BRIDGING LIKELIHOOD-RATIO OFFLINE RL AND ROBUST RL: TRACTABILITY AND ERROR BOUNDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Likelihood-based offline reinforcement learning optimizes against models that fit the data, but its statistical appeal comes with a computational obstacle. A single likelihood constraint couples the entire model, whereas row-wise robust dynamic programming minimizes transition rows independently. We develop a tractable bridge between these formulations and quantify the price of using it. For tabular product transition models, planning over the exact state–action rectangular hull reduces to scalar reverse-KL problems and value iteration. Our central contribution is a two-sided, occupancy-sensitive analysis of the resulting approximation: a convex repair constructs a likelihood-feasible adversary, and a complementary convex program lower-bounds the loss incurred by independent row minimization. The error depends on incompatibility along visited rows, not likelihood-budget overspending alone. We characterize exactness for arbitrary initialization and distinguish fixed-policy approximation from coupled policy optimality. Combining this analysis with likelihood calibration yields offline performance guarantees that separate statistical uncertainty, relaxation error, and numerical error. The construction extends to valid shared linear-mixture models. Finite-model experiments support the bridge through faster matched-objective rectangular planning, occupancy-sensitive gap bounds, and common-parameter exactness checks. The bridge therefore makes a principled approximation to likelihood-based offline planning implementable and explains the additional pessimism introduced for computation.

1 INTRODUCTION

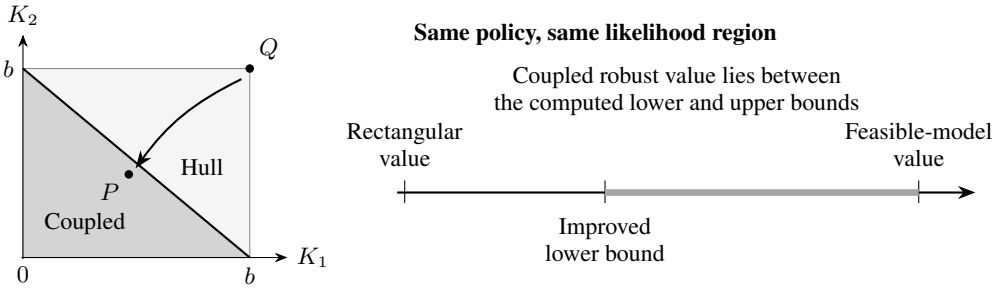
Offline reinforcement learning (RL) uses pessimism to distinguish high return from an artifact of incomplete data. Plausible transition models or consistent value functions provide lower performance estimates (Jin et al., 2021; Xie et al., 2021; Uehara & Sun, 2023); coverage and representation determine their statistical meaning (Chen & Jiang, 2019; Foster et al., 2022). A computational relaxation can nevertheless make a valid pessimistic objective *substantially more conservative*.

Constrained Pessimistic Policy Optimization with likelihood-ratio constraints (CPPO-LR) evaluates a policy against *one model that explains the entire dataset* (Uehara & Sun, 2023, extended version). A tabular likelihood region assigns one global budget across transition rows. Robust dynamic programming instead minimizes continuation expectations independently (Iyengar, 2005; Nilim & El Ghaoui, 2005). Each row can spend the full budget although their combination may not fit the data. On the same containment event, both objectives lower-bound true return: their difference is *approximation cost*, not stronger statistical protection.

How can likelihood-based pessimism be made tractable without losing its statistical meaning? We retain the coupled objective as the reference and analyze its exact state–action rectangular hull. Classical robust Bellman methods and scalar reverse-KL solvers make this hull implementable (Wiesemann et al., 2013; Adam & Mácha, 2022; Ho et al., 2022). The missing question is whether one likelihood-feasible model can realize the independent adverse rows along trajectories relevant to the policy.

Our main result expresses approximation error as an occupancy-weighted continuation residual minimized over the original likelihood region. Every nontrivial covered-row minimizer exhausts the budget. This yields **initialization-aware exactness**: one family of minimizing transitions must

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.



(a) One shared budget versus independent row budgets

(b) Convex approximation bounds and feasible-model evaluation

Figure 1: **Statistical pessimism and approximation error are distinct.** (a) With row-likelihood costs K_1, K_2 and budget b, the coupled set requires $K_1 + K_2 \le b$; its rectangular hull permits $K_1, K_2 \le b$ separately. A row-wise model Q is repaired to a coupled-feasible P. The arrow is illustrative, not a linear path in KL coordinates. (b) A lower bound on the rectangularization gap improves the lower endpoint; evaluating P gives the upper endpoint. These deterministic bounds enclose the coupled robust value, not the true-model return.

keep the initial support in a closed set reaching at most one such row (Theorems 2–3). Unreachable uncertainty can be repaired at zero cost; recurrently visited rows can instead compete for the budget.

When exactness fails, a convex occupancy-weighted repair constructs a likelihood-feasible adversary, and a complementary program lower-bounds the residual no feasible model can remove (Theorem 4, Proposition 5). These bounds enclose the coupled value without computing its minimum. A separate result bounds policy suboptimality: exact evaluation of the selected policy need not imply coupled optimality.

The statistical guarantee separates the comparator's likelihood uncertainty, rectangularization error, and numerical error. Shared-model extensions explicitly retain stochasticity. Experiments test matched-objective computation, recurrent approximation bounds, and common-parameter compatibility. The contribution is an **implementable likelihood relaxation with quantified error**, not a general coupled solver or uniform improvement over CPPO-LR.

2 The likelihood objective and its statistical target

Consider a finite discounted Markov decision process (MDP) with states $\mathcal{S}$, actions $\mathcal{A}$, known rewards $r(s,a) \in [0,1]$, discount $\gamma \in (0,1)$, and initial distribution μ. Let $\Delta(X)$ denote the probability simplex on a finite set X and $\Pi = \prod_{s\in\mathcal{S}} \Delta(\mathcal{A})$ the stationary randomized policy class. Vectors are columns; inequalities are componentwise. Write I for the identity and $\mathbf{1}$ for the all-ones vector. For a stochastic kernel P and $\pi \in \Pi$, set $P_\pi(s,x) = \sum_a \pi(a \mid s) P(x \mid s,a)$ and $r_\pi(s) = \sum_a \pi(a \mid s) r(s,a)$. With $\mathbb{P}_{P,\pi,\mu}$ denoting their trajectory law, define

$$V_P^\pi = (I - \gamma P_\pi)^{-1} r_\pi, \qquad J_P^\pi = \mu^\top V_P^\pi, \qquad d_P^\pi(s,a) = (1-\gamma) \sum_{t\ge 0} \gamma^t \mathbb{P}_{P,\pi,\mu}(s_t = s, a_t = a).$$

We use $\mathrm{TV}(p,q) = \frac{1}{2}\|p-q\|_1$ and $D_{\mathrm{KL}}(q\|p) = \sum_{x:q(x)>0} q(x)\log(q(x)/p(x))$, with value $+\infty$ if $q(x) > 0 = p(x)$. Logarithms are natural; $\mathrm{span}(v) = \max_x v_x - \min_x v_x$, $\mathrm{supp}\, q = \{x : q(x) > 0\}$, and $\mathbf{1}\{\cdot\}$ denotes an indicator. Theoretical returns are unnormalized; empirical displays use $J = (1-\gamma)J$.

A row $i = (s,a) \in \mathcal{I} := \mathcal{S} \times \mathcal{A}$ has a nonempty, predeclared successor set $X_i \subseteq \mathcal{S}$. Known rows $i \in \mathcal{I}_K$ are fixed at $P_i^0 \in \Delta(X_i)$; unknown rows range independently over $\Delta(X_i)$. Rows are embedded in $\Delta(\mathcal{S})$ with zero mass outside X_i. *An empirical zero does not remove a declared successor.* Let N_{ix} count observed successors, $N_i = \sum_x N_{ix}$, $\mathcal{I}_+ = \{i \notin \mathcal{I}_K : N_i > 0\}$, and $\mathcal{I}_0 = \{i \notin \mathcal{I}_K : N_i = 0\}$. Let $n = \sum_{i\in\mathcal{I}_+} N_i$ and $m = |\mathcal{I}_+|$. For $i \in \mathcal{I}_+$, set $\widehat{p}_i(x) = N_{ix}/N_i$ and $\widehat{\rho}_i = N_i/n$. Known-row likelihood terms cancel.

Fix the data and a finite budget $b \geq 0$. Define $\ell(P) = \sum_{i \in \mathcal{I}_+, x: N_{ix} > 0} N_{ix} \log P_i(x)$. A saturated maximum-likelihood kernel $\widehat{P}$ uses empirical covered rows, the prescribed known rows, and arbitrary admissible uncovered rows. The tabular likelihood-ratio region is

$$\mathcal{U}_b = \{P : K(P) \leq b\}, \qquad K(P) = \ell(\widehat{P}) - \ell(P) = \sum_{i \in \mathcal{I}_+} N_i D_{\mathrm{KL}}(\widehat{p}_i \| P_i). \tag{1}$$

For $n > 0$, a normalized likelihood tolerance ζ corresponds to $b = n\zeta$. We study $J^\pi_{\mathrm{LR}} = \min_{P \in \mathcal{U}_b} J^\pi_P$ and $J^\star_{\mathrm{LR}} = \max_{\pi \in \Pi} J^\pi_{\mathrm{LR}}$. Compactness and continuity give attainment. The minimizing kernel must be *shared across all states visited from* μ; minimizing separately for each initial state would change the objective. Section 4 calibrates b so this deterministic construction has a confidence-set interpretation.

3 What rectangular planning changes

3.1 One global budget becomes independently available row budgets

A set is (s, a)-rectangular if it is a Cartesian product of transition-row sets. Its rectangular hull is the product of its row projections, where the projection at i contains every P_i attained by a member of the original set. This is different from state-rectangularity, which can retain coupling among actions at one state.

Proposition 1 (Exact likelihood geometry)**.** *For $i \in \mathcal{I}_+$ and $u \geq 0$, let $\mathcal{B}_i(u) = \{p \in \Delta(X_i) : N_i D_{\mathrm{KL}}(\widehat{p}_i \| p) \leq u\}$, and let $\mathcal{Z} = (\prod_{i \in \mathcal{I}_0} \Delta(X_i)) \times (\prod_{i \in \mathcal{I}_K} \{P^0_i\})$. Then*

$$\mathcal{U}_b = \bigcup_{u \in \mathbb{R}^m_+ : \sum_i u_i \leq b} \left[\prod_{i \in \mathcal{I}_+} \mathcal{B}_i(u_i) \times \mathcal{Z} \right], \qquad \mathcal{R}_b := \mathrm{Rect}_{sa}(\mathcal{U}_b) = \prod_{i \in \mathcal{I}_+} \mathcal{B}_i(b) \times \mathcal{Z}. \tag{2}$$

The hull is the unique smallest (s, a)-rectangular superset of $\mathcal{U}_b$. For every such superset $\mathcal{V}$, policy, and initial distribution, $\inf_{P \in \mathcal{V}} J^\pi_P \leq \min_{P \in \mathcal{R}_b} J^\pi_P \leq J^\pi_{\mathrm{LR}}$.

To project row i, spend up to b there and set every other covered row to its empirical law. Thus its KL radius is b/N_i, not a newly calibrated rowwise threshold. Equation (2) also represents the coupled value as a minimum over budget allocations, but gives no convexity guarantee for that outer optimization. The hull is least conservative *among rectangular supersets of this fixed region*, not among all valid confidence sets.

For $v \in \mathbb{R}^{|\mathcal{S}|}$, write $\mathcal{R}_i = \mathrm{proj}_i \mathcal{U}_b$ and $\phi_i(v) = \min_{p \in \mathcal{R}_i} p^\top v$. Its Bellman operators are

$$(T_{\mathrm{R}} v)(s) = \max_a \{r(s, a) + \gamma \phi_{sa}(v)\}, \qquad (T^\pi_{\mathrm{R}} v)(s) = \sum_a \pi(a \mid s)\{r(s, a) + \gamma \phi_{sa}(v)\}. \tag{3}$$

These are monotone γ-contractions with fixed points $V^\star_{\mathrm{R}}, V^\pi_{\mathrm{R}}$. Row minimizers at V^π_{R} assemble into one $Q \in \mathcal{R}_b$, so $J^\pi_{\mathrm{R}} := \mu^\top V^\pi_{\mathrm{R}} = \min_{P \in \mathcal{R}_b} J^\pi_P$. If π_{R} is deterministic and greedy at $V^\star_{\mathrm{R}}$, its policy operator fixes the same vector; uniqueness gives $V^{\pi_{\mathrm{R}}}_{\mathrm{R}} = V^\star_{\mathrm{R}}$. For any other π, $T^\pi_{\mathrm{R}} V^\star_{\mathrm{R}} \leq V^\star_{\mathrm{R}}$ implies $V^\pi_{\mathrm{R}} \leq V^\star_{\mathrm{R}}$ by monotone iteration. This explains both greedy optimality and its limitation: *the assembled minimizing rows need not lie in $\mathcal{U}_b$.*

3.2 Exactness is a condition on reachable minimizing rows

Fix π, put $v = V^\pi_{\mathrm{R}}$, and choose $Q_i \in \arg\min_{p \in \mathcal{R}_i} p^\top v$. The nonnegative gap $G_\pi = J^\pi_{\mathrm{LR}} - J^\pi_{\mathrm{R}}$ measures the *change of objective*, not statistical estimation error. A simulation identity makes the obstruction explicit.

Theorem 2 (Exactness and likelihood-budget saturation)**.** *For $a_i(P; v) = P_i^\top v - \phi_i(v) \geq 0$ and $P \in \mathcal{U}_b$,*

$$G_\pi = \min_{P \in \mathcal{U}_b} \frac{\gamma}{1 - \gamma} \sum_i d^\pi_P(i) a_i(P; v). \tag{4}$$

Hence $G_\pi = 0$ exactly when some coupled-feasible model minimizes every positively occupied row. For $i \in \mathcal{I}_+$, let $c_i(v)$ be the minimum of $N_i D_{\mathrm{KL}}(\widehat{p}_i \| p)$ over $p \in \mathcal{B}_i(b)$ satisfying $p^\top v = \phi_i(v)$. For $b > 0$,

$$c_i(v) = b\,\mathbf{1}\{\widehat{p}_i^\top v > \min_{x \in X_i} v_x\}. \tag{5}$$

If μ has full support, exactness is equivalent to $\kappa_\pi(v) := \sum_{i=(s,a)\in\mathcal{I}_+ : \pi(a|s)>0} c_i(v) \le b$. For arbitrary μ this is sufficient but need not be necessary. For $b = 0$, $\mathcal{U}_b = \mathcal{R}_b$.

