SEARCH OR REFRESH? CANDIDATE-EXPOSURE CERTIFICATES FOR OFFLINE POLICY EXTRACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Increasing action-search computation changes the continuation policy that an offline critic must evaluate. We study which unresolved rankings matter for this change through candidate exposure: an error affects selection only when competing actions appear together without being screened by a higher-ranked candidate. We derive an interval envelope, relate it to fixed-vector robust regret within a factor of four, and obtain equality for two supported actions. We provide a certified threshold approximation for larger candidate budgets, while using an exact near-linear algorithm for two draws. An exact moment-tree alternative is near-linear in action count at fixed budget. Model confidence sets built from trajectories support adaptive stopping and policy replacement. Our replacement test lower-bounds shared-model baseline regret and is exact for a common uncertain prefix followed by a known suffix. A stepwise analysis separates marginal-value uncertainty, residual inflation, visitation bounds, and numerical approximation. A certificate-matched study charges every schedule for final verification and uses identical sequential replacement gates. Across 192 trajectory datasets, exposure and robust root-value checks have complementary yields; fixed-episode confidence improves graph certification but does not establish broad computational superiority. The results identify what decision-sensitive certification adds, what it costs, and where its uncertainty relaxations lose useful information.

1 INTRODUCTION

An offline reinforcement learning (RL) agent must turn a value function learned from a fixed dataset into actions. This policy-extraction step can limit performance even when the value function contains useful information (Park et al., 2024). A simple approach samples K actions from a proposal distribution and executes the one with the highest critic score. The budget K offers a simple way to control search: more candidates increase the maximum predicted score without collecting new transitions. The important question is whether this additional computation improves the policy's actual return.

The answer depends on what the critic evaluates. A policy-evaluation critic predicts the return after an action followed by a particular downstream policy. Increasing K changes that downstream behavior as well as the current decision. Consider a safe route with a modest reward and a risky route that pays well only after a good later decision. A critic for weak downstream behavior can prefer the safe route. More search improves the later decision, but also makes the agent more likely to follow the critic's stale preference at the root. Return can then decrease even though the critic remains *exact for its reference policy*. This failure is different from exploiting ordinary estimation error.

Refreshing the critic can repair this mismatch, but requiring accurate values everywhere can waste computation. An uncertain ranking has no effect when the competing actions are never drawn together. Even when they co-occur, a higher-ranked candidate can prevent their disagreement from determining the action. Conversely, an apparently small error can matter when search repeatedly exposes the affected alternatives. We call these selection events *candidate exposure*. They connect the proposal distribution, search budget, and critic uncertainty for the decision actually made.

Expected-max Q-learning (EMaQ) already provides the Bellman operator matched to K (Ghasemipour et al., 2021). We ask a different question: *which unresolved rankings must a partly refreshed critic settle before a specified search budget is reliable?* A useful answer must separate

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.

three tasks: diagnosing relevant ranking uncertainty, deciding when further updates are unnecessary, and establishing whether the resulting policy can safely replace the retained policy. The last two comparisons have different reference policies and cannot be interchanged.

We derive an interval exposure envelope, compare it with regret under one fixed uncertain value vector, and propagate it to a return bound. We give exact two-draw and fixed-budget algorithms, plus a lower-memory approximation with explicit error. We then connect trajectory confidence and policy replacement to shared-model baseline regret, including an exact shared-prefix case. Finally, we measure the losses from successive uncertainty relaxations and test the complete workflow with identical replacement gates and strong exact baselines. Matching statistical rates appear in the appendix. A faster local diagnostic need not produce a faster or better retained policy.

The theory concerns finite, time-layered Markov decision processes (MDPs) and a known fixed proposal. It does not supply confidence intervals for neural critics or an optimal joint allocator over search, refresh, and new data. Its purpose is narrower: to make clear what additional computation can establish, and when the available offline information cannot certify an improvement.

2 Related work

Policy evaluation, improvement, and action gaps. Batch methods repeatedly evaluate and improve policies using fixed data (Lagoudakis & Parr, 2003; Ernst et al., 2005). Fitted-iteration and error-propagation analyses describe how value errors affect control (Munos & Szepesvári, 2008; Farahmand et al., 2010), while approximate modified policy iteration studies the evaluation–improvement tradeoff (Scherrer et al., 2015). Action-gap results explain why accurate decisions may require less than uniformly accurate values (Farahmand, 2011; Bellemare et al., 2016). We retain these principles but weight uncertain rankings by the candidate sets that expose them, including screening by higher-ranked actions.

Candidate-based policy extraction. EMaQ defines the expected-max operator and its budget-dependent solutions (Ghasemipour et al., 2021); neither is new here. Empirical work identifies extraction as a separate bottleneck (Park et al., 2024), and decoupled extraction uses a behavior proposal with critic-based reranking (Lin et al., 2026). Best-of-N theory studies coverage and imperfect scoring, including degradation under excessive search (Huang et al., 2025). Our exact-reference-critic example isolates a sequential continuation mismatch rather than score-estimation error. The main technical object is the interval exposure envelope and its certified computation, not the observation that a search curve can decline.

Offline uncertainty and safe improvement. Support restrictions and conservative value learning address distribution shift in offline RL (Fujimoto et al., 2019; Kumar et al., 2019; 2020). Pessimistic and Bellman-consistent analyses give statistical guarantees under coverage assumptions (Jin et al., 2021; Xie et al., 2021). High-confidence improvement and baseline bootstrapping control the risk of replacing a policy (Thomas et al., 2015; Laroche et al., 2019). We use an explicit model-confidence event to support repeated queries of one dataset. The propagated-advantage gate is a sufficient test within this framework, not a new general formulation of safe policy improvement.

Shared-model baseline regret. Petrik et al. (2016) directly optimize improvement relative to a baseline under the *same* uncertain model and show why separate pessimistic/optimistic evaluations can be too conservative. They also establish hardness of the general objective. Our propagated gate is a tractable lower bound on that existing objective, not its replacement or exact general solution. We characterize a shared-uncertain-prefix case where it is exact, give a strict uncertain-model separation from both simpler tests, and compare with exact shared-model regret in restricted solvable families.

Robust regret and reusable confidence. Rectangular robust dynamic programming (DP) supplies the transition minimizations used by our gates (Iyengar, 2005; Nilim & El Ghaoui, 2005). Work on robust MDPs and randomized minimax regret also distinguish decisions from adverse uncertainty (Ahmed et al., 2013; Mastin et al., 2014). Here the distinction is local: one adverse value vector must serve all sampled sets, whereas an easier envelope allows a different vector for each set. Our factor-four comparison is for these two objectives, not for an entire uncertain MDP. Confidence that holds at every sample count builds on established sequential inference (Howard et al., 2021); our

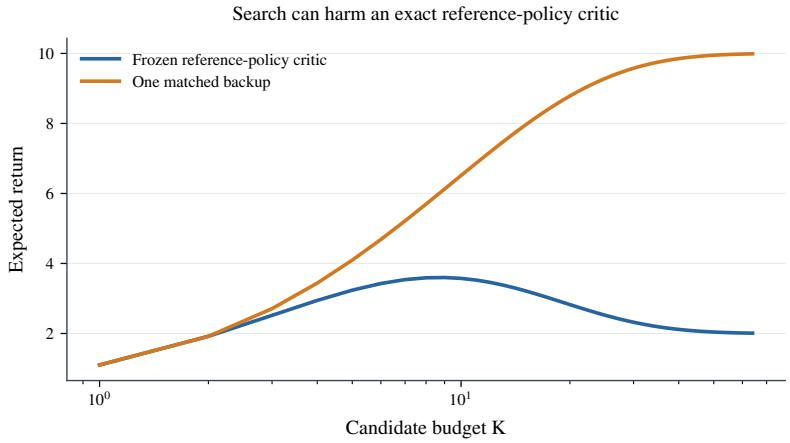


Figure 1: More search can hurt an exact reference-policy critic. A matched backup repairs the two-step continuation mismatch. This mechanism illustration does not charge the refresh cost; Table 1 pays for the complete finite-data workflow.

elementary construction makes adaptive reuse explicit. Appendix E gives further comparisons and the roles of the remaining statistical and computational tools.

3 Setup: What changes when the search budget changes?

An MDP has layers $\mathcal{S}_0, \dots, \mathcal{S}_{H-1}$, initial distribution μ, rewards in $[0, 1]$, and $\gamma \in (0, 1]$. A transition advances one layer or terminates. The proposal $\beta(\cdot \mid s)$ generates K independent candidates $C = (A_1, \dots, A_K)$. A fixed action-index priority resolves ties. Let $I_q(C)$ be the candidate selected by scores q and $\pi_{q,K}$ its marginal action distribution. Define

$$\mathcal{F}_K q(s) = \mathbb{E}_C \max_{a \in C} q(s, a), \qquad \mathcal{T}_K q(s, a) = r(s, a) + \gamma P_{s,a} \mathcal{F}_K q. \tag{1}$$

This is EMaQ's expected-max operator. Backward induction gives its solution Q_K and value $V_K = \mathcal{F}_K Q_K$. The comparator chooses optimally *from each sampled set*; it is not unrestricted Q^*. Write $J(\pi) = \mu V^\pi$ and $J_K = \mu V_K$.

An exact mismatch. Set $\gamma = 1$ for this illustration. At a root, Safe pays c and terminates; Continue leads to Good with reward B or Bad with reward zero. The proposal selects Safe with probability σ and Good with probability p. If $Bp < c$, the exact Q^β ranks Safe first, giving

$$J_{\text{fr}}(K) = c\{1 - (1 - \sigma)^K\} + (1 - \sigma)^K B\{1 - (1 - p)^K\}. \tag{2}$$

With $(c, B, p, \sigma) = (2, 10, .1, .1)$, return falls from 3.59 at $K = 8$ to 2.26 at $K = 32$. One matched backup at $K = 32$ yields 9.66. These displayed rewards and returns are in illustrative units; dividing all of them by ten gives the assumed reward range $[0, 1]$. Every frozen-policy return is at least the reference value 1.10 in the displayed units, so there is no contradiction with improvement over β.

Proposition 1 (Exact propagation cost)**.** *Insert forced-action states so that the final choice is d transitions from the root. Let $W_K = \gamma^d B[1 - (1 - p)^K]$ and assume $\gamma^d Bp < c < W_K$. Starting from $q_0 = Q^\beta$, synchronous updates $q_{m+1} = \mathcal{T}_K q_m$ retain the old root ranking for $m < d$ and correct it at $m = d$. Before the correction,*

$$J_K - J(\pi_{q_m,K}) = \{1 - \sigma^K - (1 - \sigma)^K\}(W_K - c). \tag{3}$$

Afterwards the gap is zero.

This is an exact statement about synchronous local updates, *not* a lower bound against backward planning. A known layered model can be solved in one ordered backward pass.

4 WHICH UNCERTAIN RANKINGS CAN SEARCH EXPOSE?

At one state, remove unsupported actions and sort the remaining A actions in increasing q order, including tie priority. Let $F_i = \sum_{j \le i} \beta_j$, $F_0 = 0$. The probability of selecting action i is

$$w_i(K) = F_i^K - F_{i-1}^K. \tag{4}$$

For a possible true value vector v, the local loss is

$$\ell_K(q, v) = \mathbb{E}_C[\max_{a \in C} v_a - v_{I_q(C)}] = \mathcal{F}_K v - \sum_i w_i(K) v_i. \tag{5}$$

Suppose only $l_i \le v_i \le u_i$ is known. We distinguish

$$R_K = \sup_{v \in [l,u]} \ell_K(q, v), \qquad \Gamma_K = \mathbb{E}_C \sup_{v \in [l,u]} [\max_{a \in C} v_a - v_{I_q(C)}]. \tag{6}$$

R_K uses one value vector for all candidate sets; Γ_K lets the adverse vector depend on the set. Moving the supremum inside the expectation makes the second quantity easier to compute but potentially larger. A useful certificate must therefore explain both its computational advantage and the price of relaxing a single common value vector.

Theorem 2 (Candidate-exposure envelope)**.** *For $t \ge 0$, let $B_i(t) = \sum_{j<i : u_j - l_i > t} \beta_j$. Define*

$$H_i(t) = F_i^K - F_{i-1}^K - (F_i - B_i(t))^K + (F_{i-1} - B_i(t))^K. \tag{7}$$

Then

$$\Gamma_K = \sum_i \int_0^\infty H_i(t)\, dt, \qquad \Gamma_K/4 \le R_K \le \Gamma_K. \tag{8}$$

Exact integration takes $O(A^2 + A \log A)$ arithmetic/sort work and $O(A)$ scratch space. With two supported actions,

$$R_K = \Gamma_K = (u_1 - l_2)_+ [1 - \beta_1^K - \beta_2^K]. \tag{9}$$

Here $H_i(t)$ is the probability that i wins and a lower-ranked alternative could be better by more than t. The four terms count co-occurrence while excluding higher-ranked candidates. For the factor-four bound, one random global assignment of lower/upper endpoints realizes any adverse pair with probability $1/4$. Full proofs, including the faster exact integration baseline, appear in Appendix A.

