ij}	Stationary row occupancy and flow to successor j
$g_M^\pi(s)$, $g_M^\star(s)$	Candidate and optimal average reward from state s
U, $L(\pi)$, $C(\pi)$	Optimistic gain, incumbent lower gain, and their gap
h, v; U_h, $L_{\pi,v}$	Separate potential witnesses and their checked bounds
$\mathrm{span}(w)$	$\max_s w_s - \min_s w_s$
D; $H_\star, H_\pi$	Unknown diameter; true-model witness spans
d_i, α_i	Local uncertainty price; allocation weight $d_i\sqrt{N_i}$
w_i, $\widetilde{w}_i$	Leading certificate coefficients for known and unknown rewards
$q_i^\star$	Asymptotic fraction of new requests to row i
Δ_π, $m = \varepsilon - \Delta_\pi$	True incumbent loss and verification margin
$T_m = 8m^{-2}\log(1/m)$	Reference scale for the small-margin limit
λ_i, $B^*(\lambda)$	Scaled historical count and optimal residual certificate cost
ε, δ; $\mathcal{E}$	Accuracy, failure probability, and simultaneous coverage event

Symbols suppressed in U, L, and C refer to the current confidence set. Lower bounds retained across checks stay attached to the policy that produced them. Probability-box bounds q_i, allocation proportions $q_i^\star$, and the scalar q in the two-state example have distinct roles.

WHICH SAMPLING RULE IS USED?

Rule	Purpose and guarantee
DCR	Repeated local prices with half the requests assigned to least-filled rows. It supports general completion with policy proposals and is used for the behavior-policy log experiment.
DCR-TRACK	Repeated finite-contraction prices, cumulative-target tracking, and square-numbered forced requests. It verifies a fixed incumbent and attains the small-margin constant in Theorem 4.2.
DCR-FROZEN	Prices estimated after a negligible pilot and then frozen; discrete residual allocation with square-numbered forced requests. It is the theoretical growing-history variant in Theorem G.4.
Initial-price baseline	Initial prices frozen without the theoretical pilot, with the same half-safeguard as practical DCR. This empirical baseline is distinct from DCR-FROZEN.
Plug-in baseline	Empirical occupancies and biases substituted into the theoretical weights, with the same tracking and square-numbered safeguard as DCR-TRACK. It has the same limiting constant but can behave differently at finite budgets.

Guide to the appendix. Appendices A–C establish the confidence event and the two certification programs. Appendix D proves the witness-span bound and common-potential separation. The price-dependent stopping and historical-coverage results are proved in Appendix G; Appendix I analyzes shared-model cancellations. General completion, navigation, and information arguments have separate sections. For experiments, Appendix M tests the stopping constants, Appendix N examines historical scaling, and Appendix O follows behavior-policy logs through frozen-controller certification. Earlier control and routing studies are retained afterward.

A Statistical model and simultaneous confidence

A.1 Notation, gain, and the sampling filtration

Write $d = SA$ and $i = (s, a)$. The transition vector P_i belongs to the probability simplex Δ_S; $r_i \in [0, 1]$ is a mean reward, rather than a realized reward. Rewards and successors observed on the same step may be dependent. Their joint row distribution, denoted ν_i, does not change between the historical and online phases.

For a stationary randomized policy π, let $P_\pi(s, j) = \sum_a \pi(a \mid s) P_{sa,j}$ and $r_\pi(s) = \sum_a \pi(a \mid s) r_{sa}$. Its gain vector is

$$g_M^\pi = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} P_\pi^t r_\pi. \tag{13}$$

This limit exists for every finite stochastic matrix, including periodic and multichain matrices. The gain need not be constant. A finite MDP admits a stationary deterministic policy that maximizes gain at every state; $g_M^\star(s)$ denotes that optimal value. Communication makes $g_M^\star$ constant, but that property is not used to validate the decision certificate.

The historical observations at row i are a prefix of a sequence with law ν_i. Subsequent visits reveal further elements of that sequence. Equivalently, the action at a trajectory step is chosen before the reward and successor are observed, and their conditional law given the past and the selected row is ν_i. A generative query chooses both components of i; a trajectory query chooses only the action at the current state. All observed transitions count, including those collected while traveling towards another row.

The row-stream construction can be used to couple the experiments across methods. It is not an additional independence assumption on the successive states of a trajectory. Those states are dependent. The construction says that an unused random innovation is drawn according to the appropriate row law whenever that row is visited. Sampling decisions are predictable with respect to the past innovations.

Let $N_i(t)$ count all observations of row i, including $n_{0,i}$ historical observations. Historical counts may be heterogeneous, zero, or themselves determined by a predictable logging trajectory. A uniform-in-prefix confidence event is required before conditioning on these random counts. It would generally be incorrect to apply a fixed-sample confidence statement conditional on an arbitrary observation-dependent stopping count. It would also be incorrect to treat selectively retained transitions as an unfiltered row prefix.

A.2 Explicit confidence sets for unknown rewards and transitions

For $n \geq 1$, define

$$a_n = \frac{\delta}{2dn(n+1)}, \qquad b_R(n) = \min\left\{1, \sqrt{\frac{\log(2/a_n)}{2n}}\right\},$$
$$b_P(n) = \min\left\{2, \sqrt{\frac{2[S\log 2 + \log(1/a_n)]}{n}}\right\}. \tag{14}$$

For a row with n observations use

$$I_i^R(n) = [\widehat{r}_i(n) - b_R(n), \widehat{r}_i(n) + b_R(n)] \cap [0, 1], \tag{15}$$
$$\mathcal{P}_i(n) = \{p \in \Delta_S : \|p - \widehat{P}_i(n)\|_1 \leq b_P(n)\}. \tag{16}$$

At zero observations, set $I_i^R(0) = [0, 1]$ and $\mathcal{P}_i(0) = \Delta_S$. Known rewards replace I_i^R by a singleton; known structural zeros may intersect $\mathcal{P}_i$ with their corresponding face of the simplex.

Lemma A.1 (Simultaneous row event)**.** *The preceding sets satisfy*

$$\mathbb{P}\{r_i \in I_i^R(n),\ P_i \in \mathcal{P}_i(n) \text{ for every } i \text{ and every } n \geq 0\} \geq 1 - \delta.$$

In particular, the event holds at all adaptively selected evaluation times under either access model.

Proof. For a fixed row and positive prefix length, Hoeffding's inequality gives $\mathbb{P}(|\widehat{r}_i - r_i| > b_R(n)) \leq a_n$ before clipping; clipping to the natural range cannot exclude the mean. For any zero-mass vector $z \in \mathbb{R}^S$,

$$\|z\|_1 = \max_{v \in \{-1,1\}^S} v^\top z.$$

For each sign vector v, the samples $v_{S'}$ lie in $[-1, 1]$. The one-sided Hoeffding bound and a union bound over 2^S sign vectors yield

$$\mathbb{P}\{\|\widehat{P}_i - P_i\|_1 > x\} \leq 2^S \exp(-nx^2/2).$$

When the unclipped radius is below two, substitution of that radius gives failure probability at most a_n. Otherwise the clipped radius is two and the failure event is empty, because two probability vectors have ℓ_1 distance at most two. The analogous reward clipping at one also makes its failure event empty whenever clipping is active. A sum over the two types of failure, d rows, and $n \geq 1$ is at most

$$2d \sum_{n \geq 1} a_n = \delta \sum_{n \geq 1} \frac{1}{n(n+1)} = \delta.$$

The resulting event is about all row prefixes. Looking up a random prefix on this same event requires no additional conditioning or sample splitting. □

The radii decrease to zero and are nonincreasing for the constants above. To see monotonicity before clipping, differentiate $(c + \log x + \log(x + 1))/x$ for $x \geq 1$; the derivative has the sign of $1 + x/(x + 1) - c - \log x - \log(x + 1)$. This expression is negative at $x = 1$ for the present $c \geq \log 4$, and its derivative $1/(x + 1)^2 - 1/x - 1/(x + 1)$ is negative for $x \geq 1$. The polynomial union allocation is chosen for a transparent proof, not for the smallest logarithmic factor. A valid confidence sequence with a sharper logarithm can replace it without changing the flow programs or the safety proof.

A.3 Probability boxes for practical DCR experiments

The behavior-policy log and earlier practical-DCR studies use probability boxes. The price-tracking mechanism studies instead use the ℓ_1 balls above. For coordinate j of row i, let $K_{ij}(n)$ be its successor count and choose the one-sided error level

$$\alpha_n = \frac{\delta}{2dSn(n+1)}. \tag{17}$$

The lower and upper Clopper–Pearson endpoints are

$$\ell_{ij}(n) = \begin{cases} 0, & K_{ij} = 0, \\ B^{-1}(\alpha_n; K_{ij}, n - K_{ij} + 1), & K_{ij} > 0, \end{cases} \tag{18}$$

$$q_{ij}(n) = \begin{cases} 1, & K_{ij} = n, \\ 1 - B^{-1}(\alpha_n; n - K_{ij}, K_{ij} + 1), & K_{ij} < n, \end{cases} \tag{19}$$

where $B^{-1}(\cdot; a, b)$ is a beta quantile. At $n = 0$, the interval is $[0, 1]$. The row set is

$$\mathcal{P}_i^{\mathrm{box}}(n) = \{p \in \Delta_S : \ell_{ij}(n) \leq p_j \leq q_{ij}(n),\ j \in \mathcal{S}\}.$$

Each coordinate count is binomial even though different coordinates at the same prefix are dependent. Union bounding the marginal interval failures therefore gives simultaneous coverage over all rows, coordinates, and prefix lengths. No independence across successor coordinates is used. Every empirical row lies in its probability box, so the intersection with the simplex is nonempty.

The box-based experiments have known rewards and assign the full error budget to transitions. With unknown rewards, splitting δ between boxes and reward intervals would give a joint event; the unknown-reward tracking study uses ℓ_1 balls instead. Boxes and balls can also be intersected without changing validity or polyhedrality. Theorems 4.2 and 5.1 use (16). The corresponding mechanism experiments retain those balls, whereas behavior-policy log and earlier practical-DCR counts use (19). Their confidence constants and observed counts are not interchanged.

A.4 Data reuse, nesting, and retained evidence

The confidence sets need not be nested: a newly observed successor can shift an empirical center. Suppose an evaluation at time j produces checked bounds $\overline{U}_j$ and $\underline{L}_j$ for policy π_j. Maintain

$$\overline{U}_t^{\mathrm{keep}} = \min_{j \le t} \overline{U}_j, \qquad \underline{L}_t^{\mathrm{keep}} = \max_{j \le t} \underline{L}_j, \qquad \pi_t^{\mathrm{keep}} = \pi_{j_t}, \quad j_t \in \arg\max_{j \le t} \underline{L}_j. \tag{20}$$

The retained policy and scalars satisfy the same all-state guarantee, proved in Appendix C.6. In particular, the lower scalar must remain attached to the policy that earned it. A newly selected policy cannot inherit another policy's lower bound.

The acquisition programs are evaluated at the current sets and proposals. Their local derivatives need not be derivatives of the cached, nonsmooth envelope (20). The cache serves the stopping decision; the current programs serve sampling design. Both uses are legitimate on the same event.

B Stationary-flow geometry and policy guarantees

B.1 Perspective constraints, zero occupancies, and realizability

A row polytope may be written as

$$\mathcal{P}_i = \{p \ge 0 : \mathbf{1}^\top p = 1,\ A_i p \le b_i\}.$$

The condition $y_i \in u_i \mathcal{P}_i$ becomes

$$y_i \ge 0, \quad \mathbf{1}^\top y_i = u_i, \quad A_i y_i \le b_i u_i. \tag{21}$$

For $u_i > 0$, divide by u_i. For $u_i = 0$, nonnegativity and the row-sum equality force $y_i = 0$, so no arbitrary division or artificial positive floor is required. This is why the stationary-flow programs remain meaningful before every row has been sampled. The feasible sets are closed and bounded: $0 \le u_i \le 1$ and $0 \le y_{ij} \le u_i$. They are nonempty as well. Choose any row in each nonempty $\mathcal{P}_i$, fix any stationary policy (the given π for the lower program), and take a stationary distribution of the resulting finite chain. Its occupancy supplies a feasible point. Both objective extrema are therefore attained.

Lemma B.1 (Flow realization)**.** *Every $(u, y) \in \mathcal{F}(\mathcal{C})$ is a stationary occupancy of some plausible model and stationary randomized policy. If $(u, y) \in \mathcal{F}_\pi(\mathcal{C})$, it is a stationary occupancy under the specified policy π in some plausible model.*

Proof. Set $d_s = \sum_a u_{sa}$. For $d_s > 0$ define $\mu(a \mid s) = u_{sa}/d_s$; choose any distribution at states with $d_s = 0$. For $u_i > 0$ define $Q_i = y_i/u_i \in \mathcal{P}_i$. For a zero-occupancy row choose any member of the nonempty $\mathcal{P}_i$. Then

$$\sum_{s,a} d_s \mu(a \mid s) Q_{sa,j} = \sum_i y_{ij} = d_j,$$

so d is stationary under Q_μ, with state–action occupancy u. When the fixed-policy constraints hold, $u_{sa} = \pi(a \mid s) d_s$ at all positive-mass states, so the same proof uses $\mu = \pi$ everywhere. Multiple recurrent classes may receive positive mass; the stationary distribution can be their mixture. □

The lemma is an occupancy statement, not a reachability statement from a specified initial state. A plausible model may contain a recurrent class that its policy cannot reach from some other state. Allowing such classes makes the upper program optimistic and the lower program conservative, which is precisely the direction needed for an all-initial-state guarantee. It does not manufacture free access to those classes in a trajectory experiment.

B.2 Limiting occupancy for a fixed true policy

Fix an initial distribution ρ and a policy π. Define finite-horizon expected occupancies

$$u_{sa}^{(T)} = \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{P}_\rho^\pi(S_t = s, A_t = a), \qquad y_{sa,j}^{(T)} = u_{sa}^{(T)} P_{sa,j}.$$

Their conservation error is

$$\sum_a u_{ja}^{(T)} - \sum_i y_{ij}^{(T)} = \frac{\rho_j - \mathbb{P}_\rho^\pi(S_T = j)}{T}. \tag{22}$$

The right-hand side tends to zero, and $u^{(T)}$ has a limit by finite-chain Cesaro convergence. Consequently, its limiting flow belongs to $\mathcal{F}_\pi(\mathcal{C})$ whenever $M \in \mathcal{C}$. Its objective under the true mean reward is $\rho^\top g_M^\pi$. If the policy is multichain, changing ρ changes the mixture of recurrent classes but not the validity of the constraints.

B.3 Proof of the direct certificate theorem

Proof of Theorem 3.1. Work on the event $\mathcal{E}$. At any selected time t, take an average-reward optimal stationary policy and initial state s. The preceding construction gives a feasible limiting flow of value $g_M^\star(s)$ under r. Since $r_i \le r_i^+$,

$$g_M^\star(s) \le \sum_i r_i^+ u_i^\star \le U(\mathcal{C}_t).$$

For an arbitrary policy π and the same initial state, its limiting flow is feasible in the lower program. Since $r_i^- \le r_i$,

$$L(\pi, \mathcal{C}_t) \le \sum_i r_i^- u_i^\pi \le g_M^\pi(s).$$

Subtracting proves (4). This is a deterministic statement for every π and every set containing M. There is therefore no union bound over policies or stopping times. In particular, it remains true when a policy was proposed using these same observations, when the proposal mechanism is randomized, and when stopping is observation-dependent.

For computational size, (u, y) has $SA + S^2A$ components. Probability boxes require $2S^2A$ row inequalities, SA row-sum equalities, S conservation equations, and one normalization. A fixed randomized policy requires an additional SA linear equations. An ℓ_1 row ball can be represented with nonnegative auxiliary variables z_{ij} satisfying

$$z_{ij} \ge y_{ij} - \widehat{p}_{ij} u_i, \quad z_{ij} \ge -y_{ij} + \widehat{p}_{ij} u_i, \quad \sum_j z_{ij} \le b_P(N_i) u_i.$$

This also has $O(S^2A)$ size. Standard finite-dimensional linear optimization applies; neither deterministic policies nor polytope vertices are enumerated. □

The exact safety statement is

$$\mathbb{P}\{\tau < \infty,\ \exists s : g_M^\star(s) - g_M^{\widehat{\pi}}(s) > \varepsilon\} \le \delta.$$

This bound alone does not assert that τ is finite. Theorem 5.1 supplies the separate termination statement under communication. Separating these two facts avoids importing a reachability condition into the validity of a numerical decision already supported by the data.

B.4 What the certificate measures

For exact model knowledge, the lower program satisfies

$$L(\pi, \{M\}) = \min_s g_M^\pi(s), \qquad U(\{M\}) = \max_s g_M^\star(s). \tag{23}$$

For the first identity, the stationary distributions of P_π are convex combinations of the stationary distributions of its recurrent classes. Their minimum reward is the smallest recurrent-class gain.

Starting in that class attains it; starting in a transient state gives a convex combination of class gains. The second identity follows in the same way, optimizing also over the policy. A communicating true model has constant $g^\star_M$, so an average-optimal policy has certificate zero at exact knowledge even if that policy has transient states.

With uncertainty, $U - L$ compares two possibly different plausible models. It is not the exact robust regret

$$\sup_{M' \in \mathcal{C}} \max_s \{g^\star_{M'}(s) - g^\pi_{M'}(s)\},$$

which couples the competitor and incumbent to the same model. Decoupling those two model choices is what makes both certification programs linear. It can add conservatism, but preserves a direct guarantee about the policy decision. The uncertainty prices therefore explain the present upper bound, not an exact information-theoretic optimum over all learning procedures.

In a known multichain model with different optimal gains at different starting states, $\max_s g^\star(s) - \min_s g^\pi(s)$ can be positive even for an optimal policy. The all-state scalar certificate can then fail to vanish. This is one reason the generic finite-time theorem uses communicating dynamics. The safety theorem remains valid without that restriction; no hidden communicating assumption is used in its proof.

B.5 Exact meaning of the two uncertain extrema

Proposition B.2 (The best and worst separate gain explanations)**.** *For the rectangular polyhedral set in Section 2,*

$$U(\mathcal{C}) = \max_{M' \in \mathcal{C}} \max_s g^\star_{M'}(s), \qquad L(\pi, \mathcal{C}) = \min_{M' \in \mathcal{C}} \min_s g^\pi_{M'}(s).$$

The extrema are attained, and C is the smallest gap obtainable by subtracting an all-model scalar lower bound on π from an all-model scalar upper bound on the optimum.

Proof. Every model and initial-state limiting occupancy is feasible, so the upper LP is at least the right-hand upper expression and the lower LP is at most the right-hand lower expression. To obtain the reverse inequalities, take an optimal upper flow and realize it using Lemma B.1, with reward endpoints r^+. Its stationary distribution is a mixture of recurrent-class distributions. At least one class has reward at least the mixture's reward U. Starting in that class under the realized policy gives gain at least U, so the right-hand upper expression is at least U. The same argument with an optimal lower flow, reward r^-, and a class whose reward is no larger than its mixture proves the lower equality for the fixed policy. These realized models and classes also establish attainment despite possible discontinuity of individual stationary distributions as kernels change.

Finally, any scalar u that upper-bounds all those model/state optimal gains satisfies $u \geq U$. Any scalar ℓ that lower-bounds every corresponding gain of π satisfies $\ell \leq L$. Thus $u - \ell \geq U - L$, with equality achieved by the two programs. The qualifier *separate* is essential: a comparison that couples the two policies to the same model may preserve additional cancellations. □

C Dual witnesses and strict incumbent-wise tightening

C.1 Weak duality by cancellation

For a feasible stationary flow and any potential h,

$$\sum_{i,j} y_{ij} h_j - \sum_{s,a} u_{sa} h_s = 0. \tag{24}$$

If $r_i^+ + p^\top h - h_s \leq z$ for every $p \in \mathcal{P}_i$, multiplication by u_i and summation therefore give $\sum_i r_i^+ u_i \leq z$. Taking the maximum over flows proves $U \leq U_h$.

For the fixed-policy lower program, set $d_s = \sum_a u_{sa}$. For any v, each feasible flow has

$$\sum_i r_i^- u_i = \sum_{s,a} u_{sa} \left(r_{sa}^- + Q_{sa}^\top v - v_s \right)$$
$$\geq \sum_s d_s \sum_a \pi(a \mid s) \left(r_{sa}^- + \min_{p \in \mathcal{P}_{sa}} p^\top v - v_s \right) \geq L_{\pi,v},$$

where Q_i realizes the flow. Minimizing its objective gives $L \geq L_{\pi,v}$. These inequalities do not require the selected potentials to solve a Poisson equation or to have a certified span. Any finite vectors are valid witnesses, including vectors computed approximately from the same data.

C.2 Strong duality and finite formulations

Lemma C.1 (Potential form of the flow duals)**.** *For nonempty compact polyhedral row sets,*

$$U = \inf_{h \in \mathbb{R}^S} U_h, \qquad L(\pi, \mathcal{C}) = \sup_{v \in \mathbb{R}^S} L_{\pi,v}.$$

Finite optimal dual solutions exist.

Proof. For the upper program, regard each pair (a, p) with $p \in \mathcal{P}_{sa}$ as an extended action at state s. A linear objective over $\mathcal{P}_{sa}$ attains its extrema at vertices, so there is a finite, possibly large, extended-action representation. The normalized stationary-occupancy linear program in that representation has the dual

$$\min_{z,h} z \quad \text{subject to} \quad z + h_s \geq r_{sa}^+ + p^\top h \quad (s, a, p \in \operatorname{vert}(\mathcal{P}_{sa})).$$

It is feasible (take $z = 1, h = 0$), its primal is feasible, and both objectives are bounded. Linear-program strong duality applies. Eliminating z yields $\inf_h U_h$. Projecting the extended occupancy onto (u, y) gives exactly the perspective formulation, so the value is unchanged.

For the lower program the policy's action probabilities are fixed. At state s, nature chooses a tuple $(p_a)_{a \in \mathcal{A}}$ with $p_a \in \mathcal{P}_{sa}$. Its expected reward is $r_\pi^-(s)$ and its transition law is $\sum_a \pi(a \mid s) p_a$. Rectangularity implies

$$\min_{(p_a)} \sum_a \pi(a \mid s) p_a^\top v = \sum_a \pi(a \mid s) \min_{p \in \mathcal{P}_{sa}} p^\top v.$$

Apply the occupation-measure duality to this minimizing extended MDP. Its dual maximizes z subject to the lower Bellman inequalities. Eliminating z gives $\sup_v L_{\pi,v}$. A feasible dual is $z = 0, v = 0$; the primal is bounded between zero and one. The fixed-policy perspective constraints are the projection of this occupation program. No assertion of a unique recurrent class is needed. □

The extended-action argument proves duality, but it is not the implementation. The implementation uses the compact perspective constraints, avoiding vertex enumeration. Potentials are invariant to a global additive constant, so one coordinate may be anchored to zero without changing any bound.

C.3 Checked approximate solutions

Suppose a row support routine provides certified bounds on $\max_{p \in \mathcal{P}_i} p^\top h$ and $\min_{p \in \mathcal{P}_i} p^\top v$, and the final aggregation errors are bounded. Let $\xi_U, \xi_L \geq 0$ bound the possible underestimation of U_h and overestimation of $L_{\pi,v}$. Then

$$C_{\text{checked}} = U_h + \xi_U - L_{\pi,v} + \xi_L \tag{25}$$

remains an upper bound on policy loss. The correct acceptance condition is $C_{\text{checked}} \leq \varepsilon$. A small primal–dual gap supports numerical accuracy but is not itself a proof that a floating-point solution is feasible.

For boxes, row support computation is particularly simple. Start with $p = \ell$ and remaining mass $1 - \mathbf{1}^\top \ell$. For a maximum, allocate this remaining mass in decreasing order of h_j, never exceeding q_j; reverse the order for a minimum. A swap argument proves optimality: if $h_j > h_k$, $p_j < q_j$, and $p_k > \ell_k$, moving feasible mass from k to j increases the objective. The cost is $O(S \log S)$ for sorting

and $O(S)$ for filling. Thus the returned potentials permit independent row-wise evaluation of the bound.

The experiments use double-precision support evaluation and reserve a numerical margin when interpreting a certified stop. The box implementation rejects an independently evaluated objective discrepancy above 10^{-5}; the ℓ_1 implementation uses 10^{-7}. Observed discrepancies are many orders of magnitude below the reported tolerances. These are numerical checks of the witnesses, not outward-rounded verification of every confidence endpoint and support value. Reported true gains are evaluated separately from the certificate solver.

C.4 A TRANSIENT LIMITING EXAMPLE

The full-support separation is proved in Appendix D. An exact zero-gap example helps explain the same issue at a transient state. Let state 0 have a reward-one self-loop and a reward-zero action to state 1. At 1, two actions both return to 0, with rewards zero and $1/2$. The incumbent stays at 0 and chooses the zero-reward return at 1. It earns gain one from both states, so $U = L = 1$ at exact knowledge.

Anchoring a common potential as $h = (0, x)$ gives

$$U_h = \max\{1, x, \tfrac{1}{2} - x\}, \qquad L_{\pi,h} = \min\{1, -x\}.$$

The gap is at least $1/2$, with equality at $x = -1/2$. Separate witnesses $h = (0, 0)$ and $v = (0, -1)$ give $U_h = L_{\pi,v} = 1$. Thus a one-time inferior action can be irrelevant to average reward but still obstruct a common-potential test. Proposition 3.2 shows that the distinction also persists with positive recurrent-state mass and full-support transitions.

C.5 A FINITE CANDIDATE PORTFOLIO

Let Π_t^{prop} be any nonempty finite collection of proposed policies, possibly constructed adaptively. One upper LP is shared by all proposals, while one lower LP is needed per candidate. Define

$$L_t^{\mathrm{prop}} = \max_{\pi \in \Pi_t^{\mathrm{prop}}} L(\pi, \mathcal{C}_t).$$

Selecting an argmax and using $U - L_t^{\mathrm{prop}}$ is valid by Theorem 3.1. Enlarging the proposal set can only improve this certificate. In particular, adding an empirical policy cannot invalidate the optimistic-witness guarantee used for termination. The computational cost is linear in the number of *evaluated proposals*, not in the number A^S of deterministic policies. Our implementation evaluates at most two newly proposed policies at each check and retains the best previously evaluated lower bound.

C.6 RETAINING BOUNDS ACROSS NONNESTED CONFIDENCE SETS

At evaluation j, let a checked scalar U_j upper-bound $g_M^\star(s)$ for every s, and let L_j lower-bound $g_M^{\pi_j}(s)$ for every s. All these inequalities hold simultaneously on $\mathcal{E}$. Define

$$\overline{U}_t = \min_{j \le t} U_j, \qquad j_t \in \arg\max_{j \le t} L_j, \qquad \underline{L}_t = L_{j_t}.$$

For each s, every U_j bounds the same number $g_M^\star(s)$, so $g_M^\star(s) \le \overline{U}_t$. The specific lower inequality attached to j_t gives $g_M^{\pi_{j_t}}(s) \ge \underline{L}_t$. Subtraction proves the retained certificate. The upper sequence is nonincreasing and the lower sequence nondecreasing, so their difference cannot increase. This argument concerns scalar bounds on a fixed environment, not set inclusion; the confidence sets need not be nested. Keeping the lower number while discarding its policy would lose the second inequality.