A nontrivial minimizing row must *spend all of b*: otherwise moving slightly toward a lowest-value successor improves it. This likelihood-specific fact specializes the general simultaneous-solvability perspective of Grand-Clément et al. (2026). With full-support initialization, at most one selected covered row can be nontrivial. With a concentrated start, the relevant question is instead which rows one feasible minimizing model can *avoid*.

Call i *critical* if it is covered, $\pi(a \mid s) > 0$, and the strict inequality in (5) holds; denote these rows by $\mathcal{C}_\pi(v)$. Consider a candidate $j \in \mathcal{C}_\pi(v) \cup \{\perp\}$, where $\perp$ means no critical row. Keep known rows fixed, use empirical laws at noncritical covered rows, and allow uncovered rows to choose any lowest-value successor. Candidate j may use any row minimizer in $\mathcal{R}_j$; other critical rows are disallowed. Repeatedly delete a state if some action with positive policy probability has no allowed row supported on the remaining states. Let Z_j^∞ be the final set, starting from $\mathcal{S}$.

Theorem 3 (Initialization-aware likelihood exactness)**.** *Under the tabular product-domain assumptions, for $b > 0$ and any μ,*

$$G_\pi = 0 \quad \Longleftrightarrow \quad \exists j \in \mathcal{C}_\pi(v) \cup \{\perp\} \text{ with } \operatorname{supp}\mu \subseteq Z_j^\infty. \tag{6}$$

A successful candidate yields one coupled-feasible kernel attaining the rectangular value from every state in Z_j^∞. Each deletion procedure terminates after at most $|\mathcal{S}|$ strict reductions.

Finite reverse KL keeps every observed successor positive. Replacing a reached noncritical row by its empirical law therefore introduces no new reachable state. Conversely, a closed minimizing set visits at most the one budget-saturating row, so its rows fit the global budget. Appendix B.3 supplies both proofs and support oracles. These exact equality tests are not a tolerance-free floating-point classifier.

One candidate must cover the entire initial support. Disconnected branches can each be exact from a pure start but not from their mixture: both cannot spend b. Proposition 9 expresses this common-model cost through concavity in the initial distribution and characterizes equality. The mixed-start experiment illustrates that standard concavity principle.

3.3 Bounding the loss when exactness fails

Exactness identifies a structural obstruction but does not measure its cost. To obtain an upper bound, repair Q toward the empirical model. For $\alpha \in [0,1]^m$ and $i \in \mathcal{I}_+$, set

$$P_i(\alpha) = (1-\alpha_i)Q_i + \alpha_i \widehat{p}_i, \quad f_i(\alpha_i) = N_i D_{\mathrm{KL}}(\widehat{p}_i \| P_i(\alpha)), \quad w_i = d_Q^\pi(i)\,\mathrm{TV}(Q_i, \widehat{p}_i),$$

leaving known and uncovered rows unchanged. Choose the repair through

$$B_\pi(Q) = \min_{\alpha \in [0,1]^m} \sum_{i \in \mathcal{I}_+} w_i \alpha_i \quad \text{subject to} \quad \sum_{i \in \mathcal{I}_+} f_i(\alpha_i) \le b. \tag{7}$$

Theorem 4 (Occupancy-weighted budget repair)**.** *Program* (7) *is convex and feasible. Every feasible α gives*

$$J_{\mathrm{R}}^\pi \le J_{\mathrm{LR}}^\pi \le J_{P(\alpha)}^\pi, \qquad 0 \le G_\pi \le J_{P(\alpha)}^\pi - J_{\mathrm{R}}^\pi \le \frac{\gamma}{(1-\gamma)^2} \sum_{i \in \mathcal{I}_+} w_i \alpha_i. \tag{8}$$

For $b > 0$, one scalar multiplier $\lambda \ge 0$ and separable minimizations of $w_i\alpha_i + \lambda f_i(\alpha_i)$ recover its optimum. If $K(Q) \le b$, no repair is needed and $G_\pi = 0$.

The weights keep d_Q^π *fixed*, so the objective is linear; repair-dependent occupancy would change the problem. Ordinary evaluation of a feasible repair upper-bounds the coupled minimum, not necessarily true return. Appendix C proves (8) and gives the first-order condition (28) and overspending bound (27).

A lower bound asks the complementary question: *how much residual must every coupled model retain?* Choose $0 \le \underline{d}_i \le d_P^\pi(i)$ for every $P \in \mathcal{U}_b$. One choice uses coordinate minima $\underline{P}_i(x) = \min_{p \in \mathcal{R}_i} p(x)$ and the substochastic matrix $S_\pi(s,x) = \sum_a \pi(a \mid s)\underline{P}_{sa}(x)$:

$$\underline{d}_{sa} = (1-\gamma)[\mu^\top (I - \gamma S_\pi)^{-1}](s)\pi(a \mid s).$$

Unlike the time-zero bound $(1-\gamma)\mu(s)\pi(a \mid s)$, this envelope retains unavoidable future visits.

Proposition 5 (Occupancy-weighted lower bound)**.** *For any nonnegative uniform occupancy lower envelope,*

$$A_{\underline{d}}(v) = \min_{P \in \mathcal{U}_b} \sum_i \underline{d}_i [P_i^\top v - \phi_i(v)], \qquad \frac{\gamma}{1-\gamma} A_{\underline{d}}(v) \le G_\pi. \tag{9}$$

This is a convex program with a one-multiplier dual and separable reverse-KL-penalized row problems. A dual-feasible objective gives a lower endpoint; a primal objective does not. If μ has full support and $\underline{d}_{sa} \ge (1-\gamma)\mu(s)\pi(a \mid s)$, then $A_{\underline{d}}(v) > 0$ exactly when Theorem 2's criterion fails.

Together, (8)–(9) make faithfulness checkable: a gap upper bound below a chosen value tolerance establishes that accuracy; a lower bound above it rules it out. A narrow interval far above zero instead measures substantial conservatism. *A zero lower bound alone is inconclusive.* Appendix D propagates errors in v and the dual computation.

To quantify the gap, consider a known uniform router to m uncertain binary branches, each observed to succeed N times. Take $m, N \ge 1$, $\zeta > 0$, budget $b = mN\zeta$, and known reward-one and reward-zero absorbing successors. Then

$$J_{\mathrm{LR}}^\pi = \frac{\gamma^2}{1-\gamma} e^{-\zeta}, \qquad J_{\mathrm{R}}^\pi = \frac{\gamma^2}{1-\gamma} e^{-m\zeta}. \tag{10}$$

Jensen's inequality spreads the coupled budget over branches; independent minima each spend it in full. The ratio in (10) is $e^{(m-1)\zeta}$, while the absolute normalized gap is bounded by γ^2. Both bounds recover that gap. Making only one branch reachable instead gives zero gap despite the same possible factor-m overspending (Appendix E.2). Recurrence tests the additional difficulty that visitation depends on uncertain transitions.

3.4 Evaluation error and policy loss are different quantities

A policy can have $G_{\pi_{\mathrm{R}}} = 0$ yet be suboptimal for the coupled objective. In the branching example, add a known safe action with value between the two expressions in (10). Rectangular planning selects it and evaluates it exactly; coupled planning prefers the branches. The separate control question therefore requires an upper bound *applying to every policy*. With $v^\star = V_{\mathrm{R}}^\star$, consider

$$\Delta_B = \min_{P \in \mathcal{U}_b,\, t \ge 0} t \quad \text{subject to} \quad r(s,a) + \gamma P_{sa}^\top v^\star \le v^\star(s) + t \quad \forall (s,a). \tag{11}$$

Proposition 6 (Coupled policy-suboptimality bound)**.** *Every feasible (P,t) in (11) gives $0 \le J_{\mathrm{LR}}^\star - J_{\mathrm{LR}}^{\pi_{\mathrm{R}}} \le t/(1-\gamma)$. Thus $\Delta_B = 0$ suffices for coupled optimality; the converse is not claimed.*

The same feasible P makes $v^\star + t\mathbf{1}/(1-\gamma)$ a supersolution for every ordinary policy operator, upper-bounding $J_{\mathrm{LR}}^\star$ without minimax interchange. Appendix E.3 gives the proof and example.

4 From a numerical planner to an offline performance guarantee

The preceding analysis holds *for fixed data*. Let $P^\star$ belong to the predeclared transition-model domain. Fix $\beta \in (0,1)$ before observing successors; a confidence statement requires $\mathcal{U}_b$ to contain $P^\star$ with the stated probability. For exogenous, fixed-length row labels, successors are conditionally independent with laws $P_i^\star$; this includes prescribed row counts and independently sampled row labels. Let $k_i = |X_i|$ and $D = \sum_{i \in \mathcal{I}_+} (k_i - 1)$.

Theorem 7 (Likelihood calibration under two sampling designs)**.** *For the exogenous design, the budget*

$$b_{\mathrm{iid}} = D + \sqrt{2D\log(1/\beta)} + \log(1/\beta) \tag{12}$$

gives $\mathbb{P}\{P^\star \in \mathcal{U}_{b_{\mathrm{iid}}} \mid \textit{row labels}\} \geq 1-\beta$. *For predictable row selection, each successor must have conditional law* $P_i^\star$ *given the pre-transition history. Let* $L_t(P)$ *be the transition likelihood through time* t, $\widehat{P}_t$ *its saturated maximizer, and* M_t *its integral against a fixed product Dirichlet*$(1/2,\dots,1/2)$ *measure on unknown rows. Set* $\mathcal{U}_t(u) = \{P : \log L_t(\widehat{P}_t) - \log L_t(P) \leq u\}$. *The budget*

$$b_t = \log(1/\beta) + \mathrm{pen}_t, \qquad \mathrm{pen}_t = \log L_t(\widehat{P}_t) - \log M_t \tag{13}$$

satisfies $\mathbb{P}\{P^\star \in \mathcal{U}_t(b_t)\ \forall t\} \geq 1-\beta$.

The first part uses multinomial concentration (Agrawal, 2020); the second a fixed likelihood-mixture supermartingale (Krichevsky & Trofimov, 1981; Lindon & Malek, 2022). The latter permits arbitrary history-dependent behavior, randomized resets before a transition, and stopping based on past data, provided the stated conditional Markov law holds. It needs no mixing assumption or prior law on $P^\star$.

Let $N_i(t)$ be the prefix count. Covered-row KL radii are b_{iid}/N_i and $b_t/N_i(t)$; uncovered rows remain unrestricted. At identical realized counts, their difference is

$$\frac{\mathrm{pen}_t - D_t - \sqrt{2D_t \log(1/\beta)}}{N_i(t)}, \qquad D_t = \sum_{j:N_j(t)>0,\ j\notin\mathcal{I}_K} (k_j - 1). \tag{14}$$

For m fixed k-outcome rows with equal count N and positive limiting proportions, $\mathrm{pen}_t = \frac{m(k-1)}{2}\log N + O(1)$: these particular radii scale as $\log N/N$ versus $1/N$. The exact finite-count penalty is (34); neither construction is always numerically larger. Only the trajectory result licenses successor-dependent stopping. Appendix G gives the assumptions and boundary cases together.

On either containment event, $J_{\mathrm{R}}^{\pi} \leq J_{\mathrm{LR}}^{\pi} \leq J_{P^\star}^{\pi}$ *simultaneously for every policy*. For $\epsilon_{\mathrm{alg}} \geq 0$, let $\widehat{\pi} \in \Pi$ satisfy $J_{\mathrm{R}}^{\widehat{\pi}} \geq \mu^\top V_{\mathrm{R}}^\star - \epsilon_{\mathrm{alg}}$. Any comparator $\pi^\dagger \in \Pi$ satisfies

$$J_{P^\star}^{\pi^\dagger} - J_{P^\star}^{\widehat{\pi}} \leq \epsilon_{\mathrm{alg}} + \underbrace{J_{P^\star}^{\pi^\dagger} - J_{\mathrm{LR}}^{\pi^\dagger}}_{\text{likelihood uncertainty}} + \underbrace{G_{\pi^\dagger}}_{\text{rectangularization error}}. \tag{15}$$

This is the pessimistic comparator argument (Jin et al., 2021; Xie et al., 2021; Uehara & Sun, 2023), with the cost of computation separated from likelihood uncertainty. Crucially, the gap belongs to the **comparator**; the returned policy's gap cannot replace it.

For an explicit comparison, suppose $n \geq 1$, the comparator visits no uncovered unknown row, and $C = \max_{i\in\mathcal{I}_+} d_{P^\star}^{\pi^\dagger}(i)/\widehat{\rho}_i < \infty$. Theorem 15 and (15) give

$$J_{P^\star}^{\pi^\dagger} - J_{P^\star}^{\widehat{\pi}} \leq \epsilon_{\mathrm{alg}} + \min\left\{ \frac{\gamma}{(1-\gamma)^2}\sqrt{\frac{2bCm}{n}}, \frac{\gamma}{(1-\gamma)^2}\sqrt{\frac{2bC}{n}} + G_{\pi^\dagger} \right\}. \tag{16}$$

The exact coupled optimizer has the matched bound $\gamma(1-\gamma)^{-2}\sqrt{2bC/n}$ in Appendix H. Thus (16) does **not establish statistical dominance over CPPO-LR** (Uehara & Sun, 2023): it quantifies the price of an implementable relaxation. The factor $\sqrt{m}$ compares upper bounds, not an unavoidable loss. Only the comparator needs coverage; the learner need not know C. At a trajectory stopping time, the same deterministic statements use prefix counts and b_t on the simultaneous event; no independence argument is reused. The uncovered-row term remains necessary whenever the comparator lacks coverage.

Numerical accuracy is the remaining term. For row-objective error at most δ, let $\widetilde{T}$ use those approximate objectives at v, let $\xi = \|v - \widetilde{T}v\|_\infty$, and let $\widetilde{\pi}$ be greedy for the same objectives. Proposition 13 gives

$$\|v - V_{\mathrm{R}}^\star\|_\infty \leq \frac{\xi + \gamma\delta}{1-\gamma}, \qquad \|V_{\mathrm{R}}^\star - V_{\mathrm{R}}^{\widetilde{\pi}}\|_\infty \leq \frac{2\gamma(\xi+\delta)}{(1-\gamma)^2}. \tag{17}$$

For normalized policy-loss target $\bar{\epsilon}$, choose raw tolerances $\delta, \xi \leq \bar{\epsilon}(1-\gamma)/(4\gamma)$. Objective intervals, not a fixed dual-coordinate tolerance, determine row accuracy. Appendix F treats empirical zeros and residual-corrected lower values, building on scalar divergence solvers (Adam & Mácha, 2022; Ho et al., 2022); Algorithm 1 gives the pipeline.