A numerical example. Let $q = (.40, .43)$ and let both entries have radius .03. A generic ranking-loss bound is .06. For proposal $(.1, .9)$ and $K = 2$, the exact exposure bound is only $.03(1 - .1^2 - .9^2) = 5.40 \times 10^{-3}$. The uncertain inversion matters only when both actions are drawn. Conversely, a large K can *screen* an inversion: for uniform proposals, $q = (0, 1, 3)$ and $v = (1, 0, 3)$ give

$$\ell_K(q, v) = (2/3)^K - 2(1/3)^K \longrightarrow 0. \tag{10}$$

The factor four is not claimed tight. It relates two local rectangular quantities, not worst-case regret over Bellman-consistent MDPs.

When uncertainty is harmless. If $u_j \le l_i$ for every supported pair $j < i$, no interval permits an inversion and $\Gamma_K = 0$ even when individual values are uncertain. At $K = 1$, it is zero regardless of the intervals: the policy has no candidate-selection choice. The certificate measures uncertainty about selection, not value-estimation accuracy itself.

4.1 FROM ONE DECISION TO RETURN

Proposition 3 (Loss propagation)**.** *For $\pi = \pi_{q,K}$ and $\ell(s) = \ell_K(q_s, Q_K(s, \cdot))$,*

$$V_K - V^\pi = \ell + \gamma P^\pi (V_K - V^\pi). \tag{11}$$

If all boxes contain their respective Q_K values, then

$$0 \le J_K - J(\pi) \le \mathbb{E}_\pi \sum_h \gamma^h \Gamma_K(q_{S_h}; l_{S_h}, u_{S_h}) \le \sum_h \gamma^h \max_{s \in \mathcal{S}_h} \Gamma_K(q_s; l_s, u_s). \tag{12}$$

The final bound avoids estimating deployment visitation. It can be conservative because it combines a rectangular box with a worst-state bound at each layer.

5 FAST CERTIFICATES WITHOUT ENUMERATING ACTION PAIRS

The general exact integral inspects quadratically many potential gaps. That is *not* the right baseline for two candidate draws. Distinguish $K = 2$ (two draws from arbitrarily many actions) from $A = 2$ (two supported actions with arbitrary K).

An exact near-linear special case. For $K = 2$,

$$\Gamma_2 = 2 \sum_i \beta_i \sum_{j<i} \beta_j (u_j - l_i)_+. \tag{13}$$

Each distinct pair occurs in two orders and the critic chooses its higher-ranked member. Maintaining cumulative masses and upper-endpoint-weighted masses in a Fenwick tree computes (13) exactly in $O(A \log A)$ time and $O(A)$ space. Ordered upper endpoints allow batched prefix sums instead. We include both implementations and their proof (Appendix B.1); no approximation is needed. The data structure itself is standard (Fenwick, 1994).

There is also an exact $O(K^2 A \log A)$ moment-tree construction for fixed K (Appendix B.2); approximation is not the only route to subquadratic dependence on A. It uses $O(KA)$ memory and is slower than quadratic integration at our tested $K = 8, 32$ sizes.

For larger K, our lower-memory approximation uses threshold queries: what is the probability that an exposed mistake exceeds size t? Integrating these probabilities recovers the envelope. This avoids a quadratic computation at every state and update, but still requires an empirical total-cost comparison.

Let $T(t) = \sum_i H_i(t)$, $\theta = T(0)$, $M_i = \max_{j<i}(u_j - l_i)_+$ (zero for $i = 1$), and $G = \max_i M_i$. A sweep through critic order with a cumulative-mass tree computes every $B_i(t)$ in $O(A \log A)$ time and $O(A)$ space. If upper endpoints follow critic order, as in symmetric residual intervals, binary search in one sorted array suffices.

Theorem 4 (Certified approximation)**.** *A single threshold query gives the upper bound $U_0 = \sum_i M_i H_i(0) \geq \Gamma_K$, exact when $A \leq 2$. For $\eta \in (0, 1]$ and $\tau > 0$, an upper Riemann sum at $0, \tau, (1+\eta)\tau, \ldots, G$, clipped at G and combined by taking the minimum with U_0, gives U satisfying*

$$\Gamma_K \leq U \leq (1+\eta)\Gamma_K + \tau\theta. \tag{14}$$

Its arithmetic/sort cost is

$$O\left(A \log A \left[1 + \frac{\log_+(G/\tau)}{\log(1+\eta)} \right] \right), \tag{15}$$

with $O(A)$ space. The zero-gap case returns zero directly.

The proof rounds each candidate set's worst permitted loss upward to the next geometric grid point. A positive loss below τ costs at most τ; larger losses grow by at most a factor $1 + \eta$. These are deterministic numerical bounds, not confidence levels. Both parameters may be selected after inspecting the data. The aggregate return certificate inherits the same relative slack and additive slack at most $\tau \sum_h \gamma^h$.

Numerical accuracy versus confidence. The geometric grid controls deterministic numerical slack, not statistical failure probability. Its aggregate additive slack is at most $\tau \sum_h \gamma^h$. We check the generic bound first, then the coarse bound, and refine only when needed; a midpoint tail lower bound can rule out a successful refinement. For $K = 2$, every exposure method uses the exact near-linear kernel. For $A = 2$, every method uses the same exact batched two-action formula. The remaining comparison is therefore not driven by an unnecessarily weak special-case baseline.

6 OFFLINE DATA: STOP REFRESHING OR REPLACE THE POLICY?

These are distinct questions. A stopping certificate compares a policy with its matched-K optimum. A replacement certificate compares it with the currently retained policy. Neither follows merely from an increased critic score.

6.1 One confidence event for a reused trajectory dataset

Assume stationary, fully observed Markov dynamics with fresh conditional outcomes, Bernoulli rewards, and known next-layer support including termination. Actions are chosen before their outcomes. The M state–action rows have random visit counts $N_{s,a}$ in a behavior-trajectory dataset. Let $D_{s,a}$ be the number of allowed successors, and define total variation (TV) as half the L_1 distance. For the nth local visit, spend

$$\alpha_n = \frac{\delta}{2Mn(n+1)}. \tag{16}$$

Use a two-sided Clopper–Pearson reward interval with failure α_n, and a transition total-variation radius

$$t_{s,a,n} = \min\left\{1, \sqrt{\frac{D_{s,a}\log 2 + \log(1/\alpha_n)}{2n}}\right\}. \tag{17}$$

Unsampled rows receive full intervals and simplices; singleton successor supports have zero radius. These constructions use classical binomial and finite-alphabet concentration (Clopper & Pearson, 1934; Hoeffding, 1963; Weissman et al., 2003).

Theorem 5 (Random-count model confidence)**.** *With probability at least $1-\delta$, all reward and transition sets above contain the true model simultaneously at every local sample count. The event therefore holds at random visit counts, including after an optional stop in data collection. On this same event, all subsequent data-dependent critic and policy queries have model containment.*

The proof applies a union bound to each row's entire outcome stream, *not* over a fixed count conditional on having observed that count. It is a simple count-spending construction, not a new general confidence-sequence method (Howard et al., 2021). Hidden state, drift, or filtering records by their outcomes violates the assumptions. Fixed independent episodes with a fixed behavior policy also permit $\delta/(2M)$ after conditioning on each row's episode-visit indicators (Proposition 16). Our sensitivity study uses this tighter set; it does not permit outcome-adaptive collection.

6.2 A residual certificate with an uncertainty floor

Let $\widehat{M}$ be the empirical model and $[r^-, r^+]$ its reward interval. For any q, K, define $v_q = \mathcal{F}_K q$, including terminal value zero, and

$$e_{s,a} = \max\{\widehat{r}_{s,a} - r^-_{s,a}, r^+_{s,a} - \widehat{r}_{s,a}\} + \gamma t_{s,a}\,\mathrm{span}(v_q \mid \text{allowed successors}), \tag{18}$$

$$d_h = \max_{s\in\mathcal{S}_h,\, a\in\mathrm{supp}\,\beta_s} \{|\widehat{\mathcal{T}}_K q(s,a) - q(s,a)| + e_{s,a}\}, \qquad b_H = 0, \quad b_h = d_h + \gamma b_{h+1}. \tag{19}$$

Here span is the maximum minus minimum value over the allowed successors. The radius combines disagreement with the empirical backup and uncertainty in that backup itself.

Proposition 6 (Adaptive stopping)**.** *On the model-containment event, $|Q_K(s,a) - q(s,a)| \le b_h$ on supported actions in layer h, simultaneously for all q, K. Thus* (12) *with intervals $q_s \pm b_h$, or its certified upper approximation, gives a valid stopping rule. It remains valid after adaptive reuse of the dataset. Taking its minimum with $2\sum_h \gamma^h b_h$ never worsens the generic certificate.*

The implementation intersects these boxes with $[0, \sum_{t=0}^{H-h-1}\gamma^t]$ and caps the generic bound by the maximum possible return difference. An m-update returned critic costs $m+1$ backups: its final residual must also be computed. The same backup is reused as the next iterate if stopping fails. Even a zero empirical residual leaves the $e_{s,a}$ terms. A failed certificate does not prove that improvement is impossible.

6.3 Propagating uncertain advantages

Robust backward evaluation gives $Q^{\pi_o} \in [L_o, U_o]$ for the retained policy π_o. For a proposed π_n, set $\Delta\pi = \pi_n - \pi_o$ and

$$G_o(s) = \sum_{\Delta\pi_a \ge 0} \Delta\pi_a L_o(s,a) + \sum_{\Delta\pi_a < 0} \Delta\pi_a U_o(s,a). \tag{20}$$

Requiring $G_o(s) \ge 0$ everywhere is safe but can reject a good initial-state decision because of a small loss at a later state. We instead lower-bound the accumulated advantage under the *proposed* policy.

Theorem 7 (Propagated-advantage gate)**.** *Let $\mathcal{P}_{s,a}$ be the transition confidence set. Starting from terminal $z = 0$, compute backward*

$$z(s) = G_o(s) + \gamma \sum_a \pi_n(a \mid s) \min_{P \in \mathcal{P}_{s,a}} Pz. \tag{21}$$

On the same model event, $\mu z \le J(\pi_n) - J(\pi_o)$ for all adaptively chosen policy pairs. Hence $\mu z \ge 0$ certifies replacement. This test contains the all-state test, can strictly improve on it, and is exact when the model and old-policy values are known exactly.

This is a conservative composition of robust DP and the performance-difference identity, not exact minimization of the return difference over the model set. Taking its union with the all-state test and a separate root-value interval test stays safe without another statistical penalty. Rejection does not show that the proposed policy is harmful.

Comparison with shared-model regret. Define

$$D_{\mathcal{C}}(\pi_n, \pi_o) = \min_{M \in \mathcal{C}} [J_M(\pi_n) - J_M(\pi_o)]. \tag{22}$$

This is the baseline-regret objective of Petrik et al. (2016). Both the separate root test and μz lower-bound it, and neither generally computes it exactly. If the policies agree on an uncertain prefix and can differ only inside a closed, fully known suffix, propagation *is* exact for this specified pair (Appendix B.6). Common prefix costs cancel before minimization.

A simple uncertain-model witness makes the comparison concrete. A shared root transition reaches a changed branch with probability $p \in [.2, .8]$, or an unchanged branch worth one. In the changed branch, the proposed policy gains $.4$ at the first decision and loses $.1$ at the next. The old and new branch values are .2 and .5. The all-state test rejects the later loss; the root test gives $.6 - .84 = -.24$. Propagation instead gives $\min_p .3p = .06$, equal to the exact shared-model improvement. The same strict separation holds for finite best-of-four policies, not just deterministic policies (Appendix B.7).

7 Where does conservatism enter?

The factor four compares two *local rectangular* objectives. It does not account for the statistical width of those intervals, their loss of shared-value dependence, or the later use of layer maxima and worst-state visitation. A fast and locally tight envelope can therefore still be a poor global certificate.

A known ranking with overlapping intervals. Suppose two actions have values $(C, C + g)$, where $g > 0$ is known and $C \in [0, W]$ is an uncertain common continuation value. A critic ranking the second action first makes no selection error in any model. Nevertheless, for proposal $(p, 1 - p)$, its tight marginal box gives

$$\Gamma_K = [1 - p^K - (1 - p)^K](W - g)_+. \tag{23}$$

Thus a two-action envelope can be exact for its box yet positive when the shared-model loss is zero. No multiplicative global guarantee follows. Appendix B.5 gives the Markov realization and proof.

Measuring the separate losses. We examine a nested sequence: actual return gap; tight marginal boxes with one adverse value vector per state; the menu envelope; state–action residual boxes; layerwise residual boxes; robust visitation; worst-state aggregation; and numerical approximation. The intermediate row radii obey a stronger backward recursion than the layer maxima, proved in Appendix B.3. All intervals are intersected with the known reward range. Each step can only enlarge the bound on the model event. The difference between successive steps measures that particular ordered relaxation, while the first difference combines genuine statistical uncertainty and lost model dependence. True visitation is used only for this analysis, never for learning.

This decomposition is more informative than the final width alone. The largest source of looseness depends on the problem: shared offsets can make marginalization arbitrarily loose, an irrelevant row can inflate a layer radius, and a state visited with probability ρ can make worst-state aggregation $1/\rho$ times too large. The observed dominant term depends on the problem and on the order in which relaxations are applied.

Table 1: Certificate-matched comparison on 120 shared datasets. Every method includes its final verifier and the identical sequential replacement gate. Success is out of 720 attempts; time is mean milliseconds for the three-budget sequence. Cost per certificate charges unsuccessful attempts as well.