The retained gap need not be the difference of the two programs at the current dataset, so its derivative is not the current LP price. DCR uses current programs for acquisition and retained checked bounds for the stopping decision. This separates the two operations without changing either guarantee.

D HOW MUCH DATA DOES AN INCUMBENT NEED?

This section relates the computable certificate to the difficulty of verifying a particular policy. The distinction from the generic completion theorem is useful: finding a new policy may require exploring

the whole control problem, whereas an existing policy may already admit short witnesses. The witnesses below occur only in the analysis. The learner continues to evaluate the same two flow programs.

D.1 Exact-model witnesses and row perturbations

Fix a finite communicating true MDP M and a stationary incumbent π. Set

$$G^+ = g^\star_M, \qquad G^-_\pi = \min_s g^\pi_M(s), \qquad \Delta_\pi = G^+ - G^-_\pi.$$

By Lemma C.1 applied to the singleton row sets $\{P_i\}$, there are finite vectors $h^\star, v^\pi$ such that

$$r_i + P_i^\top h^\star - h^\star_s \le G^+ \qquad (i = (s,a)), \tag{26}$$

$$\sum_a \pi(a \mid s)(r_{sa} + P_{sa}^\top v^\pi - v^\pi_s) \ge G^-_\pi \qquad (s \in \mathcal{S}). \tag{27}$$

Their respective maximum and minimum attain G^+ and G^-_π. Write $H_\star = \operatorname{span}(h^\star)$ and $H_\pi = \operatorname{span}(v^\pi)$. The second witness exists even when π has several recurrent classes; it need not solve a constant-gain Poisson equality on every state. This is why we use the lower inequality rather than assume an all-policy unichain condition.

We first give the elementary perturbation calculation used twice below. If $a, b \in \Delta_S$, then $\mathbf{1}^\top(a-b) = 0$. For $c = (\max_j w_j + \min_j w_j)/2$,

$$|(a-b)^\top w| = |(a-b)^\top (w - c\mathbf{1})| \le \tfrac{1}{2}\|a-b\|_1 \operatorname{span}(w). \tag{28}$$

Suppose both P_i and a candidate row p belong to the empirical ℓ_1 ball of radius $b_P(N_i)$. The triangle inequality gives $\|p - P_i\|_1 \le 2b_P(N_i)$. Likewise, if the empirical reward interval has half-width $b_R(N_i)$ and contains r_i, then $r_i^+ - r_i \le 2b_R(N_i)$ and $r_i - r_i^- \le 2b_R(N_i)$. Endpoint clipping can only decrease these errors. Define

$$e_i(w, N) = 2b_R(N_i) + b_P(N_i) \operatorname{span}(w). \tag{29}$$

For known rewards the first term is zero. At $N_i = 0$, use the full-uncertainty radii from Appendix A; no fictitious observation is inserted.

D.2 Proof of the incumbent verification bound

Proof of Theorem 3.3. We work throughout on $\mathcal{E}$, so every empirical confidence set contains the same true MDP.

Step 1: Upper-bound the optimistic program with the true upper witness. For every row $i = (s,a)$ and every $p \in \mathcal{P}_i$,

$$\begin{aligned} r_i^+ + p^\top h^\star - h^\star_s &= r_i + P_i^\top h^\star - h^\star_s + (r_i^+ - r_i) + (p - P_i)^\top h^\star \\ &\le G^+ + e_i(h^\star, N). \end{aligned}$$

Maximizing over p and i, then using weak duality, gives

$$U(\mathcal{C}_N) \le U_{h^\star}(\mathcal{C}_N) \le G^+ + \max_i e_i(h^\star, N). \tag{30}$$

This step does not require the learner to know $h^\star$ or to return it from the LP solver.

Step 2: Lower-bound the incumbent program with its own witness. For each state s, the same calculation in the other direction gives

$$\begin{aligned} &\sum_a \pi(a \mid s) \left[r^-_{sa} + \min_{p \in \mathcal{P}_{sa}} p^\top v^\pi - v^\pi_s \right] \\ &\quad \ge \sum_a \pi(a \mid s) \left[r_{sa} + P_{sa}^\top v^\pi - v^\pi_s - e_{sa}(v^\pi, N) \right] \\ &\quad \ge G^-_\pi - \sum_a \pi(a \mid s) e_{sa}(v^\pi, N). \end{aligned}$$

Taking the minimum over states and applying lower weak duality yields

$$L(\pi, \mathcal{C}_N) \geq G_\pi^- - \max_s \sum_a \pi(a \mid s) e_{sa}(v^\pi, N). \tag{31}$$

Notice that the lower error uses only actions selected by the incumbent, while the upper error also checks competing actions.

Step 3: Subtract the two bounds. Combining (30) and (31) proves the heterogeneous-count inequality

$$C(\pi, \mathcal{C}_N) \leq \Delta_\pi + \max_i e_i(h^\star, N) + \max_s \sum_a \pi(a \mid s) e_{sa}(v^\pi, N). \tag{32}$$

If every row has at least n observations, monotonicity of the radii and $\sum_a \pi(a \mid s) = 1$ imply

$$C(\pi, \mathcal{C}_N) \leq \Delta_\pi + 4b_R(n) + (H_\star + H_\pi) b_P(n).$$

Thus the certificate passes once $n \geq n_\pi$, with n_π defined in the theorem. Existence follows from $\Delta_\pi < \varepsilon$ and the fact that the radii tend to zero.

Step 4: Count only the missing observations. Suppose π is included among the evaluated candidates at every check. Until all rows reach n_π, a least-filled row has count below n_π. Sampling it decreases $\sum_i [n_\pi - N_i]_+$ by one. Any interleaved design request can only decrease this potential or leave it unchanged. Starting from n_0, at most $\sum_i [n_\pi - n_{0,i}]_+$ safeguard requests are therefore needed. Alternating one design and one safeguard request, with a check after each pair, gives the asserted bound of $2\sum_i [n_\pi - n_{0,i}]_+ + 2$. The retained portfolio can only stop earlier. The extra two allows a final pair and can be omitted with a check at exactly the completing request.

Step 5: Substitute the concentration radii. Let $m = \varepsilon - \Delta_\pi > 0$. With (14), it suffices to choose

$$n \geq c \, \frac{(1 + H_\star + H_\pi)^2}{m^2} \left[S + \log \frac{d}{\delta} + \log(n+1) \right]$$

for a universal numerical c. Solving this scalar inequality by increasing n yields the stated $\widetilde{O}(m^{-2})$ bound. The quantity n_π is an analysis threshold; neither it nor the witness spans are inputs to DCR. □

The proof is simultaneous over all policies with finite witnesses because it is deterministic on $\mathcal{E}$. An incumbent selected from the historical data can therefore be retained and re-evaluated without a sample split. Merely storing its old lower bound is different from re-evaluating it: the latter is required for the n_π completion guarantee. A finite portfolio of repeatedly evaluated incumbents enjoys the minimum of their verification thresholds. The guarantee is specific to verification; it does not assert a prior-free optimal-bias-only complexity for discovering an unknown optimal policy.

D.3 Rare recurrent decisions: a separation with full-support transitions

Proof of the full-support claim in Proposition 3.2. Let $0 < p, q < 1$. Both actions at state 0 have reward one and transition $(1-p, p)$. At state 1, both actions have transition $(q, 1-q)$, with rewards zero and $1/2$. The incumbent chooses the zero-reward action at state 1. Actions do not change the transition kernel, so every stationary policy induces the same irreducible, aperiodic two-state chain. Its stationary distribution is

$$d_0 = \frac{q}{p+q}, \qquad d_1 = \frac{p}{p+q}.$$

The incumbent and optimal gains are respectively $q/(p+q)$ and $(q + p/2)/(p+q)$. Therefore

$$C(\pi, \{M\}) = \Delta_\pi = \frac{p}{2(p+q)}. \tag{33}$$

For a common potential, anchor $h_0 = 0$ and put $h_1 = x$. Its upper and lower Bellman expressions are

$$U_h = \max\{1 + px, \tfrac{1}{2} - qx\}, \qquad L_{\pi,h} = \min\{1 + px, -qx\}.$$

The upper maximum includes $1/2 - qx$, while the lower minimum is at most $-qx$. Their difference is at least $1/2$ for every x. At $x = -1/(p+q)$, $1 + px = -qx = q/(p+q)$, so equality holds. Consequently the *optimized* common-potential gap is $1/2$, not just the gap of a particular numerical witness. Its ratio to (33) is $(p+q)/p$, which diverges as $p \downarrow 0$ with q fixed.

Separate witnesses are

$$h^\star = \left(0, -\frac{1}{2(p+q)}\right), \qquad v^\pi = \left(0, -\frac{1}{p+q}\right).$$

Direct substitution gives $U_{h^\star} = g^\star_M$ and $L_{\pi,v^\pi} = g^\pi_M$. This verifies the result independently of occupation-measure duality. □

Corollary D.1 (Verification need not scale with the diameter)**.** *In this family fix $q = 1/2$, $0 < \varepsilon < 1/2$, and $0 < p \le \varepsilon/2$. Uniform completion to $\widetilde{O}(\varepsilon^{-2}\log(1/\delta))$ observations per row suffices to verify the incumbent with the separate certificate. The hidden constant is independent of p, although the diameter is $\max\{1/p, 2\}$. No common-potential certificate formed from a confidence set containing the true model can verify this incumbent at tolerance $\varepsilon < 1/2$.*

Proof. Here $\Delta_\pi = p/(1+2p) \le \varepsilon/2$ and $H_\star + H_\pi = 3/[2(p+1/2)] \le 3$. Apply Theorem 3.3 with margin at least $\varepsilon/2$ and $S = 2, A = 2$. The hitting times from 0 to 1 and from 1 to 0 are geometric with means $1/p$ and two, giving the diameter. Finally, enlarging exact row knowledge to any set containing the true model increases U_h and decreases $L_{\pi,h}$ for every fixed h. The optimized common gap therefore stays at least $1/2$. □

The certificate can accept an action that loses reward on a sufficiently rare *recurrent* state, not only an action confined to a transient state. This is relevant when the objective is to verify an existing controller rather than insist that it be locally greedy everywhere. A competing learning algorithm may instead return another policy; the separation does not compare the entire learning algorithms.

D.4 A FINITE-STATE CLASS WITH AN EXACT VERIFICATION GAP

The separation is not limited to two states. The following class isolates how a common potential retains the largest local sacrifice while separate potentials average it under the recurrent dynamics.

Proposition D.2 (Action-independent transitions and constant optimal rewards)**.** *Let every action induce the same irreducible transition matrix P, with stationary distribution μ. Suppose the largest available reward at every state is the same $c \in [0,1]$. For a stationary incumbent π, write $r_\pi(s) = c - \Delta_s$, where $0 \le \Delta_s \le c$. At exact model knowledge,*

$$C(\pi, \{M\}) = \mu^\top \Delta, \qquad \inf_h \{U_h - L_{\pi,h}\} = \max_s \Delta_s.$$

If every row of P is the same full-support vector μ, exact witnesses are $h^\star = 0$ and $v^\pi = (\mu^\top\Delta)\mathbf{1} - \Delta$. Their spans satisfy $H_\star = 0$ and $H_\pi = \mathrm{span}(\Delta)$, independently of $D = 1/\min_s \mu_s$ for $S \ge 2$.

Proof. Every policy has stationary distribution μ. Choosing a reward-c action at each state gives $g^\star = c$, while $g^\pi = c - \mu^\top\Delta$. The exact-flow identities give the first equality. Put $x = Ph - h$. Action independence and the constant best reward give

$$U_h - L_{\pi,h} = \max_s x_s - \min_s (x_s - \Delta_s).$$

For each s, this difference is at least $x_s - (x_s - \Delta_s) = \Delta_s$; $h = 0$ attains their maximum. If $P = \mathbf{1}\mu^\top$, then $Pv^\pi = 0$ and $g^\pi\mathbf{1} + v^\pi = c\mathbf{1} - \Delta = r_\pi + Pv^\pi$. The witness spans follow immediately. From any different state, the hitting time of state s is geometric with mean $1/\mu_s$, giving the diameter identity. □

In the rank-one subclass, let only one state have deficit $a > 0$, hold a fixed, and send its stationary probability to zero while keeping every probability positive. The certificate ratio diverges, whereas the sum of witness spans remains a. When $\mu^\top\Delta < \varepsilon$, known rewards and Theorem 3.3 give verification once $b_P(n) \le (\varepsilon - \mu^\top\Delta)/a$, independently of the diverging diameter. This is a fixed-incumbent statement: with known rewards, a policy choosing reward c everywhere already has a zero-width certificate without transition samples, so identifying some good policy can be easier than verifying the supplied one.

E PLANNING AND COMPLETION WITH UNKNOWN DIAMETER

We prove the completion theorem in three parts. First, a finite computation produces an optimistic potential with a small observed residual. Second, that potential yields a candidate whose certificate is controlled by row uncertainty. Third, the least-filled safeguard reaches the required counts. The diameter appears in the proof of these bounds, but is not needed to perform any of the three steps.

E.1 EXISTENCE OF A BOUNDED OPTIMISTIC POTENTIAL

The true MDP is communicating in this section. Its unknown diameter is

$$D = \max_{s \neq z} \inf_{\mu} \mathbb{E}_s^{\mu}[\tau_z], \qquad \tau_z = \inf\{t \geq 0 : S_t = z\}.$$

At a fixed evaluation time on $\mathcal{E}$, the optimistic extended MDP has actions (a, p) with $p \in \mathcal{P}_{sa}$ and rewards $r_{sa}^{+} \in [0, 1]$. Each true transition row is available, so a true hitting policy is feasible in the extended MDP. Its diameter is at most D, as in the extended-MDP span argument of Jaksch et al. (2010). This comparison is used to analyze the operator T^{+}, not to supply a hitting policy to the algorithm.

Lemma E.1 (An optimistic Bellman solution with bounded span)**.** *There are a scalar g and a finite vector h^{+}, anchored at $h_{s_0}^{+} = 0$, such that*

$$g\mathbf{1} + h^{+} = T^{+}h^{+}, \qquad \operatorname{span}(h^{+}) \leq D, \qquad g = U(\mathcal{C}).$$

Proof. **Step 1: Bound discounted value differences.** For the proof only, let V_β be the optimal discounted value in the optimistic extended MDP, with $0 < \beta < 1$. It exists because the finite-state discounted Bellman operator is a contraction and the row sets are compact. Since rewards lie in $[0, 1]$, $0 \leq V_\beta \leq (1 - \beta)^{-1}\mathbf{1}$.

Starting from s, follow a true hitting policy to z and then an optimal discounted policy. Rewards before hitting are nonnegative, and hence

$$V_\beta(s) \geq \mathbb{E}_s[\beta^{\tau_z}] V_\beta(z).$$

Consequently,

$$V_\beta(z) - V_\beta(s) \leq \mathbb{E}_s\left[\frac{1 - \beta^{\tau_z}}{1 - \beta}\right] = \mathbb{E}_s\left[\sum_{j=0}^{\tau_z - 1} \beta^j\right] \leq \mathbb{E}_s[\tau_z] \leq D.$$

The inequality holds for every ordered pair, so $\operatorname{span}(V_\beta) \leq D$.

Step 2: Pass to a bounded subsequence. Define $h_\beta = V_\beta - V_\beta(s_0)\mathbf{1}$ and $g_\beta = (1 - \beta)V_\beta(s_0)$. These vectors and scalars lie in compact sets: $h_\beta(s_0) = 0$, $|h_\beta(s)| \leq D$, and $0 \leq g_\beta \leq 1$. Along a sequence $\beta \uparrow 1$, select a convergent subsequence with limit (h^{+}, g). The discounted Bellman equation becomes

$$g_\beta + h_\beta(s) = \max_a \{r_{sa}^{+} + \beta \max_{p \in \mathcal{P}_{sa}} p^{\top} h_\beta\}.$$

The support functions are continuous, so taking limits proves the displayed average-reward Bellman equation and the span bound. The discount parameter is an existence-proof device; the planning algorithm below does not use a discounted statistical bridge.

Step 3: Identify the scalar with the flow optimum. The Bellman equation implies the upper inequalities for U, so weak duality gives $U \leq g$. Choose one maximizing action and row at every state. Any stationary distribution of the resulting finite chain has average reward g, since summing its Bellman equalities cancels h^{+}. Its flow is feasible in $\mathcal{F}(\mathcal{C})$, so $U \geq g$. This establishes equality. □

E.2 A FINITE RESIDUAL SOLVER WITH AN EXPLICIT BOUND

Let $\mathcal{H} = \{h \in \mathbb{R}^S : h_{s_0} = 0\}$ and equip it with the norm $\|h\|_{\mathrm{sp}} = \operatorname{span}(h)$. Define

$$F(h) = T^{+}h - (T^{+}h)(s_0)\mathbf{1}.$$

The operator F maps $\mathcal{H}$ to itself and has the fixed point h^+ from Lemma E.1. Monotonicity and translation invariance of T^+ imply

$$\|F(x) - F(y)\|_{\mathrm{sp}} \le \|x - y\|_{\mathrm{sp}}. \tag{34}$$

Indeed, if $m\mathbf{1} \le x - y \le M\mathbf{1}$, then $m\mathbf{1} \le T^+x - T^+y \le M\mathbf{1}$; subtracting an anchoring constant does not affect the span.

Use the anchored iteration

$$h_0 = 0, \qquad h_{k+1} = \frac{k+1}{k+3} F(h_k). \tag{35}$$

We adapt the nonexpansive anchoring schedule of Lee et al. (2025) to the optimistic state-value map. We include an elementary estimate for these coefficients to make the completion argument self-contained.

Lemma E.2 (Finite optimistic planning)**.** *For every $k \ge 0$, the iterates in* (35) *satisfy*

$$\|h_k - h^+\|_{\mathrm{sp}} \le D, \qquad \mathrm{span}(h_k) \le 2D, \qquad \mathrm{span}(T^+h_k - h_k) \le \frac{8D}{k+3}.$$

Therefore the observed condition $\mathrm{span}(T^+h_k - h_k) \le \zeta$ is reached after $O(1 + D/\zeta)$ iterations. Each iteration uses one row-support computation per state–action pair.

Proof. **Step 1: Bound the distance to a fixed point.** Put $a_k = (k+1)/(k+3)$. Since $F(h^+) = h^+$, the triangle inequality and (34) give

$$\|h_{k+1} - h^+\|_{\mathrm{sp}} \le a_k \|h_k - h^+\|_{\mathrm{sp}} + (1 - a_k)\|h^+\|_{\mathrm{sp}}.$$

Starting from $\|h_0 - h^+\|_{\mathrm{sp}} \le D$, induction proves the first bound. It also implies $\|h_k\|_{\mathrm{sp}} \le 2D$ and $\|F(h_k)\|_{\mathrm{sp}} \le 2D$.

Step 2: Bound consecutive differences. For $k \ge 1$, let $d_k = \|h_k - h_{k-1}\|_{\mathrm{sp}}$. Using $a_k - a_{k-1} = 2/[(k+2)(k+3)]$,

$$\begin{aligned} d_{k+1} &\le a_k \|F(h_k) - F(h_{k-1})\|_{\mathrm{sp}} + (a_k - a_{k-1})\|F(h_{k-1})\|_{\mathrm{sp}} \\ &\le \frac{k+1}{k+3} d_k + \frac{4D}{(k+2)(k+3)}. \end{aligned}$$

The base case is $d_1 = \|F(0)\|_{\mathrm{sp}}/3 \le 2D/3$. Induction now gives $d_k \le 4D/(k+2)$: substituting this upper bound in the preceding line yields exactly $4D/(k+3)$.

Step 3: Turn successive differences into a residual. For every $k \ge 0$,

$$\begin{aligned} \|F(h_k) - h_k\|_{\mathrm{sp}} &\le \|F(h_k) - h_{k+1}\|_{\mathrm{sp}} + d_{k+1} \\ &= (1 - a_k)\|F(h_k)\|_{\mathrm{sp}} + d_{k+1} \le \frac{8D}{k+3}. \end{aligned}$$

The anchored and unanchored residuals have the same span. This proves the last assertion, including $k = 0$. The algorithm checks the residual itself; the sufficient iteration count is not an input. □

The result applies to polyhedral row uncertainty without listing its vertices. For boxes a support computation takes $O(S \log S)$ time, and an ℓ_1 ball also admits a finite LP. This is a precision-dependent planning bound, not a claim of strongly polynomial complexity independent of D. Approximate computations use certified residual upper bounds and reserve numerical slack before accepting a policy.

E.3 From an observed residual to a policy certificate

For a finite potential h, set

$$q_i^+(h) = r_i^+ + \max_{p \in \mathcal{P}_i} p^\top h, \quad q_i^-(h) = r_i^- + \min_{p \in \mathcal{P}_i} p^\top h, \quad w_i(h) = q_i^+(h) - q_i^-(h).$$

Let $\pi_h(s) \in \arg\max_a q^-_{sa}(h)$. For each state, choose an action a_s maximizing the upper row value. Then

$$\max_a q^-_{sa}(h) \geq q^-_{s,a_s}(h) = \max_a q^+_{sa}(h) - w_{s,a_s}(h) \geq \max_a q^+_{sa}(h) - \max_i w_i(h).$$

Subtracting h_s, taking the minimum over s, and using (5) gives

$$U_h - L_{\pi_h,h} \leq \operatorname{span}(T^+h - h) + \max_i w_i(h). \tag{36}$$

Lemma E.3 (An observed-width certificate)**.** *If* $\operatorname{span}(T^+h - h) \leq \zeta$*, then*

$$C(\pi_h, \mathcal{C}) \leq \zeta + \max_i \left[r_i^+ - r_i^- + \max_{p \in \mathcal{P}_i} p^\top h - \min_{p \in \mathcal{P}_i} p^\top h \right].$$

This inequality is computable from the current confidence set and witness. Its validity does not require communication.

Proof. The right side bounds $U_h - L_{\pi_h,h}$ by (36). Weak duality then gives $C(\pi_h, \mathcal{C}) \leq U_h - L_{\pi_h,h}$. □

For ℓ_1 balls of radius $b_P(N_i)$, any two rows in the same ball differ by at most $2b_P(N_i)$. Applying (28), using reward width at most $2b_R(N_i)$, and using $\operatorname{span}(h_k) \leq 2D$ proves

$$C(\pi_{h_k}, \mathcal{C}) \leq \zeta + \max_i \{2b_R(N_i) + 2Db_P(N_i)\}. \tag{37}$$

The lower-greedy proposal is essential for this argument. An empirical optimizer can be included in the portfolio as well, but does not replace the guaranteed proposal.

E.4 Proof of the residual sample-count theorem

Proof of Theorem 5.1. **Step 1: Give a sufficient count level.** Run (35) until the observed residual is at most $\zeta = \varepsilon/4$. By (37), the lower-greedy candidate is certifiable as soon as all $N_i \geq n_\star$. The portfolio's best certificate is no larger, so a check at that point stops.

Step 2: Charge only unsatisfied deficits. Let $R(N) = \sum_i [n_\star - N_i]_+$. If $R(N) > 0$, at least one count is below $n_\star$, and hence a least-filled row has count below $n_\star$. A safeguard observation reduces R by exactly one. Every design observation reduces it by zero or one. Therefore at most $R(n_0)$ safeguard observations suffice, irrespective of the design scores. With one design observation per safeguard observation and a check after each pair, the total is at most $2R(n_0) + 2$. An initially passing certificate costs zero new observations.

Step 3: Convert stopping into a policy guarantee. All the preceding arguments are on the single event $\mathcal{E}$. Theorem 3.1 applies to the selected policy and selected stopping time on this same event. Thus with probability at least $1 - \delta$ the method both stops by the displayed bound and returns an all-state ε-optimal policy. No union bound over proposed policies is needed.

Step 4: Solve the scalar count inequality. Let $L(n) = S + \log(d/\delta) + 2\log(n+1)$. For a universal c,

$$2b_R(n) + 2Db_P(n) \leq c(D+1)\sqrt{L(n)/n}.$$

Choosing n at least a sufficiently large constant times

$$\frac{(D+1)^2}{\varepsilon^2}\left[S + \log\frac{d}{\delta} + \log\frac{(D+1)(S+1)}{\varepsilon\delta}\right]$$

makes the required inequality hold. This proves (12). The monotonicity of the radii, established in Appendix A, makes every larger count sufficient as well. □

E.5 Almost-sure termination of the unrestricted count implementation

Corollary E.4 (Termination outside the designated confidence event)**.** *For the empirical ℓ_1 balls in Appendix A, with positive radius at every finite positive count and the full simplex at count zero, the direct-sampling algorithm using the uncapped anchored planner terminates almost surely on a communicating true MDP. The high-probability count bound remains the one in Theorem 5.1.*

Proof. First, every such ball contains a strictly positive probability vector: mix its empirical center with the uniform vector using a sufficiently small positive mixing weight. Choosing these rows produces a communicating model inside the confidence set at every finite dataset, even if that set happens not to contain the true model. Lemmas E.1–E.2, applied with this model's finite diameter, show that each planning call ends. Thus a planning call cannot permanently block further observations.

Second, the prefix failure bounds are summable over n for each of the finitely many rows. The first Borel–Cantelli lemma implies that, almost surely, there is a finite random n_{good} after which every prefix confidence set contains its true row. No independence of the failure events is required. If the algorithm had not stopped, infinitely many least-filled safeguard requests would force the minimum row count to diverge. Eventually all counts would exceed both n_{good} and $n_\star$. The actual confidence set would then contain the true model and the sufficient-width argument would force the next check to stop, a contradiction. □

This result concerns the mathematical algorithm without a hard numerical iteration cap. A finite experimental runtime cap must still be reported as a computational failure rather than as a certified policy.

E.6 Batches, checking frequency, and the implemented planner

The one-request presentation isolates the proof; a practical implementation checks after batches. Let B_k be the cumulative number of new observations and assign half of each increment to the safeguard. Once the cumulative safeguard count is at least $R(n_0)$, all rows reach $n_\star$. For $B_1 = 512$ and $B_{k+1} = 2B_k$, the first sufficient checkpoint obeys

$$B_{k_\star} < 4R(n_0) + 512. \tag{38}$$

If $R(n_0) \le 256$, the first batch suffices. Otherwise, the first dyadic $B_k \ge 2R(n_0)$ is below $4R(n_0)$. Integer splitting only adds the final rounding convention. This bound concerns overshoot of a sufficient envelope, not overshoot relative to an unknown optimal stopping rule.

The policy-proposal experiments solve the optimistic residual with lazy iteration $v_{k+1} = (v_k + T^+ v_k)/2$, which is often faster than (35). Every returned witness is checked. The finite implementation cap raises an error if the required residual is not reached; it is not a termination assumption in Theorem 5.1. For a guaranteed implementation, use the anchored solver directly, or interleave it with the faster solver and accept the first verified residual. The anchored path ensures finite progress even if the fast path does not converge quickly. The statistical stopping rule remains the checked pair of flow bounds.