5 PRESERVING VALID SHARED TRANSITION STRUCTURE

Sharing parameters can reduce likelihood uncertainty, but never makes stochasticity automatic. Let $P_{\theta,i} = A_i\theta$, where row x of $A_i \in \mathbb{R}^{|\mathcal{S}|\times d}$ is the fixed feature $\varphi_{ix}^\top$, zero outside X_i. All features and structural constraints are predeclared. Collect *every row's normalization and all known-row equalities* in $H\theta = h$. The parameter domain and negative log likelihood are

$$\Theta = \{\theta : H\theta = h,\ A_j\theta \geq 0\ \forall j\}, \qquad F(\theta) = -\sum_{i\in\mathcal{I}_+, x:N_{ix}>0} N_{ix}\log(\varphi_{ix}^\top\theta).$$

Assume F has a finite constrained minimizer $\widehat{\theta}$. Put $\Theta_b^{\mathrm{LR}} = \{\theta \in \Theta : F(\theta) - F(\widehat{\theta}) \leq b\}$. Its kernel image is compact, even if redundant parameters are unbounded (Appendix I). For $g_i(v) = A_i^\top v$, the exact row backup is

$$\min_\theta\{g_i(v)^\top\theta : H\theta = h,\ A_j\theta \geq 0\ \forall j,\ F(\theta) - F(\widehat{\theta}) \leq b\}. \tag{18}$$

This is a convex program over **valid complete models**, not just valid row i. Different backups may use different feasible parameters; their assembled rows define the exact rectangular hull. Stochasticity gives contraction and the numerical bounds of Section 4. If $P^\star = P_{\theta^\star}$ for $\theta^\star \in \Theta$, the saturated confidence event implies $F(\theta^\star) - F(\widehat{\theta}) \leq b$; this uses no claim of a parameter-dimension-only calibration.

An optional outer ellipsoid must retain the same validity constraints. For $\Sigma \succ 0$, $R \geq 0$, let $E = \{\theta : \|\Sigma^{1/2}(\theta - \widehat{\theta})\|_2 \leq R\}$ and $\mathcal{C} = E \cap \Theta$. The row minimum over $\mathcal{C}$ is an SOCP. The unrestricted formula $\widehat{\theta}^\top g - R\|\Sigma^{-1/2}g\|_2$ is exact *only if its optimizer is in* Θ; otherwise it is merely a relaxation of that row problem, not the valid-MDP backup. Theorem 17 and Corollary 19 give implementable tests and a likelihood-curvature sufficient condition for using the norm formula. Calibration alone does not make E stochastic, and shrinking its radius to force validity need not preserve coverage. For a matrix $L \in \mathbb{R}^{d\times q}$ and equality-aware ellipsoid $E_L = \{\widehat{\theta} + Lz : \|z\|_2 \leq R\}$ with $R > 0$, the exact whole-set test is

$$H\widehat{\theta} = h, \qquad HL = 0, \qquad \varphi_{jx}^\top\widehat{\theta} \geq R\|L^\top\varphi_{jx}\|_2 \quad \forall(j,x). \tag{19}$$

Only when (19) holds is $\widehat{\theta}^\top g - R\|L^\top g\|_2$ a valid backup for every direction. Even a boundary probability estimated as zero can prevent an untruncated positive-radius ellipsoid. Feasibility of a *particular* optimizer suffices only for that backup.

For fixed π, let Q attain its rectangular value for the exact shared likelihood set. Repair into that *same set* by

$$\min_{\theta\in\Theta_b^{\mathrm{LR}}} \frac{1}{2}\sum_i d_Q^\pi(i)\|A_i\theta - Q_i\|_1. \tag{20}$$

Every feasible solution of (20) gives an ordinary-value upper endpoint for the coupled policy value and a gap bound equal to its objective times $\gamma/(1-\gamma)^2$. The residual lower program transfers as well. If an outer ellipsoid is used, Appendix I.1 separately bounds its approximation error; a repair only into $\mathcal{C}$ does not upper-bound the exact likelihood value. The valid-kernel comparator and sharing bounds (Proposition 20) require no norm shortcut. The RAMBO comparison is restricted to a *frozen linear surrogate*, not nonlinear return or convergence of adversarial fitting (Proposition 22).

In the shared binary specialization, observed successors identify the common Bernoulli outcome. Its likelihood is one interval, computed before planning; a common minimizing endpoint gives exactness (Proposition 21). The forest experiment tests this mechanism. General mixtures require the observed-successor likelihood in (18); latent mode labels are not interchangeable with observations. These transition-model assumptions are stronger than value-function realizability (Foster et al., 2022; Amortila et al., 2026).

6 EMPIRICAL EVALUATION

The experiments test *computation, approximation fidelity, and shared-parameter compatibility separately*. They use one fixed objective per comparison, rather than attributing a change of confidence

Model	Counts	Cold setup + solve (s) Scalar	Conic	Warm solve (s) Scalar	Conic	Paired warm ratio Conic/scalar [95% CI]
Lake4	B	0.322	2.345	0.321	2.207	6.80 [6.78, 7.10]
Lake4	I	0.274	2.752	0.274	2.622	9.51 [9.34, 9.83]
Forest10	B	0.079	0.853	0.078	0.818	10.49 [10.40, 10.54]
Forest10	I	0.079	0.824	0.078	0.804	10.25 [10.00, 10.48]
Lake8	B	0.657	10.182	0.663	9.647	14.53 [14.44, 14.61]
Lake8	I	0.672	10.676	0.674	10.055	14.97 [14.91, 15.05]
Forest30	B	0.235	2.503	0.236	2.425	10.28 [10.22, 10.40]
Forest30	I	0.204	1.734	0.203	1.672	8.13 [7.94, 8.43]

Table 1: **Faster planning at a fixed objective and accuracy.** Mean count is 250 per unknown row; B/I denote balanced/imbalanced counts. Three warm repetitions are summarized within each dataset before paired ratios and bootstrap intervals over 20 datasets. Cold totals include construction and the first solve. Startup, independent verification, and approximation assessment are excluded. Ratios compare the specified scalar and cached-conic implementations; intervals describe dataset variation on the recorded machine.

set to an optimizer improvement. The sampling studies have known rewards, $\gamma = .9$, and nominal confidence level 0.95. Their returns, gaps, and planning-loss tolerances are normalized by $1-\gamma$; write $\bar{G}_\pi = (1-\gamma)G_\pi$. Appendix K gives complete conditions and dataset-level uncertainty calculations.

6.1 From likelihood geometry to faster planning

Proposition 1 turns the global likelihood constraint into explicit reverse-KL row problems. To isolate the benefit of this representation, we replace only the row solver: scalar reverse-KL and cached CVXPY/Clarabel implementations use the **same uncertainty set, Bellman loop, and accuracy target**. FrozenLake 4×4/8×8 and forest10/forest30 each contribute 20 paired datasets under each count design. All 1,280 cold/warm solves meet the normalized 10^{-3} policy-loss target under tighter verification; the largest checked control-loss upper bound is $2.79 \cdot 10^{-6}$.

The paired warm-time ratios are 6.80–14.97, with all dataset-bootstrap intervals above one (Table 1). The row geometry replaces repeated generic conic optimization without changing confidence or accuracy. This is faster *rectangular planning*, not a comparison with exact CPPO-LR optimization or subsequent approximation assessment.

6.2 Measuring and explaining the approximation cost

To measure the approximation, we use a recurrent model: X reaches a reward-one absorber with probability .65 and otherwise Y; Y reaches a reward-zero absorber with probability .35 and otherwise X. Both probabilities are unknown to the planner. Counts are (N, N) or $(N/5, 9N/5)$, for $N \in \{25, 250, 2500\}$, with 20 datasets per condition and three starts: X, Y, and their equal mixture.

Independent allocation of the *same global budget* encloses each coupled value to width at most $1.01 \cdot 10^{-7}$; the reference never enters the practical bounds.

The two-sided analysis locates the coupled value. Figure 2 compares the residual lower bound and feasible repaired upper values with the independently enclosed gap. At $N = 2500$, the mean upper slack is $2.43 \cdot 10^{-5}$ under balance and $6.61 \cdot 10^{-6}$ under imbalance. These are same-policy gaps, not true-return errors. We use the smaller ordinary value of the occupancy-weighted and common-α feasible repairs. This is valid postprocessing without truth or reference access, not pointwise dominance of the TV surrogate; Table 4 retains both constituents and all sample levels.

Visitation explains why the lower bound is informative. At a pure start, allocating the full budget to the only initially visited row makes the time-zero-envelope bound zero. Accounting for unavoidable later visits raises mean lower bounds by $.00579$–$.01911$ across the six conditions (Figure 3), without changing data, likelihood region, or numerical target. The added occupancy information exposes incompatible adverse choices along recurrent trajectories.

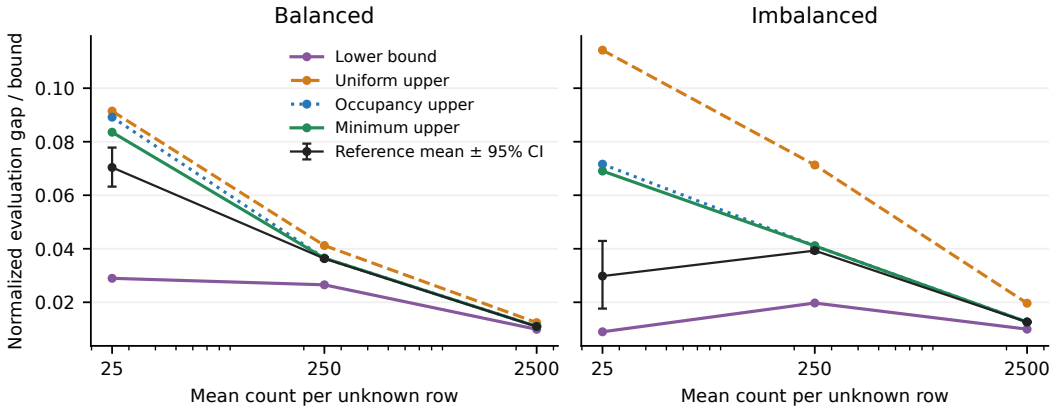


Figure 2: **Quantifying the price of one fixed likelihood relaxation.** Each point averages three initializations within each dataset and then all 20 datasets in the condition. Reference bars are dataset-bootstrap 95% intervals, not numerical-error bars. Coupled-value reference widths are at most $1.01 \cdot 10^{-7}$; gap enclosures also include rectangular-value uncertainty. The minimum upper is selected between two feasible repaired values without access to the reference.

Mixed starts reveal the same compatibility requirement across initializations. Proposition 9's premium has a positive lower endpoint in 119/120 datasets, with one sign unresolved. This *post hoc* reaggregation tests one common adversary across starts, not fresh confirmation or population inference.

6.3 When shared parameters preserve exactness

The forest model shares one unknown Bernoulli fire parameter. Its likelihood interval is computed once; row backups then compare affine endpoint values (Proposition 21). Across all 240 datasets—both state counts, both designs, and three sample levels—a feasible common endpoint gives a fixed-policy sharing-gap upper bound below $4.91 \cdot 10^{-12}$ (Table 5). This checks **one compatible adverse parameter** in an exact valid likelihood region, not return improvement across model classes. Opposing effects (Appendix L.5) show why compatibility, not sharing alone, determines exactness.

Validity requires the whole model. Separate deterministic checks, including non-target-row constraints, known rows, and continuation transforms, agree with analytic backups to $1.69 \cdot 10^{-11}$ across 36 prescribed cases in raw row-objective units (Appendix K.5). These test the implemented domain, not likelihood coverage or general conic-solver efficiency.

7 Discussion and conclusions

The bridge keeps **the statistical target separate from its computational relaxation**. Offline RL gains an implementable rectangular planner with an explicit approximation term in its comparator guarantee. Robust RL gains a likelihood-specific account of conservatism: reachability determines whether independent minima can be realized by one model, and occupancy quantifies the cost when they cannot.

The experiments support computation, approximation analysis, and a shared exactness regime, not universal performance improvement. *Evaluation fidelity does not imply coupled policy optimality*; true-return pessimism requires correct specification and calibration for the collection law. Retaining selected couplings when the bounds reveal substantial loss is a useful direction.

The contribution is an **analyzed approximation to a fixed likelihood objective**. Appendix J.5 compares related work, representation assumptions, and alternative objectives; Appendix J.4 details proof dependencies. Tractability and fidelity can thus be studied separately while remaining tied to the original offline-RL problem.

AI USE STATEMENT

Generative AI was used to polish writing and draft sections, explore literature, provide experimental-design feedback, co-develop and debug code, and assist with mathematical and numerical analysis. The authors directed the research, conducted the empirical studies, and retain responsibility for the final arguments, evidence, citations, and submission.

REPRODUCIBILITY STATEMENT

Sections 2 and 4 state the model and sampling assumptions. Appendices A–I provide proofs and boundary cases; Appendices K and L describe the complete selected empirical cohorts, numerical checks, and post hoc analyses. The supplied source package includes exact counts, policies, feasible witnesses, frozen protocols, code, and data-to-display recipes. Recorded numerical enclosures are distinct from dataset-bootstrap intervals and from formal interval proofs. The anonymous repository link is `https://github.com/FortunaAversa/bridging/tree/main/code`.

ETHICS STATEMENT

This work studies likelihood-based robust planning using finite simulated models and no human-subject data. Its guarantees require the stated transition, reward, and sampling assumptions; they do not imply safety under unmodeled errors. The reported comparisons use complete condition-wise datasets and distinguish implementation performance, approximation bounds, and policy-improvement claims.

REFERENCES

Lukáš Adam and Václav Mácha. Projections onto the canonical simplex with additional linear inequalities. *Optimization Methods and Software*, 37(2):451–479, 2022. doi: 10.1080/10556788.2020.1797023. URL https://arxiv.org/abs/1905.03488v2. Published online in 2020; theorem numbering follows arXiv:1905.03488v2.

Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018. doi: 10.1080/23307706.2017.1397554. URL https://web.stanford.edu/~boyd/papers/cvxpy_rewriting.html.

Rohit Agrawal. Finite-sample concentration of the multinomial in relative entropy. *IEEE Transactions on Information Theory*, 66(10):6297–6302, 2020. doi: 10.1109/TIT.2020.2996134. URL https://arxiv.org/abs/1904.02291v4.

Philip Amortila, Dylan J. Foster, Nan Jiang, Ayush Sekhari, and Tengyang Xie. Harnessing density ratios for online reinforcement learning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2401.09681.

Philip Amortila, Audrey Huang, Akshay Krishnamurthy, and Nan Jiang. A unifying view of coverage in linear off-policy evaluation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2601.19030.

Mohak Bhardwaj, Tengyang Xie, Byron Boots, Nan Jiang, and Ching-An Cheng. Adversarial model for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0429ececfb199efc93182990169e73bb-Abstract-Conference.html.

Blair Bilodeau, Dylan J. Foster, and Daniel M. Roy. Minimax rates for conditional density estimation via empirical entropy. *The Annals of Statistics*, 51(2):762–790, 2023. doi: 10.1214/23-AOS2270. URL https://arxiv.org/abs/2109.10461v4.