Method	Certificates	ms/sequence	ms/certificate
Generic + exposure check	144	75.9	126.5
Exact exposure stopping	144	120.8	201.3
Geometric stopping	144	103.0	171.7
One sweep + exposure check	111	66.4	143.5
H sweeps + exposure check	144	67.9	113.1
DP + exposure check	144	67.9	113.2
DP + root-value check	84	73.7	210.5

8 EXPERIMENTS

We distinguish two outputs: a retained policy with high measured return, and a policy with a matched-budget near-optimality certificate. A fast method that produces only the first is not automatically a cheaper way to obtain the second. The primary comparison charges every schedule for final certification. True models are used only for post-selection checks.

Matched-certificate protocol. The main grid has 120 trajectory datasets: a depth-four fork, three- and five-layer eight-action graphs, a three-layer 32-action graph, and an eight-action one-step control. Each uses 200, 2,000, or 20,000 episodes and eight data seeds. Fixed tolerances are $.05, .2$. All methods initialize the same empirical behavior critic, propose budgets $2, 8, 32$ sequentially, and apply the same all-state/root/propagated replacement union. An accepted policy becomes the next baseline. A certificate applies to the candidate; rejection does not transfer it to a distinct retained policy.

The common final check C_E certifies $J(\pi_K^*) - J(\pi_{q,K}) \leq \varepsilon$ using the generic residual bound or exact clipped exposure. Generic, exact, and geometric stopping, one sweep, H sweeps, and backward DP all pay for this check, reusing already paid final diagnostics. A stronger comparator, DP+C_R, instead uses robust upper matched-budget values minus lower candidate-policy values. It certifies the *same assertion*, without relying on layer residual maxima. Neither certificate uniformly dominates the other: separated one-step reward intervals can give $C_E = 0 < C_R$, whereas unreachable uncertainty can give $C_R = 0 < C_E$ (Appendix C.1).

The clock includes confidence construction, initial critic and policy bands, refresh, final verification, replacement gates, and retained-policy updates. It excludes data collection and hidden-model checks. Table 1 reports total time divided by successful certificates, charging failures as well. Zero yield is reported as undefined cost, not zero cost. Per-family returns, certificate yields, clocks, and caching conventions are in Appendix C.

Certificate yield and cost. After the same final exposure check, generic, exact/geometric, H-sweep, and DP schedules each certify 144/720 proposals; one sweep certifies 111. DP plus root intervals certifies 84. On identical DP proposals, exposure certifies 60 cases missed by root intervals, whereas root intervals add 0 cases in the main grid. The yield difference occurs in the fork and one-step control. In the broader graph sensitivity study below, the other certificate succeeds. Planning time alone omits the cost of a final DP certificate. Exposure stopping does not establish broad computational superiority over planning followed by certification. The value of the additional check is visible in its yield, not inferred from a return curve alone.

Graph and confidence sensitivity. Across 72 additional datasets from eight seeds per graph structure, DP–root certificates increase from 42/432 with count-spending confidence to 80/432 with the fixed-episode set. Mean reward-interval and transition-radius ratios are 0.63 and 0.72, respectively. The residual-exposure methods still certify no sensitivity-grid proposal at these tolerances. Thus failure of the residual-exposure certificate does not imply that finite-data certification is impossible: a different valid bound and a tighter valid model set can succeed. These are 72 additional datasets with count-specific streams, not nested trajectories. The bootstrap carries all counts and queries for each graph together. Together with the main grid, 192 datasets and 264 model sets produce 11,088

Table 2: Sensitivity over 24 random graphs and three episode counts. Each cell pools 432 candidate queries. All residual-exposure schedules have zero yield here. Root-value certification succeeds, but not every certified candidate passes the common replacement gate. Time includes failed queries and the full workflow.

Confidence set	C_E cert.	C_R cert.	C_R retained	ms/C_R cert.
Count-spending	0	42	35	392.7
Fixed episodes	0	80	60	205.6

correlated proposal stages, not independent trials. Model containment holds throughout this grid, with no false certificate or harmful accepted update observed. This is an empirical result, not a guarantee of zero risk.

The paired increase in root-value yield is 8.80 percentage points (95% graph-block bootstrap interval $[7.64, 9.95]$). Table 2 also distinguishes a certificate for a candidate from one attached to the deployed retained policy. Since the same confidence event covers both tests, a practitioner may try either or both without an extra statistical penalty, but must pay for every requested computation; an adaptive ordering of these checks is not evaluated here.

Computational and uncertainty checks. The workflow uses the exact near-linear branch for $K = 2$. The moment-tree comparison in Appendix B.2 shows why an asymptotic improvement alone need not reduce measured cost. The slack analysis reuses the 96 main-grid datasets excluding the 32-action graph, at budgets $2, 8, 32$ and initial and empirically converged critics, giving 576 correlated queries. Tight marginal uncertainty contributes 61.1% of ordered slack, row/layer residual inflation 34.1%, and numerical integration 0.02%, using the capped decomposition in Appendix B.4. These percentages are conditional on that grid and construction. The confidence sensitivity changes certificate yield without changing numerical resolution; no universal source of failure is established.

The uncertain-prefix check varies four root-probability intervals, three immediate gains, three later losses, and four policy choices (deterministic or candidate budgets $2, 4, 8$). Propagation matches exact shared-model improvement in all 144 cases and accepts while the all-state/root gate union rejects in 122. A known-transition reward-box comparison uses horizons $2, 4$, width two, four actions, 20 data seeds, three episode counts, and budgets $2, 8, 32$. All methods share the known dynamics and learned reward intervals. The all-state/root union accepts 40 of 360 proposals, adding propagation accepts 60, and exact shared regret accepts 132. These restricted comparisons do not establish general dominance. A separate fixed-count validation had one model-confidence miss among 480 datasets (Appendix C); zero observed failures in the trajectory grid do not imply zero risk.

9 Discussion and practical scope

Candidate exposure identifies which uncertain rankings matter to sampled action selection. The interval theorem and its larger-K approximation answer this local question, but cannot remove model uncertainty or restore discarded correlations. For $K = 2$, use exact near-linear computation. Stopping and replacement remain separate decisions.

Every schedule must pay for its final certificate. Exposure can certify decisions despite uncertain values, but residual and marginal relaxations can hide improvements. Robust root bounds succeed on some graphs where exposure fails. Neither check is uniformly better in time or yield.

The theory assumes finite layered states and actions and a known proposal. It does not establish neural-critic confidence, optimal allocation of search, refresh, and data, or broad superiority on offline-RL benchmarks.

AI USE STATEMENT

In this work, we used generative AI tools to draft sections of the paper and polish writing. Additionally, we used generative AI tools for editing code. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2108.13264.

Asrar Ahmed, Pradeep Varakantham, Yossiri Adulyasak, and Patrick Jaillet. Regret based robust solutions for uncertain Markov decision processes. In *Advances in Neural Information Processing Systems*, volume 26, 2013. URL https://papers.nips.cc/paper_files/paper/2013/hash/e1e32e235eee1f970470a3a6658dfdd5-Abstract.html.

Marc G. Bellemare, Georg Ostrovski, Arthur Guez, Philip S. Thomas, and Rémi Munos. Increasing the action gap: New operators for reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2016. URL https://arxiv.org/abs/1512.04860.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404. URL https://academic.oup.com/biomet/article/26/4/404/291538.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005. URL https://jmlr.org/papers/v6/ernst05a.html.

Amir-massoud Farahmand. Action-gap phenomenon in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 24, pp. 172–180, 2011. URL https://proceedings.neurips.cc/paper/2011/hash/013d407166ec4fa56eb1e1f8cbe183b9-Abstract.html.

Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*, volume 23, 2010. URL https://proceedings.neurips.cc/paper/2010/hash/65cc2c8205a05d7379fa3a6386f710e1-Abstract.html.

Peter M. Fenwick. A new data structure for cumulative frequency tables. *Software: Practice and Experience*, 24(3):327–336, 1994. doi: 10.1002/spe.4380240306. URL https://doi.org/10.1002/spe.4380240306.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2106.06860.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2052–2062, 2019. URL https://proceedings.mlr.press/v97/fujimoto19a.html.

Seyed Kamyar Seyed Ghasemipour, Dale Schuurmans, and Shixiang Shane Gu. EMaQ: Expected-max Q-learning operator for simple yet effective offline and online RL. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3682–3691, 2021. URL https://proceedings.mlr.press/v139/ghasemipour21a.html.

Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Fast Bellman updates for robust MDPs. In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1979–1988, 2018. URL https://proceedings.mlr.press/v80/ho18a.html.

Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Partial policy iteration for L_1-robust Markov decision processes. *Journal of Machine Learning Research*, 22(275):1–46, 2021. URL https://jmlr.org/papers/v22/20-445.html.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://doi.org/10.1080/01621459.1963.10500830.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991. URL https://arxiv.org/abs/1810.08240.

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J. Foster. Is best-of-N the best of them? coverage, scaling, and optimality in inference-time alignment. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25075–25126, 2025. URL https://proceedings.mlr.press/v267/huang25c.html.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005. doi: 10.1287/moor.1040.0129. URL https://doi.org/10.1287/moor.1040.0129.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 652–661, 2016. URL https://proceedings.mlr.press/v48/jiang16.html.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5084–5096, 2021. URL https://proceedings.mlr.press/v139/jin21e.html.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016. URL https://jmlr.org/papers/v17/kaufman16a.html.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2110.06169.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy Q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c2073ffa77b5357a498057413bb09d3a-Abstract.html.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/0d2b2061826a5df3221116a5085a6052-Abstract.html.

Michail G. Lagoudakis and Ronald Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 4:1107–1149, 2003. URL https://jmlr.org/papers/v4/lagoudakis03a.html.

Romain Laroche, Paul Trichelair, and Remi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3652–3661, 2019. URL https://proceedings.mlr.press/v97/laroche19a.html.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv:2005.01643, 2020. URL https://arxiv.org/abs/2005.01643.

Xuyao Lin, Yixiang Shan, Jinru Duan, Tao Yang, Xinyu Zhao, Runyu Lei, Yiming Zhao, Jiaxin Fan, Zongbao Feng, and Peng Jia. Decoupling policy extraction for offline reinforcement learning. arXiv:2608.20909, 2026. URL https://arxiv.org/abs/2608.20909.

Andrew Mastin, Patrick Jaillet, and Sang Chin. Randomized minmax regret for combinatorial optimization under uncertainty. arXiv:1401.7043, 2014. URL https://arxiv.org/abs/1401.7043.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9:815–857, 2008. URL https://jmlr.org/papers/v9/munos08a.html.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005. doi: 10.1287/opre.1050.0216. URL https://doi.org/10.1287/opre.1050.0216.

Seohong Park, Kevin Frans, Sergey Levine, and Aviral Kumar. Is value learning really the main bottleneck in offline RL? In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2406.09329.

Marek Petrik, Mohammad Ghavamzadeh, and Yinlam Chow. Safe policy improvement by minimizing robust baseline regret. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/file/9a3d458322d70046f63dfd8b0153ece4-Paper.pdf.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2103.12021.

Bruno Scherrer, Mohammad Ghavamzadeh, Victor Gabillon, Boris Lesner, and Matthieu Geist. Approximate modified policy iteration and its application to the game of Tetris. *Journal of Machine Learning Research*, 16(49):1629–1676, 2015. URL https://jmlr.org/papers/v16/scherrer15a.html.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2139–2148, 2016. URL https://proceedings.mlr.press/v48/thomasa16.html.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2380–2388, 2015. URL https://proceedings.mlr.press/v37/thomas15.html.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdú, and Marcelo J. Weinberger. Inequalities for the L_1 deviation of the empirical distribution. Technical Report HPL-2003-97 (R.1), Hewlett-Packard Laboratories, 2003. URL https://shiftleft.com/mirrors/www.hpl.hp.com/techreports/2003/HPL-2003-97R1.pdf.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2106.06926.

Andrea Zanette and Martin J. Wainwright. Bellman residual orthogonalization for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/14ecbfb2216bab76195b60bfac7efb1f-Abstract-Conference.html.

A Complete proofs and scope of the claims

A.1 Expected-max preliminaries

For any two score vectors x, y and any menu C, $|\max_{a\in C} x_a - \max_{a\in C} y_a| \le \|x-y\|_\infty$. Averaging proves nonexpansiveness of $\mathcal{F}_K$. It is also monotone and preserves constant shifts. Backward induction therefore defines a unique finite-horizon Q_K. Observing the sampled menu before acting gives the menu maximization in the Bellman optimality recursion; any measurable randomized choice from that menu is no better than its maximizer. This proves the benchmark interpretation used in the main text and is the finite-horizon counterpart of EMaQ's established fixed-point theory.

A.2 Proof of the fork and Proposition 1

The frozen root chooses Continue exactly when every sampled action is Continue, an event of probability $(1-\sigma)^K$. The final policy chooses Good whenever at least one Good sample appears, with probability $u_K = 1-(1-p)^K$. Multiplying these independent decision probabilities gives (2). With depth d the effective downstream payoff is $W_K = \gamma^d B u_K$.

At the final state, both initial critic entries already equal their immediate rewards, so every later sweep leaves them unchanged. Its matched extracted value, however, is Bu_K rather than Bp. One sweep changes the predecessor's continuation score. Inductively, after m synchronous sweeps the correction has reached exactly the states at distance at most m from the final choice; upstream scores have the same uncorrected children as in the previous sweep. Hence for $m < d$ the root Continue score remains $\gamma^d Bp$, and at $m = d$ it becomes W_K. Forced-action states introduce no other rankings.