To see that a fast-path witness retains the needed span bound, its unanchored lazy iterates are finite-horizon values in the extended MDP with rewards $r^+/2$ and transitions $(I + p)/2$. A true hitting policy now takes at most $2D$ expected steps. Truncating that policy at the finite horizon loses at most half its hitting time in reward, so ordered value differences are at most D. Thus a witness accepted from either path has span at most $2D$. The supplement implements the interleaved solver and checks its residual separately; the historical experimental curves retain their original lazy-solver implementation.

F Decision prices and residual design

F.1 Perturbing a row confidence set

Fix a proposed policy throughout this subsection. Write both certification objectives as maximizations: $V_+ = U$ and $V_- = -L$. For a box row, the relevant inequalities are

$$y_{ij} - q_{ij}(\theta_i)u_i \le 0, \qquad \ell_{ij}(\theta_i)u_i - y_{ij} \le 0.$$

The centers and rewards are held fixed, and the endpoints vary according to the contraction in the main text. Decreasing θ_i makes the set smaller; thus both U and $-L$ are nondecreasing functions of each uncertainty parameter θ_i.

Let $\lambda_{ij}^{\sigma,q}$ and $\lambda_{ij}^{\sigma,\ell}$ be nonnegative multipliers for the two inequalities in the maximization problem V_σ. With equality-constraint multipliers suppressed, its Lagrangian contains

$$-\sum_{i,j} \lambda_{ij}^{\sigma,q}(y_{ij} - q_{ij}(\theta_i)u_i) - \sum_{i,j} \lambda_{ij}^{\sigma,\ell}(\ell_{ij}(\theta_i)u_i - y_{ij}). \tag{39}$$

Proof of Theorem 4.1. **Step 1: State the regularity used for differentiation.** For each sign σ, choose primal and dual optimal solutions that are differentiable in a neighborhood of the parameter under consideration. One sufficient condition is that a nonsingular optimal basis remains optimal in that neighborhood. This is a condition on the local optimization problem, not on the true transition matrix. At a boundary parameter, the same argument applies along a differentiable one-sided selection. Without such a selection, the formula below is only a solver-dependent score.

Step 2: Remove the derivatives of the optimizers. Write all inequalities, including nonnegativity, as $g_k(x,\theta) \le 0$ and all equalities as $Ex = e$. The maximizing Lagrangian is $\mathcal{L} = c^\top x - \sum_k \lambda_k g_k(x,\theta) - \nu^\top(Ex - e)$. Along a primal–dual optimal selection, stationarity gives $\nabla_x \mathcal{L} = 0$ and complementary slackness gives $\lambda_k g_k = 0$. Terms involving the derivative of ν vanish by equality feasibility. Terms $-\lambda_k' g_k$ vanish as well: if $g_k = 0$ the product is zero, while if $g_k < 0$ continuity and complementarity imply that λ_k is identically zero locally. Thus differentiating the attained value leaves only the partial derivative of the Lagrangian with respect to the parameter.

Step 3: Differentiate the perspective row constraints. The terms containing θ_i are precisely those in (39). Since

$$q_{ij}'(1) = q_{ij} - \widehat{p}_{ij}, \qquad \ell_{ij}'(1) = \ell_{ij} - \widehat{p}_{ij},$$

we obtain

$$\frac{\partial V_\sigma}{\partial \theta_i} = u_i^\sigma \sum_j \left[\lambda_{ij}^{\sigma,q}(q_{ij} - \widehat{p}_{ij}) + \lambda_{ij}^{\sigma,\ell}(\widehat{p}_{ij} - \ell_{ij})\right].$$

Adding the two derivatives gives (8). Each summand is nonnegative because the empirical center belongs to the box.

Step 4: Treat other uncertainty descriptions. For an ℓ_1 ball, its perspective constraint is $\sum_j z_{ij} - \rho_i u_i \le 0$. Its Lagrangian term is $-\lambda_i^\sigma(\sum_j z_{ij} - \rho_i u_i)$, whose derivative in ρ_i is $u_i^\sigma \lambda_i^\sigma$. For reward endpoints, the objectives U and $-L$ have coefficients r^+ and $-r^-$. Their derivatives for increasing r_i^+ and decreasing r_i^- are respectively u_i^+ and u_i^-. A simultaneous contraction towards a reward center multiplies these derivatives by the corresponding distances to that center. □

This derivative includes occupancy even though the constraint multipliers come from a flow program: the radius and box endpoints multiply u_i in the perspective formulation. Omitting that factor would confuse a row-wise Bellman price with its contribution to the normalized long-run objective.

F.2 Degeneracy, inactive rows, and boundary effects

The differentiability qualification is substantive. Linear programs may have multiple optimal faces, and at an endpoint such as $\ell_{ij} = 0$ a row inequality can be redundant with $y_{ij} \ge 0$. Two valid dual solutions may assign different multipliers to these redundant constraints. A contraction that moves ℓ_{ij} into the interior can therefore have a one-sided derivative different from the value obtained from an arbitrary boundary multiplier.

We use the solver-returned formula as a nonnegative local score in these cases. We do not certify a predicted global decrement in C, and no sample-complexity claim depends on the score being the unique derivative. The least-filled safeguard protects termination even if all these scores are uninformative. A finite one-sided comparison provides a score even when a derivative is ambiguous, as follows.

The formula also explains why an occupancy-only score is incomplete. Positive u_i does not imply that a row constraint is binding. A row can be frequently visited but already irrelevant to the gap between the two programs. Conversely, a small occupancy can be multiplied by a substantial value price. Rewards, the state-space geometry, and the incumbent all enter through the decision LPs rather than through a prescribed external relevance vector.

A change in actual sample count is not a pure contraction. New data can move the empirical center, change which policy is proposed, and switch the active recurrent class. The derivative is conditional on those objects, while the subsequent certificate is computed from the new data. This is the same distinction between a local experimental-design objective and the realized outcome of an experiment.

F.3 A FINITE ALTERNATIVE AT DEGENERATE OPTIMA

For $0 < \kappa \leq 1$, let $\mathcal{C}^{i,\kappa}$ contract only row i from $\theta_i = 1$ to $1 - \kappa$, leaving its center, the rewards, the candidate, and every other row unchanged. Define

$$d_i^{(\kappa)} = \frac{C(\pi, \mathcal{C}) - C(\pi, \mathcal{C}^{i,\kappa})}{\kappa}.$$

The contracted set is nonempty because it contains the empirical center. Its two feasible flow sets are subsets of the original sets. The upper optimum can only decrease and the lower optimum can only increase, proving $d_i^{(\kappa)} \geq 0$. If the differentiability conditions of Theorem 4.1 hold, the definition of the one-sided derivative gives $d_i^{(\kappa)} \to d_i$ as $\kappa \downarrow 0$. Otherwise this quantity remains a well-defined finite comparison, computed with two additional flow programs per row. It measures a specified counterfactual contraction, not an expected decrement under an unknown future observation. Natural-log and earlier practical-DCR experiments use multiplier scores; the separately specified DCR-TRACK experiments use finite contractions.

The same calculation explains the residual proxy. On positive counts set $\theta_i(m) = \sqrt{N_i/(N_i + m_i)}$. At a differentiable point,

$$C(\pi, \mathcal{C}_{\theta(m)}) = C(\pi, \mathcal{C}) - \sum_i d_i + \sum_i \frac{d_i \sqrt{N_i}}{\sqrt{N_i + m_i}} + o\big(\|\theta(m) - \mathbf{1}\|\big).$$

Terms independent of m do not affect the local design, giving $\alpha_i = d_i \sqrt{N_i}$. This is a first-order justification at the current data. A large batch or a shifted empirical center need not obey the expansion, which is why a new stopping certificate is always evaluated after collection.

F.4 CONTINUOUS RESIDUAL ALLOCATION

For positive baseline counts n_i and nonnegative coefficients α_i, consider

$$\min_{x_i \geq n_i,\ \sum_i (x_i - n_i) = B} f(x) = \sum_i \alpha_i x_i^{-1/2}. \tag{40}$$

If some $\alpha_i > 0$ and $B > 0$, KKT conditions give, on each active coordinate,

$$-\frac{\alpha_i}{2x_i^{3/2}} + \lambda = 0, \qquad x_i = (\alpha_i/(2\lambda))^{2/3}.$$

With $\zeta = (2\lambda)^{-2/3}$ this is

$$x_i = \max\{n_i, \zeta \alpha_i^{2/3}\}. \tag{41}$$

The left-hand side summed over coordinates is continuous and nondecreasing in ζ, strictly increasing once a positive-weight row receives samples, and tends to infinity. Thus the budget equation determines the active target uniquely. Zero-weight rows receive no design samples unless all coefficients are zero, in which case any allocation is optimal for this objective and the implementation uses least-filled sampling.

In DCR, $\alpha_i = d_i \sqrt{n_i}$. This is the coefficient obtained by approximating the current contribution d_i as decaying according to inverse-square-root statistical uncertainty. It is not an equality between the actual future policy certificate and $f(x)$. The continuous optimizer is useful because it credits the existing counts before assigning new observations. A static proportional rule instead spends the new budget without subtracting the counts already present at each row.

F.5 Exact integer allocation and the implemented rounding rule

For integer counts, define the reduction from the next sample at row i as

$$\Delta_i(n) = \alpha_i \left(n^{-1/2} - (n+1)^{-1/2} \right).$$

It is nonnegative and nonincreasing in n. Each row provides an ordered list of such reductions. Choosing the largest available reduction, incrementing its row, and repeating B times is exactly optimal for the integer version of (40).

Proof. The objective reduction from any allocation is the sum of prefixes of these lists. If an allocation includes a smaller selected reduction while omitting a larger available reduction on another row, exchanging the corresponding final increment improves the objective. Since reductions within a row are ordered, repeatedly selecting the largest currently available reduction always produces prefixes and gives the B largest reductions consistent with those prefixes. No other feasible prefix selection has a larger sum. □

A heap implements this exact rule in $O((d+B)\log d)$ time. The practical-DCR experiments use a faster continuous solve followed by rounding: floor the targets in (41), then allocate the remaining fewer than d samples by largest discrete marginal. This conserves the batch budget and never removes an existing sample. This rounding implementation is distinct from the exact greedy integer theorem and from the cumulative tracking rule. The statistical safety of DCR depends on the final observed counts and checked certificate, not on exact optimization of the local design objective.

Lemma F.1 (A deterministic rounding bound)**.** *Let x solve* (40)*, with integer $n_i \geq 1$ and integer B, and let $\widetilde{x}$ be the implemented rounded target. Then $\widetilde{x}_i \geq n_i$, $\sum_i (\widetilde{x}_i - n_i) = B$, and*

$$f(\widetilde{x}) \leq f(x) + \frac{1}{2} \sum_i \frac{\alpha_i}{\lfloor x_i \rfloor^{3/2}}.$$

If $m = \min_i \lfloor x_i \rfloor$, then also $f(\widetilde{x}) \leq \sqrt{1 + 1/m}\, f(x)$.

Proof. Because n_i is an integer, $\lfloor x_i \rfloor \geq n_i$. The remaining budget is $\sum_i (x_i - \lfloor x_i \rfloor)$, an integer strictly less than d. Adding that many samples establishes feasibility and exact budget conservation. Since f decreases coordinatewise, $f(\widetilde{x}) \leq f(\lfloor x \rfloor)$. For $k_i = \lfloor x_i \rfloor \geq 1$,

$$k_i^{-1/2} - x_i^{-1/2} = \int_{k_i}^{x_i} \frac{dt}{2t^{3/2}} \leq \frac{1}{2k_i^{3/2}}.$$

Summing proves the additive bound. Finally $x_i/k_i < 1 + 1/k_i \leq 1 + 1/m$ gives the multiplicative bound coordinatewise. Neither argument claims that this rounding is the exact integer optimizer. □

F.6 Positive floors and arbitrary offline coverage

Inverse-square-root design is written on positive counts. In the unit-request theoretical algorithm, zero-count rows can receive a conventional design score while the least-filled safeguard eventually samples them. In the numerical design implementation, the proxy uses $n_i \vee 1$ only for evaluating a score. The true row count and its zero-data confidence set are not replaced by a fictitious observation.

The flow LPs themselves require no positive floor. For example, a known reward-one action that certainly remains in a reward-one state can certify its policy while other actions have no observations, provided the specified confidence sets already imply the required lower bound. Conversely, if a zero-count row permits a better long-run alternative or traps the incumbent in a worse class, the flow certificate exposes that unresolved decision. This is a useful distinction from a rule that acquires at least one sample of every row before considering whether to stop.

F.7 THE GENERAL SAFEGUARDED VARIANT

Algorithm 1 General DCR with policy proposals and a half-least-filled safeguard

Require: Historical row observations, accuracy ε, confidence δ

Set all counts to their historical values; initialize an optional incumbent.
while no policy has been certified **do**
    Form simultaneous row confidence sets from all observations.
    Propose an empirical policy and the optimistic-witness policy of Section 5.
    Compute U once and L for each proposal or incumbent under evaluation.
    Keep the best certified lower bound and its policy; check dual bounds.
    if the checked upper–lower gap is at most ε **then**
        return the policy associated with the lower bound.
    end if
    Use the active decision programs to compute the scores (8).
    Spend a design budget on residual targets (9) and an equal safeguard budget on least-filled rows.
    Collect observations directly or with a learned sample-collection procedure; retain observations from every visited row.
end while

G FROM UNCERTAINTY PRICES TO ACTUAL VERIFICATION TIME

This section proves Theorem 4.2. Its assumptions are stronger than those of the general completion theorem: rewards are known, every transition probability is positive, the true optimal deterministic policy is unique, and the incumbent is fixed. These conditions let us control both empirical-center motion and changes in the active decision. The acquisition rule is an explicitly modified direct sampler. We do not transfer its guarantee to the original half-least-filled implementation.

Tracking allocations is classical in best-arm identification and has also been used for MDP identification (Garivier & Kaufmann, 2016; Al Marjani & Proutiere, 2021). We give a self-contained tracking argument. The perturbation calculation follows standard sensitivity reasoning (Bonnans & Shapiro, 2000), while the allocation calculation is elementary convex optimization (Boyd & Vandenberghe, 2004). The result to establish is that these calculations describe the actual empirical decision certificate and its first passing time.

G.1 MODEL, EVENTS, AND LIMITING COEFFICIENTS

Fix a finite MDP with known rewards and $P_i(j) > 0$ for every row i and successor j. Let ψ be its unique optimal deterministic policy and π a fixed deterministic incumbent. Write

$$\Delta_\pi = g^\psi(P) - g^\pi(P), \qquad u^\nu_{sa} = \mu^\nu(s)\nu(a \mid s), \qquad g^\nu \mathbf{1} + h^\nu = r_\nu + P_\nu h^\nu, \quad \nu \in \{\psi, \pi\},$$

where μ^ν is the stationary distribution and each bias is anchored at one state. Full support implies irreducibility for every policy. Set

$$H_\nu = \operatorname{span}(h^\nu), \qquad w_i = \frac{u_i^\psi H_\psi + u_i^\pi H_\pi}{2}, \qquad J = \{i : w_i > 0\}, \qquad Z = \sum_i w_i^{2/3},$$

and assume $J \neq \varnothing$. These are true-model quantities used in the analysis, not inputs.

For fixed δ, the radii in Appendix A satisfy

$$b_P(n) = 2\sqrt{\frac{\log n}{n}}(1 + o(1)). \tag{42}$$

Let $\mathcal{E}$ be the simultaneous prefix coverage event. On another event $\mathcal{L}$ of probability one, the bounded-variable law of the iterated logarithm, applied to each successor indicator, gives

$$\|\widehat{P}_i(n) - P_i\|_1 = O\left(\sqrt{\frac{\log\log(n \vee 3)}{n}}\right) \quad \text{for every row } i. \tag{43}$$

Only finitely many rows and coordinates are involved. A sampling rule reveals prefixes of these row streams, so this event also controls adaptive counts. In particular, $\mathbb{P}(\mathcal{E} \cap \mathcal{L}) \geq 1 - \delta$. The LIL here is used for asymptotic analysis, not to construct the confidence sets or justify their coverage.

The incumbent may be selected from historical data and then fixed. The arguments hold on the same row event for every deterministic policy; there are finitely many such policies. Thus this selection does not require a sample split or an additional error allocation. We do not require a corresponding uniform statement over an uncountable family of changing randomized incumbents.

G.2 An expansion for the actual empirical certificate

Lemma G.1 (Observed certificate with unequal counts)**.** *On $\mathcal{L}$, uniformly over count vectors whose minimum tends to infinity,*

$$C(\pi, \mathcal{C}_N) - \Delta_\pi = (1 + o(1)) \sum_i w_i b_P(N_i). \tag{44}$$

The counts may be adaptive and need not be proportional. The centers and the optimal solutions are those of the observed confidence sets.

Proof. **Step 1: Establish a stable policy neighborhood.** There is a neighborhood of P in which every transition entry is positive. On this neighborhood the stationary distribution and anchored bias of a fixed policy solve nonsingular linear systems, so its gain is a smooth function of the row entries. The true gain of ψ strictly exceeds that of every other deterministic policy. There are finitely many such policies, and continuity therefore gives a smaller neighborhood throughout which ψ remains optimal. Every empirical confidence set with sufficiently large minimum count lies inside that neighborhood, uniformly in the count vector. Hence

$$U(\mathcal{C}_N) = \max_{K \in \mathcal{C}_N} g^\psi(K), \qquad L(\pi, \mathcal{C}_N) = \min_{K \in \mathcal{C}_N} g^\pi(K). \tag{45}$$

All these kernels are irreducible. Thus both optimistic policy changes and possible changes in recurrent structure have ended after a finite prefix.

Step 2: Differentiate the gains. For a row perturbation E with $\mathbf{1}^\top E_i = 0$, differentiating the Poisson equation and multiplying by μ^ν cancels the bias derivative. This gives

$$Dg^\nu(P)[E] = \sum_i u_i^\nu E_i^\top h^\nu. \tag{46}$$

Smoothness on a fixed neighborhood gives a uniform quadratic Taylor remainder. For sufficiently small radii, the empirical balls also lie in the simplex interior, where

$$\max_{\mathbf{1}^\top z = 0,\, \|z\|_1 \leq \rho} z^\top h = \frac{\rho}{2} \operatorname{span}(h). \tag{47}$$

Step 3: Expand both observed objectives. Put $e_i = \widehat{P}_i - P_i$ and $b_i = b_P(N_i)$. Equations (45)–(47) imply

$$U - g^\psi(P) = \sum_i u_i^\psi e_i^\top h^\psi + \frac{1}{2} \sum_i u_i^\psi H_\psi b_i + R_+, \tag{48}$$

$$L - g^\pi(P) = \sum_i u_i^\pi e_i^\top h^\pi - \frac{1}{2} \sum_i u_i^\pi H_\pi b_i + R_-. \tag{49}$$

The gains only depend on rows selected by their respective policies. Moreover, if $H_\nu = 0$, its Poisson equation says $r_\nu = g^\nu \mathbf{1}$, so this known-reward policy has exactly constant gain under every transition kernel in the neighborhood. It follows that rows outside J do not affect either objective once (45) holds. For a finite constant K,

$$|R_+| + |R_-| \leq K \left[\max_{i \in J} (\|e_i\|_1 + b_i) \right]^2.$$

This explains why a zero coefficient does not hide a second-order uncertainty term.

By (42)–(43), $\|e_i\|_1 = o(b_i)$ uniformly over sufficiently long prefixes. The center terms in (48)–(49) are consequently $o(\sum_i w_i b_i)$. The same holds for the remainders: $\min_{i \in J} w_i > 0$, so their ratio to $\sum_i w_i b_i$ tends to zero even when the radii are unequal. Subtracting the expansions proves the result. All neighborhood and Taylor bounds are uniform over the relevant tail of count vectors. □

The logarithm in (42) is substantive. It makes empirical-center motion smaller than the confidence radius. A fixed-time radius proportional to $n^{-1/2}$, or a confidence sequence with an LIL-size radius, does not admit this argument by direct substitution. The theorem also does not assume that an individual observation decreases the certificate.

G.3 Finite contraction prices consistently estimate the coefficients

Fix $\kappa \in (0, 1)$ and contract only row i from radius b_i to $(1 - \kappa)b_i$ around its empirical center. Write $\mathcal{C}_N^{i,\kappa}$ for this set and use

$$d_i^{(\kappa)} = \frac{C(\pi, \mathcal{C}_N) - C(\pi, \mathcal{C}_N^{i,\kappa})}{\kappa}, \qquad \widehat{w}_i = \frac{d_i^{(\kappa)}}{b_P(N_i)}.$$

At zero count one may use the uniform center and radius two, representing the full simplex. This specifies a nonempty contraction before data are available; its arbitrary initial score has no effect on the following limit.

Lemma G.2 (Consistency of finite prices)**.** *On $\mathcal{L}$, $\widehat{w}_i \to w_i$ uniformly over count vectors with diverging minimum. No unique LP basis or differentiable primal–dual selection is required.*

Proof. Work after (45) holds and the balls lie in the simplex interior. First consider $V(\rho) = \max_K g^\nu(K)$ over the product of balls. The support of the row-i gradient on the zero-sum unit ℓ_1 ball converges uniformly to $u_i^\nu H_\nu/2$ as the kernels approach P.

Take an outer optimizer and contract its row-i displacement toward the center by κ. The result is feasible for the inner problem and changes that row by at most $\kappa\rho_i$ in ℓ_1. Integrating the smooth gradient bounds the value loss above by $\kappa\rho_i[u_i^\nu H_\nu/2 + o(1)]$. Conversely, start from an inner optimizer and add $\kappa\rho_i z$, where z is a zero-sum unit vector maximizing the true row gradient. The triangle inequality puts it in the outer ball; simplex feasibility follows from the interior condition. Integrating the gradient gives a gain of $\kappa\rho_i[u_i^\nu H_\nu/2 + o(1)]$. The bounds remain valid with a zero limiting support. Apply them to g^ψ and $-g^\pi$, add, and divide by $\kappa\rho_i$. □

The argument also controls a genuine one-sided radius derivative where it exists. We use finite prices to avoid attributing a unique derivative to arbitrary multipliers of a degenerate finite LP. Once the stable neighborhood is reached, $w_i = 0$ actually implies $d_i^{(\kappa)} = 0$.

G.4 Cumulative tracking with sparse forced observations

Let t count new requests, $M_i(t) = N_i(t) - n_{0,i}$, and $Q(0) = 0$. At a checkpoint, form

$$\widehat{q}_i = \widehat{w}_i^{2/3} / \sum_j \widehat{w}_j^{2/3},$$

using uniform proportions if the denominator is zero or before the first price update. Hold these proportions fixed until the next checkpoint. At every request $t \geq 1$, first add the current $\widehat{q}$ to Q. If $t = j^2$, request row $1 + (j \bmod d)$, with rows indexed from one; otherwise request a row maximizing $Q_i - M_i(t-1)$. Increment its count after observing the requested row. At most $\lfloor \sqrt{t} \rfloor$ of the first t requests are forced, and each row is forced infinitely often.

For batched implementation, the counter operations between two checkpoints may be performed before collecting the resulting row batch. They depend only on the previous checkpoint's weights. Thus no LP is needed at individual requests. The choices are still nonanticipatory.

Lemma G.3 (Tracking limit)**.** *If checkpoints tend to infinity, this rule samples every row infinitely often and $N_i(t)/t \to q_i^\star$ almost surely for finite historical counts.*

Proof. Forced observations imply diverging counts. Lemma G.2 then gives $\widehat{q}(t) \to q^\star$. Let $e_i(t) = M_i(t) - Q_i(t)$, so $\sum_i e_i(t) = 0$. At a nonforced step the deficits before the increment sum to one, and the chosen deficit is therefore at least $1/d$. Its new error is at most one; errors of unselected rows weakly decrease. A forced observation increases the maximum error by at most one. If F_t is the number of forced requests,

$$\max_i e_i(t) \le 1 + F_t, \qquad \min_i e_i(t) \ge -(d-1)(1+F_t).$$

Since $F_t \le \sqrt{t}$, $(M-Q)/t \to 0$. Cesaro convergence gives $Q/t \to q^\star$, and $n_0/t \to 0$. □

The original half-least-filled rule does not have this conclusion. For example, if one desired proportion exceeds one half and all the other rows have smaller limiting counts, least-filled requests eventually go to those other rows. The largest row can then receive at most the design half of the requests. Replacing this safeguard is an explicit algorithmic change in the present result.

G.5 Proof of the stopping-time theorem

Proof of Theorem 4.2. We work on $\mathcal{E} \cap \mathcal{L}$. For any allocation with limiting proportions $q_i > 0$ on J and diverging counts everywhere, Lemma G.1 yields

$$C_t - \Delta_\pi \sim 2\left(\sum_{i \in J} \frac{w_i}{\sqrt{q_i}}\right)\sqrt{\frac{\log t}{t}}. \tag{50}$$

The deterministic buffer $(1 + \sum_i N_i(t))^{-1}$ is smaller order. By Lemma G.3, this applies to DCR-TRACK with $q = q^\star$. Minimizing $\sum_i w_i/\sqrt{q_i}$ on the probability simplex gives $q_i \propto w_i^{2/3}$ on J and zero elsewhere; the minimum is $Z^{3/2}$.

Let $A_t = C_t + (1 + \sum_i N_i(t))^{-1}$. Coverage gives $C_t \ge \Delta_\pi$, hence $A_t > \Delta_\pi$ at every finite checkpoint. The first passing checkpoint at tolerance $\Delta_\pi + m$ must therefore tend to infinity as $m \downarrow 0$. This rules out a finite-time equality that would otherwise invalidate inversion of (50).

Set $T_m = 8m^{-2}\log(1/m)$. For every fixed $z > 0$,

$$\frac{2\sqrt{\log(zT_m)/(zT_m)}}{m} \longrightarrow z^{-1/2}.$$

For any $\eta > 0$, equation (50) bounds $A_t - \Delta_\pi$ above and below, at all sufficiently large times, by $2(Z^{3/2} \pm \eta)\sqrt{\log t/t}$. The comparison functions are eventually decreasing. There can be no passing checkpoint before a time asymptotic to $(Z^{3/2} - \eta)^2 T_m$, and a checkpoint by a time asymptotic to $(Z^{3/2} + \eta)^2 T_m$ passes. Since $t_{k+1}/t_k \to 1$, the extra wait for a checkpoint is $o(T_m)$. Letting $\eta \downarrow 0$ proves the first limit.

For uniform sampling, $q_i = 1/d$ and the same reasoning gives the second limit. Strict convexity on the positive-weight coordinates, together with the waste of probability mass on zero-weight coordinates, proves that the ratio is strictly below one unless all d weights are equal. The common coverage event applies to both samplers under the same row-stream coupling, and it establishes the usual policy guarantee at every passing stop. □

The limit holds with fixed M, fixed retained π, and fixed δ, as the tolerance approaches the true incumbent loss from above. The sampler is given ε, not Δ_π or m. A truly optimal incumbent has $m = \varepsilon$. We do not infer convergence of expected stopping times from this pathwise statement. We also do not claim finite-sample dominance or information-theoretic optimality among all identification procedures. The optimality here concerns the leading certificate constant among limiting direct-sampling proportions.