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

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1042–1051, 2019. URL https://proceedings.mlr.press/v97/chen19e.html.

Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained actor critic for offline reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3852–3878, 2022. URL https://proceedings.mlr.press/v162/cheng22b.html.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, second edition, 2006. ISBN 9780471241959.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016. URL https://jmlr.org/papers/v17/15-408.html.

Dylan J. Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. In *Proceedings of the 35th Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 3489, 2022. URL https://proceedings.mlr.press/v178/foster22a.html. Extended abstract; full paper arXiv:2111.10919v2.

Paul J. Goulart and Yuwen Chen. Clarabel: An interior-point solver for conic programs with quadratic objectives. *arXiv preprint arXiv:2405.12762*, 2024. URL https://arxiv.org/abs/2405.12762.

Julien Grand-Clément, Nian Si, and Shengbo Wang. Dynamic programming-compatible uncertainty sets in robust Markov decision processes. *arXiv preprint arXiv:2411.08435*, 2026. URL https://arxiv.org/abs/2411.08435v2. Version 2; first circulated in 2024.

Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Robust ϕ-divergence MDPs. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/d2fe3a5711a6d488da9e9a78b84ee24c-Abstract-Conference.html.

Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Efficient algorithms for robust Markov decision processes with s-rectangular ambiguity sets. *arXiv preprint arXiv:2602.05591*, 2026. URL https://arxiv.org/abs/2602.05591.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321. URL https://arxiv.org/abs/1808.03204v7.

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

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *Proceedings of the 38th 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.

R. E. Krichevsky and V. K. Trofimov. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207, 1981. doi: 10.1109/TIT.1981.1056331.

Mengmeng Li, Daniel Kuhn, and Tobias Sutter. Policy gradient algorithms for robust MDPs with nonrectangular uncertainty sets. *SIAM Journal on Optimization*, 36(1):120–151, 2026. doi: 10.1137/24M1631250. URL https://arxiv.org/abs/2305.19004v4.

Michael Lindon and Alan Malek. Anytime-valid inference for multinomial count data. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/12f3bd5d2b7d93eadc1bf508a0872dc2-Abstract-Conference.html.

Zhishuai Liu and Pan Xu. Linear mixture distributionally robust Markov decision processes. *arXiv preprint arXiv:2505.18044*, 2025. URL https://arxiv.org/abs/2505.18044.

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.

Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Bridging distributionally robust learning and offline RL: An approach to mitigate distribution shift and partial data coverage. In *Proceedings of the 7th Annual Learning for Dynamics & Control Conference*, volume 283 of *Proceedings of Machine Learning Research*, pp. 619–634, 2025. URL https://proceedings.mlr.press/v283/panaganti25a.html.

Python MDP Toolbox contributors. Python Markov decision process toolbox: Forest example, n.d. URL https://pymdptoolbox.readthedocs.io/en/latest/api/example.html. Documentation; accessed September 23, 2026.

Marc Rigter, Bruno Lacerda, and Nick Hawes. RAMBO-RL: Robust adversarial model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6691c5e4a199b72dffd9c90acb63bcd6-Abstract-Conference.html.

Anirudh Satheesh, Ziyi Chen, Furong Huang, and Heng Huang. Provably efficient algorithms for s- and non-rectangular robust MDPs with general parameterization. *arXiv preprint arXiv:2602.11387*, 2026. URL https://arxiv.org/abs/2602.11387.

Laixi Shi and Yuejie Chi. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *Journal of Machine Learning Research*, 25(200):1–91, 2024. URL https://www.jmlr.org/papers/v25/22-1482.html.

Cheng Tang, Zhishuai Liu, and Pan Xu. Robust offline reinforcement learning with linearly structured f-divergence regularization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 58842–58882, 2025. URL https://proceedings.mlr.press/v267/tang25f.html.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standardized interface for reinforcement learning environments. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/file/d7ff1795e8527f6443371c3933bdb52b-Paper-Datasets_and_Benchmarks_Track.pdf. Datasets and Benchmarks Track.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2023. URL https://arxiv.org/abs/2107.06226v4. Version 4; extended version of the ICLR 2022 paper, including CPPO-LR.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013. doi: 10.1287/moor.1120.0566.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11404–11413, 2021. URL https://proceedings.mlr.press/v139/xie21d.html.

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://proceedings.neurips.cc/paper_files/paper/2021/hash/34f98c7c5d7063181da890ea8d25265a-Abstract.html.

Tengyang Xie, Dylan J. Foster, Yu Bai, Nan Jiang, and Sham M. Kakade. The role of coverage in online reinforcement learning. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.04157.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Proceedings of the 35th Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 2730–2775, 2022. URL https://proceedings.mlr.press/v178/zhan22a.html.

A LIKELIHOOD GEOMETRY AND ROBUST DYNAMIC PROGRAMMING

This appendix fixes the product model class, successor sets, known rows, rewards, and stationary policy class of Section 2. *No sampling assumption is needed* until Appendix G. Our dynamic-programming arguments instantiate classical rectangular robust MDP theory (Iyengar, 2005; Nilim & El Ghaoui, 2005; Wiesemann et al., 2013); the proofs are included to specify the exact uncertainty set and policy class.

A.1 LIKELIHOOD IDENTITY, COMPACTNESS, AND PROJECTIONS

For a covered row, the saturated multinomial log likelihood is maximized by $\widehat{p}_i$. Thus

$$\ell(\widehat{P}) - \ell(P) = \sum_{i \in \mathcal{I}_+} \sum_{x: N_{ix} > 0} N_{ix} \log \frac{N_{ix}/N_i}{P_i(x)} = \sum_{i \in \mathcal{I}_+} N_i D_{\mathrm{KL}}(\widehat{p}_i \| P_i).$$

This remains an extended-real identity when a positively observed outcome receives probability zero. Uncovered rows have no likelihood contribution. Known rows have the same likelihood under every feasible kernel, so their contribution cancels. All simplex constraints use the predeclared X_i; unobserved points of X_i are retained.

Reverse KL is convex and lower semicontinuous in its second argument (Cover & Thomas, 2006). Hence $\mathcal{U}_b$ and $\mathcal{R}_b$ are nonempty compact sets: the empirical covered rows have cost zero, and the ambient product of finite simplexes and known singleton rows is compact. For every stochastic P_π,

$$(I - \gamma P_\pi)^{-1} = \sum_{t \geq 0} \gamma^t P_\pi^t, \qquad \|(I - \gamma P_\pi)^{-1}\|_{\infty \to \infty} = \frac{1}{1-\gamma}.$$

Consequently V_P^π and J_P^π are continuous jointly in P and π. Minima over either uncertainty set are attained, and the minimum value is continuous in π by compactness. Since Π is a compact product of action simplexes, the coupled stationary-policy maximum is attained as well. This argument does not assert that a coupled optimum is deterministic.

Proof of Proposition 1. For a coupled-feasible P, choose $u_i = N_i D_{\mathrm{KL}}(\widehat{p}_i \| P_i)$. Then $u \geq 0$, $\sum_i u_i \leq b$, and $P_i \in \mathcal{B}_i(u_i)$. Conversely, membership in a product with budgets summing to at most b implies $K(P) \leq b$. The common factor $\mathcal{Z}$ contains the uncovered simplexes and known singleton rows on both sides. This proves the union identity.

If $P \in \mathcal{U}_b$, each nonnegative likelihood contribution is at most b, so its ith covered row belongs to $\mathcal{B}_i(b)$. For the reverse projection inclusion, fix any $p \in \mathcal{B}_i(b)$, put $P_i = p$, and set every other covered row to its empirical distribution. Choose arbitrary feasible uncovered rows and the prescribed known rows. The resulting kernel lies in $\mathcal{U}_b$. Every uncovered simplex and every known singleton is also its own projection. Taking their product proves the hull formula.

Let $\mathcal{V} = \prod_i \mathcal{V}_i$ be any (s,a)-rectangular superset of $\mathcal{U}_b$. Each $\mathcal{V}_i$ must contain the corresponding projection of $\mathcal{U}_b$. Therefore $\mathcal{R}_b \subseteq \mathcal{V}$. This proves unique minimality by set inclusion, and $\inf_{P \in \mathcal{V}} J_P^\pi \leq \min_{P \in \mathcal{R}_b} J_P^\pi$ for every π, reward function, and initial distribution. Minimality is relative to *supersets of this fixed* $\mathcal{U}_b$, not to all statistically valid confidence sets. □

A.2 EVALUATION AND GREEDY OPTIMALITY

For every row set $\mathcal{R}_i = \mathrm{proj}_i \mathcal{U}_b$ and vectors v, w,

$$|\phi_i(v) - \phi_i(w)| \leq \sup_{p \in \mathcal{R}_i} |p^\top (v - w)| \leq \|v - w\|_\infty.$$

Also $v \leq w$ implies $\phi_i(v) \leq \phi_i(w)$, and $\phi_i(v + c\mathbf{1}) = \phi_i(v) + c$. Convexity of the row sets is not needed for these statements; validity as probability distributions is essential. Taking action averages or maxima proves monotonicity and γ-contraction of both operators in (3). Banach's fixed-point theorem gives unique fixed points.

Write $T_P^\pi v = r_\pi + \gamma P_\pi v$ for ordinary policy evaluation. For a fixed π, select a row minimizer Q_i at $v = V_{\mathrm{R}}^\pi$. Because the uncertainty set is a product, these minimizers form a feasible kernel $Q \in \mathcal{R}_b$.

Then $T_Q^\pi v = T_{\mathrm{R}}^\pi v = v$, so $v = V_Q^\pi$. For any $P \in \mathcal{R}_b$, $T_P^\pi v \geq v$. Monotone iteration of the ordinary policy operator yields $V_P^\pi \geq v$. Thus one stationary kernel Q attains the componentwise rectangular worst value, and $J_{\mathrm{R}}^\pi = \mu^\top v$. This common minimizer is why the scalar initial-distribution objective agrees with the componentwise rectangular fixed point.

Let $v^\star$ be the fixed point of T_{R}, and select at each state an action attaining its finite maximum. The resulting deterministic π_{R} satisfies $T_{\mathrm{R}}^{\pi_{\mathrm{R}}} v^\star = v^\star$, hence $V_{\mathrm{R}}^{\pi_{\mathrm{R}}} = v^\star$. For any stationary randomized π, $T_{\mathrm{R}}^\pi v^\star \leq T_{\mathrm{R}} v^\star = v^\star$. Iterating the monotone contraction T_{R}^π gives $V_{\mathrm{R}}^\pi \leq v^\star$. This proves greedy optimality over the full stated policy class. The fixed-point identity establishes the greedy policy's value; monotone iteration establishes its optimality.

A.3 Scope of the time-consistency statement

In this paper, row-wise stagewise planning means that nature may minimize each current state–action expectation over its row projection independently of the other choices. This is exactly (3), so its uncertainty set is exactly $\mathcal{R}_b$. A coupled likelihood budget does not generally permit all of these minimizers simultaneously. An alternative planner could preserve state-level coupling, augment information about a budget, or perform global non-rectangular optimization. Those are different solution concepts or state descriptions. State-rectangular robust planning and simultaneous solvability are well-developed alternatives (Wiesemann et al., 2013; Grand-Clément et al., 2026). No necessity claim about all possible dynamic programs follows from Proposition 1.

B Simulation identities and compatibility

B.1 A simulation lemma with an explicit reference occupancy

The resolvent argument is the standard simulation-lemma proof used in model-based offline RL; see, for example, Uehara & Sun (2023, Lemma 9). We state both orientations because *fixing the rectangular rather than repaired occupancy* is what makes the subsequent repair objective convex. The identities themselves are not new.

Lemma 8 (Simulation identities)**.** *For stochastic kernels P, Q and a stationary π,*

$$J_P^\pi - J_Q^\pi = \frac{\gamma}{1-\gamma} \sum_i d_P^\pi(i)(P_i - Q_i)^\top V_Q^\pi, \tag{21}$$

$$= \frac{\gamma}{1-\gamma} \sum_i d_Q^\pi(i)(P_i - Q_i)^\top V_P^\pi. \tag{22}$$

In particular, either reference occupancy yields an absolute bound by $\gamma(1-\gamma)^{-2}$ times its occupancy-weighted row TV distance.

Proof. Subtract the Bellman equations in the two ways

$$V_P^\pi - V_Q^\pi = \gamma P_\pi (V_P^\pi - V_Q^\pi) + \gamma (P_\pi - Q_\pi) V_Q^\pi$$

and

$$V_P^\pi - V_Q^\pi = \gamma Q_\pi (V_P^\pi - V_Q^\pi) + \gamma (P_\pi - Q_\pi) V_P^\pi.$$

Multiply by the appropriate resolvent and by $\mu^\top$. The identity $(1-\gamma)[\mu^\top (I - \gamma P_\pi)^{-1}](s)\pi(a \mid s) = d_P^\pi(s,a)$ then gives (21)–(22). Since $r \in [0,1]$, every value vector has span at most $1/(1-\gamma)$. The variational characterization of TV gives $|(p-q)^\top v| \leq \mathrm{span}(v)\, \mathrm{TV}(p,q)$, proving the bounds. □

B.2 Exact compatibility and the role of initialization

Proof of Theorem 2. Let $v = V_{\mathrm{R}}^\pi$, and for each state define $a_\pi(P; v)(s) = \sum_a \pi(a \mid s)[P_{sa}^\top v - \phi_{sa}(v)]$. Because $P \in \mathcal{U}_b \subseteq \mathcal{R}_b$, this vector is nonnegative. Subtracting $v = T_{\mathrm{R}}^\pi v$ from the Bellman equation for P gives

$$V_P^\pi - v = \gamma P_\pi (V_P^\pi - v) + \gamma a_\pi(P; v).$$

The resolvent and the occupancy identity yield

$$J_P^\pi - J_{\mathrm{R}}^\pi = \frac{\gamma}{1-\gamma} \sum_i d_P^\pi(i) a_i(P; v).$$

Minimizing over the compact set $\mathcal{U}_b$ proves (4). Each summand is nonnegative, so the attained minimum is zero if and only if every positively occupied row has zero residual under some feasible P.

For a covered row, the face $F_i(v) = \{p \in \mathcal{B}_i(b) : p^\top v = \phi_i(v)\}$ is nonempty and compact. Its likelihood-cost minimum is attained because KL is lower semicontinuous and finite on $\mathcal{B}_i(b)$. Suppose $\kappa_\pi(v) \le b$. On covered rows whose actions have positive π-probability, choose minimum-cost members of $F_i(v)$; on covered rows with zero action probability choose $\widehat{p}_i$. On uncovered positive-action rows choose any row minimizer, and elsewhere choose any feasible row. This assembled kernel belongs to $\mathcal{U}_b$ and makes all policy-relevant residuals vanish. Hence $G_\pi = 0$ for every μ.