Before the switch the return is $c(1-(1-\sigma)^K)+(1-\sigma)^K W_K$. Afterwards it is $c\sigma^K+(1-\sigma^K)W_K$, because Safe is selected only when every root candidate is Safe. Subtraction gives (3). The sign assumptions make the two rankings strict. This argument proves an equality for the prescribed iterates, not an adversarial oracle-complexity theorem.

A.3 Proof of Theorem 2

If $I_q(C) = i$, the menu contains i and no action ranked above i. The event has probability $F_i^K - F_{i-1}^K$, proving (4). For that fixed menu,

$$\sup_{v\in[l,u]} \{\max_{a\in C} v_a - v_i\} = \max\big(\{0\} \cup \{(u_j - l_i)_+ : j \in C,\ j < i\}\big). \tag{24}$$

The upper bound follows coordinatewise. It is attained by setting $v_i = l_i$ and setting the other present coordinates to their upper endpoints. The selected action is determined by q, not by this adverse v.

Fix $t \ge 0$, and let $B_i(t)$ denote the proposal mass of the lower-ranked actions with $u_j - l_i > t$. The event that the selected action is i and (24) exceeds t requires at least one i, at least one such threatening action, and no action above i. Starting from the event that i is selected and subtracting the event containing no threat gives exactly $H_i(t)$ in (7). Events for different selected i are disjoint. The nonnegative tail-integral identity yields (8).

Sort all upper endpoints once, using an order ρ with $u_{\rho_1} \le \cdots \le u_{\rho_A}$. For winner i, filter this order to $j < i$ with $u_j > l_i$. The resulting gaps $u_j - l_i$ are already sorted; compute suffix proposal masses in that order and integrate the constant tail between consecutive gaps. Duplicate gaps contribute zero-length intervals. Filtering and scanning cost $O(A)$ for each i, giving $O(A^2 + A\log A)$ total work and $O(A)$ scratch space. All timing comparisons use this quadratic integrator. Powers are counted as arithmetic operations; this is not a bit-complexity statement for unbounded K.

To prove the upper part of (8), move the supremum inside the expectation. For the lower part, draw one global random vector Z by independently selecting each coordinate's lower or upper endpoint with probability $1/2$. For any fixed menu with positive envelope, choose a maximizing threatening action $j \ne i$. On the event $Z_j = u_j$, $Z_i = l_i$, which has probability $1/4$, the menu's actual loss is at least its envelope. On other events the loss is nonnegative. Thus

$$\mathbb{E}_Z \mathbb{E}_C \{\max_{a\in C} Z_a - Z_{I_q(C)}\} \ge \Gamma_K/4.$$

Every realization Z is one fixed vector in the box, so the left side is at most R_K. The argument uses neither Bellman realizability nor independence of coordinates in an actual critic; it only constructs admissible vectors for the rectangular relaxation.

With two actions, loss is possible only when both appear. On that event the selected action is 2 and the loss is $(v_1 - v_2)_+$. The same fixed choice $v_1 = u_1, v_2 = l_2$ maximizes all nonzero-loss menus simultaneously. The probability of both actions is $1 - \beta_1^K - \beta_2^K$, proving equality. For verification with small A, the function $\ell_K(q, v)$ is convex in v (a maximum averaged over menus minus a linear term). Maximizing a convex function on a rectangle has a vertex maximizer: holding all other coordinates fixed, moving a coordinate to one of its endpoints cannot decrease the maximum. Repeating over coordinates justifies our exhaustive corner comparator.

A.4 Proof of Proposition 3

Use $Q_K = r + \gamma P V_K$ and $Q^\pi = r + \gamma P V^\pi$. Adding and subtracting $\sum_a \pi(a \mid s) Q_K(s, a)$ gives

$$\begin{aligned} V_K(s) - V^\pi(s) &= \mathcal{F}_K Q_K(s) - \sum_a \pi(a \mid s) Q_K(s, a) + \gamma \sum_a \pi(a \mid s) P_{s,a}(V_K - V^\pi) \\ &= \ell(s) + \gamma P_s^\pi (V_K - V^\pi). \end{aligned}$$

The first term is the expected regret of using the q-selected action in the same random menu whose Q_K maximizer defines V_K. This common-menu coupling is essential. Unroll to termination and average under μ to obtain an exact sum of local losses along π. Apply Theorem 2 at every state. At each layer, the probability of still being nonterminal is at most one; bounding its contribution by the largest local envelope yields the last expression in (12). Nonnegativity follows either from the local losses or the matched-budget optimality interpretation.

A.5 Fixed-count model confidence

For outcome-independent fixed counts, assign $\alpha = \delta/(2M)$ to the reward interval and to the transition set at each row. The transition radius is

$$t_{s,a} = \min\{1, \sqrt{[D_{s,a} \log 2 + \log(2M/\delta)]/[2n(s, a)]}\}. \tag{25}$$

Use full sets for zero counts and zero transition radius for singleton support. Let $\mathcal{E}$ be simultaneous model containment.

For a fixed transition row with D possible successors, total variation between $\widehat{P}$ and P equals $\max_{B \subseteq [D]} \{\widehat{P}(B) - P(B)\}$. For each fixed subset B, Hoeffding's one-sided inequality for the sample indicators gives $\mathbb{P}\{\widehat{P}(B) - P(B) > t\} \le e^{-2nt^2}$. A union bound over at most 2^D subsets and the radius in (25) gives failure at most $\delta/(2M)$. Clipping the radius at one covers every simplex point. When $D = 1$ there is no transition uncertainty. The binomial interval has failure at most $\delta/(2M)$ by inversion of its two equal-tail binomial tests. Union bounding both kinds of intervals over M rows gives $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$.

This fixed-count proof alone does not justify conditioning on an arbitrary random visit count. Theorem 5, proved below, supplies the additional count-uniform event for trajectories. For fixed counts, independent binomial reward and multinomial successor draws give the sufficient statistics directly, with the same distribution as the individual observations.

A.6 Proof of Proposition 6

On $\mathcal{E}$, for any finite vector v on a row's support,

$$|(P - \widehat{P})v| \le \mathrm{TV}(P, \widehat{P}) \operatorname{span}(v).$$

To see this, subtract the midpoint of the minimum and maximum of v, use that $(P - \widehat{P})\mathbf{1} = 0$, and apply the ℓ_1–ℓ_∞ inequality. Reward containment and this inequality imply $|\mathcal{T}_K q - \widehat{\mathcal{T}}_K q| \le e(q, K)$ simultaneously for every q, K. The implication is deterministic on a model-containment event; q need not be independent of the data.

At the final layer, there is no continuation, so the residual triangle inequality gives $|Q_K - q| \le d_{H-1} = b_{H-1}$ on supported actions. Assuming the bound at layer $h+1$, nonexpansiveness of $\mathcal{F}_K$ on the proposal support implies

$$|Q_K(s,a) - q(s,a)| \le |\mathcal{T}_K q(s,a) - q(s,a)| + \gamma b_{h+1} \le d_h + \gamma b_{h+1} = b_h.$$

Earlier termination contributes zero and cannot increase the bound. Induction proves the stated value-radius containment. Apply Proposition 3. Finally, if i is selected and j is in its menu, $q_j \le q_i$, so $(u_j - l_i)_+ = (q_j - q_i + 2b_h)_+ \le 2b_h$. Hence $\Gamma_K \le 2b_h$, which proves comparison with the generic certificate. Every implication holds for every query on the one event $\mathcal{E}$, including an adaptively selected last query. No union bound over iterates, budgets, or candidate policies is required.

A.7 Robust policy bands and baseline replacement tests

Proposition 8 (Baseline safe gates)**.** *On simultaneous model containment, either $G_o(s) \ge 0$ at every state or $\mu V^-_{\pi_n} \ge \mu V^+_{\pi_o}$ implies $J(\pi_n) \ge J(\pi_o)$. The union of these tests is valid for every adaptively selected pair of policies on the same event.*

Start terminal lower and upper values at zero. For a fixed policy π, at each layer compute

$$L^\pi(s,a) = r^-_{s,a} + \gamma \min_{p \in \mathcal{C}_{s,a}} p^\top V^-_{h+1}, \qquad V^-_h(s) = \sum_a \pi(a \mid s) L^\pi(s,a),$$

$$U^\pi(s,a) = r^+_{s,a} + \gamma \max_{p \in \mathcal{C}_{s,a}} p^\top V^+_{h+1}, \qquad V^+_h(s) = \sum_a \pi(a \mid s) U^\pi(s,a).$$

Containment and monotonicity prove $L^\pi \le Q^\pi \le U^\pi$ by backward induction. The minimum and maximum over a simplex TV ball are elementary mass-transport linear programs: shift up to the permitted TV mass from the lowest- to highest-value coordinates, reversing direction for the minimum. The implementation is independently tested against a general linear programming (LP) solver.

The objective in (20) is linear and coordinate-separable; minimizing each coordinate at its appropriate endpoint proves the exact box formula. On $\mathcal{E}$, $\langle \Delta\pi, Q^{\pi_o} \rangle \ge G_o$. The finite-horizon performance-difference identity expresses $J(\pi_n) - J(\pi_o)$ as the expected discounted sum of these advantages along π_n. All nonnegative G_o therefore suffice. For the scalar gate, $J(\pi_n) \ge \mu^\top V^-_{\pi_n} \ge \mu^\top V^+_{\pi_o} \ge J(\pi_o)$. Both statements are deterministic for every policy pair on $\mathcal{E}$. Induction over the adaptively accepted policies proves the whole-sequence guarantee.

A.8 A two-point lower bound

Here KL denotes Kullback–Leibler divergence.

Lemma 9. *For risky proposal probability $p \in (0,1)$ and $\chi_K(p) = 1 - p^K - (1-p)^K$, consider risky means $1/2 \pm \Delta$, $0 < \Delta \le 1/4$, and known safe mean $1/2$. If $n \le 1/(32\Delta^2)$, every rule choosing from a fresh K-sample menu has expected regret at least $\chi_K(p)\Delta/4$ in one of the two environments, even with unlimited computation.*

Let P_+ and P_- denote the laws of the offline risky-action samples. A single observation has divergence

$$\mathrm{KL}(\mathrm{Ber}(1/2+\Delta) \| \mathrm{Ber}(1/2-\Delta)) = 2\Delta \log \frac{1/2+\Delta}{1/2-\Delta} \le 16\Delta^2$$

for $\Delta \le 1/4$. For example, applying $\log((1+x)/(1-x)) \le 2x/(1-x^2)$ with $x = 2\Delta$ gives the stronger constant $32/3$. Independence gives $\mathrm{KL}(P_+ \| P_-) \le 16n\Delta^2$. Pinsker's inequality bounds their total variation by $\sqrt{8n\Delta^2} \le 1/2$ in the stated regime. The average error of any test under equal prior probabilities is at least $(1 - \mathrm{TV}(P_+, P_-))/2 \ge 1/4$.

If the menu contains only one action, the extractor and its matched comparator must choose the same action. If it contains both, selecting between them is exactly a test of the sign of the risky advantage; choosing the wrong action loses Δ. The menu distribution and all its multiplicities are independent of the environment sign and cannot supply statistical information. The probability of this mixed menu is χ_K. Thus average regret is at least $\chi_K \Delta/4$, and one of the two environments has at least

that regret. Internal randomization or additional computation cannot improve the testing lower bound. Our numerical experiment computes the exact equal-prior Bayes error $\frac{1}{2}\sum_{x=0}^{n}\min\{p_+(x), p_-(x)\}$, rather than simulating the sign test.

A.9 Proof of Theorem 4: threshold queries and approximation

Fix the critic order once, excluding zero-probability actions. For a threshold t, computing $H_i(t)$ reduces to the prefix query

$$B_i(t) = F_{i-1} - \sum_{j<i:\, u_j \le l_i + t} \beta_j.$$

Coordinate-compress the upper endpoints. Before processing winner i, a Fenwick tree stores the masses β_j for exactly $j < i$, indexed by u_j. Query the cumulative mass through $l_i + t$, compute $B_i(t)$ and then $H_i(t)$, and insert β_i at u_i. Binary search and a prefix-sum/update each take $O(\log A)$; preprocessing sorts in $O(A \log A)$ and space is $O(A)$. Repeated equal upper endpoints are inserted into the same compressed bin. This construction is valid for arbitrary interval order. If the u_i already increase in critic order, the prefix mass is obtained directly from F at the binary-search cutoff, clipped to the first $i-1$ actions. It has the same stated worst-case cost but avoids explicit tree updates.

For a random candidate set, define the nonnegative variable

$$D(C) = \sup_{v \in [l,u]} \{\max_{a \in C} v_a - v_{I_q(C)}\}.$$

Theorem 2 gives $\mathbb{P}\{I_q(C) = i, D(C) > t\} = H_i(t)$, hence $\mathbb{P}\{D(C) > t\} = T(t)$. If $I_q(C) = i$, then $D(C) \le M_i$. Therefore

$$\mathbb{E}D(C) \le \sum_i M_i \mathbb{P}\{I_q(C) = i, D(C) > 0\} = U_0.$$

With at most two supported actions, every positive-loss set contains both, has the same selected action, and has $D(C) = M_2$. The preceding inequality is then equality. A simple useful lower bound is $\mathbb{E}D(C) \ge t\mathbb{P}\{D(C) > t\}$; setting $t = G/2$ proves the refinement screen used by the algorithm.