G.6 HISTORICAL DATA ON THE CRITICAL SCALE

Fixed finite historical counts vanish from a first-order infinite-sample limit. To express a nontrivial residual effect, consider a family of historical prefixes indexed by $m \downarrow 0$, with the same true model and retained incumbent, such that

$$T_m = 8m^{-2}\log(1/m), \qquad \frac{n_{0,i}(m)}{T_m} \longrightarrow \lambda_i \in [0,\infty), \qquad H(m) := \sum_i n_{0,i}(m) \longrightarrow \infty. \tag{51}$$

The row-stream event controls the entire family of prefixes. This scaling is distinct from the fixed-history theorem. Related offline–online identification work likewise studies how historical observations change allocation constraints (Agrawal et al., 2023); here the constraint is derived for the present flow certificate.

Define $F(x) = \sum_{i\in J} w_i/\sqrt{x_i}$, taking $F(x) = \infty$ if a positive-weight coordinate is zero, and let

$$B^*(\lambda) = \min_{x\geq\lambda}\left\{\sum_i (x_i - \lambda_i) : F(x) \leq 1\right\}. \tag{52}$$

If $F(\lambda) \leq 1$, this value is zero. Otherwise, the positive-weight target is

$$x_i^* = \max\{\lambda_i, cw_i^{2/3}\} \quad (i \in J), \qquad x_i^* = \lambda_i \quad (i \notin J), \tag{53}$$

where $c > 0$ makes $F(x^*) = 1$. These are the KKT conditions for the convex program; the water-filling form itself is standard.

Use a separate variant, DCR-FROZEN. Given historical total H, first collect a pilot until every row has at least $\lceil\sqrt{H}\rceil$ observations. Then compute the finite-price estimates $\widehat{w}_i$ and freeze them. At subsequent nonforced requests, choose the largest available discrete reduction

$$\widehat{w}_i\big[N_i^{-1/2} - (N_i+1)^{-1/2}\big].$$

Retain the square-time round-robin forced requests and use the current certificate with the same positive buffer. Check immediately after the pilot and after each subsequent request. Alternatively, check at the end of the pilot and on the fixed grid $t_k = k^4$ of cumulative new requests. This grid has $o(T_m)$ gaps at every positive scaled residual budget, uniformly over the historical family. The algorithm uses H and data, not m, Δ_π, λ, or the true weights.

Theorem G.4 (Residual stopping with growing historical coverage)**.** *Under the assumptions of Theorem 4.2 and* (51)*, the total new observations of* DCR-FROZEN*, including the pilot, satisfy on* $\mathcal{E}\cap\mathcal{L}$

$$\lim_{m\downarrow 0}\frac{\tau_{\mathrm{Frozen}}(\Delta_\pi + m)}{T_m} = B^*(\lambda).$$

The conclusion includes $B^*(\lambda) = 0$.

Proof. **Step 1: The pilot consistently estimates prices at negligible cost.** Its cost is at most $d\lceil\sqrt{H}\rceil = o(T_m)$, since $H = O(T_m)$. Its minimum count diverges. The uniform tail statements in Lemmas G.1 and G.2 therefore apply to all subsequent count vectors: the frozen estimates converge to w. Once the stable neighborhood is reached, zero-weight rows have exactly zero finite prices, and all positive-weight estimates are positive. The pilot changes the scaled historical lower bounds by $o(1)$.

Step 2: Compare the safeguarded greedy objective with its optimum. For fixed weights, each row's sequence of discrete reductions is nonincreasing. Greedy selection is therefore exactly optimal for the integer objective $\sum_i \widehat{w}_i/\sqrt{N_i}$ at every budget, as proved in Appendix F. Interleaved forced selections do not worsen the objective relative to pure greedy with the same number of design selections: after D greedy selections, the combined selected reductions contain the D largest initial reductions, because each greedy step selects the largest one still available. The objective is also no better than the optimum at the total budget $D + F$. At budgets of order T_m, there are only $F = O(\sqrt{T_m}) = o(T_m)$ forced selections. Thus the safeguarded objective lies between optima whose scaled budgets differ by $o(1)$. Continuous targets and integer targets have the same limit,

since all positive-weight target counts have order T_m and rounding a finite number of coordinates is negligible.

Step 3: Identify the certificate at any positive scaled budget. For fixed $b > 0$, the optimal scaled design value tends to

$$f_\lambda(b) = \min_{x \geq \lambda,\, \sum_i (x_i - \lambda_i) = b} F(x).$$

This follows from compactness, convergence of the frozen coefficients and historical lower bounds, and the KKT conditions. Every positive-weight optimal coordinate is bounded away from zero. The objective comparison in Step 2 gives the same limit for the actual counts, and bounds their positive-weight coordinates away from zero as well. Uniformly on compact subsets of $x_i > 0$,

$$\frac{b_P(N_i)}{m} \longrightarrow \frac{1}{\sqrt{x_i}} \quad \text{when } N_i/T_m \to x_i. \tag{54}$$

Consequently Lemma G.1 gives $(C - \Delta_\pi)/m \to f_\lambda(b)$ at this budget. The buffer is $o(m)$. This step transfers the allocation objective to the actual empirical certificate, including its moving centers.

Step 4: Control the first passing checkpoint. For every $b > B^*(\lambda)$, $f_\lambda(b) < 1$, so the certificate passes by $bT_m + o(T_m)$. Letting $b \downarrow B^*(\lambda)$ proves the upper limit. If $B^*(\lambda) = 0$, this argument already proves the result.

Suppose instead that $B^*(\lambda) > 0$ and a subsequence passes using at most bT_m new observations for some $b < B^*(\lambda)$. The uniform expansion applies after the pilot. Passing bounds every $w_i b_P(N_i)/m$. Monotonicity of b_P and (54) then force N_i/T_m to be bounded away from zero for $i \in J$. The budget bound supplies a further subsequence with $N/T_m \to x$, where $x \geq \lambda$ and $\sum_i (x_i - \lambda_i) \leq b$. The expansion and passing imply $F(x) \leq 1$, contradicting (52). This proves the lower limit. □

Rows whose historical coverage is strictly above the unconstrained level $cw_i^{2/3}$ in (53) eventually receive no design requests before completion; the negligible safeguard can still touch them. At equality, only a statement on the scaled allocation is implied. The condition $H(m) \to \infty$ enables consistent frozen estimates without giving the unknown verification margin to the pilot. Fixed finite history is handled by DCR-TRACK instead. Neither result implies a uniform finite-sample advantage for arbitrary historical datasets.

G.7 UNKNOWN REWARDS WITH INTERIOR MEANS

The same argument permits unknown rewards, provided their true means are away from the clipping boundaries. This extension requires prices that contract both types of uncertainty.

Corollary G.5 (Joint reward and transition prices)**.** *Replace the known-reward assumption by bounded row rewards with unknown means $r_i \in (0, 1)$, and use both confidence constructions in Appendix A. Rewards and successors in a row observation may be dependent. Retain full-support transitions, a unique optimal deterministic policy, and a fixed deterministic incumbent. Define*

$$\widetilde{w}_i = \frac{u_i^{\psi}(1 + H_\psi) + u_i^{\pi}(1 + H_\pi)}{2}.$$

Form a finite price by contracting both the reward interval about $\widehat{r}_i$ and the transition ball about $\widehat{P}_i$ by the same fixed factor $1 - \kappa$, then divide that price by $b_P(N_i)$. The conclusions of Theorems 4.2 and G.4 hold with w replaced by $\widetilde{w}$, including the scale $T_m = 8m^{-2}\log(1/m)$.

Proof. The empirical reward means obey the bounded-variable LIL on a probability-one event, jointly with (43). Independence of the reward and successor coordinates is unnecessary. Since $r_i \in (0, 1)$, interval clipping is eventually inactive. Gains are smooth jointly in the rewards and transitions on the same full-support neighborhood, and

$$Dg^{\nu}(P, r)[E, z] = \sum_i u_i^{\nu}\big(z_i + E_i^{\top} h^{\nu}\big).$$

The strict optimal-policy gap persists in a neighborhood of (P, r). Thus the analogue of (44) has leading uncertainty term

$$\sum_i \left[(u_i^{\psi} + u_i^{\pi}) b_R(N_i) + \frac{u_i^{\psi} H_{\psi} + u_i^{\pi} H_{\pi}}{2} b_P(N_i) \right].$$

The manuscript's radii satisfy $b_R(n)/b_P(n) \to 1/2$, uniformly over long prefixes. Reward-center and transition-center errors are smaller than these radii, and the joint Taylor remainder is quadratic. The leading term is consequently $(1 + o(1)) \sum_i \widetilde{w}_i b_P(N_i)$. Here $\widetilde{w}_i = 0$ only if both policies assign zero occupancy to row i, so zero-weight rows remain exactly irrelevant after policy stability.

For the finite-price argument, the row perturbation set is now the product of a symmetric reward interval and a zero-sum transition ball. Its gradient support is $u_i^{\nu} b_R(N_i) + u_i^{\nu} H_{\nu} b_P(N_i)/2$. Contracting and expanding both components in the sandwich proof of Lemma G.2 therefore proves convergence of the normalized joint price to $\widetilde{w}_i$. The tracking, allocation, buffer, and first-passage arguments are unchanged. □

This corollary does not cover boundary means $r_i = 0$ or $r_i = 1$, where clipping changes the local uncertainty geometry. It also does not apply if the acquisition price contracts only transitions while ignoring unknown reward widths. The known-reward experiments and the joint-uncertainty variant are distinct specifications.

G.8 An empirical-model alternative

Finite contractions are one way to estimate the leading coefficients; they are not necessary for the limiting constant. For example, add a fixed positive pseudocount to each successor count, compute an optimal deterministic policy in this smoothed empirical MDP, and evaluate its stationary occupancy and bias together with those of the incumbent. Substitute these quantities into (10), or the joint-reward formula in Corollary G.5. For an unobserved reward row use any fixed value in $[0, 1]$, and use uniform target proportions when all estimated weights vanish. Use the same tracking, forced requests, actual certificate, and stopping buffer.

Corollary G.6 (Consistent plug-in weights)**.** *Under the corresponding known- or unknown-reward assumptions above, this plug-in tracking rule has the same stopping-time limit as* DCR-Track *in Theorem 4.2.*

Proof. Forced requests make all row counts diverge. Fixed pseudocounts vanish, so the empirical model converges to the true model. Unique optimality implies that its optimal deterministic policy eventually equals ψ. Full support makes the stationary occupancies and anchored biases continuous nearby. Hence the substituted coefficients converge to w or $\widetilde{w}$, and the tracking proof gives the same allocation limit. Since the actual confidence certificate is unchanged, its expansion and first-passage argument give the stated limit. □

The two rules can nevertheless differ before the whole confidence set has a common optimal policy. An action absent from the empirical optimum and incumbent receives zero plug-in weight, even when its confidence interval still permits a strong competitor. Finite contractions price that unresolved competitor through the actual programs. Appendix M.1 compares both rules, including the extra computation needed by finite contractions; neither the common limit nor this distinction implies universal finite-budget dominance.

G.9 Boundaries of the result

The positivity and unique-optimum assumptions are substantive. With a tie between optimal policies, the first-order optimistic objective can be the maximum of several support expressions, rather than the expression for one selected policy. With structural zeros or changing recurrent classes, the simplex tangent set and gain regularity require a different analysis. The theorem concerns a retained incumbent and the current LP certificate; moving candidates or cached bounds from different datasets require separate arguments. These restrictions do not affect the general validity or completion results elsewhere in the paper.

H A QUANTITATIVE NEIGHBORHOOD FOR THE CERTIFICATE EXPANSION

The asymptotic argument in Appendix G eventually enters a neighborhood in which the optimistic policy is stable. Here we give an explicit sufficient neighborhood, finite expansion and price errors, and a conservative count bound for reaching it. These bounds explain the initial regime; they do not establish finite-budget dominance of price tracking.

We retain the known-reward assumptions of Theorem 4.2. There are at least two deterministic policies, the unique optimum is ψ, and π is the retained deterministic incumbent. Define

$$p_{\min} = \min_{i,j} P_i(j) > 0, \qquad \alpha = S p_{\min}, \qquad \gamma = \min_{\sigma \neq \psi} \{g^{\psi}(P) - g^{\sigma}(P)\} > 0.$$

Every true policy kernel minorizes α times the uniform distribution. Rewards in $[0,1]$ imply $\gamma \leq 1$. For a policy ν, retain the notation $u^{\nu}, h^{\nu}, H_{\nu}$ from Appendix G. None of these true-model quantities is supplied to the learner.

At a particular dataset, let b_i denote the empirical ℓ_1 radius and set

$$E_i = \widehat{P}_i - P_i, \quad e_i = \|E_i\|_1, \quad \eta_i = e_i + b_i, \quad \eta = \max_i \eta_i, \quad \eta_{\nu} = \max_{i: u_i^{\nu} > 0} \eta_i. \tag{55}$$

These definitions allow arbitrary unequal counts. The actual errors e_i can be measured in a simulator with known truth; they are generally analysis quantities in deployment.

H.1 FINITE PERTURBATION OF A FIXED-POLICY GAIN

Lemma H.1 (Gain, stationary distribution, and bias perturbations)**.** *Fix a deterministic policy ν. If a stochastic kernel K satisfies $\max_{i: u_i^{\nu} > 0} \|K_i - P_i\|_1 \leq q < \alpha$ and K_{ν} is irreducible, then*

$$H_{\nu} \leq \alpha^{-1}, \qquad \|\mu_K^{\nu} - \mu^{\nu}\|_1 \leq q/\alpha, \qquad |g^{\nu}(K) - g^{\nu}(P)| \leq H_{\nu} q/2. \tag{56}$$

$$\left| g^{\nu}(K) - g^{\nu}(P) - \sum_i u_i^{\nu} (K_i - P_i)^{\top} h^{\nu} \right| \leq \frac{H_{\nu} q^2}{2\alpha}, \tag{57}$$

$$\operatorname{span}(h_K^{\nu} - h^{\nu}) \leq \frac{H_{\nu} q}{\alpha - q}. \tag{58}$$

The last bound does not depend on how the two biases are anchored.

Proof. Write $Q = P_{\nu}$, $R = K_{\nu}$, $\mu = \mu^{\nu}$, and $\mu_R = \mu_K^{\nu}$. The Doeblin minorization contracts zero-mass vectors in ℓ_1 by at most $1 - \alpha$, and contracts spans by the same factor. Applying this to the Poisson equation gives $H_{\nu} \leq \operatorname{span}(r_{\nu})/\alpha \leq 1/\alpha$. Stationarity implies

$$\mu_R - \mu = Q^{\top}(\mu_R - \mu) + (R - Q)^{\top} \mu_R.$$

The second term has ℓ_1 norm at most q, so contraction proves $\|\mu_R - \mu\|_1 \leq q/\alpha$. Each coordinate difference is at most $q/(2\alpha)$ because the difference vector has zero total mass.

Substitution of the true Poisson equation yields the exact identity

$$g^{\nu}(K) - g^{\nu}(P) = \mu_R^{\top}(R - Q) h^{\nu}. \tag{59}$$

Every component of $(R - Q)h^{\nu}$ has absolute value at most $H_{\nu} q/2$, proving the gain bound. Subtracting the linear term leaves $(\mu_R - \mu)^{\top}(R - Q)h^{\nu}$. The zero-mass span inequality bounds this by $\frac{1}{2}\|\mu_R - \mu\|_1 \operatorname{span}((R - Q)h^{\nu})$, which is at most $H_{\nu} q^2/(2\alpha)$.

Finally, the Dobrushin coefficient of R is at most that of Q plus q, hence at most $1 - \alpha + q$. The difference of the Poisson equations gives

$$(I - R)(h_K^{\nu} - h^{\nu}) = (R - Q)h^{\nu} - (g^{\nu}(K) - g^{\nu}(P))\mathbf{1}.$$

Taking spans and contracting by R proves (58). □

The denominator $\alpha - q$ in the bias bound matters. In the two-state chain with both cross-transition probabilities $1/4$ and rewards $(0.8, 0.2)$, $\alpha = 1/2$ and $H = 1.2$. Reducing both cross-transition probabilities to 0.2 changes each row by ℓ_1 distance 0.1 and increases the bias span to 1.5. Its change 0.3 equals $Hq/(\alpha - q)$ and exceeds $Hq/\alpha = 0.24$.

H.2 Policy stability and the observed certificate

The following sufficient condition controls every row, including rows whose eventual price is zero:

$$\eta < \min\{2p_{\min}, \alpha\gamma\}. \tag{60}$$

It implies $\eta < \alpha$, since $\gamma \le 1$. For any plausible row K_i, a coordinate can differ from P_i by at most half its ℓ_1 distance, so $K_i(j) \ge p_{\min} - \eta/2 > 0$. The same argument applies to every zero-sum displacement in the entire radius-b_i ball around the empirical center. Thus simplex clipping is inactive. For every deterministic $\sigma \ne \psi$, Lemma H.1 also gives

$$g^{\psi}(K) - g^{\sigma}(K) \ge \gamma - \eta/\alpha > 0.$$

Consequently ψ remains uniquely optimal throughout the confidence set, and the actual flow objectives equal $\max_K g^{\psi}(K)$ and $\min_K g^{\pi}(K)$. This is stronger than stability only at the empirical center.

Set

$$\Gamma(E) = \sum_i E_i^{\top}(u_i^{\psi}h^{\psi} - u_i^{\pi}h^{\pi}), \qquad W(b) = \sum_i w_i b_i, \qquad R = \frac{H_{\psi}\eta_{\psi}^2 + H_{\pi}\eta_{\pi}^2}{2\alpha},$$

where $w_i = (u_i^{\psi}H_{\psi} + u_i^{\pi}H_{\pi})/2$.

Proposition H.2 (Finite observed-certificate expansion)**.** *Under* (60)*,*

$$|C(\pi,\mathcal{C}) - \Delta_{\pi} - \Gamma(E) - W(b)| \le R, \qquad |C(\pi,\mathcal{C}) - \Delta_{\pi} - W(b)| \le \sum_i w_i e_i + R. \tag{61}$$

If $W(b) > 0$ and $e_i \le \theta b_i$ on all positive-weight rows, then

$$\left|\frac{C(\pi,\mathcal{C}) - \Delta_{\pi}}{W(b)} - 1\right| \le \theta + R/W(b).$$

Proof. Apply (57) uniformly over the confidence set, with $q = \eta_{\nu}$ for each policy. The linear gain perturbation contains its center term and the row support terms. Since clipping is inactive,

$$\max_{\mathbf{1}^{\top}z=0,\ \|z\|_1 \le b_i} z^{\top}h^{\nu} = b_i H_{\nu}/2.$$

Taking a maximum or minimum changes the linear optimum by no more than its uniform remainder. Subtracting the two expansions proves the first inequality. The zero-mass span inequality gives $|\Gamma(E)| \le \sum_i w_i e_i$, proving the rest. □

The center term cancels exactly when $\pi = \psi$. In general, coverage alone only gives $e_i \le b_i$; it does not make the center contribution lower order. This separates the local perturbation argument from the additional probabilistic control of center errors.

Zero weights do not conceal an omitted second-order term here. Full support gives positive occupancy to every row used by a deterministic policy. A zero contribution therefore means either that the policy does not use the row or that $H_{\nu} = 0$. In the latter case its known state reward is constant, and its gain is exactly independent of transitions. If all weights vanish, the stable-neighborhood certificate equals Δ_{π} exactly; no division by $W(b)$ is made.

H.3 A finite error bound for contraction prices

For $\nu \in \{\psi, \pi\}$, define

$$\varepsilon_{\nu}^{\mu} = \frac{\eta_{\nu}}{2\alpha}, \qquad \varepsilon_{\nu}^{h} = \frac{H_{\nu}\eta_{\nu}}{\alpha - \eta_{\nu}}, \qquad \Lambda_i^{\nu} = \begin{cases} \frac{1}{2}\left[H_{\nu}\varepsilon_{\nu}^{\mu} + \min\{1, u_i^{\nu} + \varepsilon_{\nu}^{\mu}\}\varepsilon_{\nu}^{h}\right], & a = \nu(s), \\ 0, & a \ne \nu(s). \end{cases}$$

Proposition H.3 (Finite-price error without inverse contraction dependence)**.** *Under* (60)*, for every $b_i > 0$ and $\kappa \in (0, 1]$,*

$$\left|\frac{d_i^{(\kappa)}}{b_i} - w_i\right| \le \Lambda_i^{\psi} + \Lambda_i^{\pi}. \tag{62}$$

The bound does not require a unique LP basis and has no factor $1/\kappa$.

Proof. The row gradient of a fixed-policy gain, restricted to zero-sum directions, is represented by $u_i^\nu(K)h_K^\nu$. Lemma H.1 gives, throughout the confidence set,

$$|u_i^\nu(K) - u_i^\nu| \le \varepsilon_\nu^\mu, \qquad u_i^\nu(K) \le \min\{1, u_i^\nu + \varepsilon_\nu^\mu\}, \qquad \operatorname{span}(h_K^\nu - h^\nu) \le \varepsilon_\nu^h$$

on a used row. Splitting $u_i^\nu(K)h_K^\nu - u_i^\nu h^\nu$ into occupancy and bias differences, for every zero-sum z we obtain

$$\left|D_i g^\nu(K)[z] - u_i^\nu z^\top h^\nu\right| \le \Lambda_i^\nu \|z\|_1.$$

The derivative is identically zero on an unused row.

For $\max_K g^\nu(K)$, contract the row of an outer optimizer toward its center by κ to obtain an inner-feasible point. The displacement has norm at most κb_i. Integrating the gradient bound controls the value loss above by $\kappa b_i(u_i^\nu H_\nu/2 + \Lambda_i^\nu)$. Conversely, from an inner optimizer add κb_i times a unit zero-sum direction attaining the support of the true gradient. It is outer-feasible by the triangle inequality and the interior condition. The value increase is at least $\kappa b_i(u_i^\nu H_\nu/2 - \Lambda_i^\nu)$. Apply the same argument to $\max g^\psi$ and $\max(-g^\pi)$ and add. Both the error and the displacement contain κb_i before division, proving the claim. □

When $\eta_\nu \le \alpha/2$, a simpler bound is $\Lambda_i^\nu \le \mathbf{1}\{a = \nu(s)\}5H_\nu\eta_\nu/(4\alpha)$. If $w_i = 0$, both error terms vanish and the finite price is exactly zero in this neighborhood. If $b_i = 0$, the displayed normalized finite price is unnecessary and undefined, while the certificate expansion remains valid. These statements concern finite counterfactual contractions, not the decrement produced by one future observation.

H.4 A SUFFICIENT TIME TO ENTER THE NEIGHBORHOOD

Corollary H.4 (Explicit entry under square-time forcing)**.** *Let*

$$r_0 = \tfrac{1}{2}\min\{2p_{\min}, \alpha\gamma\}, \qquad n_{\mathrm{reg}} = \min\{n \ge 1 : b_P(n) \le r_0/2\}, \qquad n_{0,\min} = \min_i n_{0,i}.$$

On the simultaneous coverage event, every count vector with $\min_i N_i \ge n_{\mathrm{reg}}$ *satisfies* (60) *and* $\eta \le \alpha/2$*. The square-time safeguard guarantees this after at most*

$$t_{\mathrm{reg}} = d^2[n_{\mathrm{reg}} - n_{0,\min}]_+^2$$

new direct requests; the first subsequent evaluation therefore lies in the stated neighborhood.

Proof. Coverage gives $e_i \le b_P(N_i)$, so monotonicity of the radius implies $\eta \le 2b_P(n_{\mathrm{reg}}) \le r_0$. Also $r_0 \le \alpha/2$. By time t, every row has received at least $\lfloor\lfloor\sqrt{t}\rfloor/d\rfloor$ forced requests. Substituting t_{reg} proves the count claim. □

The scalar threshold n_{reg} can be computed numerically if its analysis parameters are specified; these parameters are not algorithm inputs. The bound controls all rows and uses only the safeguard, so it may be much larger than the actual entry time. In particular, the square-forcing bound can grow as the fourth inverse power of the neighborhood radius, up to logarithms. No practical finite-budget superiority conclusion follows from it.

H.5 A FINITE TAIL BOUND FOR RELATIVE CENTER ERRORS

An optional quantitative alternative to the LIL controls the ratio $e_i/b_P(N_i)$. Write the un-clipped radius as

$$b_P(n) = \sqrt{2[c + \log n + \log(n+1)]/n},$$

and take $n_0 \ge 2$ large enough that clipping is inactive for all $n \ge n_0$. For every $\theta > 0$,

$$\mathbb{P}\Big\{\exists i,\ n \ge n_0 : \|\widehat{P}_i(n) - P_i\|_1 > \theta b_P(n)\Big\} \le \frac{d\,2^S e^{-\theta^2 c/2}}{1 - 2^{-\theta^2}} n_0^{-\theta^2}. \tag{63}$$

To prove this, for each sign vector in $\{-1, 1\}^S$ consider its centered successor partial sums. Hoeffding's exponential supermartingale gives $\mathbb{P}(\max_{k \le 2L} S_k \ge x) \le e^{-x^2/(4L)}$. On a block $L \le n < 2L$, a violation of the desired bound requires a sign-vector partial sum of at least $x = \theta L b_P(L)$, because

$nb_P(n)$ is increasing. Union bounding over signs gives $2^S e^{-\theta^2 c/2} L^{-\theta^2}$. Sum over rows and the dyadic blocks $L = 2^j n_0$ to obtain (63). The event is uniform in prefix length and therefore also holds under adaptive sampling.

Choosing n_0 to make this bound at most ζ and forcing every row to $\max\{n_{\mathrm{reg}}, n_0\}$ gives the relative expansion bound with explicit finite confidence at least $1 - \delta - \zeta$. The constants can be poor when θ is small. This quantifies, rather than removes, the extra requirement for center motion to be a small part of the uncertainty width.

All results in this section concern known rewards. Unknown reward intervals require additional reward-perturbation terms in both the stability condition and the remainders; transition condition (60) alone does not control a changing reward ordering.

I What the separate certificate can miss

The flow certificate compares two plausible explanations of the data. It is safe, but the two explanations can use different reward or transition models. A direct policy comparison uses the same model on both sides, allowing shared uncertainty to cancel. Maximum-regret policy evaluation has long used this distinction in discounted MDPs with known transitions and uncertain rewards (Regan & Boutilier, 2009; 2010). We give an exact specialization to show what the relaxation costs here.