Conversely, suppose $\mu(s) > 0$ for every s and $G_\pi = 0$. A minimizing P in (4) has zero residual on every row with positive occupancy. For every $\pi(a \mid s) > 0$,

$$d_P^\pi(s, a) \ge (1 - \gamma)\mu(s)\pi(a \mid s) > 0.$$

Its row therefore belongs to $F_i(v)$, and $\kappa_\pi(v) \le \sum_{i \in \mathcal{I}_+} N_i D_{\mathrm{KL}}(\widehat{p}_i \| P_i) \le b$. Without full-support initialization, this argument does not force minimization on unreachable states; the converse need not hold.

To prove (5), fix $b > 0$ and a covered row. If $\widehat{p}_i^\top v = \min_{X_i} v$, the empirical row minimizes the objective with zero cost. Otherwise every finite-KL row assigns positive mass to an observed successor whose value exceeds $\min_{X_i} v$, so its objective is strictly greater than this minimum. If a row minimizer p had $N_i D_{\mathrm{KL}}(\widehat{p}_i \| p) < b$, mix it by a sufficiently small positive amount with a point mass on a minimum-value successor. The objective strictly decreases, while continuity of KL at p on the empirical support keeps the likelihood cost below b. This contradicts optimality. Every minimizer therefore has cost exactly b. Summing over covered rows with positive action probability proves the cardinality criterion. At $b = 0$, each covered row must equal its empirical distribution; the known and uncovered factors remain independent, hence $\mathcal{U}_b = \mathcal{R}_b$. □

Relation to simultaneous solvability. Simultaneous solvability relates compatible row optimizers to Bellman representations (Grand-Clément et al., 2026). Here (4) is for one reward function, policy, and initial distribution. The likelihood-cost test gives a concrete characterization under full-support initialization, and its reverse-KL specialization reduces to the explicit budget-saturation property. It is not a characterization of universal tractability across all rewards or of optimality among policies. Its exact form is also not automatically a numerically robust zero test: an approximate face $p^\top v \le \phi_i(v) + \eta$ defines a different, approximate compatibility problem. Residual-corrected bounds are preferable when exact ties cannot be resolved numerically.

B.3 A REACHABILITY-COMPLETE EXACTNESS TEST

The full-support assumption in Theorem 2 makes every policy-positive row relevant. For an arbitrary initial distribution, relevance depends on the transitions chosen by the adversary. The following construction resolves that dependence without enumerating all subsets of states. It uses the likelihood-specific saturation identity (5); the general simultaneous-solvability perspective is due to Grand-Clément et al. (2026). Our statement concerns one fixed reward, policy, budget, and initial distribution, rather than tractability for all rewards or coupled policy optimization.

Fix $b > 0$, a stationary policy π, and $v = V_{\mathrm{R}}^\pi$. Write

$$\mathcal{A}_\pi(s) = \{a : \pi(a \mid s) > 0\}, \qquad \mathcal{C}_\pi(v) = \{i = (s, a) \in \mathcal{I}_+ : a \in \mathcal{A}_\pi(s),\ \widehat{p}_i^\top v > \min_{x \in X_i} v_x\}.$$

Call these covered rows *critical*: every row minimizer spends the entire likelihood budget b. For a row i, let $F_i(v) = \arg\min_{p \in \mathcal{R}_i} p^\top v$. For each candidate $j \in \mathcal{C}_\pi(v) \cup \{\perp\}$, where $\perp$ means that no critical row is allowed to be reached, define the following row families only for $a \in \mathcal{A}_\pi(s)$:

$$\mathcal{F}_i^j = \begin{cases} \{P_i^0\}, & i \in \mathcal{I}_K, \\ \{\widehat{p}_i\}, & i \in \mathcal{I}_+ \setminus \mathcal{C}_\pi(v), \\ F_i(v), & i = j \in \mathcal{C}_\pi(v), \\ \varnothing, & i \in \mathcal{C}_\pi(v) \setminus \{j\}, \\ \Delta(\arg\min_{x \in X_i} v_x), & i \in \mathcal{I}_0. \end{cases} \tag{23}$$

Rows of zero-probability actions impose no reachability constraint. Starting with $Z_j^{(0)} = \mathcal{S}$, delete states according to

$$Z_j^{(k+1)} = \left\{ s \in Z_j^{(k)} : \text{for every } a \in \mathcal{A}_\pi(s),\ \exists p \in \mathcal{F}_{sa}^j \text{ with } \operatorname{supp} p \subseteq Z_j^{(k)} \right\}. \tag{24}$$

The decreasing sequence stabilizes after at most $|\mathcal{S}|$ strict deletions; denote its final set by Z_j^∞.

The greatest-set construction above supplies the sets used in Theorem 3. The following proof establishes both directions of (6).

Proof of Theorem 3. First, (24) computes the greatest set closed under the admissible minimizing-row choices in (23). To see this directly, suppose $W \subseteq \mathcal{S}$ has an admissible row supported on W for every policy-positive action at every state in W. Then $W \subseteq Z_j^{(0)}$. If $W \subseteq Z_j^{(k)}$, the same rows show $W \subseteq Z_j^{(k+1)}$. Induction gives $W \subseteq Z_j^\infty$. At stabilization, each state of Z_j^∞ itself has all required row choices supported on Z_j^∞.

For necessity, suppose $G_\pi = 0$. Compactness supplies a kernel $P^\circ \in \mathcal{U}_b$ attaining the minimum in (4). Let W be the set of states reachable with positive probability from $\operatorname{supp} \mu$ under (P°, π). This set contains $\operatorname{supp} \mu$ and is closed: if $s \in W$, $\pi(a \mid s) > 0$, and $P_{sa}^\circ(x) > 0$, then $x \in W$. Every such state–action row has positive discounted occupancy and therefore zero residual by (4).

A critical row reached in W belongs to its minimizing face $F_i(v)$, so its likelihood contribution equals b by (5). Since $b > 0$ and the sum of all nonnegative likelihood contributions is at most b, at most one critical row is reached. Choose this row as j, or choose $j = \perp$ if none is reached. Known rows already belong to $\mathcal{F}_i^j$. An uncovered minimizing row is supported on $\arg\min_{X_i} v$, and its existing support is contained in W.

For a noncritical covered row reached in W, replace its row by $\widehat{p}_i$. This row minimizes v with zero likelihood cost. Furthermore, finite reverse KL implies $\operatorname{supp} \widehat{p}_i \subseteq \operatorname{supp} P_i^\circ \subseteq W$: a positive empirical probability cannot receive zero model probability. Thus this replacement does not introduce a new reachable state. The existing row of j, if present, lies in $F_j(v)$ and is supported on W. Consequently W is a closed set for (23). The preceding induction gives $\operatorname{supp} \mu \subseteq W \subseteq Z_j^\infty$.

For sufficiency, choose a successful j and, for every $s \in Z_j^\infty$ and $a \in \mathcal{A}_\pi(s)$, choose one row in $\mathcal{F}_{sa}^j$ supported on Z_j^∞. Set all remaining covered rows to their empirical distributions, keep known rows fixed, and choose any admissible remaining uncovered rows. The result is a stochastic kernel P. Every covered row has likelihood cost zero except possibly j, whose cost is at most b. Hence $P \in \mathcal{U}_b$.

Initialization lies in Z_j^∞, and the selected rows keep the process there under π. On every positively occupied row, the chosen transition minimizes v: this is immediate for the known and uncovered cases, follows from noncriticality for empirical rows, and from membership in $F_j(v)$ for the special row. Equation (4) now gives $J_P^\pi = J_{\mathrm{R}}^\pi$ and $G_\pi = 0$. Starting instead from any state of Z_j^∞ gives the same argument with the same kernel. Finally, $b = 0$ forces each covered row to be empirical, while the known and uncovered factors are identical in $\mathcal{U}_b$ and $\mathcal{R}_b$. □

Implementing the support tests. Known and noncritical covered rows pass exactly when their prescribed supports lie in the current Z. An uncovered row passes exactly when $Z \cap \arg\min_{X_i} v$ is nonempty. Other critical rows fail. For $i = j$, the test is equivalent to

$$\operatorname{supp} \widehat{p}_j \subseteq Z, \qquad \min_{\substack{p \in \Delta(X_j \cap Z) \\ N_j D_{\mathrm{KL}}(\widehat{p}_j \| p) \le b}} p^\top v = \phi_j(v). \tag{25}$$

The restricted problem is feasible under the support condition, because it contains $\widehat{p}_j$. Its minimum is no smaller than $\phi_j(v)$; equality gives a row in $F_j(v)$ supported on Z, and the converse is immediate. This restriction is a support-feasibility test inside the proof procedure, not removal of unseen successors from the original statistical model. The same reverse-KL row oracle solves it.

With exact row/support oracles, there are $|\mathcal{C}_\pi(v)| + 1$ candidates, at most $|\mathcal{S}| + 1$ deletion passes per candidate, and only one potentially nonlinear row test per pass. All other tests are finite support inclusions or intersections. This is a *finite-oracle reduction*, not a polynomial bit-complexity claim: it requires the exact rectangular value, exact criticality classification, and exact equality in (25). Floating-point agreement within a tolerance does not establish exactness. Numerical implementations must either provide sufficient witnesses and error bounds or report an unresolved equality.

Two consequences and an important quantifier. For full-support μ, a successful set must equal $\mathcal{S}$. Two critical selected rows prevent this for every candidate; zero or one allows it. Thus (6) recovers Theorem 2. In the unreachable-branch construction, the candidate for the sole reachable critical row succeeds although other critical rows exist outside its invariant set. For the two-row recurrent construction with a positive observed success from X and a positive observed return from Y to X, a minimizing X row assigns positive probability to Y, whereas finite likelihood forces a positive return probability from Y to X. A zero-gap kernel would therefore have to reach both critical rows from either transient state. Every candidate fails, so the gap is strictly positive by compactness. The empirical lower bounds quantify a gap whose positivity has this structural explanation.

Finally, $\operatorname{supp}\mu \subseteq \bigcup_j Z_j^\infty$ is not sufficient. Consider initialization on two disconnected uncertain branches that terminate in known sinks, each with an observed successful transition and a nontrivial row minimum. Each pure initialization allows its own row to spend the full budget, but their mixture cannot use both allocations in one kernel. Requiring *one common* j in (6) preserves the scalar coupled objective's quantifier.

B.4 Initialization mixtures and coupled-model incompatibility

The gap also has a useful concavity property. This is a direct consequence of minimization of affine functions, not a new general concavity principle. It clarifies why the empirical mixed-start reference cannot be obtained by averaging pure-start minima.

Proposition 9 (The initialization-mixture premium)**.** *Fix π and a nonempty compact coupled set $\mathcal{U}$ of stochastic kernels, with rectangular hull $\mathcal{R}$. Let $V_{\mathcal{R}}^\pi$ be the rectangular policy-value vector and write $G(\mu) = \min_{P\in\mathcal{U}} \mu^\top V_P^\pi - \mu^\top V_{\mathcal{R}}^\pi$. For initial distributions $\mu_1, \dots, \mu_k$ and weights $\lambda \in \Delta(\{1,\dots,k\})$, put $\mu_\lambda = \sum_j \lambda_j \mu_j$. Then*

$$G(\mu_\lambda) - \sum_j \lambda_j G(\mu_j) = \min_{P\in\mathcal{U}} \sum_j \lambda_j \left(\mu_j^\top V_P^\pi - \min_{Q\in\mathcal{U}} \mu_j^\top V_Q^\pi \right) \geq 0. \tag{26}$$

Equality holds if and only if one $P \in \mathcal{U}$ minimizes the policy value for every j with $\lambda_j > 0$.

Proof. The rectangular value is linear in initialization because one stationary rectangular minimizer attains its value vector. Subtracting those linear terms gives (26). Each term inside the attained minimum is nonnegative. Their positive-weight sum vanishes exactly when each term does, proving the equality condition. Compactness and value continuity guarantee attainment. Equivalently, G is concave as the minimum of affine functions minus a linear function; see Boyd & Vandenberghe (2004, Section 3.2.3) for the underlying convex-analysis rule. □

C Occupancy-weighted likelihood repair

C.1 Convexity and the value bounds

Proof of Theorem 4: feasibility and value bound. Since $Q \in \mathcal{R}_b$ and $b < \infty$, $K(Q) < \infty$ and each covered Q_i is positive on the support of $\widehat{p}_i$. The mixture path therefore has finite continuous likelihood throughout $[0,1]$. Convexity of KL in its second argument makes each f_i convex, and $f_i(1) = 0$. The point $\alpha = \mathbf{1}$ is feasible, so the compact convex program has a minimizer.

Every feasible mixture $P(\alpha)$ is a stochastic kernel in $\mathcal{U}_b$. Therefore $J_Q^\pi = J_{\mathrm{R}}^\pi \leq J_{\mathrm{LR}}^\pi \leq J_{P(\alpha)}^\pi$. By (22),

$$J_{P(\alpha)}^\pi - J_Q^\pi \leq \frac{\gamma}{(1-\gamma)^2} \sum_i d_Q^\pi(i)\, \mathrm{TV}(P_i(\alpha), Q_i).$$

The TV distance on a covered row is exactly $\alpha_i\, \mathrm{TV}(\widehat{p}_i, Q_i)$, and all other rows are unchanged. This proves (8). Notice that d_Q^π is fixed when α is optimized; replacing it by $d_{P(\alpha)}^\pi$ would destroy the stated convex formulation. □

For $b > 0$ and $\tau = K(Q)/b > 1$, the following consequence of Theorem 4 gives a generic upper bound:

$$G_\pi \le \frac{\gamma}{(1-\gamma)^2} B_\pi(Q) \le \frac{\gamma}{(1-\gamma)^2} \sqrt{\frac{b}{2n}} \frac{\tau - 1}{\sqrt{\tau}} \left(\sum_{i \in \mathcal{I}_+} \frac{d_Q^\pi(i)^2}{\widehat{\rho}_i} \right)^{1/2}. \tag{27}$$

Uniform repair and (27). If $K(Q) \le b$, Q is coupled-feasible and the interval collapses. Otherwise $b > 0$ and $\tau = K(Q)/b > 1$. Convexity gives

$$f_i(1 - 1/\tau) \le \tau^{-1} N_i D_{\mathrm{KL}}(\widehat{p}_i \| Q_i), \qquad \sum_i f_i(1 - 1/\tau) \le b.$$