For the geometric bound, take $a_0 = 0$, $a_1 = \min\{\tau, G\}$ and $a_{j+1} = \min\{(1+\eta)a_j, G\}$ until G is reached. The case $G = 0$ is immediate. On each interval $[a_j, a_{j+1})$, monotonicity of T gives

$$(a_{j+1} - a_j)T(a_{j+1}) \le \int_{a_j}^{a_{j+1}} T(t)\, dt \le (a_{j+1} - a_j)T(a_j).$$

The sum of the right endpoints is a lower bound L and the sum of the left endpoints is an upper bound U_{grid}. To control slack, define the rounded variable $D^+(C)$ to be zero if $D(C) = 0$, and otherwise the smallest grid endpoint not below $D(C)$. The upper sum is exactly $\mathbb{E}D^+(C)$: for a positive D in $(a_j, a_{j+1}]$, all upper-sum bins through j contribute and total a_{j+1}. If $0 < D \le \tau$, this is at most τ; if $D > \tau$, the preceding positive grid point is smaller than D and rounding increases it by at most $1 + \eta$. Consequently, pointwise,

$$D \le D^+ \le (1+\eta)D + \tau \mathbf{1}\{D > 0\}.$$

Taking expectations proves $\Gamma_K \le U_{\mathrm{grid}} \le (1+\eta)\Gamma_K + \tau\theta$. Both U_0 and U_{grid} dominate Γ_K, so $U = \min\{U_0, U_{\mathrm{grid}}\}$ preserves the upper guarantee. It is also at least L. The number of positive grid steps is at most $1 + \lceil \log_+(G/\tau)/\log(1+\eta) \rceil$, up to the last clipped endpoint. Combining this with the threshold-query complexity gives (15). Neither this argument nor the coarse bound requires random sampling of candidate sets.

Aggregate approximation. At state s, let the exact and approximate local bounds be Γ_s, U_s and exposed-ambiguity probability θ_s. Then

$$\sum_h \gamma^h \max_{s \in \mathcal{S}_h} U_s \le (1+\eta) \sum_h \gamma^h \max_{s \in \mathcal{S}_h} \Gamma_s + \tau \sum_h \gamma^h \max_{s \in \mathcal{S}_h} \theta_s.$$

Each $\theta_s \le 1$. State-specific η_s, τ_s are equally valid if their corresponding local slack is retained; the experiments use common $\eta = .25$ and $\tau = .05\varepsilon/\sum_h \gamma^h$. All inequalities hold for every input, so adaptive numerical resolution does not spend additional statistical confidence.

Numerical qualification. The theorem is in real arithmetic. The implementation uses stable differences of powers and clips small cancellation artifacts to the mathematically valid probability range. The tests compare independent formulas and enumeration within declared tolerances. It is not a directed-rounding or interval-arithmetic proof of the floating-point output. A safety-critical implementation would need that additional numerical verification.

A.10 Proof of Theorem 5: random visit counts

One can generate a stationary finite Markov environment using an infinite independent stream of outcome pairs for each state–action row. Each pair has that row's joint reward/successor law. On the nth visit to the row, reveal its nth pair. Actions and the decision to continue data collection are measurable with respect to previously observed outcomes and fresh policy randomization, not unrevealed pairs. This construction induces the stated Markov trajectory distribution. Rewards and successors in the *same* pair need not be independent. Marginal reward indicators and marginal successor indicators are each independent and identically distributed along a row stream.

For row (s, a) and each fixed $n \geq 1$, the two-sided exact binomial interval has coverage at least $1 - \alpha_n$. For transitions, apply Hoeffding to the n indicators of every subset of the known successor support. Since total variation is the maximum positive subset probability discrepancy, the union of at most $2^{D_{s,a}}$ one-sided events fails with probability at most

$$2^{D_{s,a}} \exp(-2nt_{s,a,n}^2) \leq \alpha_n$$

when the unclipped radius is below one; a radius of one trivially contains the whole simplex. As before a singleton support is known exactly. The bounds hold for the *first n outcomes in the potential row stream*, whether or not that many visits eventually occur.

Union both kinds of failures over rows and all n:

$$\mathbb{P}\{\text{any row interval fails at any count}\} \leq \sum_{s,a} \sum_{n \geq 1} 2\alpha_n = \sum_{s,a} \frac{\delta}{M} \sum_{n \geq 1} \frac{1}{n(n+1)} = \delta.$$

The last sum telescopes to one. On its complement, selecting the actually observed, possibly random count cannot create a failure. This includes outcome-dependent *stopping of collection*, because the entire stream event already holds. It does not justify outcome-dependent *removal of selected records*: retained records need not be a stream prefix. Unknown initial distribution, misspecified layer support, nonstationarity, and partial observability are also not covered by this theorem.

Finally, model containment is a pathwise statement about every row, independent of which critic is later computed. Residual containment, robust backward evaluation, and all gate inequalities are deterministic implications of that statement. Thus an arbitrarily long adaptive sequence of offline queries does not require a new union bound over the queries. This is distinct from adaptively querying an ordinary holdout value estimate without protecting its reuse.

A.11 Proof of Theorem 7: propagated advantages

Let $a_o(s) = \sum_a [\pi_n(a \mid s) - \pi_o(a \mid s)] Q^{\pi_o}(s, a)$. The old-policy Bellman equation and a telescoping sum along the proposed policy give

$$J(\pi_n) - J(\pi_o) = \mathbb{E}_{\pi_n} \sum_{h=0}^{H-1} \gamma^h a_o(S_h). \tag{26}$$

For completeness, subtract $V^{\pi_o}(s)$ from $V^{\pi_n}(s)$, add and subtract $\sum_a \pi_n(a \mid s) Q^{\pi_o}(s, a)$, and use $\sum_a \pi_o(a \mid s) Q^{\pi_o}(s, a) = V^{\pi_o}(s)$. This yields the recursion with local term a_o and continuation under P^{π_n}; unrolling proves the identity.

On model containment, minimizing the linear form in the old-value box gives $G_o(s) \leq a_o(s)$. Let y be the true expected discounted sum of G_o under π_n, with terminal zero. We show $z \leq y$ by backward induction. At the final layer, both equal G_o. If $z \leq y$ at later states, the true row $P_{s,a}$ belongs to $\mathcal{P}_{s,a}$, and nonnegative policy weights imply

$$z(s) = G_o(s) + \gamma \sum_a \pi_n(a \mid s) \min_{P \in \mathcal{P}_{s,a}} Pz \leq G_o(s) + \gamma \sum_a \pi_n(a \mid s) P_{s,a} y = y(s).$$

Average under μ and use (26) to obtain the bound. If every $G_o \geq 0$, induction gives every $z \geq 0$, so the new test contains the all-state test. If the model and old values are exact, the interval minimum gives $G_o = a_o$ and the transition minimum uses the true row only; the same recursion is exactly the performance difference. These statements hold for every pair of policies on one model event, establishing adaptive validity and monotonic return along accepted replacement sequences.

Strict containment witness. Consider an undiscounted two-step model. At the root the two actions have the same deterministic transition to the second state; their rewards are 0 and $.4$. At the second state their rewards are $.5$ and $.4$. The old policy selects the zero-reward root action and the $.5$ terminal action; the proposed policy selects the $.4$ root action and the $.4$ terminal action. With an exact model, G_o is $.4$ at the root and $-.1$ at the second state. The all-state test fails, while the propagated test certifies the exact return gain $.3$. The test is not asserted to dominate the separate root-value interval test. In an uncertain model, independently pessimizing boxes and transitions may combine effects that no single model realizes, which is why we claim a sufficient bound rather than exact robust improvement.

Scope of the preceding witness. The exact-model witness above separates propagation only from the all-state test; the separate root-value test already accepts it. The uncertain-model separation from their actual union is Appendix B.7.

A.12 Statistical rate statement

Consider one decision, safe mean $1/2$, risky mean $1/2 \pm \Delta$, n risky Bernoulli samples, and risky proposal probability p. Only a mixed candidate set requires distinguishing the signs. Its probability is $\chi_K(p) = 1 - p^K - (1-p)^K$.

Theorem 10 (Two-action minimax rate)**.** *For $n \geq 1$, $p \in (0,1)$, and fixed $K \geq 1$, the minimax regret over the two signs and $0 \leq \Delta \leq 1/4$, relative to the matched-K comparator, satisfies*

$$\frac{\chi_K(p)}{4\sqrt{32n}} \;\leq\; \inf_{\mathcal{A}} \sup_{\pm,\Delta} \mathbb{E}\,\mathrm{Regret}(\mathcal{A}) \;\leq\; \frac{\chi_K(p)}{2\sqrt{en}}. \tag{27}$$

Comparing the empirical risky mean with $1/2$ satisfies the upper guarantee; the lower bound allows unlimited computation and randomized menu-dependent decisions.

Both bounds are zero at $K = 1$, where no action choice is possible. The proof specializes classical two-point testing and Hoeffding's inequality (Kaufmann et al., 2016; Hoeffding, 1963). The useful conclusion is the *same exposure factor* in the certificate and statistical obstruction, not a new bandit identification method. This worst-case result does not turn a particular rejected dataset into an impossibility certificate.

What the rate means. The quantity n counts observations of the risky action, not total episodes. Missing coverage and poor estimation are therefore distinct from a small search budget. When Kp is small and the risky action is rare, $\chi_K(p) \leq Kp$ makes its ranking uncertainty hard to expose. For fixed $p \in (0,1)$, increasing K eventually exposes the comparison almost surely, leaving the usual $n^{-1/2}$ worst-case rate. At fixed nonzero exposure, the minimax sample requirement therefore scales quadratically with inverse target regret. More Bellman updates cannot supply those observations.

A.13 Proof of Theorem 10: matching upper and lower rates

The lower bound is Lemma 9 with $\Delta_n = (32n)^{-1/2}$. For $n \geq 1$, $\Delta_n \leq 1/4$, so the instance belongs to the claimed parameter family. The decision rule may depend arbitrarily on the candidate multiplicities, but conditional on a mixed menu those multiplicities are independent of the unknown sign. They cannot reduce the testing error below the same total-variation lower bound. Thus the minimax loss is at least $\chi_K(p)/(4\sqrt{32n})$.

For the upper bound, select the risky action in a mixed menu when its empirical mean exceeds $1/2$, select the safe action when it is smaller, and randomize equally at a tie. Nonmixed menus impose no regret. Under either sign, Hoeffding's inequality bounds the probability of selecting the inferior

action by $e^{-2n\Delta^2}$. Therefore

$$\mathbb{E}\,\mathrm{Regret} \le \chi_K(p)\Delta e^{-2n\Delta^2} \le \frac{\chi_K(p)}{2\sqrt{en}}.$$

The second inequality follows by differentiating xe^{-2nx^2} on $x \ge 0$; its maximum occurs at $x = 1/(2\sqrt{n})$. This upper maximization on all nonnegative x also covers the restricted interval $[0, 1/4]$. The same calculation works for $\Delta \le 1/2$, though our matching lower statement only needs the smaller family. Both sides vanish for $K = 1$. Data are n observations of the risky arm, not n arbitrary behavior episodes; the latter count depends on coverage.

B Stronger exact computation and a conservatism analysis

B.1 Exact computation with two candidate draws

For $K = 2$, a repeated action incurs no selection loss. A distinct pair $\{j, i\}$ with $j < i$ occurs in two orders, each with probability $\beta_i\beta_j$, and the critic chooses i. The worst permitted loss is $(u_j - l_i)_+$. Summing these disjoint events proves

$$\Gamma_2 = 2\sum_i \beta_i \sum_{j<i} \beta_j (u_j - l_i)_+.$$

This argument remains valid with tied critic scores, because a fixed tie priority supplies the total order. Unsupported actions can be removed. The ambiguity probability is similarly $2\sum_i \beta_i \sum_{j<i:u_j>l_i} \beta_j$.

Sort the upper endpoints and compress their ranks. Before processing winner i, a Fenwick tree contains all $j < i$, with one field storing β_j and another storing $\beta_j u_j$. A strict suffix query above l_i gives mass B_i and weighted mass C_i. Its contribution is $2\beta_i(C_i - l_i B_i)$. Query before inserting i to exclude self-comparisons. Each query and insertion costs $O(\log A)$; sorting costs $O(A \log A)$; memory is $O(A)$. This is an application of a standard data structure, not a new Fenwick-tree result. When the upper endpoints follow critic order, prefix sums and batched binary searches compute the same expression, avoiding explicit tree queries and updates. Both branches are compared against independent enumeration of all ordered pairs in the tests. Timing reports $K = 2$ separately from larger K.

B.2 An exact moment-tree algorithm for fixed candidate budget

The two-draw identity suggests a stronger general comparator than quadratic integration. The following construction is exact in real arithmetic and near-linear in the action count for *fixed* K. Its dependence on K and its implementation constants matter.

Proposition 11 (Fixed-budget exact integration)**.** *For integer $K \ge 2$, the envelope can be computed in $O(A \log A + K^2 A \log A)$ arithmetic operations and $O(KA)$ space. For $K = 2$ the direct Fenwick algorithm is preferable. This complexity treats integer powers and floating-point coefficients as arithmetic operations; it is not a bit-complexity or rounding-error guarantee.*

Proof. Process actions in increasing critic order. Before winner i, let $F = \sum_{j<i} \beta_j$, $b = \beta_i$, $U_i = \max_{j<i} u_j$, and

$$C_i(x) = \sum_{j<i:u_j \le x} \beta_j.$$

For $x \in [l_i, U_i]$, the probability that i wins and a predecessor has upper endpoint exceeding x is

$$(F + b)^K - F^K - \{(C_i(x) + b)^K - C_i(x)^K\}.$$

The term of degree zero in C_i cancels, as does the degree-K term in the finite difference. Therefore, for $U_i > l_i$, winner i contributes

$$\sum_{d=1}^{K-1} \binom{K}{d} b^{K-d} \left[F^d (U_i - l_i) - \int_{l_i}^{U_i} C_i(x)^d \, dx \right]. \tag{28}$$

For $U_i \le l_i$ the contribution is zero. Every bracket in (28) is nonnegative since $C_i(x) \le F$.