Proposition I.1 (Exact comparison under known action-independent dynamics)**.** *Suppose every action induces the same known irreducible transition matrix P, with stationary distribution μ. Let rewards vary independently in intervals $[\ell_{sa}, q_{sa}]$, and fix a deterministic incumbent π. With the maximum of an empty set taken as $-\infty$, its separate certificate and shared-model maximum regret are*

$$C_{\mathrm{sep}}(\pi) = \sum_s \mu_s \big[\max_a q_{sa} - \ell_{s\pi(s)}\big], \tag{64}$$

$$R_{\mathrm{shared}}(\pi) := \sup_{r \in [\ell, q]} [g_r^\star - g_r^\pi] = \sum_s \mu_s \max\big\{0, \max_{a \neq \pi(s)} (q_{sa} - \ell_{s\pi(s)})\big\}. \tag{65}$$

Writing $\omega_s = q_{s\pi(s)} - \ell_{s\pi(s)}$ and $k_s = \max_{a \neq \pi(s)}(q_{sa} - \ell_{s\pi(s)})$, the exact excess is

$$C_{\mathrm{sep}}(\pi) - R_{\mathrm{shared}}(\pi) = \sum_s \mu_s [\omega_s - \max\{0, k_s\}]_+.$$

Proof. Every policy has stationary distribution μ, so $g_r^\star = \sum_s \mu_s \max_a r_{sa}$ and $g_r^\pi = \sum_s \mu_s r_{s\pi(s)}$. Optimizing these gains separately gives (64). In their difference, comparison of the incumbent action with itself is exactly zero. Each distinct action has contrast $r_{sa} - r_{s\pi(s)}$, maximized by its upper endpoint and the incumbent's lower endpoint. The reward set factors across states, so these statewise maxima can be attained together. This proves (65). The excess follows by subtracting $\max\{0, k_s\}$ from $\max\{\omega_s, k_s\}$ at each state. □

For a randomized incumbent, the same box support calculation gives

$$R_{\mathrm{shared}}(\pi) = \sum_s \mu_s \max_a \left[(1 - \pi(a \mid s)) q_{sa} - \sum_{b \neq a} \pi(b \mid s) \ell_{sb} \right].$$

Both formulas take $O(SA)$ time once μ is known. On the simultaneous confidence event, they are valid for a data-selected incumbent. They are an exact correction in this specified class; unknown action-dependent transitions do not admit this statewise argument.

An unbounded ratio with multiple uncertain actions. Fix any full-support distribution μ, $S \geq 2$, and $0 < \rho < 1$. Set $P = (1 - \rho) I + \rho \mathbf{1} \mu^\top$. This known chain has distinct rows, full support, and stationary distribution μ. At every state, give incumbent action zero interval $[b, b + W]$ and every challenger interval $[b - \eta, b + \eta]$, where $0 < \eta < \min\{W/2, b\}$ and $b + W < 1$. All intervals have positive width. Each action can be uniquely optimal in a plausible model. Nonetheless,

$$C_{\mathrm{sep}} = W, \qquad R_{\mathrm{shared}} = \eta, \qquad C_{\mathrm{sep}} / R_{\mathrm{shared}} = W/\eta \longrightarrow \infty \quad \text{as } \eta \downarrow 0.$$

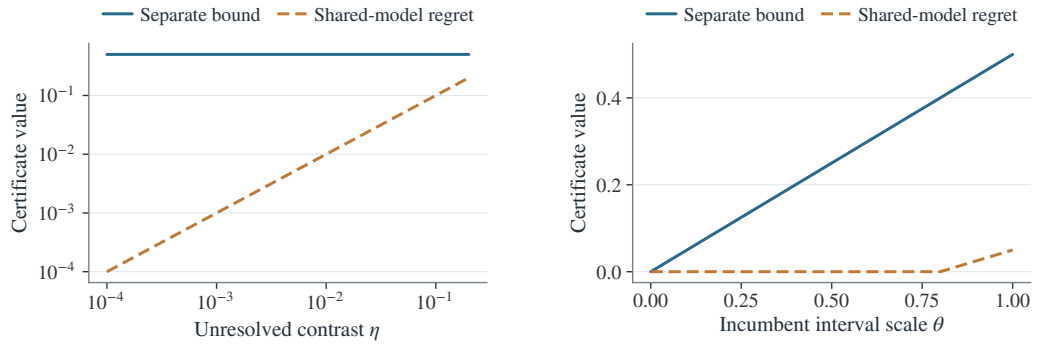
For example, true rewards $r_{s0} = b + \eta/4$ and $r_{s1} = b + \eta/2$, with other challengers below these values, give a unique optimal action and positive incumbent loss $\eta/4$. At tolerance 2η, the shared-model bound already accepts this incumbent whereas the separate certificate fails. The distinction therefore persists when every recurrent state has unresolved action choices.

When a positive price pays for unnecessary precision. Consider instead the true incumbent reward $b + W/2$, with challenger rewards b. Contract only the incumbent intervals toward that midpoint by a factor $\theta \in [0, 1]$. Every resulting set contains this true model, and direct substitution gives

$$C_{\text{sep}}(\theta) = W\theta, \qquad R_{\text{shared}}(\theta) = [\eta - (1-\theta)W/2]_+.$$

Once $1 - \theta \geq 2\eta/W$, the policy comparison is resolved, but the separate certificate still assigns a positive price to reducing the incumbent interval. These are exact counterfactual contractions, not a model of future random observations. They explain why our allocation theorem optimizes the specified certificate rather than all statistically valid decisions.

Independent checks. The accompanying exact-arithmetic script enumerates reward-box vertices and competitor policies on small instances. It checks 55 deterministic and randomized cases, including positive true loss, unique true optima, and both formulas above. The parameter and contraction curves are analytic evaluations, not additional stochastic trials.



(a) Varying η at fixed $W = 0.5$.

(b) Contracting the incumbent interval.

Figure 3: Uncertainty that cancels in a direct comparison can remain costly for the separate certificate. Left: with $W = 0.5$, its value stays at W while the shared-model regret equals η. Right: with $W = 0.5$ and $\eta = 0.05$, contracting the incumbent interval resolves shared-model regret before the separate bound vanishes. Both panels evaluate the exact formulas above.

J Unknown Navigation and Trajectory Realization

J.1 A Sample-Collection Reduction with Learned Dynamics

Statistical requests and interaction steps are different resources. To realize requests without a supplied navigator, one can use an existing unknown-MDP collector. We use the following result as an external ingredient, not as a new theorem of this paper.

Theorem J.1 (Sample collection; Tarbouriech et al. (2021))**.** *In a communicating MDP with finite unknown diameter D, GOSPRL satisfies predictable sampling requirements bounded by a deterministic coordinatewise envelope $\bar{b}_i$. With failure probability at most δ_{nav}, its collection time is*

$$\widetilde{O}\left(D\sum_i \bar{b}_i + D^{3/2}S^2A\right).$$

The algorithm learns navigation from observed transitions and does not require D as input.

The coordinatewise qualification matters. A bound on the total number of adaptive requests does not automatically provide an envelope with that same total: different histories can concentrate their

requests on different rows. We therefore state a frozen-demand construction that applies the collection result conditionally at each phase.

J.2 Doubling floors with frozen design demands

At phase $k = 0, 1, \ldots$, let the current sampling floor be $q_k = 2^k$. At the start of the phase, compute the certificate and stop if it passes. Otherwise, with current counts N, let

$$b_{k,i}^{\mathrm{guard}} = [q_k - N_i]_+, \qquad B_k^{\mathrm{guard}} = \sum_i b_{k,i}^{\mathrm{guard}}.$$

If this total is zero, advance the floor without collecting data. Otherwise use the current decision scores to define a design vector b_k^{design} of nonnegative integer requests whose total is at most B_k^{guard}. Freeze the sum

$$b_k = b_k^{\mathrm{guard}} + b_k^{\mathrm{design}}$$

and invoke a fresh collector for these additional counts. All interactions, including incidental visits, are added to N. Check the certificate again at completion; checking it more often is also safe. The collector's starting state is the actual current state, not a reset state.

Allocate the statistical event probability $\delta/2$ to the row confidence sequence and allocate

$$\delta_{\mathrm{nav},k} = \frac{\delta}{2(k+1)(k+2)}$$

to collection phase k. Conditional on the history at that phase, b_k is fixed and the unused transition innovations have their original row laws. The fixed-demand collection guarantee therefore applies with $\overline{b} = b_k$. A conditional union bound gives success for all invoked collectors with probability at least $1 - \delta/2$.

Corollary J.2 (Residual trajectory envelope)**.** *Let $n_\star$ be the count threshold of Theorem 5.1, with its confidence budget replaced by $\delta/2$. The phase construction returns an ε-optimal policy with probability at least $1 - \delta$, and its interaction time is*

$$\widetilde{O}\left(D \sum_i [2n_\star - n_{0,i}]_+ + D^{3/2} S^2 A\right). \tag{66}$$

The hidden logarithms include the number of doubling phases. If the historical certificate already passes, the interaction cost is zero.

Proof. On the statistical event, any completed floor at least $n_\star$ permits stopping by Theorem 5.1. Let K be the first integer with $q_K \geq n_\star$; then $q_K < 2n_\star$ except for the harmless equality convention at $n_\star = 1$. Before phase k, the current counts dominate n_0. Hence

$$\sum_k B_k^{\mathrm{guard}} \leq \sum_i \sum_{k=0}^{K} [q_k - n_{0,i}]_+ \leq 2 \sum_i [q_K - n_{0,i}]_+.$$

For the last inequality, if $n_{0,i} \geq q_K$ both sides vanish. Otherwise, let k_i be the first active dyadic floor for row i. For one active floor the inequality is immediate. For at least two, the geometric sum of q_k is at most $2q_K$, while the repeated subtraction of $n_{0,i}$ is at least $2n_{0,i}$. The same inequality follows directly.

The total frozen demand is at most twice the total guard demand. Applying Theorem J.1 conditionally in each phase, summing costs, and union bounding failures gives

$$\widetilde{O}\left(4D \sum_i [q_K - n_{0,i}]_+ + (K+1) D^{3/2} S^2 A\right).$$

Use $q_K < 2n_\star$ and absorb $K + 1 = O(\log(n_\star + 1))$ in the logarithms. Incidental observations can only reduce later deficits. The decision certificate gives the returned policy guarantee. □

Equation (66) is a composition with an existing learning-based collector. It is not obtained by assuming free navigation or by multiplying a lower bound by D. The GOSPRL paper already develops the general decoupling of sampling requirements and collection, including policy-learning applications; our use specializes its interface to retained offline counts and the direct decision test.

J.3 A DATA-CHECKABLE HITTING-TIME WITNESS

A proposed router can also be evaluated directly. Fix a target state z and a stationary routing policy ν. Set $H(z) = 0$ and seek $H(s) \geq 0$ satisfying

$$H(s) \geq 1 + \sum_a \nu(a \mid s) \max_{p \in \mathcal{P}_{sa}} p^\top H, \qquad s \neq z. \tag{67}$$

For polyhedral row sets, support-function dualization gives a finite LP in H and the row dual variables. The objective can minimize $\sum_{s \neq z} H(s)$ or a specified starting-state value. The policy ν is fixed during this verification; jointly optimizing ν and H is not asserted to be the same LP.

Proposition J.3 (Verified routing cost)**.** *If $M \in \mathcal{C}$ and a finite H satisfies* (67)*, then $\mathbb{E}_s^\nu \tau_z \leq H(s)$ for all s. The expected cost of obtaining one observation of (z, a) is at most $H(s) + 1$, by routing to z and then choosing a.*

Proof. Under the true rows, on $\{t < \tau_z\}$,

$$\mathbb{E}[H(S_{t+1}) \mid \mathcal{F}_t] \leq H(S_t) - 1.$$

Multiply this inequality by the past-measurable indicator $\mathbf{1}\{t < \tau_z\}$. Because $H(z) = 0$, the stopped increment satisfies

$$\mathbb{E}[H(S_{(t+1)\wedge\tau_z}) - H(S_{t\wedge\tau_z}) \mid \mathcal{F}_t] \leq -\mathbf{1}\{t < \tau_z\}.$$

Sum for $t = 0, \ldots, T-1$ and take expectations. All terms are integrable at this finite horizon because H is a finite vector. Telescoping gives

$$\mathbb{E}_s(T \wedge \tau_z) \leq H(s) - \mathbb{E}_s H(S_{T\wedge\tau_z}) \leq H(s).$$

Monotone convergence gives the hitting-time claim and implies almost-sure hitting. The last sample costs one additional transition. □

If the LP is infeasible, the router has not been certified for all currently plausible rows. That is not evidence that the true target is unreachable. Sparse confidence sets can include spurious closed classes. More interaction or a different routing policy may be required. A verified witness may guide collection or provide an online cost diagnostic, but DCR's statistical correctness does not rely on the witness being feasible.

J.4 THE IMPLEMENTED ROUTER AND ITS QUALITATIVE SAFEGUARD

The trajectory experiment implements a simpler router than GOSPRL. It forms a smoothed empirical transition model, approximately solves a unit-cost stochastic shortest-path problem to the selected state, and follows its greedy routing policy. Actions whose computed routing scores lie within 10^{-10} of the row minimum are treated as tied; the lowest action index is selected. Every action is replaced by an independent uniform action with probability $\xi = 0.15$. At the target state, the specified target action is selected unless this random-action replacement occurs. Targets alternate between decision-price priorities and least-filled rows; the exact switching rule is given in Appendix Q.

This implementation does not inherit the polynomial bound (66). It does have a general exploration property. In a finite communicating MDP, the directed graph containing every positive-probability action transition is strongly connected. Let $p_{\min} > 0$ be its smallest positive transition probability. From any state to a requested state there is a simple path of at most $S - 1$ transitions. Forced random actions follow the path and take the requested final action with conditional probability at least

$$p_\circ = (\xi/A)^S p_{\min}^{S-1} > 0$$

within a block of S steps. The bound does not require the empirical router to be correct.

Conditional failure probabilities across blocks are at most $1 - p_\circ$, so every row is visited infinitely often almost surely. More explicitly, repeated conditional-expectation bounds give a geometric tail for failing to visit a fixed row over consecutive blocks. Repeating after each visit and union bounding over the finite rows implies that every fixed finite count target is reached almost surely. Thus the row confidence widths vanish along the trajectory, and the finite-width argument in Appendix E eventually permits stopping on the confidence event. The resulting crude dependence on $p_{\min}$ can be much worse than a diameter-based collector bound. We use the trajectory experiment to measure actual interaction efficiency, not to claim GOSPRL's rate for this implementation.

K AN ADAPTIVE INFORMATION CONSTRAINT AFTER A WARM START

The uniform residual envelope is sufficient; it need not be necessary for a specific decision. A complementary lower bound can be stated directly in terms of alternative models that require different answers. This is a standard change-of-measure principle specialized to historical observations and adaptive additional counts, rather than a claim of lower-bound matching for DCR. Offline-data best-arm identification already credits historical observations in this way and develops asymptotically matching tracking procedures under its stated offline-data conditions (Agrawal et al., 2023). The result below records the corresponding transcript accounting for our MDP observation model.

K.1 TRANSCRIPT DIVERGENCE WITH HISTORICAL OBSERVATIONS

Fix two MDPs M, M' on the same state and action alphabets. Let ν_i, ν_i' be their reward–successor row laws. Let the initial historical counts $n_{0,i}$ be deterministic and the same under both models. The learner receives the corresponding row-stream prefixes and can thereafter choose queries adaptively. Let τ be its stopping time and $m_i(\tau)$ its additional counts. In a trajectory model, the current state is determined by preceding transitions and is included in the transcript. Any initial-state draw not already determined by the recorded history has the same conditional law under both models; otherwise its relative-entropy contribution must be added.

Lemma K.1 (Adaptive likelihood accounting)**.** *Suppose the stopped transcript laws are well-defined, the procedure terminates almost surely under both models, and the displayed expectation is finite. Then*

$$\mathrm{KL}(\mathbb{P}_M^{\mathrm{tr}} \| \mathbb{P}_{M'}^{\mathrm{tr}}) = \sum_i \big(n_{0,i} + \mathbb{E}_M m_i(\tau)\big)\, \mathrm{KL}(\nu_i \| \nu_i'). \tag{68}$$

If the right side is infinite, the resulting lower-bound inequality remains trivial and valid.

Proof. **Step 1: Use a fixed-length representation of a stopped experiment.** Append a distinguished symbol † at every position after the algorithm stops. The full transcript is now an infinite sequence: the finite historical observations, the algorithm's actions and observations up to τ, the reported policy, and the padding symbols. Include the algorithm's random seed, whose law is the same under both models. Given the seed and preceding observations, every sampling and stopping decision is a fixed function. The post-stop padding law is identical under M and M'.

Step 2: Apply the finite chain rule. Write $\mathcal{H}_T$ for the information in the random seed, the historical data, and the first T online positions of this padded sequence. Action and stopping kernels have zero conditional divergence because the same algorithm is run under both models. At an active observation position with selected row i, the conditional divergence is $\mathrm{KL}(\nu_i \| \nu_i')$. At a padded position it is zero. Consequently,

$$\mathrm{KL}(\mathbb{P}_M|_{\mathcal{H}_T} \| \mathbb{P}_{M'}|_{\mathcal{H}_T}) = \sum_i n_{0,i}\, \mathrm{KL}(\nu_i \| \nu_i') + \mathbb{E}_M \sum_{t=1}^{T} \mathbf{1}\{t \le \tau\}\, \mathrm{KL}(\nu_{I_t} \| \nu_{I_t}').$$

The current state in trajectory sampling is already determined by the recorded past. It contributes no additional independent observation, but its transition into the next state is part of the row law ν_i.

Step 3: Pass to the full transcript without optional-stopping shortcuts. Relative entropy on increasing information fields satisfies

$$\mathrm{KL}(\mathbb{P}_M^{\mathrm{tr}} \| \mathbb{P}_{M'}^{\mathrm{tr}}) = \sup_T \mathrm{KL}(\mathbb{P}_M|_{\mathcal{H}_T} \| \mathbb{P}_{M'}|_{\mathcal{H}_T}).$$

One way to verify this identity is to use the variational formula $\mathrm{KL}(P\|Q) = \sup_f \{\mathbb{E}_P f - \log \mathbb{E}_Q e^f\}$ over bounded measurable f. Conditional expectations of any such f with respect to $\mathcal{H}_T$, under the dominating probability $(P+Q)/2$, remain uniformly bounded and converge to f in $L^1(P+Q)$. Thus their two variational terms converge. Taking the supremum over f proves the reverse inequality to data processing and hence the identity.

Step 4: Count visits. Every conditional divergence in Step 2 is nonnegative. Monotone convergence, followed by Tonelli's theorem, therefore gives

$$\mathbb{E}_M \sum_{t \ge 1} \mathbf{1}\{t \le \tau\}\, \mathrm{KL}(\nu_{I_t} \| \nu_{I_t}') = \sum_i \mathbb{E}_M m_i(\tau)\, \mathrm{KL}(\nu_i \| \nu_i').$$

This proves (68). It also proves the identity in the extended nonnegative reals; no unjustified interchange of signed log-likelihood terms is needed. When the right side is infinite the subsequent lower bound is immediate. □

This is an unconditional statement under the model. If the historical counts were selected by an observation-dependent logging rule, the corresponding expected historical counts and full logging transcript must be included in the likelihood calculation. One cannot condition on the realized offline statistics and continue to treat $n_{0,i}\,\mathrm{KL}(\nu_i\|\nu_i')$ as their conditional information contribution.

K.2 Alternatives that require different decisions

Let

$$\Pi_\varepsilon(M) = \{\pi : \max_s [g_M^\star(s) - g_M^\pi(s)] \le \varepsilon\}.$$

Assume $\Pi_\varepsilon(M) \cap \Pi_\varepsilon(M') = \varnothing$ and an algorithm is (ε,δ)-correct and terminates almost surely on both models, where $\delta < 1/2$. The event $A = \{\widehat{\pi} \in \Pi_\varepsilon(M)\}$ has probability at least $1-\delta$ under M and at most δ under M'. Mapping the transcript to the indicator of A cannot increase relative entropy. Moreover, $\mathrm{kl}(p,q)$ is increasing in p and decreasing in q when $p > q$: its derivatives have signs determined by $\log[p(1-q)/(q(1-p))]$ and $q-p$, respectively. Therefore $p \ge 1-\delta$ and $q \le \delta$ imply $\mathrm{kl}(p,q) \ge \mathrm{kl}(1-\delta,\delta)$. Combining this with Lemma K.1 gives

$$\sum_i \big(n_{0,i} + \mathbb{E}_M m_i(\tau)\big)\,\mathrm{KL}(\nu_i\|\nu_i') \ge \mathrm{kl}(1-\delta,\delta). \tag{69}$$

Thus every such alternative imposes a constraint on additional sampling after the actual warm-start allocation. A lower-bound optimization can minimize $\sum_i z_i$ subject to these inequalities and $z_i \ge 0$, across all relevant alternatives. This is an information-theoretic benchmark, not a computable replacement for the two certification LPs.

Equation (69) explains why a scalar offline sample count cannot generally characterize completion. Samples from rows whose laws agree under a hard alternative contribute no information for distinguishing that alternative, no matter how many such samples are present. It also explains why a sufficient-radius frontier and a best-policy-identification lower bound need not coincide: the latter depends on alternatives that actually change the acceptable policy set.

K.3 A two-action example with an explicit residual law

Consider one state with two Bernoulli-reward actions. Under M, the means are $1/2+\Delta$ and $1/2$. Under M', they are $1/2+\Delta$ and $1/2+2\Delta$, where $0 < \Delta \le 1/8$. Take $\varepsilon < \Delta/2$. Even allowing randomized output policies, the acceptable policy sets are disjoint: under M, the probability of choosing action 1 must be at least $1-\varepsilon/\Delta$; under M' it must be at most ε/Δ.

Only the second row changes. Equation (69) becomes

$$(n_{0,2} + \mathbb{E}_M m_2)\,\mathrm{kl}(1/2, 1/2+2\Delta) \ge \mathrm{kl}(1-\delta,\delta).$$

Moreover,

$$\mathrm{kl}(1/2, 1/2+2\Delta) = -\tfrac{1}{2}\log(1-16\Delta^2) \le \frac{8\Delta^2}{1-16\Delta^2} \le 12\Delta^2.$$

Consequently,

$$\mathbb{E}_M m_2 \ge \left[\frac{\mathrm{kl}(1-\delta,\delta)}{12\Delta^2} - n_{0,2}\right]_+. \tag{70}$$

Putting the same number of historical samples on action 1 instead of action 2 cannot remove this necessary term. This example embeds in finite communicating control problems by making the informative decision recur. Its role is to demonstrate a genuine decision-level residual constraint, not to establish that the dual acquisition rule attains every such constraint.

K.4 Why navigation is not multiplied into this lower bound

The transcript identity counts observed row transitions regardless of how they were reached. An upper bound on a collector's hitting time cannot be multiplied into (69) to obtain a lower bound: one trajectory may collect several informative rows along a single path, and an informed route can be much faster than a worst-case diameter. An interaction-time lower bound requires an additional argument about reachability in the hard model family. Theorem J.1 and Corollary J.2 are upper bounds, with no claim of a matching navigation lower bound for DCR.

L Transient effects, exact examples, and interpretation

L.1 A rare gateway with constant decision relevance

Consider a three-state system with gateway o, productive state G, and unproductive state B. A gateway action a moves to G with probability p_a and to B otherwise. From either G or B, the chain returns to o with probability $1/L$ and otherwise remains in the current state. Rewards are zero at o and B, and one at G. The expected excursion length is L. Taking a gateway action once per regenerative cycle gives

$$d^{\pi}(o) = \frac{1}{1+L}, \qquad g^{\pi} = \frac{L p_{\pi(o)}}{1+L}. \tag{71}$$

For gateway actions with probabilities $p_{\pm} = 1/2 \pm \Delta$, using the inferior action loses

$$g^{\star} - g^{\pi_-} = \frac{2L\Delta}{1+L}.$$

Thus the decision remains of constant importance as the gateway's stationary mass tends to zero. The bias difference between G and B is L: subtracting their two Poisson equations gives $h(G) - h(B) = L$. The product of gateway occupancy and this successor difference is $L/(1+L)$, precisely the sensitivity in (71).

Suppose only the gateway probabilities are unknown, with valid intervals $p_a \in [\ell_a, q_a]$. The other rows and rewards are known in this explanatory example. The two decision programs have the explicit values

$$U = \frac{L}{1+L} \max_a q_a, \qquad L(\pi) = \frac{L}{1+L} \ell_{\pi(o)}, \qquad C(\pi) = \frac{L}{1+L} \left(\max_a q_a - \ell_{\pi(o)} \right). \tag{72}$$

The long excursion is accounted for by flow conservation, without a discounted bridge or a separately supplied span multiplier. For large L, the certificate does not mistakenly downweight the gateway by $1/L$. The row's potential price compensates for its small stationary frequency.

Equation (72) is an interpretation of the flow mechanism, not the uncertainty model supplied in the maintenance, queueing, or RiverSwim experiments. In those experiments every successor coordinate is initially uncertain. Known excursion rows are used here only to obtain an exact closed form.

L.2 Transient mistakes that do not change the gain

The opposite phenomenon is equally important. In the transient example in Appendix C, the inferior return action is used at most once before entering a reward-one recurrent state. It lowers finite-horizon reward and bias, but not average gain. A blanket statement that all transient errors have constant average-reward impact would therefore be false. The gateway example concerns a recurrent decision followed by long excursions; the transient two-state example concerns a one-time decision under the incumbent.

The stationary-flow programs distinguish these cases through their feasible recurrent behaviors. In (71), the gateway has positive recurrent occupancy and its transition choices change the productive fraction of every cycle. In that transient incumbent example, the inferior-return state has zero stationary occupancy. This is why neither stationary frequency alone nor the worst one-step Bellman residual is a complete proxy for decision relevance.

L.3 A LIMITATION FROM DECOUPLING THE PLAUSIBLE MODELS

Two LPs also have a clear cost. Take a one-state, one-action bandit whose mean reward is only known to lie in $[\ell, q]$. There is only one policy, so its regret is zero for every plausible model. Nevertheless,

$$U = q, \qquad L = \ell, \qquad C = q - \ell.$$

The certificate compares the optimistic model against the pessimistic model even though there is no policy choice to resolve. A trivial one-action problem can of course be returned immediately by recognizing that structure. More generally, two policies can share uncertain components whose effects cancel in their regret but not in $U - L$.

This example identifies a remaining statistical gap without adding a new modeling restriction: the flow certificate is a computable sufficient decision bound, not exact shared-model regret. Coupling the incumbent and competitor within one plausible model can be tighter, but generally loses the separability that makes the two programs linear. Our experiments and claims concern the stated computable certificate. The local acquisition rule optimizes its uncertainty, which need not always coincide with the most useful direction for finite-budget policy improvement.

M TESTING THE PRICE-DEPENDENT STOPPING THEOREM

This study checks the mechanism in Theorem 4.2 and Corollary G.5. It uses fully observed samples and recomputes the actual confidence sets; it does not replace random observations by a deterministic radius-contraction curve. The theoretical reference is specific to the flow certificate and confidence radii being tested.