Hence $B_\pi(Q) \le (1 - 1/\tau) \sum_{i \in \mathcal{I}_+} d_Q^\pi(i) \, \mathrm{TV}(Q_i, \widehat{p}_i)$. Pinsker's inequality, with natural-logarithm constant $\mathrm{TV}(p, q) \le \sqrt{D_{\mathrm{KL}}(q\|p)/2}$ (Cover & Thomas, 2006), and weighted Cauchy–Schwarz imply

$$\begin{aligned} \sum_{i \in \mathcal{I}_+} d_Q^\pi(i) \, \mathrm{TV}(Q_i, \widehat{p}_i) &\le \sum_{i \in \mathcal{I}_+} \frac{d_Q^\pi(i)}{\sqrt{\widehat{\rho}_i}} \sqrt{\frac{\widehat{\rho}_i D_{\mathrm{KL}}(\widehat{p}_i \| Q_i)}{2}} \\ &\le \left(\sum_{i \in \mathcal{I}_+} \frac{d_Q^\pi(i)^2}{\widehat{\rho}_i} \right)^{1/2} \sqrt{\frac{K(Q)}{2n}}. \end{aligned}$$

Multiplying by $1 - 1/\tau$ proves (27). This comparison uses the same Q-occupancy on both sides. Bounds computed with different reference occupancies need not be ordered. All gap upper bounds may also be capped by $1/(1-\gamma)$. □

C.2 A SCALAR-MULTIPLIER ALGORITHM AND PRIMAL–DUAL BOUNDS

In this subsection, all row sums are over $\mathcal{I}_+$. For $b > 0$, taking all α_i sufficiently close to one gives strict likelihood feasibility in the relative interior of the box. Slater's condition therefore holds (Boyd & Vandenberghe, 2004, Section 5.2.3). Define

$$h_i(\lambda) = \min_{0 \le a \le 1} \{ w_i a + \lambda f_i(a) \}, \qquad D(\lambda) = \sum_i h_i(\lambda) - \lambda b, \qquad \lambda \ge 0.$$

Convex duality gives $B_\pi(Q) = \max_{\lambda \ge 0} D(\lambda)$. The Karush–Kuhn–Tucker (KKT) conditions (Boyd & Vandenberghe, 2004, Section 5.5.3) give, at an interior minimizer,

$$w_i = -\lambda f_i'(\alpha_i), \qquad f_i'(\alpha) = -N_i \sum_{x : \widehat{p}_i(x) > 0} \frac{\widehat{p}_i(x)(\widehat{p}_i(x) - Q_i(x))}{(1-\alpha) Q_i(x) + \alpha \widehat{p}_i(x)}. \tag{28}$$

The second derivative is

$$f_i''(a) = N_i \sum_{x : \widehat{p}_i(x) > 0} \frac{\widehat{p}_i(x)(\widehat{p}_i(x) - Q_i(x))^2}{[(1-a) Q_i(x) + a \widehat{p}_i(x)]^2} \ge 0.$$

The function f_i is nonincreasing: convexity and the endpoint minimum $f_i(1) = 0$ imply this, or it follows from the derivative formula. It is strictly convex unless $Q_i = \widehat{p}_i$. If $w_i = 0$, set $\alpha_i(\lambda) = 1$ without cost. For $w_i > 0$ and $\lambda > 0$, minimize using the signs of $w_i + \lambda f_i'(0)$ and $w_i + \lambda f_i'(1)$; if neither endpoint is optimal, bisect the unique interior derivative root. An endpoint optimum is allowed even when $f_i'(1) < 0$, which occurs when Q_i puts mass outside empirical support.

The total likelihood $\sum_i f_i(\alpha_i(\lambda))$ is nonincreasing in λ. To see this, write the optimality inequalities at $\lambda_1 < \lambda_2$, add them, and cancel the linear costs. As λ grows, the likelihood tends to its attainable minimum zero. Thus a doubling search followed by bisection finds a multiplier with feasible likelihood, except in the already solved zero-cost case. If repairing zero-weight rows alone meets the budget, the optimum is zero and no multiplier search is needed.

Any feasible $\bar{\alpha}$ gives the primal upper bound $U = \sum_i w_i \bar{\alpha}_i$ on $B_\pi(Q)$. An exact scalar minimization gives $D(\lambda) \le B_\pi(Q)$. There is also a rigorous lower-bound formula for approximate scalar minimizers: with $H_i(a) = w_i a + \lambda f_i(a)$ and any $a \in [0,1]$, convexity implies

$$h_i(\lambda) \ge H_i(a) + \min\{-aH_i'(a), (1-a)H_i'(a)\}.$$

The right side is the minimum of the supporting affine function over the box. Summing these bounds and subtracting λb gives a valid dual lower bound L. Thus $U - L$ establishes the accuracy of the *repair optimization*. Multiplying U by $\gamma/(1-\gamma)^2$ bounds the robust-value gap; $U - L$ does not directly measure the difference between the repaired value and the true coupled minimum.

Feasibility is the required direction. A solver status flag or a small likelihood violation is not enough to make $P(\alpha) \in \mathcal{U}_b$. Given a stochastic kernel $\widetilde{P}$ in the product model class with finite $K(\widetilde{P})$, define covered rows

$$P_i' = (1-\eta)\widetilde{P}_i + \eta \widehat{p}_i, \qquad \eta = \begin{cases} 0, & K(\widetilde{P}) \le b, \\ 1 - b/K(\widetilde{P}), & K(\widetilde{P}) > b, \end{cases}$$

and keep other rows unchanged. Convexity ensures $K(P') \le b$ in exact arithmetic. *Re-evaluate the final likelihood and return value* after this correction; do not keep the objective from the uncorrected kernel. In floating-point implementations a small inward margin is appropriate. A machine-arithmetic guarantee additionally requires directed rounding or interval bounds on logarithms and linear solves.

C.3 A STRONGER FULL-KERNEL PROJECTION BOUND

The row-mixture family was chosen for low-dimensional separability, not because it solves the best possible repair problem. The next observation extends the construction to any convex coupling set of valid kernels.

Proposition 10 (General feasible-kernel projection)**.** *Let $\mathcal{U}$ be a nonempty compact convex set of finite stochastic kernels and let $\mathcal{R} = \mathrm{Rect}_{sa}(\mathcal{U})$. Fix π and a rectangular worst kernel Q. Then*

$$B_{\mathrm{full}}(Q) = \min_{P \in \mathcal{U}} \sum_i d_Q^\pi(i)\, \mathrm{TV}(P_i, Q_i)$$

is a convex program, and every feasible P establishes

$$J_Q^\pi \le \min_{P' \in \mathcal{U}} J_{P'}^\pi \le J_P^\pi, \qquad J_P^\pi - J_Q^\pi \le \frac{\gamma}{(1-\gamma)^2} \sum_i d_Q^\pi(i)\, \mathrm{TV}(P_i, Q_i).$$

For the tabular likelihood set, $B_{\mathrm{full}}(Q) \le B_\pi(Q)$. For $\mathcal{U} = \{P_\theta : \theta \in \mathcal{C}\}$, where $\mathcal{C}$ is a compact convex parameter set with valid rows $P_{\theta,i} = A_i\theta$, it reduces to

$$\min_{\theta \in \mathcal{C}} \ \frac{1}{2} \sum_i d_Q^\pi(i) \|A_i\theta - Q_i\|_1. \tag{29}$$

Proof. The objective is a nonnegative weighted sum of convex norms of affine arguments, and the feasible set is compact and convex. Attainment and convexity follow. The value interval uses set inclusion and the TV bound is (22). The row mixtures form a subset of the coupled likelihood set, so their minimum cost cannot be smaller than the full projection cost. Substituting $P_i = A_i\theta$ gives (29), which admits a second-order-cone representation when $\mathcal{C}$ is the validity-constrained ellipsoid defined in Appendix I. □

This is a *convex surrogate* for finding a feasible adversary close to Q, not a convex reformulation of $\min_P J_P^\pi$. Li et al. (2026) measure non-rectangularity through a return-gradient optimization gap between a coupled set and its state-rectangular hull, anchored at a coupled-feasible model. Here Q is computed in the state–action rectangular hull and need not be coupled-feasible; minimizing its fixed-occupancy transition distance explicitly returns a feasible coupled kernel. Thus the reference model, hull, and bound being optimized differ. Zero distance is sufficient for exactness but is not necessary when different transitions attain the same continuation value. The face criterion of Theorem 2 handles that distinction.

D A CONVEX LOWER BOUND ON RECTANGULARIZATION ERROR

D.1 UNIFORM OCCUPANCY LOWER ENVELOPES

For fixed π, define $\underline{P}_i(x) = \min_{p \in \mathrm{proj}_i \mathcal{U}_b} p(x)$ for $x \in \mathcal{S}$, with zero outside X_i. For known rows this is the known probability. Each component is nonnegative and no greater than the corresponding component of every row in its projection; in particular $\sum_x \underline{P}_i(x) \le 1$. The matrix S_π defined in the main text is therefore substochastic. For every $P \in \mathcal{U}_b$, $P_\pi \ge S_\pi$ entrywise. Nonnegative matrix multiplication and induction imply $P_\pi^t \ge S_\pi^t$ for all $t \ge 0$. Both discounted resolvents exist, and summing their nonnegative power series proves

$$d_P^\pi(s,a) \ge (1-\gamma)[\mu^\top (I - \gamma S_\pi)^{-1}](s)\pi(a \mid s) \ge (1-\gamma)\mu(s)\pi(a \mid s).$$

A truncated nonnegative series through any finite $t = T$ is also a valid lower envelope. The truncated sum avoids a matrix inverse, but its entries still require downward error bounds in a directed-rounding implementation. The resulting occupancy *lower* bound need not sum to one, and the coordinatewise transition minima need not form a stochastic kernel.

Binary coordinate oracles. For a covered row with empirical coordinate q_x, let $\varepsilon = b/N_i$. If $q_x = 0$, the minimum coordinate is zero. If $q_x = 1$ and $|X_i| > 1$, it is $e^{-\varepsilon}$; a singleton alphabet instead has fixed coordinate one. For $0 < q_x < 1$, it is the smaller solution $p_x \in (0, q_x]$ of

$$q_x \log(q_x/p_x) + (1-q_x)\log((1-q_x)/(1-p_x)) = \varepsilon.$$

The log-sum inequality (Cover & Thomas, 2006), grouping successor x against its complement, gives the binary KL lower bound. The remaining mass can be distributed proportionally to the other empirical coordinates, attaining the binary lower bound. Uncovered nonsingleton rows have all coordinate minima zero; an uncovered singleton row is fixed. For rigorous floating-point lower envelopes, use lower endpoints of valid binary-root brackets, not upward-rounded probability estimates.

D.2 PROOF AND COMPUTABLE DUAL

Proof of Proposition 5. The row residuals $a_i(P; v)$ are nonnegative for all $P \in \mathcal{U}_b$. Thus (4) gives

$$G_\pi \ge \frac{\gamma}{1-\gamma} \min_{P \in \mathcal{U}_b} \sum_i \underline{d}_i a_i(P; v) = \frac{\gamma}{1-\gamma} A_{\underline{d}}(v) \ge 0.$$

The objective is affine in P and the feasible set is compact and convex. A lower envelope has $\underline{d}_{sa} = 0$ whenever $\pi(a \mid s) = 0$. If it is positive on every row with $\pi(a \mid s) > 0$, the objective vanishes exactly when one feasible kernel minimizes all those rows. The construction in Theorem 2 then gives the stated equivalence under full-support initialization.

Known-row residuals are identically zero, and uncovered rows can independently minimize v with no likelihood cost. It remains to optimize over covered rows. For a multiplier $\lambda > 0$, define

$$\begin{aligned} \psi_i(\lambda) &= \min_{p \in \Delta(X_i)} \{\underline{d}_i p^\top v + \lambda N_i D_{\mathrm{KL}}(\widehat{p}_i \| p)\}, \\ D_{\underline{d}}(\lambda) &= \sum_{i \in \mathcal{I}_+} \psi_i(\lambda) - \lambda b - \sum_{i \in \mathcal{I}_+} \underline{d}_i \phi_i(v). \end{aligned} \tag{30}$$

At $\lambda = 0$, define $\psi_i(0) = \underline{d}_i \min_{X_i} v$ and use the same expression for $D_{\underline{d}}(0)$. Weak duality gives $D_{\underline{d}}(\lambda) \le A_{\underline{d}}(v)$ for every $\lambda \ge 0$. For $b > 0$, the empirical kernel has zero likelihood cost, and sufficiently small mixtures with uniform rows give relative-interior points with cost strictly below b. Slater's condition gives $A_{\underline{d}}(v) = \max_{\lambda \ge 0} D_{\underline{d}}(\lambda)$ (Boyd & Vandenberghe, 2004, Section 5.2.3). The case $b = 0$ is already exact and needs no dual search. □

A lower-bound direction at both levels. To bound $A_{\underline{d}}(v)$ from below, lower-bound each $\psi_i(\lambda)$, and *upper*-bound each subtracted $\phi_i(v)$. Evaluating the penalized objective at a feasible row only upper-bounds ψ_i and is not a valid substitute. The following scalar dual gives rowwise lower bounds without a general-purpose cone solver.

Fix a covered row, $a = \lambda N_i > 0$, $z_x = \underline{d}_i v_x$, and $q = \widehat{p}_i$. Define $\mathcal{D}_i$ to consist of normalization multipliers ν satisfying $z_x + \nu > 0$ when $q_x > 0$ and $z_x + \nu \geq 0$ when $q_x = 0$. Then

$$\psi_i(\lambda) = \max_{\nu \in \mathcal{D}_i} \left\{ a + a \sum_{x: q_x > 0} q_x \log(z_x + \nu) - a \log a - \nu \right\}, \tag{31}$$

To derive (31), minimize the Lagrangian $\sum_x (z_x + \nu) p_x + a \sum_{q_x > 0} q_x \log(q_x / p_x) - \nu$ coordinate by coordinate: on the observed support the minimizer is $p_x = a q_x / (z_x + \nu)$, while an unobserved coordinate is zero unless its coefficient vanishes. Substitution proves the displayed dual. Its strict concavity in ν gives either the unique normalization root

$$a \sum_{x: q_x > 0} \frac{q_x}{z_x + \nu} = 1,$$

or a boundary solution at an unseen minimum, in which case the remaining mass is placed there. The derivative of the scalar objective is the left side minus one and is decreasing. If the derivative at an admissible lower boundary is nonpositive, that boundary is optimal; otherwise bisection brackets the root. For $\underline{d}_i = 0$, $p = q$ gives $\psi_i = 0$. At $\lambda = 0$, use $\psi_i(0) = \underline{d}_i \min_{X_i} v$.

Every admissible ν in (31), even before inner convergence, gives a lower bound on ψ_i. Summing these row bounds in (30) therefore gives a valid global lower bound. To optimize the bound, increase or decrease λ according to whether the associated row solutions overspend or underspend b. The total KL of penalized minimizers is nonincreasing in λ: adding the two optimality inequalities at multipliers $\lambda_1 < \lambda_2$ proves this directly. Doubling and bisection can thus approach a dual maximizer when the budget is active. *Every dual-feasible point is valid; convergence is needed only for tightness.* Since $A_{\underline{d}}(v) \geq 0$, a negative lower bound can be replaced by zero. In a flat case where a lowest-value kernel already meets the budget, the optimum may be at $\lambda = 0$.