Compress the sorted lower and upper endpoints into at most $2A$ knots. The cumulative distribution function (CDF) is constant between consecutive knots. A segment tree stores the integral moments $M_d = \int C_i(x)^d dx$ on each node interval, for $d = 0, \dots, K-1$. At initialization M_0 is the interval length and higher moments are zero. Inserting action i adds b to C on the suffix $x \ge u_i$. For a whole node interval under an addition b, update

$$M'_d = \sum_{t=0}^{d} \binom{d}{t} b^{d-t} M_t.$$

A lazy scalar addition records the change for descendants. Updating moments in descending degree, generating binomial coefficients and powers rowwise, needs only $O(K)$ temporary space. Each node update costs $O(K^2)$. A suffix update and range query visit $O(\log A)$ tree nodes; querying $[l_i, U_i]$ returns the moments needed in (28). Summing the A contributions proves correctness and the stated complexity. Endpoints themselves have zero integration measure, so either consistent convention at ties is valid. □

This construction limits the claim for geometric integration: approximation is *not necessary* for subquadratic dependence on A at every fixed K. It trades exact polynomial moment updates and $O(KA)$ memory for $O(A)$ memory and threshold-resolution dependence. Our vectorized implementation uses dense transforms, allocating an additional $O(K^2)$ workspace; it does not implement the space-minimal rowwise variant. It is compared with the quadratic integrator on 144 cases: symmetric intervals $q \pm .2$ and general random intervals, $A \in \{32, 128, 512, 2048\}$, six seeds, and $K \in \{2, 8, 32\}$. Both methods use the same boxes and three timing repeats; these repeats are not independent problem instances. At $A = 2{,}048$, symmetric $K = 8/32$ moment timings are approximately $162/269$ ms versus $102/100$ ms for the quadratic kernel. For general intervals they are $369/778$ ms versus $125/124$ ms. Thus the stronger asymptotic exact method is not faster at the tested sizes. The integrated comparison uses the direct exact $K = 2$ kernel and quadratic general-K integration, rather than a hindsight minimum selected separately on each instance. These timings describe the tested sizes and implementations; they do not establish a universal runtime ordering.

B.3 State–action residual radii before taking layer maxima

Let $\delta_{s,a} = |\widehat{\mathcal{T}}_K q(s,a) - q(s,a)| + e_{s,a}$, with e from (18). Set $a = 0$ at termination and compute backwards

$$a(s,b) = \delta_{s,b} + \gamma \max_{P \in \mathcal{P}_{s,b}} P \mathcal{F}_K a. \tag{29}$$

Here b denotes an action, not the layer radius b_h. This recurrence is an additional robust dynamic program used in the analysis; its cost is not hidden in the fast stopping routine.

Proposition 12 (Nested residual boxes)**.** *On the model event, all supported coordinates satisfy*

$$|Q_K(s,b) - q(s,b)| \le a(s,b) \le b_h, \qquad s \in \mathcal{S}_h.$$

The coordinatewise model hull $[L_K, U_K]$ returned by robust backward evaluation is contained in $[q-a, q+a]$, which is contained in $[q-b_h, q+b_h]$.

Proof. For any nonnegative error vector a and any sampled set C, $|\max_C Q - \max_C q| \le \max_C |Q - q| \le \max_C a$ whenever $|Q-q| \le a$. Averaging gives $|\mathcal{F}_K Q - \mathcal{F}_K q| \le \mathcal{F}_K a$. The triangle inequality, the simultaneous model event, and backward induction give the first inequality in the proposition from (29). For the second, $\delta_{s,b} \le d_h$ on supported actions and $\mathcal{F}_K a(s') \le b_{h+1}$ by induction, while termination has value zero. Thus $a(s,b) \le d_h + \gamma b_{h+1} = b_h$. The argument holds for every model in the confidence set, not just the generating model, proving the hull containment.

For completeness, in a finite layered model with independent reward intervals and row-rectangular transition sets, robust backward evaluation returns the actual extrema of each Q_K coordinate. At the last layer choose all reward lower (or all upper) endpoints simultaneously. Inductively, a single choice of later rows attains every lower continuation value; each preceding row can independently choose its reward endpoint and minimizing transition against that continuation vector. Monotonicity of $\mathcal{F}_K$ completes the lower construction; the upper construction is identical with maxima. This does not imply that arbitrary mixtures of lower and upper coordinates are jointly attainable. That missing correlation is precisely what a marginal box discards. □

B.4 AN ORDERED DECOMPOSITION OF CERTIFICATE SLACK

Fix a generating model $M \in \mathcal{C}$, a critic q, and its extracted policy π. Write $\mathcal{O}_M(f) = \mathbb{E}_{\pi,M}\sum_h \gamma^h f(S_h)$ and $\mathcal{O}_{\mathcal{C}}(f) = \max_{N\in\mathcal{C}} \mathcal{O}_N(f)$. In the following analysis M is used *only for evaluating slack*, never for choosing a critic or accepting a policy. Define the eight scalars

$$
\begin{aligned}
B_0 &= J_K^M - J^M(\pi), & B_1 &= \mathcal{O}_M(R_K(q; L_K, U_K)),\\
B_2 &= \mathcal{O}_M(\Gamma_K(q; L_K, U_K)), & B_3 &= \mathcal{O}_M(\Gamma_K(q; q-a, q+a)),\\
B_4 &= \mathcal{O}_M(\Gamma_K(q; q-b_h, q+b_h)), & B_5 &= \mathcal{O}_{\mathcal{C}}(\Gamma_K(q; q-b_h, q+b_h)),\\
B_6 &= \sum_h \gamma^h \max_{s\in\mathcal{S}_h} \Gamma_K(q_s; q_s-b_h, q_s+b_h), & B_7 &= \sum_h \gamma^h \max_{s\in\mathcal{S}_h} U(q_s; q_s-b_h, q_s+b_h).
\end{aligned}
$$

Every displayed box is intersected with the known value range $[0, \sum_{t=0}^{H-h-1}\gamma^t]$ at layer h. This common intersection preserves all containments. All functions inside $\mathcal{O}$ are statewise. Then

$$B_0 \le B_1 \le B_2 \le B_3 \le B_4 \le B_5 \le B_6 \le B_7, \qquad B_2 \le 4B_1. \tag{30}$$

The first step uses the exact performance-difference recursion and the fact that the actual Q_K vector belongs to the hull. The second is the local rectangular theorem. The next two follow from nested boxes. The true transition model is feasible for B_5; worst-state aggregation bounds every visitation measure for B_6. Finally $U \ge \Gamma_K$. The additional factor-four inequality applies only to the second step. These comparisons require no independence between successive critic queries.

$B_1 - B_0$ includes uncertainty about the true model and loss of shared value structure; it is *not* a pure measurement of correlation loss. $B_3 - B_2$ includes symmetric recentering and the row model-error bound. The later increments isolate the explicitly ordered relaxations. Let B_g be the generic residual bound, capped by the maximum possible return difference. The deployed certificate is $\min\{B_7, B_g\}$. Applying the same cap to every stage, $\widetilde{B}_i = \min\{B_i, B_g\}$, preserves the chain and leaves B_0 unchanged. The main-text percentages divide sums of $\widetilde{B}_{i+1} - \widetilde{B}_i$ by the sum of $\widetilde{B}_7 - B_0$ across the stated analysis grid. Capping can erase late increments, so these percentages describe only the capped chain. These are ordered attributions, not Shapley decompositions, causal effects, or a universal ranking of all possible bounds.

For replacement, let $G^M(s) = \langle \pi_n - \pi_o, Q_M^{\pi_o}\rangle(s)$ and let G_o be its marginal-box lower bound. The exact improvement is $\mathcal{O}_{M,\pi_n}(G^M)$. The analysis records the two nonnegative losses

$$J_M(\pi_n) - J_M(\pi_o) - \mathcal{O}_{M,\pi_n}(G_o), \qquad \mathcal{O}_{M,\pi_n}(G_o) - \mu z.$$

The first combines statistical value uncertainty and its rectangular relaxation; the second isolates worst-case propagation of those already-lowered advantages. No calculation here claims to solve exact common-model baseline regret for a general uncertain transition system.

B.5 SHARED CONTINUATION UNCERTAINTY CAN MAKE A RECTANGLE VACUOUS

Proposition 13 (Common-offset obstruction)**.** *Let $v(C) = (C, C+g)$, $C \in [0, W]$, $g > 0$, and let q rank the second action higher. Put $\beta = (p, 1-p)$. The worst loss over this shared-value family is zero for every K. Its tight coordinate hull gives*

$$R_K(q; [0, W] \times [g, g+W]) = \Gamma_K = [1 - p^K - (1-p)^K](W - g)_+. \tag{31}$$

Thus even an exact two-action envelope has no multiplicative comparison with regret over the shared model family.

Proof. The true gap is always g, so every selected action maximizes the true value within the candidate set. The coordinate hull allows the incompatible assignment (W, g) and the two-action formula gives (31). This family is realized by two root actions with known immediate rewards $0, g$, identical known transitions, and a common uncertain downstream reward C. It is a standard fully observed Markov construction. The uncertainty set for the underlying reward is one interval; it is the projection to separate action values that discards dependence. □

Neither layerwise inflation nor worst-state aggregation has a universal constant-loss guarantee either. An uncertain row at a state with zero initial and subsequent visitation can enlarge a common layer radius while the actual visited rows have exact values. Separately, if the only local loss L occurs at a state visited with probability ρ, the occupancy bound is ρL but the worst-state bound is L; their ratio is $1/\rho$. These constructions show why one should measure the different relaxations rather than infer global tightness from the local factor four.

B.6 Exactness for a shared uncertain prefix

Proposition 14 (Common-prefix exactness)**.** *Partition a layered MDP into a prefix and a suffix that is closed under future transitions. Rewards and dynamics within the suffix are known. The two policies agree at every prefix state, where rewards and row-rectangular transitions may be uncertain. Then the propagated-advantage bound equals the shared-model robust improvement:*

$$\mu z = \min_{M \in \mathcal{C}} [J_M(\pi_n) - J_M(\pi_o)].$$

The statement is for a specified policy pair and uncertainty set, not optimization over all policy pairs.

Proof. In the suffix, old-policy values and advantages are known exactly. In the prefix, the coefficient vector $\pi_n - \pi_o$ is zero, so the advantage is exactly zero regardless of uncertain old-policy values. Thus the local advantage function G is the same for every model and the marginal formula recovers it exactly. For each M, the performance-difference recursion expresses $J_M(\pi_n) - J_M(\pi_o)$ as the return of fixed rewards G under π_n and M's transitions. Because the prefix transitions are row-rectangular, robust backward evaluation minimizes that fixed-policy return exactly, including when suffix improvements have either sign. This evaluation is exactly the propagated gate. Uncertain prefix rewards cancel before minimization, since both policies incur them with the same prefix behavior. No assertion is made that arbitrary uncertain-model advantage functions are fixed across the model set. □

B.7 A strict uncertain-model separation from both baseline gates

At a forced-action root, the probability of entering branch s is $p \in [.2, .8]$; the other branch t returns one under both policies. The transition set is a TV ball of radius .3 around $(.5, .5)$ on the two known successors. In branch s, the retained policy obtains reward zero then .2. The proposed policy obtains .4 then .1, with certain progression between these decisions. All quantities below the root are known. Consequently

$$V_o(s) = .2, \quad V_n(s) = .5, \quad J_o(p) = 1 - .8p, \quad J_n(p) = 1 - .5p.$$

The all-state advantage test fails at the final changed state, where the advantage is $-.1$. The separate root test also fails:

$$\inf_p J_n(p) - \sup_p J_o(p) = .6 - .84 = -.24.$$

The two root policies agree, so the uncertain root action has exactly zero advantage. Propagation gives $.4 - .1 = .3$ in branch s and zero in branch t, hence

$$\mu z = \min_{p \in [.2,.8]} .3p = .06 = \min_{p \in [.2,.8]} [J_n(p) - J_o(p)] > 0.$$

This separates propagation from the *union* of the all-state and separate root-value gates under actual model uncertainty. It does not make propagation an exact robust-regret solver in general.

The example also admits finite candidate-sampling policies. Use a uniform two-action proposal at each changed state; let old and new critics have opposite rankings and use $K = 4$. The probabilities of the preferred actions are $15/16$ under each critic. The net branch improvement becomes $.3(14/16) = 21/80$. The shared robust and propagated bounds are $21/400$, while the root bound is $-207/800$ and the final-state advantage is $-7/80$. Root and unchanged-branch proposals have singleton support. Thus the separation is not an artifact of requiring deterministic policies unattainable by finite candidate search. The grid varies interval endpoints, local gain/loss, and K; its exact common-model minimum is attained at an interval endpoint because the return difference is affine in p.