Prespecified tasks. There are three states and two actions. For a distribution μ and state s, define

$$P_{s0} = 0.9\mu + 0.1e_s, \qquad P_{s1} = 0.95\mu + 0.05e_s, \qquad r = \begin{pmatrix} 0.8 & 0.77 \\ 0.5 & 0.7 \\ 0.3 & 0.8 \end{pmatrix}.$$

The three distributions are $(0.6, 0.3, 0.1)$, $(0.85, 0.12, 0.03)$, and $(0.95, 0.04, 0.01)$, labeled Moderate, Skewed, and Rare. The learner is not supplied this transition structure: it estimates six independent rows with full unknown support. The fixed incumbent chooses action zero at every state. Exact evaluation confirms a unique optimum $(0, 1, 1)$ and incumbent losses approximately 0.110968, 0.039541, and 0.013201. The smallest difference between the optimum and another deterministic policy is respectively 0.018968, 0.015077, and 0.005009. These are evaluation quantities, not learner inputs.

Observations, confidence, and allocation. Every row begins with 16 observations. The known-reward study uses seeds 120–127; the unknown-reward extension uses disjoint seeds 160–167. Reward observations in the extension are Bernoulli with the displayed means, independently generated from the successors in this experiment. The corollary itself permits dependence. A common row-stream prefix couples the two methods within each task and seed.

Transition confidence sets are empirical ℓ_1 balls. In the known-reward case the full error budget $\delta = 0.05$ goes to transitions, using

$$b_P(n) = \min\left\{2, \sqrt{\frac{2[S\log 2 + \log(d/\delta) + \log n + \log(n+1)]}{n}}\right\}.$$

The unknown-reward case splits δ equally between transitions and the Hoeffding reward intervals of Appendix A. This changes finite constants but preserves the theorem's leading constants.

DCR-TRACK uses finite half-radius contractions: it re-solves both flow programs after shrinking one row, normalizes twice the certificate decrease by that row's transition radius, and raises these coefficients to power $2/3$ to form its target proportions. Unknown rewards require simultaneous contraction of the corresponding reward interval. We use cumulative-target tracking with square-numbered requests forced cyclically through the rows. Within each batch the targets are fixed; a counter loop determines the requests before their outcomes are revealed. Uniform uses equal target

proportions through the same loop. Both stop on the current complete certificate, with the theoretical buffer $(1 + \sum_i N_i)^{-1}$ and a finite-experiment numerical reserve of 10^{-8}.

Checks occur at zero and k^4 new observations for $k = 4, \dots, 64$, giving 62 checks and a cap of 2^{24}. Every path is run to the cap, regardless of first passage. Each reward-information setting therefore contains 48 paths and 2,976 records. We evaluate margins $m \in \{0.1, 0.05, 0.025, 0.0125, 0.00625\}$ after collection. The tolerance $\Delta_\pi + m$ normalizes this mechanism test by true policy loss; the acquisition process is independent of these thresholds and never receives Δ_π. This is distinct from the fixed deployment tolerance used in the logged-controller experiment.

All margins and censored cases. With known rewards, all task–seed runs stop for every tested margin. At $m = 0.00625$, the ratios of mean added counts for DCR-TRACK over Uniform are 0.5499, 0.3616, and 0.2704; the predicted limits are 0.5576, 0.3643, and 0.2592. With unknown rewards, all runs of both methods stop for margins at least 0.0125. At that margin the ratios are 0.6333, 0.3881, and 0.2739, versus limits 0.6099, 0.3836, and 0.2664. At 0.00625, all 24 tracking runs pass and all 24 Uniform runs are right-censored. The corresponding capped ratios, 0.7546, 0.4706, and 0.3300, are reported as capped quantities, not estimates of uncensored stopping ratios. The cap and all margins were fixed before each study generated observations.

The unknown-reward extension was specified after the known-reward experiment and retains its three task definitions. It is not independent evidence across new task families. Pilot seeds 9120 and 9160 are separate from their respective formal seeds. No pilot or formal outcome was used to select a task, seed, or subset for the reported comparisons.

Evidence for the expansion. At the terminal budget, the mean normalized excess $(C - \Delta_\pi)/\sqrt{\log N/N}$ in the known-reward study is approximately 1.275, 1.036, and 0.889 for tracking, versus predicted constants 1.212, 0.978, and 0.824. Uniform gives 1.697, 1.670, and 1.692, versus 1.623, 1.620, and 1.619. Finite logarithmic terms and center noise remain visible; the theorem does not predict exact equality at this budget. Independent two-state calculations enumerate deterministic policies and analytically optimize their interval endpoint gains, providing a separate check without the flow solver.

Numerical and statistical interpretation. All 5,952 formal records contain the true reward/transition parameters in their specified confidence sets and have certificates above the true loss within the stated numerical tolerance. Maximum discrepancies between flow objectives and independently evaluated dual support bounds are below 5×10^{-15}. The row support routine was separately compared with 200 direct row linear programs. The compiled tracking counter matches an independent scalar implementation over 20,000 requests. These checks establish reproducible numerical evidence, not verified-arithmetic certificates. In particular, the fixed 10^{-8} reserve is a finite experimental convention; the mathematical $m \downarrow 0$ theorem uses only its vanishing buffer.

Figure 2 uses ratios of seed-averaged costs. Its intervals resample the eight paired seeds within each model 20,000 times; they describe variation on that fixed model, not uncertainty across application domains. The three deliberately structured tasks test the proposed asymptotic mechanism. The separate logged-controller study addresses a broader operational question.

Computation. At a checkpoint, Uniform solves two flow programs. Tracking solves these and twelve additional programs for the six one-row contractions. Across all formal checkpoints, measured mean certificate-and-price times are 22.7 ms for tracking versus 3.53 ms for Uniform with known rewards, and 22.3 ms versus 3.39 ms with unknown rewards. These wall-clock measurements exclude sampling, count tracking, and file writing; they describe the recorded host and six-row problems only. Finite-contraction prices therefore trade additional computation for fewer observations. The per-check times do not establish an end-to-end runtime advantage.

M.1 A SIMPLER EMPIRICAL-MODEL ALLOCATION BASELINE

Corollary G.6 shows that directly substituting empirical occupancies and biases attains the same asymptotic constant. We therefore add this comparator rather than attributing that constant uniquely to finite contractions. This study was declared after reviewing the original tracking results; it is a

Table 3: Plug-in and finite-contraction sampling at the reported small margins. Costs are mean capped added observations, with cap 2^{24}. Finite-contraction tracking passes in all eight seeds in every row. Counts in parentheses are plug-in passes out of eight; its censored costs remain in the means. These margins match the main mechanism comparisons; all five margins are supplied in the raw and summary records.

Rewards	Task	Margin	Finite price	Plug-in	Plug-in / price
Known	Moderate	0.00625	606,607	606,607 (8/8)	1.00
	Skewed	0.00625	398,919	398,919 (8/8)	1.00
	Rare	0.00625	292,825	292,825 (8/8)	1.00
Unknown	Moderate	0.0125	2,932,987	15,033,284 (1/8)	5.13
	Skewed	0.0125	1,874,161	14,914,334 (1/8)	7.96
	Rare	0.0125	1,336,336	13,177,304 (2/8)	9.86

paired follow-up comparison, not independent validation on newly selected tasks. Its source and protocol were frozen before a disjoint pilot and its formal execution. The original records remain unchanged.

Matched construction. The baseline receives only row counts, the fixed incumbent, and known rewards or observed reward sums. It adds a pseudocount of 0.5 per successor, ensuring a full-support empirical transition model. Unknown reward means use $(\text{successes} + 0.5)/(N_i + 1)$. On these three-state tasks it enumerates the eight deterministic policies, chooses the empirically optimal one with lexicographic tie handling, and solves for its stationary distribution and bias and those of the incumbent. It substitutes these estimates into w_i or $\widetilde{w}_i$. Policy enumeration is a transparent small-state implementation, not a scalability claim.

The tracking counter, square-numbered forced requests, initial counts, row streams, confidence sets, current-certificate checks, numerical reserve, cap, and all five margins match the finite-contraction study exactly. The known- and unknown-reward trials reuse its respective eight formal seeds per task; pilot seeds 9220 and 9260 are disjoint. There are 24 paths and 1,488 checks in each reward setting. No task, seed, or margin is excluded.

Equal limits do not imply equal finite-budget behavior. With known rewards, all plug-in runs pass at every margin. At $m = 0.00625$, the two rules have identical first passing checkpoints in every paired case. With unknown rewards, plug-in passes at $m = 0.0125$ in only 4 of 24 cases, against 24 of 24 for finite-contraction tracking. The model-specific capped cost ratios are 5.13, 7.96, and 9.86 (Table 3); their paired seed-bootstrap intervals are $[3.94, 5.86]$, $[5.97, 8.95]$, and $[6.11, 12.55]$. These are capped ratios, not uncensored stopping-time ratios. The same four plug-in cases pass at $m = 0.00625$. All 24 pass at $m = 0.1$, so the failure is specific to the requested precision and available budget.

At the terminal budget, every empirical optimum equals the true optimum $(0, 1, 1)$. Plug-in therefore gives zero design weight to row $(0, 1)$, which is unused by either this optimum or the incumbent. Yet this row has only 699–116,599 observations across the unknown-reward trials and can still support an optimistic competitor. A post-hoc calculation at seed 160 finds zero plug-in weight but positive finite-contraction coefficients 0.395, 0.540, and 0.483 for that row in the three models. Replacing only its reward and transition uncertainty by their empirical centers reduces the three certificates from 0.1867, 0.1359, and 0.1046 to 0.1161, 0.0437, and 0.0168. All three then pass at $m = 0.0125$. This counterfactual is an evaluation diagnostic, not an observation or an input to the learner. An accurate empirical best policy need not make its confidence-set competitors irrelevant; this explains the observed transient behavior without implying dominance over every plug-in exploration schedule.

The computation tradeoff. Mean certificate-and-weight time for plug-in is 3.66 ms per checkpoint in both reward settings, compared with about 22 ms for finite contractions. These separate-run timings exclude sampling, counter operations, and writing, and apply only to the eight-policy problems. All 2,976 additional checks contain the true model in their confidence sets and pass the numerical safety check, subject to the same double-precision qualification as the other studies. A cheaper plug-in rule with an uncertainty-dependent exploration floor remains untested.

N A BOUNDED TEST OF RESIDUAL HISTORICAL COVERAGE

This diagnostic tests the distinct DCR-FROZEN rule in Theorem G.4. Its histories follow the critical scaling of that theorem. The experiment was specified before generating its observations and retains finite regimes in which frozen prices are inaccurate. It is not a comparison with the practical Initial-price baseline from the earlier experiments.

Oracle-designed historical fixtures. We use the known-reward Moderate model and fixed incumbent of Appendix M. Let $Z = \sum_i w_i^{2/3}$, $x_i^0 = Z^2 w_i^{2/3}$, and $B_0 = Z^3 \simeq 0.367102$. Thus $F(x^0) = 1$, where $F(x) = \sum_{i:w_i>0} w_i/\sqrt{x_i}$. We fix four historical limits: subcritical history, $\lambda = 0$; partial coverage, $\lambda = 0.4x^0$; heterogeneous overcoverage, $\lambda = x^0 \odot (2, 0, 0.25, 0, 0.25, 0)$ with its second entry then replaced by $0.2B_0$; and boundary coverage, $\lambda = x^0$. The second row has zero true weight, so the heterogeneous fixture deliberately includes irrelevant historical coverage. Their residual constants are respectively 0.367102, 0.220261, 0.083476, and zero.

At each $m \in \{0.1, 0.05, 0.025, 0.0125, 0.00625\}$, set

$$n_{0,i}(m) = \max\{16, \lceil T_m^{1/3} \rceil, \lceil \lambda_i T_m \rceil\}, \qquad T_m = 8m^{-2}\log(1/m).$$

This ensures diverging historical total even for $\lambda = 0$. These are explicitly *oracle-designed simulator fixtures*: the true weights and evaluation margin determine the historical counts. The learner receives only their observed row counts, the known rewards, and the incumbent. This construction tests the specified historical scaling; it does not learn favorable histories without model knowledge or represent a sample of deployment histories. The boundary condition $F(\lambda) = 1$ concerns the leading limiting expression, not whether a finite historical certificate already passes.

Table 4: Mean capped new cost divided by T_m for the historical diagnostic, across all eight seeds. Parentheses show passing counts when below eight. Capped values in censored cells are lower bounds on the unobserved stopping costs, not observed stopping times.

Historical profile	$B^*(\lambda)$	$m = .1$	.05	.025	.0125	.00625
Subcritical	.3671	53.727	111.118 (6/8)	2.579	3.714	1.255 (7/8)
Partial	.2203	1.003	.652	.605	.500	.442
Heterogeneous	.0835	.612	.405	.314	.255	.238
Boundary	0	.658	.405	.342	.272	.214

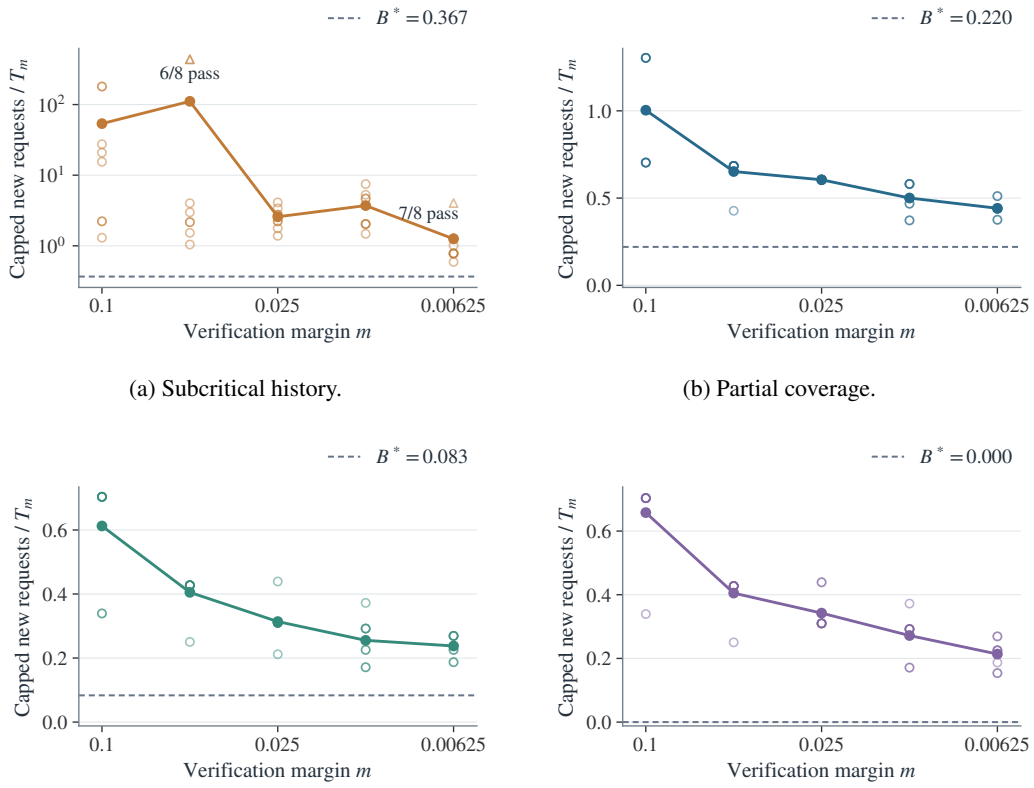
Frozen prices and actual stopping. The pilot tops every row up to $\lceil\sqrt{H}\rceil$, with $H = \sum_i n_{0,i}$, and its observations count toward the new-sample budget. We then estimate the coefficients once using finite half-radius contractions, $\widehat{w}_i = 2(C - C_i^{\text{contracted}})/b_i$, and freeze them. Subsequent nonforced requests maximize $\widehat{w}_i[N_i^{-1/2} - (N_i+1)^{-1/2}]$, with N_i including history and pilot. Ties use the smallest row index. At post-pilot request $t = j^2$, row $1 + (j \bmod 6)$ is forced. The implementation rationalizes the square-root difference to avoid cancellation; it does not replace the discrete marginal by its derivative.

The current complete certificate is checked immediately after the pilot, at cumulative-new-request times k^4 above the pilot cost, and at a cap of 2^{22} new requests. We use the vanishing buffer $(1 + \sum_i N_i)^{-1}$ and the disclosed finite numerical reserve 10^{-8}. All paths continue to the cap, regardless of passing, and first passage of $C + \text{buffer} + 10^{-8} \le \Delta_\pi + m$ is extracted afterward. Neither m, Δ_π, true weights, nor oracle residual targets enter acquisition. As elsewhere, the fixed numerical reserve is a finite experimental convention, not part of the mathematical $m \downarrow 0$ limit.

Seeds 200–207 give eight paired paths per profile and margin; seed 9200 is a disjoint pilot. The protocol and source snapshots were frozen before either stage generated samples. There are 160 formal paths and 7,184 formal checkpoint records, plus 20 pilot paths. All profiles, margins, seeds, and censored cases are retained without changing the cap or tuning the design after observing results.

Results and the finite regime. Table 4 and Figure 4 show all margins. Of 160 formal stopping assessments, 157 pass and three are right-censored: subcritical seeds 205 and 206 at $m = .05$, and

seed 203 at $m = .00625$. The disjoint pilot also has a censored subcritical case at .05, which did not trigger a design change. The three critically scaled profiles have decreasing normalized cost over these margins, but still exceed their limiting constants materially. At the smallest margin, the mean new counts are 459,004.5 for partial, 247,104.75 for heterogeneous, and 222,116.375 for boundary coverage. In particular, the boundary cost decreasing from .658 to .214 does not mean its zero limiting constant has been reached.



(a) Subcritical history.

(b) Partial coverage.

(c) Heterogeneous overcoverage.

(d) Boundary coverage.

Figure 4: All costs are capped new-request counts divided by T_m. Solid lines show seed means; open symbols show individual seeds and triangles denote right censoring. Dashed lines are the residual constants. The subcritical panel has a logarithmic vertical axis. These marks describe variation across eight seeds on one model, not confidence intervals for an expected stopping time.

Subcritical histories expose a more severe transient. At $m = .1, .05, .025$, all eight frozen estimates assign weight below 10^{-12} to at least one row with positive true weight; the counts are seven of eight at .0125 and one of eight at .00625. This cutoff is a post-hoc diagnostic, not an allocation threshold. Such rows then receive essentially only forced observations. At the smallest margin, the mean coefficient ℓ_1 error is .1792 for subcritical history, versus .0097, .0128, and .0064 for the other three profiles. Diverging pilot counts guarantee eventual consistency under the theorem's assumptions, but do not guarantee accurate prices at these finite counts.

The covered profiles reveal a separate finite correction. At their first passing checkpoints for $m = .00625$, the mean signed relative errors of the actual certificate excess against $\sum_i w_i b_P(N_i)$ are approximately .0021, .0182, and .0034 for partial, heterogeneous, and boundary coverage. Yet this finite-radius expression is, on average, about 1.23 times $mF(N/T_m)$ in all three cases. The local expansion is already informative while the logarithmic terms in the confidence radius remain visible. These are descriptive measurements at observed passing checkpoints, not confidence bounds or a causal decomposition.

Verification and scope. A separate audit implementation rebuilt all 7,184 formal transition prefixes directly from inverse-CDF uniforms and rederived every stopping assessment and summary. All recorded confidence sets cover the true rows and all certificates exceed the true loss within the stated numerical tolerance. The maximum flow objective/support discrepancy is 2.45×10^{-15}. A distinct vertex-occupancy LP recomputed pilot and cap certificates on four fixed slices, with value discrepancies below 9×10^{-16} and price discrepancies below 4.6×10^{-14}. The residual constants were also recovered by a general constrained optimizer. Before any observations, the compiled marginal counter matched a scalar implementation over 480,000 requests, including square boundaries, zero weights, and unequal historical counts; twenty complete formal count paths were subsequently replayed exactly.

These results are a bounded mechanism diagnostic. They do not establish finite-budget dominance, convergence of expected stopping costs, or benefits for arbitrary finite histories. Historical totals differ across profiles, so their cross-profile contrasts are not a matched-budget comparison of sampling policies. The subcritical delays, all censored cases, and the visibly nonzero boundary cost remain part of the reported result.

O CONTROLLERS LEARNED FROM BEHAVIOR-POLICY LOGS AND THEN FROZEN

This study evaluates a complete logged-data workflow: collect a behavior-policy log, learn an incumbent from that history alone, freeze it, acquire supplementary observations, and certify the frozen incumbent. Unlike the oracle-defined fixtures in Appendix S.5, no policy or trial is selected using the true transition model or its policy loss. The design, source and dependency hashes were frozen before a disjoint pilot and the complete execution. No pilot outcome changed the task set, historical mechanisms, seeds, thresholds or inclusion rule.

Tasks and naturally visited histories. The factorial design has four families, variants $0, 1$, and $S \in \{5, 12\}$ states with two actions, giving 16 tasks. The first three families are maintenance, queueing and RiverSwim. The fourth, *cyclic patrol*, is independently specified on a ring. For variant v, action 0 moves clockwise with probability $0.64 + 0.04v$, counterclockwise with probability 0.08, and otherwise stays; action 1 moves counterclockwise with probability $0.58 + 0.04v$, clockwise with probability 0.10, and otherwise stays. Each transition row is then mixed with a uniform successor distribution with weight 0.03. With $\theta_s = 2\pi s/S$, known rewards are

$$r(s, 0) = 0.15 + 0.65\frac{1 + \sin\theta_s}{2}, \qquad r(s, 1) = 0.12 + 0.65\frac{1 + \cos(\theta_s + 0.2v)}{2}.$$

The true models instantiate the simulators and are used for evaluation only after the entire acquisition experiment finishes.

For each task and formal seed $80, \ldots, 87$, collect a continuing trajectory starting at state 0. At each visited state, choose a uniformly random action with probability 0.1 or 0.5; otherwise choose the lowest-index action maximizing the known immediate reward. There are no resets, prescribed row counts, coverage repair or outcome-based exclusions. Historical lengths $H = 8,192$ and $65,536$ use prefixes of the same trajectory. The two behavior probabilities share primitive random numbers, while their state paths evolve naturally. The 256 long trajectories contain 16,777,216 transitions and yield 512 incumbent cases, 256 at each H. Changing the behavior or prefix length changes both coverage and the learned policy; this is not a coverage-only causal contrast.

Learning and the information boundary. The learner receives transition counts and the known reward table, not the true kernel. It adds a total transition pseudocount 0.5 per row, spread uniformly across all successor states, and runs the existing deterministic empirical policy iteration from the myopic policy. Its output remains fixed for all subsequent acquisition and certificate checks. The simulator's true transitions are absent from the learning, allocation and certificate interfaces. Supplementation uses the existing *generative row oracle*: a requested state–action row is sampled directly. Historical data are trajectories; supplementary sample costs are not online travel costs. Unknown rewards are outside this experiment's scope.

Table 5: Paired ratios of mean capped added observations, DCR/comparator, at $\varepsilon = 0.1$. Parentheses give 95% task-cluster bootstrap intervals. Ratios below one favor DCR. All 256 cases per historical length remain included.

H	Checker	DCR/Uniform	DCR/Initial price
8,192	Separate	$0.8929\ (0.7178, 0.9741)$	$0.9011\ (0.7410, 0.9782)$
	Common	$1.1087\ (1.0249, 1.2325)$	$1.0130\ (0.9969, 1.0467)$
65,536	Separate	$0.8815\ (0.6739, 0.9648)$	$0.9123\ (0.7421, 0.9744)$
	Common	$1.1500\ (1.0462, 1.2975)$	$1.1189\ (1.0074, 1.2972)$

Three paired acquisitions and two matched checkers. Uniform fills the least-observed terminal row counts. DCR allocates half of each batch by residual water filling with coefficients $d_i\sqrt{\max\{N_i, 1\}}$ from the current separate-certificate price, and half by uniform terminal-count filling. Initial price uses the same rule and safeguard but keeps its initial coefficients throughout. Within each task and seed, all methods and historical conditions use the same separately seeded supplemental row streams. Every path begins with its case's identical historical counts and frozen incumbent.

At each checkpoint, both checkers receive the same current counts, known rewards and frozen policy: the upper/lower flow LPs with separate potentials, and the globally optimized common-potential LP. Both retain the minimum *complete instantaneous gap*; the separate checker does not combine different-time upper and lower endpoints in this comparison. Confidence boxes have full unknown successor support and the existing time-uniform coordinate construction with $\delta = 0.05$. Checks are at zero and rounded $2^{9+k/2}$, $k = 0, \ldots, 24$. A first passage requires the retained checked gap plus 10^{-8} to be at most the requested precision. The primary precision is 0.1; 0.05 and 0.2 are prespecified secondary summaries. All 1,536 paths continue to the 2^{21} cap, producing 39,936 checkpoints. An unpassed case remains right-censored and contributes the cap to the displayed cost mean. These are finite-budget summaries, not unrestricted expected stopping times. Failed cases would remain in the denominator; no learning, acquisition or checker exception occurred.

Controller quality is measured afterward, without selecting cases. At $H = 8,192$, 193/256 learned incumbents have true worst-state loss at most 0.1 and 63/256 exceed it; at $H = 65,536$ the counts are 206/256 and 50/256. The corresponding mean losses are 0.16378 and 0.12690, with maximum 0.71057 at both lengths. All losses above 0.1 occur in RiverSwim. Mean observed row counts are 13.715 and 14.332 across the mixed state sizes; the minimum is four and five rows, respectively. These failures of natural coverage and policy learning are retained in all primary comparisons. Additional verification data cannot improve a controller that is deliberately frozen. The shorter-history group also has nine incumbents with actual loss at most 0.1 but exact-model common-potential gap above 0.1; the longer-history group has none. These are descriptive post-execution diagnostics, not inclusion criteria.

Positive and negative acquisition results. Table 2 reports all cases. For the separate checker, DCR uses 10.7% and 11.8% fewer capped additional observations than Uniform at the two historical lengths, and 9.9% and 8.8% fewer than Initial price. However, the DCR acquisition is constructed from *separate-certificate* prices. Applying the common checker along those same data paths gives *larger* capped costs than Uniform by 10.9% and 15.0%, and fewer common-checker passes. Thus the acquisition advantage is specific to its target certificate; refreshing its prices is not a general improvement for every verifier.

Each of 20,000 bootstrap replicates resamples the 16 task identities while retaining, within the fixed historical-length group, both behaviors, all eight seeds and all methods of each selected task. The random seed is 20260926. Intervals are computed separately for each historical length, precision and checker: they are marginal intervals, not joint inference across those groups. The prespecified paired differences are also supplied. For the separate checker, DCR minus initial-price mean capped cost is $-68,736$ with interval $[-132,244, -15,241]$ at the short history, and $-47,175$ with interval $[-84,389, -15,016]$ at the long history. The synthetic task clusters are not a random sample of application domains.

Table 6: Mean algorithm CPU seconds to first passage or cap at $\varepsilon = 0.1$, with the component accounting stated in the text. All 256 cases per historical length are included.

H	Acquisition	Separate checker	Common checker
8,192	Uniform	0.1239	0.1494
	Initial price	0.1245	0.1607
	DCR	0.1192	0.2999
65,536	Uniform	0.1107	0.1419
	Initial price	0.1120	0.1540
	DCR	0.1070	0.2891

Matched certificate comparison. Across all 1,536 primary paths, the separate checker passes on 1,194 paths and the common checker on 1,091. There are 103 separate-only passes and no common-only pass. This comparison fixes the policy and data and matches complete-gap retention on every path. At $H = 8,192$, the separate-only counts under DCR, Initial price and Uniform are 39, 24 and 13; at $H = 65,536$ they are 17, 10 and zero. The equal pass count in the final Uniform group does not imply equal sample costs: its mean capped cost is 556,875 for separate checking and 796,887 for common checking.