D.3 Approximate values and valid lower-bound directions

Proposition 11 (Residual-corrected lower gap)**.** *Let v be any value vector, $\eta \geq \|T_{\mathrm{R}}^{\pi} v - v\|_{\infty}$, and let $\underline{A} \leq A_{\underline{d}}(v)$ be any dual lower bound, with $A_{\underline{d}}(v)$ defined using the exact row objectives at v. Then*

$$G_{\pi} \geq \max \left\{ 0, \frac{\gamma \underline{A} - 2\eta}{1 - \gamma} \right\}.$$

Proof. For any $P \in \mathcal{U}_b$, subtracting v from its Bellman equation yields

$$J_P^{\pi} - \mu^{\top} v = \mu^{\top} (I - \gamma P_{\pi})^{-1} (T_{\mathrm{R}}^{\pi} v - v) + \frac{\gamma}{1 - \gamma} \sum_i d_P^{\pi}(i) [P_i^{\top} v - \phi_i(v)].$$

The first term is at least $-\eta/(1-\gamma)$; replacing d_P^{π} by $\underline{d}$ in the second term is valid because its row residuals remain nonnegative for arbitrary v. Minimizing over P gives $J_{\mathrm{LR}}^{\pi} - \mu^{\top} v \geq -\eta/(1-\gamma) + \gamma A_{\underline{d}}(v)/(1-\gamma)$. Contraction gives $J_{\mathrm{R}}^{\pi} - \mu^{\top} v \leq \eta/(1-\gamma)$. Subtraction proves the result. □

With approximate row values of absolute error δ and computed policy residual $\widetilde{\eta}$, one can use $\eta = \widetilde{\eta} + \gamma\delta$. If the oracle provides feasible row upper objectives $u_i \geq \phi_i(v)$, substitute u_i in the subtracted term of (30); the result remains a lower bound. Feasibility, root-domain checks, and directed rounding are separate from solver stopping tolerances. Ordinary floating-point residuals alone do not establish directed-rounding error bounds.

E Policy-selection bounds and analytical examples

E.1 Proof and numerical extension of the Bellman bound

Proof of Proposition 6. We use the standard monotone Bellman supersolution argument (Iyengar, 2005; Nilim & El Ghaoui, 2005), applied to a single coupled-feasible kernel. The likelihood-constrained minimization of this upper bound does not require the nonlinear coupled return to be

convex. Fix any feasible (P, t). The ordinary optimality operator $(T_P v)(s) = \max_a\{r(s,a) + \gamma P_{sa}^\top v\}$ is monotone and obeys $T_P(v + c\mathbf{1}) = T_P v + \gamma c\mathbf{1}$. Set $u = v^\star + t\mathbf{1}/(1-\gamma)$. Then

$$T_P u \le v^\star + t\mathbf{1} + \frac{\gamma t}{1-\gamma}\mathbf{1} = u.$$

Monotone iteration gives $V_P^\star \le u$, where $V_P^\star$ is the fixed point of T_P and the ordinary optimal value under P. For every stationary π, $\min_{P' \in \mathcal{U}_b} J_{P'}^\pi \le J_P^\pi \le \mu^\top V_P^\star$. Taking the maximum over π shows $J_{\mathrm{LR}}^\star \le \mu^\top v^\star + t/(1-\gamma)$. Meanwhile $J_{\mathrm{LR}}^\star \ge J_{\mathrm{R}}^{\pi_{\mathrm{R}}} = \mu^\top v^\star$. Subtracting proves the result. The feasible set of (11) is convex because $v^\star$ is fixed. A feasible kernel and sufficiently large finite t always exist; the minimum is attained after restricting to a compact sublevel set. We used weak max–min inequality, not minimax equality. □

Corollary 12 (Arbitrary approximate value vector)**.** *Let $v \in \mathbb{R}^{|\mathcal{S}|}$ be arbitrary and let $L_\pi \le J_{\mathrm{LR}}^\pi$ be a valid lower bound. If $P \in \mathcal{U}_b$, $t \in \mathbb{R}$, and $T_P v \le v + t\mathbf{1}$, then*

$$J_{\mathrm{LR}}^\star - J_{\mathrm{LR}}^\pi \le \mu^\top v + \frac{t}{1-\gamma} - L_\pi.$$

The same supersolution proof establishes this statement. It permits residual-corrected numerical lower values in place of an exact rectangular fixed point. The left-hand side is nonnegative; a negative computed upper bound therefore indicates inconsistent numerical error bounds.

E.2 Uniform branching: an exact likelihood-budget calculation

Fix integers $m, N \ge 1$ and $\zeta > 0$. There are states $s_0, s_1, \dots, s_m, g, z$. State s_0 has a *known* transition to each s_i with probability $1/m$. Each uncertain s_i has admissible successor set $\{g, z\}$ and success probability $p_i = P(g \mid s_i)$. States g, z are known absorbing states, with rewards one and zero respectively. All other rewards are zero, μ is concentrated at s_0, and there is one action. Each uncertain row has N recorded successes and no failures. No uncertainty is assigned to the router or sink transitions.

Suppress the unique action and policy in this construction. The return is $J_P = \gamma^2[(1/m)\sum_i p_i]/(1-\gamma)$. Let $x_i = -\log p_i \ge 0$. The coupled likelihood condition with $b = mN\zeta$ is $\sum_i x_i \le m\zeta$. Since e^{-x} is convex and decreasing,

$$\frac{1}{m}\sum_i e^{-x_i} \ge e^{-(1/m)\sum_i x_i} \ge e^{-\zeta},$$

and equality is attained at $x_i = \zeta$. In the rectangular hull each $x_i \le m\zeta$ separately; setting every $x_i = m\zeta$ is feasible. These computations prove (10). The ratio of rectangular to coupled value is $e^{-(m-1)\zeta}$, but the absolute gap is bounded by the finite return range. Neither quantity is an information-theoretic sample-complexity lower bound.

The rectangular worst kernel has $p_i = e^{-m\zeta}$ and $K(Q) = mb$, so $\tau = m$. On the symmetric repair path all α_i coincide at an optimum, and the budget forces repaired successes $p_i' = e^{-\zeta}$. Thus the repaired kernel's *actual value* attains the coupled minimum in this example. The TV upper bound is also exact here: each uncertain row is visited once, so $d_Q(s_i) = (1-\gamma)\gamma/m$ and $\sum_i d_Q(s_i)\,\mathrm{TV}(P_i', Q_i) = (1-\gamma)\gamma(e^{-\zeta} - e^{-m\zeta})$. Multiplication by $\gamma/(1-\gamma)^2$ gives exactly the unnormalized gap in (10).

For the unreachable-row variant, change only the known router to select s_1 with probability one. Keep the same data on all uncertain rows. The coupled adversary can spend its entire budget on row one, and both robust objectives equal $\gamma^2 e^{-m\zeta}/(1-\gamma)$. The rectangular optimizer may still spend b on every row. Set $\alpha_1 = 0$ and $\alpha_i = 1$ for $i > 1$. This repair is feasible and has zero cost because $d_Q(s_i) = 0$ for $i > 1$. It proves that $\tau = m$ is compatible with exactness. It also illustrates why full-support initialization is needed for the necessity direction of the likelihood-cost criterion.

Matching lower and repaired upper bounds. At each uncertain row in the uniform router, the occupancy is exactly $(1-\gamma)\gamma/m$, independently of its success probability. The substochastic

lower envelope therefore equals the true occupancy on every row with a nonzero residual. With $v_g = 1/(1-\gamma)$ and $v_z = 0$, the lower program is

$$A_{\underline{d}}(v) = \frac{\gamma}{m} \min_{p:\, \sum_j -N \log p_j \le b} \sum_j (p_j - e^{-m\zeta}) = \gamma(e^{-\zeta} - e^{-m\zeta}).$$

Consequently $\gamma A_{\underline{d}}(v)/(1-\gamma)$ equals the gap in (10). All repair weights are equal and the row likelihood functions are identical. Averaging any feasible repair vector preserves feasibility and objective by convexity; the optimum can therefore be chosen uniform. Its success probability is $e^{-\zeta}$, so direct evaluation of the repaired kernel also attains J^{π}_{LR}. The TV upper bound in (8), the evaluated repaired value minus J^{π}_{R}, and the lower bound therefore all equal the gap in this symmetric construction. This equality does not imply tightness for recurrent models; Section 6 evaluates that distinction.

E.3 An exactly evaluated policy can still be coupled-suboptimal

Add a safe action at s_0 that enters a new known absorbing state with reward

$$c = \frac{\gamma}{2}\left(e^{-\zeta} + e^{-m\zeta}\right) \in [0, 1],$$

while the risky action uses the uniform router above. For a common two-action set, make the additional action at every uncertain branch state a known transition to the zero-reward sink; at known sinks, duplicate the absorbing action. These additional actions cannot improve on the risky continuation. The safe value is

$$J_{\mathrm{safe}} = \frac{\gamma}{1-\gamma} c = \frac{1}{2}\left(J^{\mathrm{risky}}_{\mathrm{LR}} + J^{\mathrm{risky}}_{\mathrm{R}}\right).$$

For $m > 1$ and $\zeta > 0$, rectangular planning strictly prefers safe, while coupled planning strictly prefers risky. Randomizing at s_0 does not improve either objective: for a fixed risky probability $\omega \in [0, 1]$, the objective is the affine combination $\omega J^{\mathrm{risky}} + (1-\omega) J_{\mathrm{safe}}$, since the safe branch has no uncertain transitions. Therefore the endpoints are optimal. The rectangular policy has *zero evaluation gap* because its entire trajectory is known, but positive coupled policy-selection loss

$$J^{\star}_{\mathrm{LR}} - J^{\pi_{\mathrm{R}}}_{\mathrm{LR}} = \frac{\gamma^2}{2(1-\gamma)}\left(e^{-\zeta} - e^{-m\zeta}\right).$$

This is a direct counterexample to inferring coupled optimality from $G_{\pi_{\mathrm{R}}} = 0$ or from a coupled-feasible worst kernel for that policy alone.

F The reverse-KL row oracle and numerical planning guarantees

Divergence-constrained backups and their scalar Burg-entropy solvers are established tools (Nilim & El Ghaoui, 2005; Adam & Mácha, 2022; Ho et al., 2022). We handle empirical-zero successors and propagate row-objective accuracy to policy loss. For a covered row and a continuation vector v, write $q = \widehat{p}_i$, $\varepsilon = b/N_i$, and $S_+ = \operatorname{supp} q$. For $\varepsilon > 0$, when the optimizer puts no mass outside S_+ and v is nonconstant there, define $A(\nu) = \sum_{x \in S_+} q_x/(v_x + \nu)$. The rational tilt satisfies

$$p_x(\nu) = \frac{q_x}{(v_x + \nu) A(\nu)}, \qquad g(\nu) = \sum_{x \in S_+} q_x \log(v_x + \nu) + \log A(\nu) = \varepsilon. \tag{32}$$

On $\nu > -\min_{S_+} v$, g is decreasing and $f_{\mathrm{row}}(\nu) = 1/A(\nu) - \nu$ is increasing. A bracket $L < U$ in the valid branch with $g(L) \ge \varepsilon \ge g(U)$ gives an objective enclosure $f_{\mathrm{row}}(L) \le \phi_i(v) \le f_{\mathrm{row}}(U)$. Stop when *the objective width*, not the dual-variable width, is at most δ.

When $X_i \setminus S_+ \neq \varnothing$, set $v_0 := \min_{x \in X_i \setminus S_+} v_x$. If $v_0 < \min_{S_+} v$ and $\varepsilon \ge g(-v_0)$, the optimizer assigns

$$p_x = c \frac{q_x}{v_x - v_0}\ (x \in S_+), \qquad c = \exp\left(\sum_{x \in S_+} q_x \log(v_x - v_0) - \varepsilon\right),$$

and places the remaining mass on unseen minimizers. Its value is $v_0 + c$. The following subsection treats constant-support values, zero radius, and uncovered rows. Removing unseen states from the optimization changes the ambiguity set and can invalidate pessimism.

F.1 Complete solution of the row problem

Fix a finite successor set X, $q \in \Delta(X)$, $v \in \mathbb{R}^{|X|}$, and $\varepsilon \geq 0$. Consider

$$\phi(q, v, \varepsilon) = \min_{p \in \Delta(X)} p^\top v \quad \text{subject to} \quad D_{\mathrm{KL}}(q\|p) \leq \varepsilon.$$

For strictly positive q, the rational tilt and scalar equation are the Burg-entropy projection of Adam & Mácha (2022, Theorem 3.3, arXiv version), with their objective vector set to $-v$; see also the robust-MDP treatments of Nilim & El Ghaoui (2005); Ho et al. (2022). Their positive-reference assumption does not cover empirical zeros on the predeclared alphabet. We give the full KKT calculation below to retain such successors and obtain the objective enclosures needed for planning. Neither the interior scalar reduction nor convex duality is claimed as a new principle. Let $S_+ = \operatorname{supp} q$, $S_0 = X \setminus S_+$, $v_+ = \min_{S_+} v$, and $v_0 = \min_{S_0} v$ (with $v_0 = +\infty$ if S_0 is empty).

If $\varepsilon = 0$, KL vanishes only at $p = q$, so the answer is $q^\top v$. If the row is uncovered, there is no q and the answer is $\min_X v$. If v is constant on S_+ and v_0 is no smaller, q is optimal for every finite radius. If it is constant on S_+ with value $c_+ > v_0$, put mass $e^{-\varepsilon} q$ on S_+ and all remaining mass on an unseen minimizer; the answer is $v_0 + (c_+ - v_0)e^{-\varepsilon}$. These statements include ties and the zero-radius endpoint.

For the remaining case, v is nonconstant on S_+ and $\varepsilon > 0$. The likelihood constraint is active at the optimum. Otherwise, moving a sufficiently small amount of probability from a positively observed higher-value state to a globally lower-value state would reduce the objective while remaining feasible. An optimum must assign positive probability to every member of S_+, so this perturbation is available. Stationarity on the nonconstant observed support excludes a zero likelihood multiplier. The Lagrangian with likelihood multiplier $\lambda > 0$ and normalization multiplier ν yields

$$p_x = \frac{\lambda q_x}{v_x + \nu} \quad (x \in S_+), \qquad v_x + \nu \geq 0 \quad (x \in S_0),$$

with equality in the second relation whenever $p_x > 0$. Strong duality follows by a relatively strictly feasible positive distribution sufficiently close to q, which has KL less than ε.