B.8 AN EXACT RESTRICTED SHARED-MODEL COMPARATOR

The appropriate general objective is the common-model difference $\inf_{M\in\mathcal{C}}\{J_M(\pi_n) - J_M(\pi_o)\}$, studied by Petrik et al. (2016). Our gate is only a tractable sufficient lower bound on this objective. In particular, it can still discard dependence between the old-policy values and the transitions used to propagate their advantages.

A useful exactly solvable special case has known transitions and independent reward intervals. Let $d_\pi(s,a) = \mathbb{E}_\pi \sum_h \gamma^h \mathbf{1}\{S_h = s, A_h = a\}$ and $c = d_{\pi_n} - d_{\pi_o}$. Then

$$\inf_{r\in[r^-,r^+]}(J_r(\pi_n) - J_r(\pi_o)) = \sum_{c_{s,a}\geq 0} c_{s,a} r^-_{s,a} + \sum_{c_{s,a}<0} c_{s,a} r^+_{s,a}. \tag{32}$$

Indeed the objective is $\sum c_{s,a} r_{s,a}$ and independent endpoints minimize each term. This classical occupancy calculation preserves cancellation of shared rewards; it is included as a comparator, not claimed as a new general safe-improvement formulation. The experiment explicitly gives every method the same known dynamics in this restricted setting. It never silently replaces uncertain transitions in the integrated trajectory experiment by their generating values.

B.9 SEPARATING GENUINE MODEL UNCERTAINTY FROM PROJECTION LOSS

The first increment in (30) combines two effects. To separate them in a tractable class, consider known transitions and at most six rewards uncertain in independent intervals. For a fixed extracted policy π, $J_K(r)$ is a maximum over finitely many menu-dependent policies, each of which has return linear in r. Hence $J_K(r) - J_\pi(r)$ is convex in r and its maximum on a reward box is attained at a vertex. Exhaustively enumerating the 2^4 or 2^6 vertices therefore computes *exact* shared-model worst regret, not a sampled lower bound. The difference between that quantity and actual regret is genuine uncertainty under the specified model set. The difference between the marginal fixed-vector bound and that exact shared-model maximum is projection/independent-local-adversary loss. Both are nonnegative on the model event. This exact calculation does not determine their separate magnitudes in the larger uncertain-transition workflow.

C CERTIFICATE-MATCHED EVALUATION

This section compares the cost of returning a policy *with a specified guarantee*. A time–return comparison in which fixed schedules do not pay for a final near-optimality certificate measures a different objective. Lower time alone cannot establish a cheaper way to obtain that certificate. We therefore evaluate the complete workflow with a paid final check for every schedule. No true-model evaluation enters the check.

C.1 A COMMON STATISTICAL TARGET, TWO VALID CERTIFICATE CONSTRUCTIONS

The requested output at budget K and tolerance ε is

$$J_M(\pi^*_K) - J_M(\pi_{q,K}) \leq \varepsilon \quad \text{on the simultaneous model-containment event.} \tag{33}$$

The comparator is the optimal policy for the *same* candidate budget and proposal, not the full-action optimum. The confidence level is $1-\delta$ per dataset. Candidate certification and replacement are separate: rejecting a certified candidate does not automatically certify the retained policy.

Our common final verifier C_E first evaluates the clipped generic residual bound. If that bound exceeds the tolerance, it evaluates the exact clipped exposure bound and takes their minimum. For two actions this is the endpoint formula in Appendix D; for $K = 2$ and arbitrary action count it uses the near-linear exact kernel. Otherwise it uses the optimized quadratic integrator. Already computed final residual bands are cached. Exact exposure stopping reuses its last exact check rather than recomputing it. Fixed-sweep and backward-DP schedules construct their final bands and pay for the verification. Numerical approximation is not used in this final common check, even when the stopping rule used it.

For an additional comparator, backward DP is followed by a root-value certificate C_R:

$$C_R(\pi, K) = \max\{0, \mu\overline{V}_K - \mu\underline{V}^\pi\}. \tag{34}$$

The upper matched-budget values and lower candidate-policy values come from two robust backward evaluations on the same model set. The lower candidate values are reused by the replacement gate. This comparator can exploit local model information without passing through layerwise residual radii; omitting it would favor the exposure certificate. Both C_E and C_R certify (33), although they need not succeed on the same datasets.

Proposition 15 (Comparable root-value certification)**.** *On any model-containment event, $J_M(\pi_K^*) - J_M(\pi) \le C_R(\pi, K)$ for every adaptively chosen candidate policy $\pi = \pi_{q,K}$. Thus $C_R \le \varepsilon$ is a valid alternative certificate for* (33)*. It need not be more or less conservative than the exposure certificate.*

Proof. Monotonicity of backward expected-max evaluation gives $V_{K,M} \le \overline{V}_K$ at every state. Fixed-policy robust evaluation gives $V_M^\pi \ge \underline{V}^\pi$. Multiply by the nonnegative initial distribution and subtract. Both inequalities hold simultaneously for all queried policies on the same model event, so adaptive queries incur no new statistical penalty. Taking a maximum with zero does not decrease the bound. Neither this argument nor the robust root computation claims novelty beyond standard robust dynamic programming; their role is to match the comparator's statistical target. □

Why either certificate can be useful. Two one-step examples establish non-dominance without unknown transition effects. First, two actions have reward intervals $[0, .1]$ and $[.8, 1]$, uniform proposal and $K = 2$. Their midpoint critic has the correct ranking throughout the model set. Its residual exposure certificate is zero, whereas $C_R = 7/40$: separate root intervals fail to cancel the shared reward uncertainty even though the policy is budget-optimal in every model. Second, take two states in the same layer, initial distribution $(1, 0)$, and known rewards $(0, 1)$ at the reachable state. At the unreachable state both rewards are in $[0, 1]$. A midpoint critic and uniform proposal at $K = 2$ give $C_E = .5$ because of the layer/worst-state bound, but $C_R = 0$ because the uncertain state is never visited. Both examples have zero shared-model worst regret. These values follow directly from the two-action formula and robust one-step evaluation; enumerating all reward-box vertices independently verifies them. The 18-case grid varies proposal mass and K. This explains why certificate yield must accompany timing: a faster test that does not certify the requested assertion is not an interchangeable output.

C.2 A TIGHTER VALID SET FOR THE FIXED-EPISODE PROTOCOL

The count-spending set protects optional stopping of collection. That protection is useful, but the recorded experiments actually stop after a prespecified number N of independent episodes. In a time-layered model with a fixed behavior policy, there is a less conservative valid choice. Its stronger sampling assumptions are indispensable.

Proposition 16 (Conditioning on counts in fixed independent episodes)**.** *Suppose that N is deterministic, episodes are independent, the behavior policy is fixed independently of their outcomes, and every row can be visited at most once per episode. At row (s, a), condition on its vector of episode-visit indicators. Conditional on that vector, the observed outcome pairs are independent and identically distributed from the row's true joint reward/successor distribution. Consequently, using a Clopper–Pearson reward interval and the finite-alphabet total-variation radius with failure allocation $\alpha = \delta/(2M)$ per component at the* observed random count *gives a simultaneous model set of coverage at least $1 - \delta$. It is pathwise contained in the count-spending set constructed from the same statistics.*

Proof. Fix a row, not all rows simultaneously. In episode e, pre-generate its potential reward/successor pair Y_e independently of the upstream trajectory and of policy randomization before that row. The fixed-policy, time-layered Markov construction permits this: deciding whether to visit the row cannot depend on an outcome revealed only after that decision, and the episode cannot return to that row. Different episodes have independent randomness. Hence the vector $(Y_e)_{e=1}^N$ is independent of the vector of visit indicators $(I_e)_{e=1}^N$, with common row law. Conditional on any indicator vector containing n ones, the selected n pairs are therefore independent and identically distributed. Reward and successor components of the same pair need not be independent.

For $n > 0$, the binomial interval's conditional failure probability is at most α. The same conditional independent and identically distributed property gives the total-variation bound used in Theorem 5, also with failure at most α. For $n = 0$, both sets are full and cannot fail. Average the conditional

bound over the visit indicators; it remains at most 2α for this row. Union over M rows gives $2M\alpha = \delta$. Different rows need not be independent. Later adaptive policy selection is covered by the same deterministic model-containment argument as before.

For every observed $n \geq 1$, $\delta/(2M)$ is at least $\delta/[2Mn(n+1)]$. Increasing the allowed failure level shrinks the two-sided Clopper–Pearson interval and decreases the displayed total-variation radius. Both sets have the same empirical center and support; zero-count sets coincide. Thus the fixed-episode model set is contained in the count-spending model set for every dataset, not merely on average. □

This conditioning argument is a specialization of classical fixed-sample confidence ingredients (Clopper & Pearson, 1934; Hoeffding, 1963; Weissman et al., 2003), not a new general confidence-sequence method. It fails if collection is stopped based on outcomes, if behavior adapts between episodes, or if a row can be revisited and its earlier outcomes affect its later visit count. The count-spending construction remains the appropriate one under those broader conditions. "Fixed episodes" does not mean fixed state–action counts: visitation is still random. The tighter construction does not validate selecting the episode count after examining the data.

For a fixed candidate policy, set containment can only decrease C_R. For a fixed critic it also shrinks the residual intervals and their exact exposure envelope. This monotonicity statement does not imply monotone final retained return across two complete adaptive pipelines: changing the set can change earlier acceptance decisions and later baselines.

C.3 Experiment design, clocks, and success-normalized cost

The matched comparison uses 120 main-grid datasets and tolerances $.05, .2$. Each dataset is shared across methods, tolerances, and budgets. Seven schedules are compared: generic, exact, and geometric stopping; one sweep; H sweeps; backward DP plus C_E; and backward DP plus C_R. The first six receive C_E on their returned critic. Every method uses the same replacement union. A sequence starts with the behavior policy and proposes $K = 2, 8, 32$ in order; accepted candidates become the retained policy. Critic initialization at each budget is the common empirical behavior critic, not a method-specific warm start.

All clocks start before confidence-set construction from sufficient statistics. They include initial critic and old-policy intervals, paid stopping diagnostics, final near-optimality verification, the identical three replacement tests, and any update to retained-policy intervals. Existing diagnostic bands and candidate bounds are reused where applicable, and their original cost is charged. Data collection and true-model evaluation remain outside the timer. The timing analysis separates setup, refresh, final certificate, and gate. A fixed seeded permutation randomizes method order independently of outcomes. One full timed pass is recorded per dataset/tolerance/construction. Timing repeats are not additional independent datasets. These cost comparisons are restricted to the specified algorithms and experimental grid.

For each fixed reporting stratum we report certificate yield, mean full-sequence time, nontrivial acceptance, final retained return, and

$$\text{charged time per certificate} = \frac{\sum_d T_d}{\sum_d \sum_{K\in\{2,8,32\}} \mathbf{1}\{\text{final certificate succeeds at } (d,K)\}}. \tag{35}$$

Here T_d includes all three budgets and their gates, including unsuccessful attempts. This is an amortized throughput statistic for the recorded workload, not a causal expected waiting time to the first certificate. A zero denominator is reported as "–", not as zero cost. We separately report certified *retained* candidates: the candidate certificate transfers only if the gate accepts it, or if the retained policy is identical. A successful certificate with a rejected distinct candidate is not a deployment certificate. No timing is silently conditioned on success while discarding the cost of failures.

The sensitivity grid adds eight independently seeded graphs for each of three structures: $(H, A) = (3, 8), (5, 8), (3, 32)$, always width two and proposal concentration .7. It uses $N = 2{,}000, 20{,}000, 200{,}000$, both confidence constructions at $\delta = .05$, and the same tolerances, methods, budgets, and gates. There are 72 additional trajectory datasets and 144 paired model-set constructions. Together, the two grids contain 192 datasets, 264 model sets, and 11,088 policy-proposal stages.

These are correlated repeated queries, not 11,088 independent experimental replicates. The structures were chosen after examining the five-family main grid. This is a sensitivity study, not an unbiased estimate over all possible RL tasks.

Certificate-matched main-grid results. After the same final exposure check, generic, exact/geometric, H-sweep, and DP schedules each certify 144/720 proposals; one sweep certifies 111. DP plus root intervals certifies 84. On identical DP proposals, exposure certifies 60 cases missed by root intervals, whereas root intervals add 0 cases in the main grid. The yield difference occurs in the fork and one-step control. In the broader graph sensitivity study below, the other certificate succeeds. Fixed-sweep and DP runtimes that omit final verification cannot answer this comparison. The current evaluation charges it. DP plus the common exposure verifier obtains the same final certificates as exposure stopping without requesting certificates along the way. The one-sweep schedule remains a different return/certificate compromise rather than a superior solution to a fixed output requirement.

Sensitivity to geometry and confidence. Across 72 additional datasets from eight seeds per graph structure, DP–root certificates increase from 42/432 with count-spending confidence to 80/432 with the fixed-episode set. Mean reward-interval and transition-radius ratios are 0.63 and 0.72, respectively. The residual-exposure methods still certify no sensitivity-grid proposal at these tolerances. Thus failure of the residual-exposure certificate does not imply that finite-data certification is impossible: a different valid bound and a tighter valid model set can succeed. The sensitivity-grid success differences use paired graph blocks rather than treating counts or budgets as independent replicates. Every confidence set contained its generating model in the sensitivity grid, and no false certificate or harmful accepted replacement was observed. This does not make the nominal failure probability zero. A separate fixed-count validation observed one model-confidence failure among 480 datasets. Its binomial reward and multinomial successor statistics use the fixed-count construction in Appendix A.5; this failure is not part of the 192-dataset trajectory grid.