Computation. Table 6 reports declared implementation accounting to first passage or cap. It includes learning, confidence intervals, allocation and the named checker. DCR/common additionally charges the flow solves used by the implemented allocation path, including at its terminal check; initial-price/common charges its initial flow solve. Uniform/common excludes flow solves unnecessary for its procedure. These costs are constructed from measured components, not claimed minimal execution costs. Historical collection and count reconstruction are recorded separately and excluded. A long trajectory is generated once and its collection time is repeated in the metadata for both prefixes; those entries must not be summed or interpreted as separately timed short-history collection. The complete execution of both checkers at all checkpoints took 623.97 wall-clock and 614.38 process-CPU seconds in the pinned macOS arm64 Python environment, with BLAS/OMP thread counts one. The figures are host- and implementation-specific, not an algorithmic-complexity claim.

Independent verification and provenance. A disjoint seed-9000 pilot covered all four families and both sizes at variant 0, with both behaviors and prefix lengths, giving 288 checkpoints on 96 paths. The full audit independently replays all 256 formal natural trajectories, checks all 39,936 row-count budgets, fixed policies, allocation recurrences, confidence intervals, complete-gap running minima and CPU accumulation, and reconstructs all 9,216 stopping records across the three precisions. A preselected 384-endpoint audit covers every task, seed 80, both behaviors, both histories, every method, and added counts zero and 2^{21}. It exactly reconstructs supplemental row streams and production LP prices, and separately constructs occupancy LPs and a primal cutting-plane common-potential solver. Maximum discrepancies are 7.69×10^{-14} for the separate occupancy values and 1.11×10^{-10} for the common optimum. All observed confidence boxes contain the true model and both retained bounds cover actual loss on every recorded path. The largest production flow primal/support discrepancy is 1.06×10^{-10}. These floating-point checks are empirical validation, not interval-arithmetic proofs.

The supplement contains the immutable protocol and executed source, raw trajectories, counts and policies, complete traces, failures census, full first-passage/censoring records, marginal bootstrap ratios and paired differences, family/state/behavior breakdowns, CPU components, and independent audit scripts and outputs under `results/logged_incumbent/`. The frozen experiment source, protocol and original dependencies were not changed after the pilot. This study supports a known-reward, generative-supplement verification result on synthetic 5- and 12-state tasks; it makes no unknown-reward or trajectory-supplement claim.

P FINITE IMPLEMENTATIONS AND NUMERICAL OBJECTIVES

This section describes the compact certificate programs and the earlier box-based policy-proposal implementation. The ℓ_1 tracking and plug-in implementations are specified separately in Appendices M and M.1.

P.1 COMPACT PROGRAMS WITHOUT POLICY ENUMERATION

The upper program uses SA occupancy variables and S^2A successor-flow variables. One normalization, SA row sums, and S balance equations define the flow equalities. For boxes, two inequalities per successor coordinate enforce the transition uncertainty. The upper objective is $\sum_i r_i^+ u_i$.

For a deterministic proposed policy, only its S selected rows need variables in the lower program. This reduces the lower problem to $S + S^2$ variables and the corresponding equalities and box inequalities. For a randomized policy, all action rows remain and the linear equations $u_{sa} = \pi(a \mid s) \sum_b u_{sb}$ are imposed. Our code evaluates deterministic proposals, while the theorems allow randomized ones.

The flow certification programs do not enumerate A^S policies. Ground-truth evaluation of the five-state tasks enumerates $2^5 = 32$ policies; the separate scaling study uses up to 32 states without enumeration. The plug-in baseline separately enumerates eight empirical policies on its three-state tasks. A flow program solves one certification query, not the globally best lower bound over all policies.

P.2 POLICY PROPOSALS AND THE ROLE OF SMOOTHING

At each check, the policy-proposal implementation considers two candidates. The first is lower-greedy with respect to the optimistic relative-value-iteration witness from Appendix E. The second is average-optimal for a smoothed empirical model. Its proposal transition is

$$\widetilde{P}_{ij}^{\mathrm{prop}} = \frac{K_{ij} + 1/(2S)}{N_i + 1/2}. \tag{73}$$

Positive transition probabilities make each candidate policy in this proposal model irreducible. Policy iteration solves its gain–bias equations with a fixed anchoring coordinate and improves the Bellman actions until the policy stabilizes. The smoothing is a computational device for proposing a policy, not a statistical prior asserted to contain the true transition.

Each proposed policy is then evaluated by the lower flow program using the original, unsmoothed confidence intervals (19). The upper flow program is shared. When both proposals are identical, only one lower program is necessary. If they differ, the better lower bound determines the instantaneous selected candidate. A previously evaluated policy can remain the returned incumbent through (20). The proposal family thus affects efficiency, but cannot invalidate an accepted certificate.

P.3 HOW DUAL PRICES ARE EXTRACTED

The upper problem is passed to a minimization solver with objective $-r^+u$. The lower problem minimizes r^-u. The reported inequality marginals are converted to nonnegative maximization multipliers before applying (8). Occupancies are read from the primal variables of the same solve. The two sets of row contributions are added and small negative numerical residues are clipped to zero.

Potential witnesses can be read from the multipliers of the state conservation equations, with a sign change depending on the upper or lower objective convention. We independently evaluate the upper and lower support expressions at these witnesses. An inconsistent objective sign would violate either the primal–dual comparison or the inequality between the flow and common-potential certificates. The analytic two-state example gives an additional check that is sensitive to swapping an optimistic and a pessimistic row optimization.

No part of the score computation uses the true optimal policy, true bias, true transition probabilities, or a ground-truth policy loss. The true model is accessed only by the simulated environment and the external evaluation routines. For the static allocation ablation, the initial scores and initial counts are stored once; the dynamic version recomputes its scores after every batch.

P.4 Numerical termination and fallback behavior

The optimistic witness uses tolerance 10^{-7} on the span of its Bellman residual and a finite implementation iteration cap. Hitting this cap raises an error rather than silently claiming a valid proposal. The anchored solver in Appendix E supplies a guaranteed alternative without a fixed iteration cap. In all reported lazy-solver tasks the cap was not reached. A separate implementation interleaves the lazy and anchored paths, accepting only a checked residual; that implementation is tested independently rather than retroactively attributed to the original traces.

The flow solver uses primal and dual feasibility tolerances of 10^{-9}. Returned values are checked by separate row support computations. A discrepancy above 10^{-5} raises an error. Reported stopping counts use a further 10^{-8} threshold margin; the raw traces retain the unrounded numerical certificate. These tolerances are far below the plotted accuracy levels. A deployment requiring machine-certified probability bounds would need outward-rounded confidence endpoints and certified support arithmetic, as made explicit in (25).

The continuous water-level search uses a feasible upper bracket based on the sum of positive weights, followed by 80 bisection iterations. Target counts are floored and remaining samples assigned by marginal gain. When all design weights vanish, the batch uses least-filled sampling. The design budget and safeguard budget are kept separate, so the safeguard is not lost when some rows have zero or numerically tiny prices.

P.5 A finite LP for the hitting witness

For a box $\mathcal{P} = \{p : \ell \le p \le q, \mathbf{1}^\top p = 1\}$, its support function has the finite dual

$$\max_{p \in \mathcal{P}} p^\top H = \min_{\theta \in \mathbb{R}, z \ge 0} \left\{ \ell^\top H + (1 - \mathbf{1}^\top \ell)\theta + (q - \ell)^\top z : \; z_j \ge H_j - \theta \right\}. \tag{74}$$

To derive it, write $p = \ell + x$, $0 \le x \le q - \ell$, and $\mathbf{1}^\top x = 1 - \mathbf{1}^\top \ell$, then dualize the mass equality and upper bounds. Substitution of one copy of (74) per routing row into (67) yields an LP in H, θ, z. Existence of dual variables satisfying the substituted inequalities is equivalent to the support inequality, by finite LP duality.

The numerical routing-witness experiment fixes the empirical routing policy, solves this LP, and separately computes its true hitting times where finite. It records infeasibility as failure to certify, not as an infinite true hitting time. This experiment checks the usefulness of the witness as confidence improves; it is separate from the continuous-trajectory comparison and does not grant that trajectory an oracle router.

Q Experimental tasks and evaluation protocol

Q.1 Three continuing-control families

The earlier policy-proposal task grid has states $\{0, \dots, S-1\}$ with $S = 5$, actions $a \in \{0, 1\}$, and four variants $v \in \{0, 1, 2, 3\}$ per family. All rewards are known functions of the state and action. Transition probabilities, structural zeros, optimal policies, diameter, and navigation policies are not given to the learner. Each confidence row initially allows all S successor states. The problems are synthetic application-motivated control models, not measurements from a deployed system.

Maintenance. The state is a degradation level. Operation ($a = 0$) increases the level to $\min(S - 1, s + 1)$ with probability

$$p_s = 0.12 + 0.035v + 0.025s$$

and otherwise leaves it unchanged. Preventive service ($a = 1$) returns to state 0 with probability $q = 0.72 + 0.04v$ and otherwise moves to $\max(0, s - 1)$. Coincident destinations have their probabilities added. The rewards are

$$r(s, 0) = 1 - \left(\frac{s}{S-1}\right)^{1.7}, \qquad r(s, 1) = 0.40 - 0.10\frac{s}{S-1}.$$

Operation earns more in a healthy state but can trap an always-operate policy at the most degraded state. Service trades current reward against future condition. The MDP is communicating because

operation can move upwards and service can return to lower states. Some stationary policies are not irreducible; no all-policy ergodicity assumption is imposed.

Queueing. The state is a finite queue length. Set the arrival probability $\lambda = 0.27 + 0.045v$ and service probabilities

$$\mu_0 = 0.24 + 0.025v, \qquad \mu_1 = 0.70 + 0.025v.$$

A step increases the queue with probability $\lambda(1 - \mu_a)$, decreases it with probability $(1 - \lambda)\mu_a$, and otherwise leaves it unchanged. At either boundary, attempted outward movements are clipped to that boundary. The reward balances congestion and faster service:

$$r(s, a) = 1 - 0.62\frac{s}{S - 1} - 0.22a.$$

This family produces state-dependent service decisions. Its nearest-neighbor structure also prevents a common positive reset component from being available in every row. The learner is not told that nearest-neighbor transitions are the only possible transitions.

RiverSwim. Action 0 moves deterministically to $\max(0, s-1)$. Action 1 moves left with probability 0.08, right with probability $0.24 + 0.035v$, and otherwise stays. Destinations are clipped at the boundaries. The only nonzero rewards are $r(0, 0) = 0.05$ and $r(S - 1, 1) = 1$. This is a specified RiverSwim-type variant, rather than a claim of exact replication of every parameter in the original benchmark (Strehl & Littman, 2008). Reaching the right end requires repeated exploration against the locally available left reward.

Q.2 Historical counts and generative completion

For each task, the historical log has exactly 4,000 observations. Before normalization, action weights are $(8, 1)$ when v is even and $(1, 8)$ when v is odd; state weights decrease linearly from 1.5 at state 0 to 0.5 at state $S - 1$. Multiply the state and action weights, allocate 3,990 observations proportionally after flooring, add one observation per row, and distribute the remaining count deterministically in row order. The same counts are used by all methods. They do not depend on the true optimal policy or a measured treatment effect.

Each task is run with 24 row-stream seeds. Methods share the same independently generated stream for each state–action pair, so the first n observations of a given row are identical across methods when their counts agree. Allocation choices still lead to different prefixes at different rows. The error level is $\delta = 0.05$ for the complete uniform-in-count confidence event of (17).

All methods are evaluated at additional budgets

$$0, 2^9, 2^{10}, \ldots, 2^{21}.$$

The primary stopping threshold in this study is $\varepsilon = 0.1$; thresholds 0.2 and 0.05 are sensitivity evaluations of the same traces. The first checkpoint whose cached certificate plus numerical margin is at most ε is the observed stopping count. Continuing the trace after this point permits other thresholds and fixed-budget comparisons, but post-stop observations are not charged to that stopping time.

Q.3 Allocation methods and fair policy comparison

Uniform directs every new observation to a currently least-filled row, including the historical counts. This is a residual uniform baseline, not a deliberately weakened equal-new-count allocation. *Static* computes the initial decision price $d_i^{(0)}$, forms $\alpha_i^{(0)} = d_i^{(0)}\sqrt{n_{0,i}}$, and allocates each design batch proportionally to $(\alpha_i^{(0)})^{2/3}$ without refreshing the profile. *DCR* refreshes prices and computes the residual targets from the current counts.

Both Static and DCR spend half of each budget increment on their design rule and the remainder on least-filled sampling. Uniform spends the entire increment on least-filled sampling. The half-budget safeguard is held fixed across all families. Every method uses the same two-proposal portfolio, the same flow-based evaluator, the same confidence intervals, and the same retained-evidence cache. Thus the primary stopping comparison changes allocation, not the final planner or confidence level.

An additional proposal ablation omits the empirical proposal and evaluates only the optimistic-witness proposal. It retains the same sampling rules and task grid. This ablation was run before finalizing the two-proposal implementation and is reported in Appendix R; the final comparison uses the same expanded portfolio for every method. We do not describe the sequential development of the implementation as a preregistered study.

Q.4 True policy loss and multichain evaluation

For each of the 32 deterministic policies in a five-state task, its true transition matrix is decomposed into recurrent classes and transient states. A stationary distribution is solved within each recurrent class. Transient-state gains are obtained by solving absorption probabilities into those classes. The resulting vector is g_M^π, including cases in which it depends on the initial state. The optimum $g_M^\star$ is computed by policy enumeration for evaluation only.

At each checkpoint the recorded policy is the policy associated with the best retained lower bound, not necessarily the most recently proposed policy. Its loss is

$$\max_s \{g_M^\star(s) - g_M^{\pi_t^{\mathrm{keep}}}(s)\}.$$

This calculation uses the true transition model as an evaluation instrument. The learner's acquisition and certificate computations do not access it. A separate invariant-subspace calculation can evaluate the same gain vectors without using the recurrent-class decomposition.

Q.5 A continuous-trajectory comparison

The trajectory grid uses variants 0 and 3 from each family, with eight paired seeds per task. Historical data are obtained by running 2,000 steps from state 0, choosing action 0 with probability 0.9 and action 1 otherwise. The two online methods start from the same final historical state and historical observations. There is no reset.

Random actions chooses uniformly at every step. *Learned routing* uses the empirical shortest-path router described in Appendix J. At the start of each 1,024-step interval, odd-numbered intervals target a least-filled row, whereas even-numbered intervals target the maximum of

$$d_i \left(1 - \sqrt{\frac{N_i \vee 1}{(N_i \vee 1) + 128}}\right).$$

This is the proxy reduction for a batch of 128 observations. At the target state the router chooses its target action. With probability 0.15 an independent uniform action replaces the selected action. After 64 target-row observations, the target and empirical router are updated using the current counts and the last certificate's prices. At every 1,024-step checkpoint the complete certificate and prices are recomputed.

The router's empirical model is smoothed as $(\widehat{P} + 0.02/S)/(1.02)$ and its unit-cost value iteration runs for at most 300 iterations. This limit affects only the proposed routing policy. It does not assert a known navigation bound and it does not alter the confidence sets. All methods run for 65,536 additional interaction steps; every step, including travel and forced exploration, updates the relevant row count and is charged to the online budget.

Q.6 Evaluation metrics and uncertainty

The experimental unit for comparing methods is a base MDP task, not an individual checkpoint or a method–seed record. For this grid's stopping comparison, first average stopping counts over the 24 seeds within each task, preserving the paired methods. The reported relative reduction is one minus the ratio of the equally task-weighted means. A paired cluster bootstrap resamples the 12 tasks with replacement, keeping every method and seed of a selected task together, and recomputes this ratio. Intervals are the 2.5th and 97.5th percentiles of 20,000 bootstrap replicates.

Fixed-budget summaries average loss over the 13 positive log-spaced checkpoints within each seed and then use the same task aggregation. This is a mean over a specified logarithmic grid, not an

integral over all budgets. Absolute differences are used when losses are zero or small. Trajectory summaries resample the six trajectory tasks. Family-specific intervals have only four or two distinct tasks, respectively, and are correspondingly limited. The intervals summarize heterogeneity in this finite synthetic task grid; they are not guarantees of performance on an unspecified population of real control systems.

A run that has not passed its threshold by the largest budget is right-censored. We report its stopping fraction and a capped count summary when needed. We do not average only the successful runs and present that number as an unconditional stopping cost. At $\varepsilon = 0.1$, all methods stop in every count-model run, so the primary ratio uses no censoring convention.

R Additional empirical results

R.1 Stopping counts by task and accuracy

Table 7 expands the earlier policy-proposal study at $\varepsilon = 0.1$ by task. The baseline is already residual Uniform: it fills the smallest terminal counts, rather than giving equal new counts to unequal historical rows. DCR and Static receive only half the new budget for design, with the other half assigned to this same safeguard. Thus the stopping improvement is not obtained by removing protection from a weak baseline.

Table 7: Mean additional observations to certify $\varepsilon = 0.1$ for each task, averaged over 24 paired row streams. All 288 runs per method stop before the cap.

Task	DCR	Uniform	Static
Maintenance 0	555	555	512
Maintenance 1	2,048	4,096	2,987
Maintenance 2	512	512	512
Maintenance 3	2,048	3,755	2,304
Queueing 0	32,768	65,536	65,536
Queueing 1	43,691	65,536	65,536
Queueing 2	65,536	65,536	65,536
Queueing 3	65,536	73,728	68,267
RiverSwim 0	524,288	1,048,576	1,026,731
RiverSwim 1	524,288	764,587	524,288
RiverSwim 2	262,144	524,288	502,443
RiverSwim 3	262,144	524,288	262,144

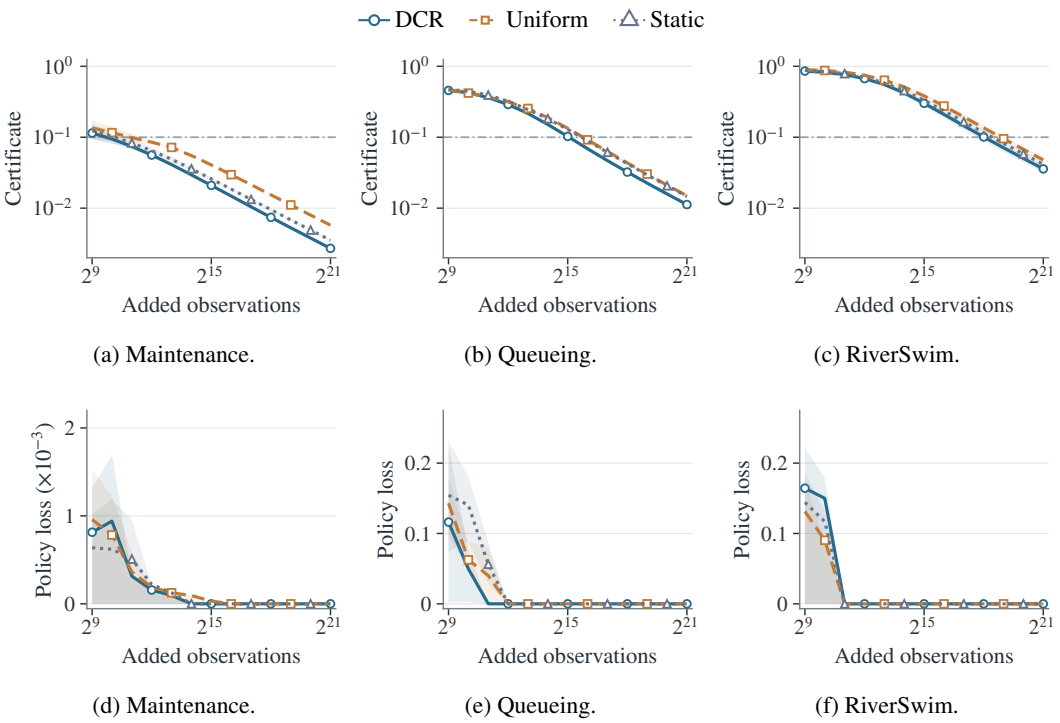
Table 8: Accuracy sensitivity. Counts are capped at 2^{21} additional observations for unstopped runs; their fractions are displayed rather than discarded.

ε	Allocation	Mean capped count	Stopped
0.20	DCR	37,986	288/288
0.20	Uniform	70,782	288/288
0.20	Static	54,171	288/288
0.10	DCR	148,796	288/288
0.10	Uniform	261,749	288/288
0.10	Static	215,566	288/288
0.05	DCR	581,675	288/288
0.05	Uniform	797,184	240/288
0.05	Static	703,090	264/288

At $\varepsilon = 0.05$, averaging only runs that stopped would misleadingly make the harder baselines look inexpensive. Table 8 instead gives both stopping fractions and mean counts capped at the common

Table 9: Policy loss averaged over 13 positive budget checkpoints, four tasks, and 24 seeds per family. All methods return the policy with their best retained certified lower bound. Lower is better.

Family	DCR	Uniform	Static
Maintenance	0.00018	0.00020	0.00016
Queueing	0.01276	0.01890	0.02700
RiverSwim	0.02418	0.01705	0.02009



(a) Maintenance. (b) Queueing. (c) RiverSwim.

(d) Maintenance. (e) Queueing. (f) RiverSwim.

Figure 5: Direct one-step sampling: decision certificates (top) and policy loss (bottom). Every panel retains all four tasks and 24 seeds in its family. Curves average seed means within tasks and then average tasks; shading is the interquartile range of the four task means. Policy-loss panels use different vertical scales. The horizontal certificate reference is $\varepsilon = 0.1$. All allocation methods use the same proposals and evaluator.

maximum budget. Those capped values are not estimates of the unrestricted mean stopping time. For $\varepsilon = 0.1$ and 0.2, all runs stop and no capping affects the means or bootstrap effects.

R.2 Certificate contraction and fixed-budget policy loss

Figure 5 gives the complete positive-budget curves. Lines average over task means after averaging seeds; the shaded band is the interquartile range of the four task means at each checkpoint, not a confidence interval. This display retains between-task variation without treating repeated checkpoints as independent observations.

The residual allocation substantially lowers certificate values at a given budget in the more difficult tasks. The loss curves tell a different story. On the queueing family DCR improves over the baselines on average, while on RiverSwim Uniform has a lower mean loss over the complete budget grid. An upper bound can continue shrinking after the selected policy has stabilized, and the policy with the strongest certified lower bound need not be the best empirical policy at each checkpoint. These effects reinforce the distinction between stopping cost and policy loss.

Table 10: Mean policy loss on the positive log-spaced budget grid. The witness-only and two-proposal variants use the same allocation protocols but may follow different data-dependent sampling paths.

Family	Allocation	Witness only	Two proposals
Maintenance	DCR	0.00076	0.00018
Maintenance	Uniform	0.00028	0.00020
Maintenance	Static	0.00129	0.00016
Queueing	DCR	0.03221	0.01276
Queueing	Uniform	0.02706	0.01890
Queueing	Static	0.03610	0.02700
RiverSwim	DCR	0.02418	0.02418
RiverSwim	Uniform	0.01705	0.01705
RiverSwim	Static	0.02009	0.02009

Table 11: Terminal means after 65,536 real interaction steps, with two tasks and eight paired seeds per family.

Family	Collector	Policy loss	Certificate
Maintenance	Learned routing	0.0000	0.0139
Maintenance	Random actions	0.0000	0.0164
Queueing	Learned routing	0.0000	0.0770
Queueing	Random actions	0.0000	0.0788
RiverSwim	Learned routing	0.2278	0.4645
RiverSwim	Random actions	0.1161	0.8615

R.3 Proposal-pool ablation

The two-proposal portfolio matters independently of allocation. Table 10 compares the final implementation to the ablation retaining only the certificate-oriented proposal. Adding the empirical proposal improves the lower-bound portfolio monotonically at a fixed dataset, by the argument in Appendix C. Its eventual policy-loss effects also depend on the changed acquisition path. For this reason the table is an end-to-end ablation, not a pointwise causal decomposition of planner and sampler effects. Every allocation method receives the same proposal expansion.

R.4 Trajectory results and censoring

Table 11 gives terminal values after the common real-interaction budget. At $\varepsilon = 0.1$, both methods stop in 32 of 48 trajectory runs; all unstopped runs belong to RiverSwim. At $\varepsilon = 0.2$, learned routing stops in 36 runs and random actions in 32. These stopping fractions and the capped costs are reported in Table 12; no unstopped run is discarded.

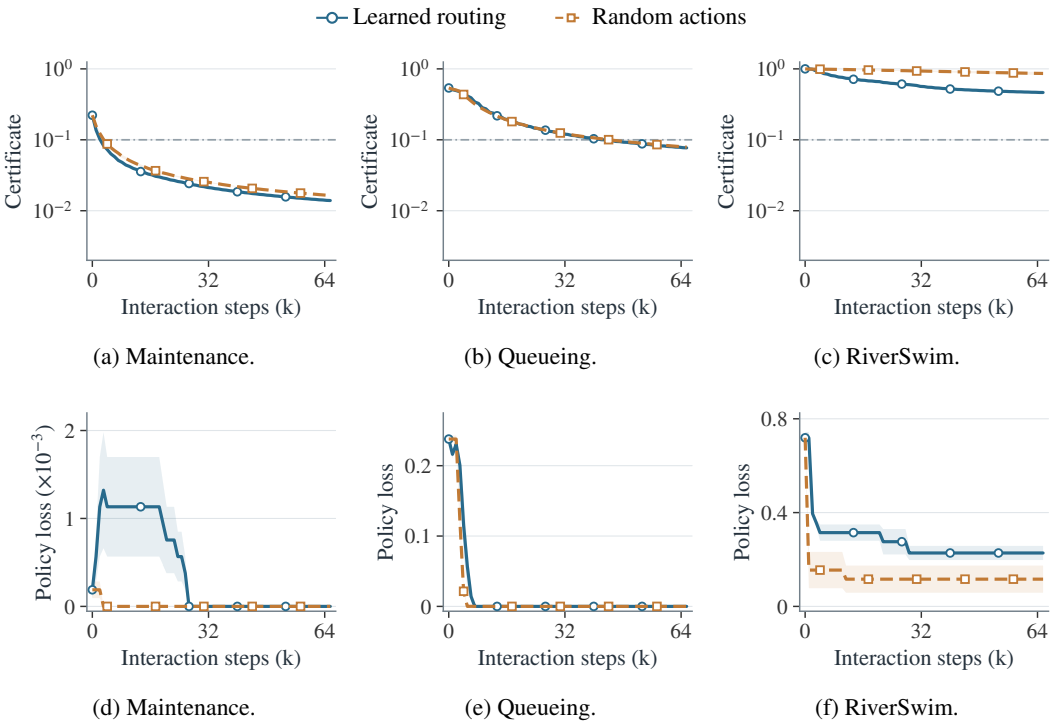
The trajectory curves in Figure 6 use the true number of environment steps on the horizontal axis. Travel, overshoot, random actions, and repeated visits are all counted. The two methods start with the same logged trajectory and current state, but choose different subsequent actions. Learned routing lowers the mean RiverSwim certificate, but its mean terminal policy loss is worse than random actions. Neither collector reaches the stricter certificate target there by the finite cap.

R.5 Numerical tie handling and trajectory sensitivity

The empirical shortest-path router can assign essentially identical values to two actions. A raw floating-point argmin may then choose different actions across numerical environments, changing subsequent visits, estimated rows, and acquisition priorities. We make the convention explicit: among actions with scores within 10^{-10} of the minimum, take the smallest action index. The value-iteration tolerance remains 10^{-7}; neither the task definitions nor the forced-exploration probability is changed. This routing convention affects the empirical heuristic, not the validity of a subsequently checked policy certificate.

Table 12: Trajectory stopping fractions and counts capped at 65,536 steps. Unstopped RiverSwim runs are retained.

ε	Collector	Mean capped steps	Stopped
0.1	Learned routing	37,248	32/48
0.1	Random actions	38,165	32/48
0.2	Learned routing	25,621	36/48
0.2	Random actions	27,328	32/48



(a) Maintenance. (b) Queueing. (c) RiverSwim.

(d) Maintenance. (e) Queueing. (f) RiverSwim.

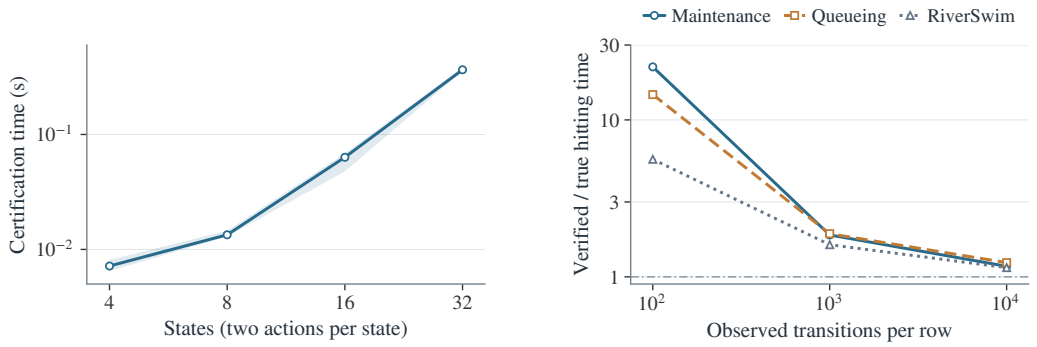
Figure 6: Continuous trajectories: retained certificates (top) and policy loss (bottom). All interaction steps are counted, including travel and repeated visits. Each family contains two tasks with eight seeds each; shading is the interquartile range of the two task means. Policy-loss panels use different vertical scales. The certificate reference is $\varepsilon = 0.1$. No reset is supplied.

We regenerated all eight seeds, six tasks, and both collectors for the two-proposal trajectory study and its witness-only ablation. The current tables and curves use these new traces. A repeated seed-zero run reproduced all 780 rows byte for byte in the pinned environment. With an older numerical stack, seed-zero returned policies agreed at every checkpoint, but counts and certificate values still differed (maximum certificate difference 0.0383). The rule therefore improves reproducibility without establishing platform-independent trajectories. Metadata records Python, NumPy, SciPy, platform, and source hashes; the original supplied traces are retained separately for provenance.

The performance change is material. The original raw-argmin traces had mean terminal RiverSwim loss 0.0480 with learned routing, versus 0.1161 with random actions. With the explicit tie convention these values are 0.2278 and 0.1161, while mean certificates are 0.4645 and 0.8615. Thus the earlier apparent policy-loss advantage does not survive this implementation perturbation. Across six tasks, the random-minus-learned terminal loss difference is -0.0372, with a paired task-bootstrap 95% interval $[-0.1441, 0.0325]$. These observations support neither a reliable policy-loss advantage nor general dominance of the empirical router. The count-level experiments use direct row requests and are unaffected by this routing change.

Table 13: CPU time for one complete proposal-and-certification step, using five seeds at each size. Timing includes all proposal and LP evaluations in that step.

States	Median seconds	Range	Median VI iterations
4	0.0072	[0.0065, 0.0081]	82
8	0.0134	[0.0129, 0.0145]	122
16	0.0633	[0.0476, 0.0685]	213
32	0.3655	[0.3563, 0.3823]	553



(a) Median certification time; seed range.

(b) Median verified/true hitting-time ratio.

Figure 7: Finite computation and a data-dependent navigation witness. The timing panel does not enumerate policies. The hitting-time panel evaluates a fixed learned router using the hitting-time inequality (67), implemented through the support-function dual (74); the displayed bound, unlike a proposed route, is checked against all currently plausible rows.

R.6 Finite computation and routing witnesses

Figure 7 and Table 13 show the measured cost of one proposal-and-certification step for queueing systems with 4–32 states and two actions. Each row has 4,000 observations; five seeds are used at each size. The operation includes the optimistic witness, empirical proposal, upper program, and lower evaluations. These are sequential CPU measurements of the implemented solver, not theoretical polynomial exponents or comparisons against a specialized competing implementation.

All 540 routing-witness queries were feasible on the stated uniform-per-row datasets. Each query fixes the empirical router and computes the hitting LP for one of five target states. The inputs cover 12 tasks, three seeds, and per-row counts 100, 1,000, and 10,000. True hitting times are evaluated separately. Figure 7 shows the ratio of the verified maximum bound to the true maximum hitting time. Feasibility on this grid does not assert feasibility for arbitrary sparse historical logs; the confidence-set ambiguity may prevent a finite witness in other cases.

R.7 Reading the empirical evidence

Three conclusions follow at different levels. The primary count experiment demonstrates fewer observations to the same valid stopping decision with all inference components held fixed. The fixed-budget experiment does not establish an aggregate policy-loss gain over residual Uniform. The continuous-trajectory experiment separates tighter certificates from policy quality: learned routing lowers the RiverSwim gap but has worse terminal loss, with unresolved stops and sensitivity to routing ties.

The study does not compare a complete DCR implementation against complete implementations of DFE, SAVIA+, or BESPOKE. Uniform and Static are controlled allocation ablations under a common evaluator. Likewise, the empirical router is not a GOSPRL implementation. The theoretical relationships to those algorithms are given in the main text and Appendix J; empirical dominance over them is not inferred from these ablations.

S Additional verification and allocation studies

We examine checking frequency, optimized certificate comparisons, operational stopping, and fixed-incumbent verification. Each study states its policy-selection, confidence, and retention conventions; additional seeds are disjoint from those in the principal comparison.

S.1 A finer batching and checking schedule

We use the same twelve tasks, historical count profiles, and 4,000 historical observations as in Appendix Q. For each task, we generate eight additional row streams using seeds $24, \dots, 31$. Every method receives the same stream for a given task and seed. The first positive budget is 512. Subsequent budgets multiply the preceding budget by either two or 1.25, round up to the next integer, and terminate at a common cap of 2,097,152. The endpoint is included even when it is not a power of the multiplier. The initial-data-only check is also included.

DCR and Static continue to reserve half of every batch for least-filled sampling; Residual Uniform spends the full batch on least-filled rows. Static keeps the initial acquisition profile, whereas DCR recomputes it. Every method uses the same policy proposals, row confidence intervals, and retained stopping rule. Changing the grid changes both when a decision is checked and when the allocation is recomputed. The comparison therefore tests the complete batching schedule, not just more frequent inspection of an unchanged sampling path.

At tolerance $\varepsilon = 0.1$, all 96 task–seed runs of every method stop before the cap for both schedules. Table 14 reports the mean additional count. For each bootstrap replicate, we sample twelve task identities with replacement and retain all eight paired seeds and methods within each sampled task. We recompute one minus the ratio of the two overall mean stopping costs. The reported 95% percentile interval uses 20,000 such replicates. Thus the twelve MDP tasks, rather than checkpoints or repeated seeds, are the units of resampling. The intervals describe this task grid; they do not quantify generalization to an unspecified population of control systems.

Table 14: Additional observations until the same 0.1 decision certificate passes, using the eight previously unused seeds per task. Each entry is a mean over 96 task–seed runs. The final columns compare DCR with Residual Uniform and Static, respectively; brackets give paired task-bootstrap 95% intervals.

Multiplier	DCR	Uniform	Static	vs. Uniform	vs. Static
2	148,907	268,373	202,645	44.5% $[35.2, 49.2]$	26.5% $[2.6, 42.1]$
1.25	113,777	207,241	155,107	45.1% $[41.1, 47.5]$	26.6% $[23.4, 33.6]$

The finer schedule lowers the observed stopping count of all three methods, while DCR's relative improvement persists. Fixed-budget loss answers a different question. Averaging each task's loss over its positive budget checkpoints and paired seeds, the Uniform-minus-DCR difference for the 1.25 schedule is 0.00067, with interval $[-0.00383, 0.00585]$. Hence this test strengthens the evidence for certified stopping, but still does not resolve aggregate fixed-budget policy superiority.

S.2 Optimizing the common-potential baseline

For the current candidate π and current confidence set, define

$$C_{\mathrm{com}}(\pi, \mathcal{C}) = \inf_h \{U_h - L_{\pi,h}\}.$$

This is not the value of a chosen optimistic potential. It is the global optimum over *all* common potentials. It is computed for the same instantaneous candidate and confidence set as the two flow programs, without mixing retained bounds from different times. In the finer-grid DCR runs, the mean instantaneous values over positive checkpoints are 0.189 for the separate certificate, 0.277 for the optimized common certificate, and 0.313 for the particular common witness returned by the proposal routine. The separate value is strictly smaller at every such checkpoint. Figure 8 presents the separate/optimized-common comparison averaged across tasks.

Table 15: Full-support incumbent verification at 65,536 observations per row. The last two columns are means over sixteen seeds, not policy-loss improvements. The exact optimized common-potential value is $1/2$ in every row.

p	Exact incumbent loss	Separate certificate	Optimized common
0.005	0.00495	0.01168	0.52670
0.020	0.01923	0.03208	0.52605
0.050	0.04545	0.06446	0.52476
0.100	0.08333	0.10666	0.52282
0.200	0.14286	0.16831	0.51968

For completeness, the optimized baseline is itself a finite LP. For a box $\{p \in \Delta_S : \ell \le p \le q\}$ and a vector a, its support function has the dual representation

$$\max_{p\in\Delta_S,\, \ell\le p\le q} p^\top a = \min_{\theta\in\mathbb{R},\, b,c\ge 0} \{\theta + q^\top b - \ell^\top c : \theta\mathbf{1} + b - c \ge a\}. \tag{75}$$

The equality follows from linear-program duality applied to the simplex equality and the two sets of box inequalities. Introduce upper and lower gain variables z, w and anchor $h(s_0) = 0$. For each upper row, use (75) with $a = h$ and require $r^+_{sa} + \theta + q^\top b - \ell^\top c - h_s \le z$. For each row selected by the deterministic candidate, use an independent support dual with $a = -h$ and require $r^-_{sa} - \theta - q^\top b + \ell^\top c - h_s \ge w$. Minimizing $z - w$ gives exactly C_{com}. All these constraints are linear. Re-evaluating the returned h with direct row support calculations checks the objective independently of the LP's internal multipliers.

This is a certificate ablation with a fixed candidate. It is not an implementation or performance comparison of complete algorithms such as DFE or SAVIA+, which can select different policies and use different exploration rules.

S.3 Finite-data verification in the full-support example

We also sample the two-state family from Proposition 3.2. Fix $q = 1/2$ and use $p \in \{0.005, 0.02, 0.05, 0.1, 0.2\}$. The incumbent always takes the zero-reward action at state one. Rewards are known. The learner estimates each transition row separately with the full probability boxes of (19); it is not given the equality of the two actions' transition rows. For each of sixteen seeds, sample $n \in \{64, 256, 1024, 4096, 16384, 65536\}$ observations per row. Compute both the separate and optimized common certificates from the same prefixes.

Table 15 gives the final-prefix mean over the sixteen seeds. The exact loss and optimized common gap are known analytically, so this experiment also checks the direction and scale of the finite-data bounds. The separate certificate approaches the actual incumbent loss, whereas the common certificate remains above $1/2$. The $p = 0.1$ case has a small margin below the target 0.1 and still has a mean certificate above that target at the final prefix. For $p = 0.2$ the incumbent is genuinely worse than 0.1, so a valid certificate should not accept it. Both regimes are included rather than reporting only cases that certify quickly.

S.4 Operational stopping and an isolated price-refresh ablation

We add a controlled operational comparison on the same twelve five-state tasks with 4,000 historical observations, full successor uncertainty, and known rewards. We use eight previously unused seeds $32, \dots, 39$, a budget multiplier of 1.25, and cap 2^{21}. The primary threshold is 0.1; 0.05 and 0.2 are secondary. The protocol, all methods, and seeds were recorded before the smoke test and full execution. All 384 method runs continue to the cap, producing 15,360 checkpoints; stopping costs use the first checked gap plus 10^{-8} that is at most the threshold.

Four specified procedures. DCR and Uniform flow use the same two-proposal portfolio and retained separate bounds as the principal study. Initial price uses DCR's residual water-filling rule and the same half-budget least-filled safeguard, but freezes $\alpha_i = d_i\sqrt{N_i}$ at its initial value. The current counts still enter the residual allocation. It therefore isolates price refresh from the change in allocation formula present in the original Static comparison.

Table 16: Accuracy sensitivity for the operational comparison. Unstopped runs remain capped at 2^{21} observations. Capped means at $\varepsilon = 0.05$ are not unrestricted expected stopping times.

ε	Method	Mean capped observations	Stopped
0.2	DCR	29,725	96/96
0.2	Uniform flow	54,077	96/96
0.2	Initial price	38,124	96/96
0.2	Uniform + Bellman	145,392	96/96
0.1	DCR	113,634	96/96
0.1	Uniform flow	207,100	96/96
0.1	Initial price	147,116	96/96
0.1	Uniform + Bellman	597,365	96/96
0.05	DCR	458,306	96/96
0.05	Uniform flow	716,859	81/96
0.05	Initial price	546,555	92/96
0.05	Uniform + Bellman	1,104,443	64/96

Uniform + Bellman samples exactly the same rows as Uniform flow but uses its own lower-greedy recommendation π_h and the checked common-potential gap $U_h - L_{\pi_h,h}$ from the optimistic witness. It retains the smallest *complete* gap and its attached policy. It does not independently retain upper and lower endpoints, which would relax the shared-witness restriction. This is a concrete instantiation of the common-potential stopping/recommendation principle in Tuynman et al. (2024, Theorems 5–6). We use our coordinatewise confidence boxes, reward estimates, and optimistic-witness solver in place of their known-reward construction with KL confidence sets. This is not a reproduction of its DFE algorithm, its full implementation, or SAVIA+. Uniform flow versus Uniform + Bellman compares complete evaluators and recommendations on shared data; it is not a fixed-candidate certificate ablation.

At $\varepsilon = 0.1$, all 96 runs of every method stop. DCR uses 22.8% fewer observations than Initial price, with paired task-bootstrap 95% interval $[16.4\%, 30.2\%]$, and 81.0% fewer than this Uniform + Bellman instantiation ($[79.2\%, 85.2\%]$). Relative to Uniform flow the reduction is 45.1% ($[40.6\%, 47.6\%]$). Each of 20,000 bootstrap replicates resamples twelve task identities and retains all paired seeds and methods within each chosen task. These intervals describe this synthetic grid; they do not establish dominance over every possible Bellman bias, sampling rule, or identification algorithm. At $\varepsilon = 0.05$, the capped-cost ratios in the supplementary records are restricted-horizon summaries, not unrestricted mean stopping-time ratios.

Computation and policy quality are separate costs. Mean measured algorithm CPU time to the primary stop is 0.1346 seconds for DCR and 0.0863 seconds for Uniform + Bellman. The latter also has zero mean policy loss at its later stop, versus 0.000207 for DCR. The sample advantage therefore does not imply a computational or policy-loss advantage. New timing measurements use one Apple M2 process with BLAS and OMP threads set to one. They exclude true-model evaluation and sample/count reconstruction; shared work is charged to both uniform procedures. They are implementation measurements on this host and are not compared numerically with the archived Linux scaling benchmark.

A diagnostic on the actual confidence boxes. For each task at seed 32, at 0 and 32,768 added uniform observations, we freeze the current candidate and empirical centers and contract one row at a time. We compare d_i with $[C(\mathbf{1}) - C(\mathbf{1} - \eta e_i)]/\eta$ for $\eta \in \{10^{-3}, 10^{-4}\}$, reoptimizing both LPs after each contraction. There are 240 row comparisons and 24 top-row comparisons per η. The top row agrees in all 24 cases at both contraction sizes. Median absolute differences are 2.1×10^{-9} and 2.1×10^{-10}, respectively, but the maximum remains about 0.00264 at an initial RiverSwim row. Thus the diagnostic supports the tested leading priorities while also exhibiting a persistent discrepancy; it does not establish differentiability or convert arbitrary degenerate multipliers into exact one-sided derivatives. All queried rows, rather than only matching cases, are retained in the supplement.

Table 17: Fixed-incumbent verification at $\varepsilon = 0.1$. Each historical size has 384 adequate cases (12 policies, four coverage profiles, eight seeds). Final passes are 384/384 for separate potentials and 128/384 for common potentials at every size. All inadequate cases remain uncertified. Capped means retain common-verifier failures at 2^{21}.

Historical	Initial passes / 384		Mean capped additions	
observations	Separate	Common	Separate	Common
4,000	0	0	637,952	2,009,771
65,536	91	0	555,728	2,009,771
1,048,576	241	37	86,251	1,507,541

Every recorded confidence box covers the true transition kernel on this grid, and independently evaluated policy losses do not exceed the retained checked bounds. The maximum LP objective/support discrepancy is 2.01×10^{-14}. The complete trace, first-passage records, CPU measurements, failures file, environment, and source hashes are supplied.

S.5 Fixed incumbents and historical coverage in the control tasks

We isolate fixed-policy verification from policy discovery on all twelve control tasks. For each task, enumerate its 32 deterministic policies *only to construct evaluation fixtures*. Choose an adequate incumbent with the largest true loss at most 0.075, and an inadequate incumbent with the smallest true loss at least 0.15; losses within 10^{-10} are tied lexicographically. Every task has an eligible policy of both types. Neither the true transition matrix nor these losses is supplied to either verifier. This is an oracle-defined diagnostic of verification, not evidence that a learner can discover these incumbents from the historical data.

Use eight new seeds $40, \ldots, 47$ and four equal-total historical profiles: balanced; incumbent-aligned with action weights $8{:}1$; incumbent-opposed with weights $1{:}8$; and sparse, with observations only of the incumbent's selected action. The initial study has 4,000 observations, 800 per state. The two actions receive $(400, 400)$, $(711, 89)$, $(89, 711)$, or $(800, 0)$ observations, ordered as incumbent/alternative. Profiles and methods share row streams for the same task and seed. Subsequent sampling always fills the least-observed rows. Checks occur at zero and $2^9, \ldots, 2^{21}$ additional observations.

At every check, the same fixed incumbent and dataset are evaluated by the separate LPs and the *globally optimized* common-potential LP. For the primary comparison, both methods retain the minimum *complete instantaneous gap* seen so far. Thus the separate method receives no extra advantage from combining different-time endpoints. We also retain those endpoints as a separately labeled secondary diagnostic. A stop requires the retained gap plus 10^{-8} to be at most 0.1. The initial execution used a 10^{-10} permissive comparison; an independent complete-trace audit applied the paper's stricter numerical margin, changed zero decisions, and recorded this amendment without replacing the original protocol or executed source. The closest retained gap to the threshold was 3.23×10^{-6}.

The original protocol was recorded before execution. After observing that 4,000-sample coverage differences had vanished before any stop, we declared an *exploratory* full-factorial extension with historical totals 65,536 and 1,048,576. It retains all incumbents, profiles, seeds, sampling rules, and thresholds. Historical observations are allocated equally across states, with integer remainders assigned by state index; action weights use largest-remainder rounding. Each extension has its own pre-execution protocol and source snapshot, but it was not predeclared before seeing the initial null result.

Verification can change the decision. At each historical size, the separate verifier certifies all 12 adequate incumbents across their 384 task–seed–profile cases. The common verifier certifies 128 cases, all belonging to RiverSwim. Its 256 unresolved maintenance/queueing cases have an exact-model common-potential gap above 0.1, so additional data cannot remove their obstruction. Neither verifier certifies any of the 384 inadequate cases at any historical size. These are repeated data

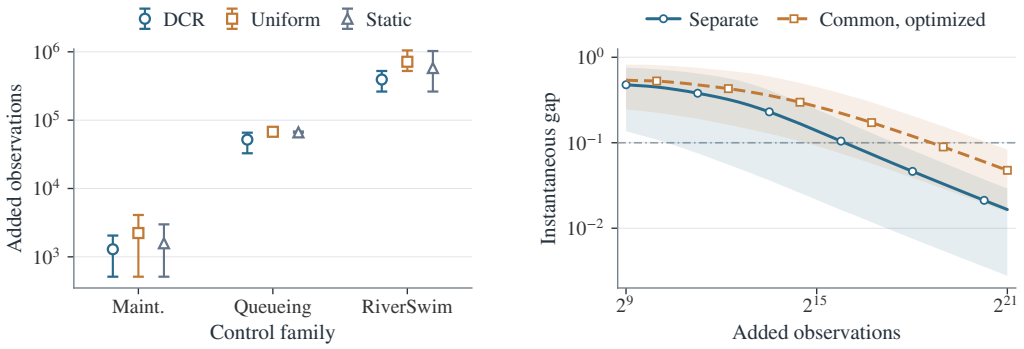
Table 18: Exploratory coverage comparison with 1,048,576 historical observations. Every row contains four task fixtures and eight seeds. Initial counts indicate immediate acceptance; costs include all 32 cases and cap unresolved common-verifier cases.

Family	Coverage	Initial passes / 32		Mean capped additions	
		Separate	Common	Separate	Common
Maintenance	Balanced	32	0	0	2,097,152
Maintenance	Aligned	32	0	0	2,097,152
Maintenance	Opposed	32	0	0	2,097,152
Maintenance	Sparse	32	0	0	2,097,152
Queueing	Balanced	24	0	262,144	2,097,152
Queueing	Aligned	25	0	172,032	2,097,152
Queueing	Opposed	0	0	380,928	2,097,152
Queueing	Sparse	0	0	39,040	2,097,152
RiverSwim	Balanced	32	7	0	458,880
RiverSwim	Aligned	32	30	0	65,536
RiverSwim	Opposed	0	0	180,224	786,432
RiverSwim	Sparse	0	0	640	2,432

instances for 12 policies of each type, not 384 independent control tasks. All recorded confidence sets cover the true rows; all checked policy bounds remain safe.

Coverage matters only when it survives subsequent collection. At 4,000 historical observations, the four profiles produce identical first-passage costs for each task and seed: least-filled sampling equalizes the counts by 4,096 additional observations, before either verifier accepts any incumbent. This null result limits claims about the benefit of historical allocation in that regime. At 65,536 historical observations, the adequate maintenance incumbents certify immediately in 32/32 aligned cases and 32/32 sparse cases, 27/32 balanced cases, and 0/32 opposed cases. The latter require 15,296 additional observations on average. At 1,048,576, coverage also changes initial acceptance in queueing and RiverSwim (Table 18). No one profile uniformly minimizes residual cost: concentrating samples on the incumbent helps its lower bound, while unknown alternative rows can still obstruct an upper explanation.

All cases continue until both primary verifiers pass or the common cap is reached. Unpassed first-passage times are right-censored and included at the cap in the displayed capped means; these means are not unrestricted expected stopping times. Complete per-profile traces, true-loss fixtures, exact-model gaps, numerical-margin checks, independent stream replays, and the separate exploratory protocols are provided in the supplement.



(a) Observations to certify $\varepsilon = 0.1$.

(b) The same current candidate and dataset.

Figure 8: Stopping cost and certificate choice. Left: original 24-seed study, with means over four tasks per family and whiskers spanning their seed-averaged means. Right: eight additional seeds on the finer grid; separate and optimized common-potential gaps are evaluated for the same DCR candidate at each check. Shading is the interquartile range of the 12 task means, not a confidence interval.

T LIMITATIONS AND RESEARCH DIRECTIONS

The decision bound is sufficient, not exact robust regret. The two flow programs optimize the competitor and incumbent in potentially different plausible models. This permits compact linear programs and can tighten common-potential certificates, but does not retain all cancellations in a shared-model policy comparison. Appendix I proves an unbounded gap on recurrent, multiple-action problems and gives an exact correction for known action-independent dynamics. Unknown action-dependent transitions still require a more general coupling argument. Consequently, optimal allocation for this certificate is not information-theoretically optimal decision making.

The allocation theorem has specific conditions. Theorem 4.2 treats full-support transitions, a unique optimal deterministic policy, a fixed incumbent, and specified polynomial-union confidence radii. It is a fixed-confidence, shrinking-verification-margin limit for DCR-TRACK; it neither bounds expected stopping time nor guarantees finite-budget dominance. Corollary G.5 permits unknown rewards only with interior means and joint reward/transition prices. Structural zeros, tied optimal policies, moving candidates, sharper confidence sequences, and general shared-model regret need different analyses. The historical extension uses a separate pilot-and-freeze rule and a growing-history regime. Its diagnostic includes long delays from inaccurate frozen prices and does not reach the predicted residual constants at the tested margins. These results do not apply verbatim to the original half-least-filled implementation, whose general completion guarantee remains independent of price quality. The latter discovery bound is more conservative than the best known prior-free worst-case rates.

General control still has a reachability cost. Certification is valid for arbitrary finite MDPs, whereas our generic termination argument assumes communication. The learner does not know the diameter. Generic discovery and trajectory costs depend on it, even though a particular incumbent can have short verification witnesses. The polynomial trajectory realization invokes an existing learned collector; the lighter empirical router tested here is distinct and does not inherit that rate. Extending the most efficient certificate-guided collection to weakly communicating dynamics with inaccessible transient rows requires a more localized termination argument.

Statistical and empirical scope. The model assumes a fixed finite state–action representation and unchanged row laws between history and supplementation. Drift, hidden-state confounding, and outcome-filtered logs require different confidence models. The behavior-policy log study includes 16 synthetic tasks with five or twelve states and known rewards. It retains all learned controllers, but does not establish application-domain generality; its supplemental queries are generative, not trajectory travel. The three-state known/unknown-reward studies test an asymptotic mechanism and do not substitute for large-scale control experiments. Flow-directed allocation performs worse when

assessed through a common-potential checker on the same data. Empirical routing also yields worse RiverSwim policy loss than random actions despite a tighter certificate, and is sensitive to numerical tie handling. Larger systems, function approximation, and complete prior-free identification baselines remain useful evaluations. All numerical stopping checks use double precision with an explicit reserve, rather than outward-rounded arithmetic.