The paired increase in DP–root certificate yield is 8.80 percentage points (95% graph-block bootstrap interval [7.64, 9.95]). This interval describes the chosen graph distribution and fixed grid, not all offline RL tasks.

D Algorithms and computational accounting

Stopped refresh for a fixed budget. Inputs are a critic q_0, empirical model, a valid model confidence set (or an exact model), K, tolerance ε, and an update cap. At iteration m, compute $z = \widehat{\mathcal{T}}_K q_m$, build the residual radii from that same z, and evaluate the generic bound. If it passes, stop. Otherwise evaluate the coarse exposure bound; if it passes, stop. For the geometric variant, compute the midpoint lower bound when refinement might matter. If that lower bound already exceeds ε, further integration cannot make the exact envelope pass, so retain the valid current upper bound and update. Otherwise refine geometrically and compare the resulting upper bound with the tolerance. On failure, use z as q_{m+1}. At the cap return the policy with its actual bound and an explicit certified/uncertified flag. Exact and generic comparators use the same backup recurrence and final diagnostic charge.

For critic order $q(s,1) \le q(s,2)$, with larger index winning a tie, the exact two-action formula is

$$\Gamma_s = (u(s,1) - l(s,2))_+[1 - \beta(1 \mid s)^K - \beta(2 \mid s)^K]. \tag{36}$$

The endpoints are *already clipped*: $l(s,a) = \max\{0, q(s,a) - b_h\}$ and $u(s,a) = \min\{V_{\max,h}, q(s,a) + b_h\}$. For the reverse critic order, exchange the two action indices. Only without clipping does the gap reduce to $(2b_h - |q(s,1) - q(s,2)|)_+$. For $q = (0, .1)$, $b_h = 1$, range $[0,1]$, uniform proposal, and $K = 2$, the implemented clipped branch gives .5, not the untrimmed .95. All stopping methods use this endpoint form. Independent boundary checks and tests of the stopping rule verify the clipped calculation. The expression is zero for one supported action and for one draw.

Cost scope. A Bellman sweep includes candidate-order calculation, expected maxima, and transition products. The certificate-matched timers include initial critic construction, confidence-set construction from observed statistics, old-policy intervals, every paid diagnostic, final verification, gates, and retained-policy interval updates. Only data collection and hidden-model evaluation are excluded. Appendix C.3 specifies the timing protocol.

Table 3: Main grid, count-spending confidence. All costs include the common gate. Successes count candidate certificates, not guaranteed improvements. "–" means no successful certificate.

Family	Method	Cert.	ms/seq.	ms/cert.	Gain
Fork, depth four	Generic+C_E	56	22.2	19.0	0.82
Fork, depth four	Exact stop	56	23.0	19.7	0.82
Fork, depth four	Geometric stop	56	23.5	20.1	0.82
Fork, depth four	One+C_E	23	13.6	28.5	0.17
Fork, depth four	H+C_E	56	17.0	14.6	0.82
Fork, depth four	DP+C_E	56	17.0	14.6	0.82
Fork, depth four	DP+C_R	40	20.0	24.0	0.82
Three-layer graph	Generic+C_E	0	58.6	–	0.07
Three-layer graph	Exact stop	0	90.4	–	0.07
Three-layer graph	Geometric stop	0	88.8	–	0.07
Three-layer graph	One+C_E	0	52.8	–	0.07
Three-layer graph	H+C_E	0	53.5	–	0.07
Three-layer graph	DP+C_E	0	52.7	–	0.07
Three-layer graph	DP+C_R	0	56.2	–	0.07
Five-layer graph	Generic+C_E	0	95.7	–	0.00
Five-layer graph	Exact stop	0	183.0	–	0.00
Five-layer graph	Geometric stop	0	168.7	–	0.00
Five-layer graph	One+C_E	0	81.0	–	0.00
Five-layer graph	H+C_E	0	82.2	–	0.00
Five-layer graph	DP+C_E	0	82.5	–	0.00
Five-layer graph	DP+C_R	0	90.8	–	0.00
32-action graph	Generic+C_E	0	189.6	–	0.00
32-action graph	Exact stop	0	294.0	–	0.00
32-action graph	Geometric stop	0	219.7	–	0.00
32-action graph	One+C_E	0	171.7	–	0.00
32-action graph	H+C_E	0	174.0	–	0.00
32-action graph	DP+C_E	0	174.7	–	0.00
32-action graph	DP+C_R	0	190.0	–	0.00
One-step control	Generic+C_E	88	13.3	7.3	0.47
One-step control	Exact stop	88	13.4	7.3	0.47
One-step control	Geometric stop	88	14.3	7.8	0.47
One-step control	One+C_E	88	12.7	6.9	0.47
One-step control	H+C_E	88	12.6	6.9	0.47
One-step control	DP+C_E	88	12.7	6.9	0.47
One-step control	DP+C_R	44	11.4	12.4	0.47

Table 4: Sensitivity grid, count-spending confidence. All costs include the common gate. Successes count candidate certificates, not guaranteed improvements. "–" means no successful certificate.

Family	Method	Cert.	ms/seq.	ms/cert.	Gain
Three-layer graph	Generic+C_E	0	61.8	–	0.16
Three-layer graph	Exact stop	0	92.4	–	0.16
Three-layer graph	Geometric stop	0	89.8	–	0.16
Three-layer graph	One+C_E	0	53.9	–	0.16
Three-layer graph	H+C_E	0	55.4	–	0.16
Three-layer graph	DP+C_E	0	55.7	–	0.16
Three-layer graph	DP+C_R	23	56.1	117.1	0.16
Five-layer graph	Generic+C_E	0	100.7	–	0.19
Five-layer graph	Exact stop	0	188.4	–	0.19
Five-layer graph	Geometric stop	0	173.9	–	0.19
Five-layer graph	One+C_E	0	86.2	–	0.16
Five-layer graph	H+C_E	0	89.4	–	0.19
Five-layer graph	DP+C_E	0	88.6	–	0.19
Five-layer graph	DP+C_R	14	92.4	316.7	0.19
32-action graph	Generic+C_E	0	205.6	–	0.15
32-action graph	Exact stop	0	309.5	–	0.15
32-action graph	Geometric stop	0	235.6	–	0.15
32-action graph	One+C_E	0	188.2	–	0.15
32-action graph	H+C_E	0	190.0	–	0.15
32-action graph	DP+C_E	0	192.8	–	0.15
32-action graph	DP+C_R	5	195.1	1873.1	0.15

Table 5: Sensitivity grid, fixed-episode confidence. All costs include the common gate. Successes count candidate certificates, not guaranteed improvements. "–" means no successful certificate.

Family	Method	Cert.	ms/seq.	ms/cert.	Gain
Three-layer graph	Generic+C_E	0	63.5	–	0.22
Three-layer graph	Exact stop	0	93.8	–	0.22
Three-layer graph	Geometric stop	0	93.3	–	0.22
Three-layer graph	One+C_E	0	57.0	–	0.23
Three-layer graph	H+C_E	0	58.9	–	0.22
Three-layer graph	DP+C_E	0	58.2	–	0.22
Three-layer graph	DP+C_R	32	56.7	85.1	0.22
Five-layer graph	Generic+C_E	0	103.2	–	0.26
Five-layer graph	Exact stop	0	189.9	–	0.26
Five-layer graph	Geometric stop	0	178.6	–	0.26
Five-layer graph	One+C_E	0	87.7	–	0.26
Five-layer graph	H+C_E	0	90.5	–	0.26
Five-layer graph	DP+C_E	0	91.1	–	0.26
Five-layer graph	DP+C_R	24	91.3	182.6	0.26
32-action graph	Generic+C_E	0	208.2	–	0.21
32-action graph	Exact stop	0	314.3	–	0.21
32-action graph	Geometric stop	0	238.6	–	0.21
32-action graph	One+C_E	0	192.6	–	0.23
32-action graph	H+C_E	0	194.1	–	0.21
32-action graph	DP+C_E	0	193.8	–	0.21
32-action graph	DP+C_R	24	194.6	389.2	0.21

For small finite tables, one can precompute $w_i(K)$ and draw directly from the resulting categorical policy. Consequently LK candidate evaluations over L deployment decisions are not an unavoidable lower bound for every implementation. No amortized optimal-allocation claim is made.

Policy replacement. For each proposed policy, all methods compute the same union of all-state, separate root-value, and propagated-advantage gates on the same model set. Every accepted policy becomes the next baseline. These tests share a statistical event but have computational costs, all included. No wrapper is claimed to optimize over all budget allocations.

E ADDITIONAL RELATED-WORK COMPARISONS

Section 2 presents the main comparisons. This appendix gives additional context for the statistical primitives, computational tools, and scope of the claims.

Batch RL and repeated evaluation. Least-squares policy iteration and tree-based batch RL established practical ways to learn control policies by repeatedly evaluating and improving them using fixed data (Lagoudakis & Parr, 2003; Ernst et al., 2005). Finite-time fitted-iteration and error-propagation analyses explain how approximation errors affect later decisions (Munos & Szepesvári, 2008; Farahmand et al., 2010). Approximate modified policy iteration makes the evaluation/improvement tradeoff explicit (Scherrer et al., 2015). Our recurrence and performance-difference arguments belong to this lineage. We do not claim that stale evaluation or multistep error propagation was previously unknown. The exact fork isolates the continuation mismatch for candidate-budget transfer; the main additional analysis weights interval ranking uncertainty by the sampled alternatives that can expose it.

Action gaps and extraction. Action-gap theory shows why value error need not translate directly into policy error (Farahmand, 2011); gap-increasing Bellman operators pursue a related robustness goal (Bellemare et al., 2016). EMaQ defines the expected-max operator and analyzes candidate-budget fixed points (Ghasemipour et al., 2021). Empirical work separates value learning from policy extraction (Park et al., 2024), while the recent decoupled-extraction preprint studies frozen-proposal reranking (Lin et al., 2026). Best-of-N theory characterizes coverage and imperfect scoring in inference-time alignment (Huang et al., 2025). We distinguish three claims: more-search degradation is not new; the EMaQ operator is not new; the particular rectangular envelope, its fixed-vector comparison, and its certified threshold approximation are the technical objects advanced here. The proofs include screening by higher-ranked actions, which is absent from a uniform minimum-gap condition.

Offline distribution shift and pessimism. The broad challenges of offline RL are reviewed by Levine et al. (2020). Batch-constrained Q-learning and bootstrapping error reduction restrict extrapolation beyond the data (Fujimoto et al., 2019; Kumar et al., 2019). Conservative Q-learning, implicit Q-learning, and simple behavior-regularized baselines offer different practical treatments of offline policy extraction (Kumar et al., 2020; Kostrikov et al., 2022; Fujimoto & Gu, 2021). Their use of learned critics should not be taken as evidence that the tabular confidence set here extends automatically to neural networks. Theoretical pessimism results and Bellman-consistent approaches address coverage and statistical efficiency in much broader learning settings (Jin et al., 2021; Rashidinejad et al., 2021; Xie et al., 2021; Zanette & Wainwright, 2022). We do not improve their general sample-complexity guarantees. Our finite-data role is narrower: support repeated adaptive queries of explicit value intervals and identify how a fixed candidate mechanism exposes unresolved decisions.

Safe policy improvement and evaluation. High-confidence improvement and baseline bootstrapping use offline evidence to control replacement risk (Thomas et al., 2015; Laroche et al., 2019). Doubly robust and data-efficient off-policy evaluation offer alternative routes to policy-value uncertainty (Jiang & Li, 2016; Thomas & Brunskill, 2016). Our gate is model based; it does not claim to be statistically superior to these estimators. Its advantage is a shared model-containment event that covers all subsequent policy queries. The new propagated-advantage test is a refinement of a specific all-state sufficient condition, with a strict witness and a measured improvement in acceptance. We neither claim universal dominance over root-value intervals nor optimal power among all safe tests.

Robust control and regret. Robust DP over rectangular uncertainty sets is established (Iyengar, 2005; Nilim & El Ghaoui, 2005), as are efficient L_1-robust backups and partial policy iteration (Ho et al., 2018; 2021). We use these primitives for confidence evaluation and propagation; we do not originate robust Bellman minimization. Robust-MDP regret and randomized minmax regret also study adverse choices under uncertainty (Ahmed et al., 2013; Mastin et al., 2014). Our R_K is a local fixed-vector objective under an exogenous random candidate set, while Γ_K exchanges its supremum and expectation. This is not the same as choosing a minimax-regret policy for an entire uncertain MDP. The factor-four proof specifically compares these two local objectives and does not transfer without qualification to joint model uncertainty.

Statistical and computational tools. Clopper–Pearson intervals and Hoeffding/finite-alphabet deviation inequalities provide the confidence ingredients (Clopper & Pearson, 1934; Hoeffding, 1963; Weissman et al., 2003). The count-spending proof is deliberately elementary and looser than modern confidence-sequence constructions may be (Howard et al., 2021); its purpose is to repair random-count reuse without a false independence argument. The two-action testing reduction draws on classical identification lower bounds (Kaufmann et al., 2016). The Fenwick-tree query is a standard data structure (Fenwick, 1994); the contribution is identifying the exposure tail that it computes and the deterministic integration guarantee. Uncertainty summaries use independent datasets as clusters rather than treating reused policies as independent runs, consistent with careful RL evaluation practice (Agarwal et al., 2021).