If no unseen mass is used, normalization gives $\lambda = 1/A(\nu)$, and the active constraint gives (32). The root must satisfy $\nu > -v_+$ and, when unseen states exist, $\nu \geq -v_0$. If $v_0 \geq v_+$, the latter constraint is automatic and the support root exists for every positive finite radius. If $v_0 < v_+$, define

$$\varepsilon_0 = g(-v_0) = \sum_{x \in S_+} q_x \log(v_x - v_0) + \log \sum_{x \in S_+} \frac{q_x}{v_x - v_0}.$$

This is nonnegative, because it is the KL divergence of the normalized rational tilt at the boundary. For $0 < \varepsilon < \varepsilon_0$, the supported root is strictly larger than $-v_0$. At equality, the root is the boundary and all mass remains observed. For $\varepsilon > \varepsilon_0$, stationarity requires $\nu = -v_0$ and

$$\lambda = c = \exp\left(\sum_{x \in S_+} q_x \log(v_x - v_0) - \varepsilon \right).$$

The observed mass equals $cA(-v_0) = e^{\varepsilon_0 - \varepsilon} \leq 1$; the remaining mass is assigned to any unseen global minimizer. The resulting expected value is

$$v_0 + \sum_{x \in S_+} p_x (v_x - v_0) = v_0 + c.$$

It satisfies the active KL constraint and the KKT conditions, so it is optimal. The constant-support escape formula is the same calculation with $\varepsilon_0 = 0$.

F.2 Monotonicity and objective enclosures

For the functions A and g in (32), put $B = \sum_{x \in S_+} q_x/(v_x + \nu)^2$. Then $f_{\mathrm{row}}(\nu) = 1/A(\nu) - \nu$, and

$$g'(\nu) = A - B/A \leq 0, \qquad f'_{\mathrm{row}}(\nu) = B/A^2 - 1 \geq 0,$$

by Jensen's inequality $B \geq A^2$. The inequalities are strict when v is nonconstant on S_+. As $\nu \to \infty$, $g(\nu) \to 0$. As $\nu \downarrow -v_+$, $g(\nu) \to \infty$ in the nonconstant-support case. Hence the supported scalar root is unique in its valid branch. For a valid bracket $g(L) \geq \varepsilon \geq g(U)$ containing that root,

$$f_{\mathrm{row}}(L) \leq \phi(q, v, \varepsilon) \leq f_{\mathrm{row}}(U).$$

The distribution $p(U)$ is feasible because its KL is $g(U) \leq \varepsilon$. Therefore an upper objective and a feasible row are returned together. The lower endpoint's distribution need not be feasible; only its monotone objective is used as a lower bound.

One evaluation costs $O(|S_+|)$ arithmetic operations. After a finite initial bracket is found, bisection contracts its width geometrically. The implementation stops on $f_{\mathrm{row}}(U) - f_{\mathrm{row}}(L)$, not on a universal dual-width tolerance. There is no asserted radius-independent Lipschitz constant converting a fixed dual tolerance into objective error near a singular endpoint.

Stable shortcuts and scaling. Translation of v changes only the objective offset, and positive scaling changes the objective and desired tolerance proportionally. Shifting the global minimum to zero is useful: an unseen-minimum escape boundary then occurs at $\nu = 0$, while the supported branch lies at positive ν. For a large radius, take $x_{\min} \in \arg\min_{x \in X} v_x$ and let e_x be the point mass at x. The mixture

$$p = tq + (1-t)e_{x_{\min}}, \qquad t \in (0, 1], \quad -\log t \leq \varepsilon$$

satisfies $D_{\mathrm{KL}}(q\|p) \leq -\log t$ and has objective at most $\min_X v + t\,\mathrm{span}(v)$. Thus choosing $t\,\mathrm{span}(v) \leq \delta$ gives a δ-wide enclosure without representing probabilities as small as $e^{-\varepsilon}$. For a tiny radius, Pinsker gives the enclosure $[q^\top v - \mathrm{span}(v)\sqrt{\varepsilon/2}, q^\top v]$ with feasible row q. Neither shortcut changes the ambiguity set; each supplies a lower bound and a feasible upper witness.

F.3 Approximate dynamic programming

Proposition 13 (Numerical error to policy loss)**.** *Let $\delta, \xi \geq 0$. Suppose each approximate row objective at v has absolute error at most δ, and let $\widetilde{T}$ be the optimality backup formed from those objectives. If $\|v - \widetilde{T}v\|_\infty \leq \xi$ and $\widetilde{\pi}$ is greedy for the same approximate objectives, then*

$$\|v - V_{\mathrm{R}}^\star\|_\infty \leq \frac{\xi + \gamma\delta}{1-\gamma}, \qquad \|V_{\mathrm{R}}^\star - V_{\mathrm{R}}^{\widetilde{\pi}}\|_\infty \leq \frac{2\gamma(\xi+\delta)}{(1-\gamma)^2}.$$

Thus $\delta, \xi \leq \epsilon(1-\gamma)^2/(4\gamma)$ suffice for ϵ-optimal rectangular planning. A residual correction converts the approximate value into a valid lower bound (Lemma 14).

Proof of Proposition 13. Every approximate row objective differs from its exact value by at most δ, so $\|\widetilde{T}v - T_{\mathrm{R}}v\|_\infty \leq \gamma\delta$. The residual condition and contraction give

$$\|v - V_{\mathrm{R}}^\star\|_\infty \leq \|v - T_{\mathrm{R}}v\|_\infty + \gamma\|v - V_{\mathrm{R}}^\star\|_\infty \leq \xi + \gamma\delta + \gamma\|v - V_{\mathrm{R}}^\star\|_\infty.$$

Rearranging proves the value bound. Approximate greediness for the same row values implies $T_{\mathrm{R}}v \leq T_{\mathrm{R}}^{\widetilde{\pi}}v + 2\gamma\delta\mathbf{1}$. Set $e = \|v - V_{\mathrm{R}}^\star\|_\infty$ and $z = \|V_{\mathrm{R}}^\star - V_{\mathrm{R}}^{\widetilde{\pi}}\|_\infty$. Insert and subtract the two operators at v and at $V_{\mathrm{R}}^\star$ to get

$$z \leq 2\gamma e + 2\gamma\delta + \gamma z.$$

Therefore

$$z \leq \frac{2\gamma}{1-\gamma}\left(\frac{\xi + \gamma\delta}{1-\gamma} + \delta\right) = \frac{2\gamma(\xi+\delta)}{(1-\gamma)^2}.$$

Substitution of the stated tolerances proves the ϵ guarantee. Here ϵ is a raw discounted-return error. For a normalized target $\bar{\epsilon} = (1-\gamma)\epsilon$, the sufficient condition is instead $\delta, \xi \leq \bar{\epsilon}(1-\gamma)/(4\gamma)$, exactly as in (17). A row interval of width at most δ permits any reported point in that interval; choosing an upper endpoint does not by itself make the value iterate pessimistic. □

For value iteration $v_{k+1} = \widetilde{T}_k v_k$ with uniform row error δ at each iterate and $v_0 = 0$,

$$\|v_k - V_{\mathrm{R}}^\star\|_\infty \leq \frac{\gamma^k + \gamma\delta}{1-\gamma}.$$

A policy greedy for approximate row objectives at v_k then has rectangular loss at most $2\gamma(\gamma^k + \delta)/(1-\gamma)^2$. Thus finite iterations give an approximate policy, not an exact optimizer. A sufficient iteration count has $\gamma^k \le \epsilon(1-\gamma)^2/(4\gamma)$ together with the same upper bound on δ. Residual stopping is preferable in practice because it can terminate earlier.

Lemma 14 (Residual-corrected lower values)**.** *For a fixed policy π, let $\widetilde{T}^\pi$ use row objectives with absolute error at most δ. If $\|v - \widetilde{T}^\pi v\|_\infty \le \eta$, then*

$$L_\pi := \mu^\top v - \frac{\eta + \gamma\delta}{1-\gamma} \le J^\pi_{\mathrm{R}}.$$

The lower bound can be replaced by $\max\{0, L_\pi\}$ under nonnegative rewards. If $Q \in \mathcal{R}_b$ is an approximate adversarial kernel and $P \in \mathcal{U}_b$ is any feasible repair, then

$$L_\pi \le J^\pi_{\mathrm{LR}} \le J^\pi_P, \qquad J^\pi_P - L_\pi \le \frac{\gamma}{(1-\gamma)^2} \sum_i d^\pi_Q(i)\, \mathrm{TV}(P_i, Q_i) + (J^\pi_Q - L_\pi).$$

Proof. Apply the contraction residual argument to T^π_{R}. The interval follows from uncertainty-set inclusion and feasibility. Add and subtract J^π_Q, and apply (22), for the final inequality. Inexact ordinary policy evaluation additionally requires a linear-system residual correction; it cannot be silently treated as exact. □

For a policy greedy for the same approximate objectives used in $\widetilde{T}$, its policy residual at that vector equals the reported optimality residual. Thus the lemma applies directly to the output of Proposition 13. Approximate upper row values alone do not automatically make the uncorrected iterate a pessimistic bound.

G Likelihood calibration under two sampling models

G.1 Independent-transition design and a finite-sample budget

Fix the sample length T, structural row domains, known rows, and $\beta \in (0,1)$ before observing successors. The row-label vector $(I_1, \dots, I_T)$ may be deterministic or jointly random, but is exogenous to successor randomness. Conditional on that *complete vector*, successors are independent with $X_t \sim P^\star_{I_t}$. This includes i.i.d. label draws; independence among the labels themselves is unnecessary. Conditioning fixes the row counts, the number n of unknown-row observations, and D. It does not justify conditioning on a future trajectory's labels, or a length chosen using observed successors. The predictable regime below replaces these assumptions rather than assuming dependent data become independent after counting.

For a multinomial empirical distribution $\widehat{q}_N$ from $q^\star$ on an alphabet of size k, Agrawal (2020, Theorem I.3) prove

$$\mathbb{E}\exp\{\lambda N D_{\mathrm{KL}}(\widehat{q}_N \| q^\star)\} \le (1-\lambda)^{-(k-1)}, \qquad 0 \le \lambda < 1. \tag{33}$$

The theorem does not require replacing k by the empirically observed support. If the true distribution has structural zeros in a predeclared alphabet, the same bound holds, possibly conservatively.

Proof of Theorem 7: independent-transition case. Conditional on the row labels, independence and (33) yield

$$\mathbb{E}[e^{\lambda K(P^\star)} \mid \text{row labels}] \le (1-\lambda)^{-D}, \qquad D = \sum_{i \in \mathcal{I}_+} (k_i - 1).$$

For $0 \le \lambda < 1$,

$$\log \mathbb{E}[e^{\lambda(K(P^\star) - D)} \mid \text{row labels}] \le D[-\log(1-\lambda) - \lambda] \le \frac{D\lambda^2}{2(1-\lambda)},$$

where the last inequality follows by summing the power series and using $1/j \le 1/2$ for $j \ge 2$. The Chernoff bound with this sub-gamma moment inequality gives $\mathbb{P}\{K(P^\star) > D + \sqrt{2Dx} + x \mid$

row labels$\} \le e^{-x}$. For completeness, when $D > 0$ choose $\lambda = \sqrt{2x/D}/(1+\sqrt{2x/D})$ and substitute into Markov's inequality. When $D = 0$, all relevant rows have singleton support and $K(P^\star) = 0$. Taking $x = \log(1/\beta)$ and averaging over the row labels proves the result even though D is data-dependent through the counts. □

The proof uses a *finite-sample exponential moment*, not a limiting chi-square approximation. It does not assert that this convenient budget is the smallest available confidence threshold. Exactly computing a multinomial tail or using a different valid confidence construction may improve constants.

G.2 Predictable trajectory sampling and a mixture supermartingale

For $t \ge 1$, let $\mathcal{F}_t$ contain the information after transition t, and let $\mathcal{G}_t$ be the information immediately before that transition, with $\mathcal{F}_{t-1} \subseteq \mathcal{G}_t \subseteq \mathcal{F}_t$. The selected row $i_t = (s_t, a_t)$ is $\mathcal{G}_t$-measurable, and

$$\mathbb{P}(X_t = x \mid \mathcal{G}_t) = P^\star_{i_t}(x).$$

The pre-transition information may include a randomized action or episode reset. Policies and stopping rules may depend on past observations, but the displayed conditional law must hold. Initial-state and behavior-policy factors are excluded from the transition likelihood; known-row factors cancel. Let Π_0 be the fixed product of Dirichlet$(1/2, \dots, 1/2)$ integration measures on all unknown rows, including unvisited rows. Write

$$L_t(P) = \prod_{u \le t:\, i_u \text{ unknown}} P_{i_u}(X_u), \qquad M_t = \int L_t(P)\, d\Pi_0(P), \qquad Z_t = \frac{M_t}{L_t(P^\star)}.$$

For every finite t, the denominator is positive almost surely under each fixed $P^\star$. The integration measure is fixed before observing data. It is a device for constructing a nonnegative test process, not an assumption that $P^\star$ is drawn from that measure.

Proof of Theorem 7: trajectory case. The mixture's predictive law is the probability distribution $\widetilde{p}_t(\cdot \mid \mathcal{G}_t)$, obtained from the past counts and current row. On an unknown row,

$$\mathbb{E}\left[\frac{Z_t}{Z_{t-1}}\middle|\mathcal{G}_t\right] = \sum_{x:P^\star_{i_t}(x)>0} \widetilde{p}_t(x \mid \mathcal{G}_t) \le 1.$$

On a known row the ratio is one. Since Z_{t-1} is $\mathcal{G}_t$-measurable, this proves $\mathbb{E}[Z_t \mid \mathcal{G}_t] \le Z_{t-1}$. The tower property then yields $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] \le Z_{t-1}$, so Z_t is a nonnegative $\mathcal{F}_t$-supermartingale with $Z_0 = 1$. It is a martingale under suitable full-support conditions, but that stronger property is unnecessary and need not hold when the true kernel has zeros. Ville's maximal inequality (Howard et al., 2020, Lemma 1) gives $\mathbb{P}\{\sup_t Z_t \ge 1/\beta\} \le \beta$. Finally,

$$K_t(P^\star) - \mathrm{pen}_t = [\log L_t(\widehat{P}_t) - \log L_t(P^\star)] - [\log L_t(\widehat{P}_t) - \log M_t] = \log Z_t.$$

Outside a probability-β event, $K_t(P^\star) \le \mathrm{pen}_t + \log(1/\beta)$ simultaneously at every time. This proves the anytime guarantee without conditioning on future row labels and without a mixing assumption. □

This instantiates the standard mixture-likelihood route to anytime validity, not a new concentration principle. The Dirichlet half measure is the Krichevsky–Trofimov construction (Krichevsky & Trofimov, 1981); Lindon & Malek (2022, Theorems 2.1 and 2.4) give mixture-based anytime inference for multinomial counts. Here the filtration argument permits predictable selection among transition rows and explicitly handles zero true probabilities. The resulting guarantee holds for each fixed kernel, not merely on average under the integration measure.

G.3 Exact count-dependent penalty and a safe alternative

Write $N_{ix} = N_{ix}(t)$ and $N_i = N_i(t)$ for counts through time t on unknown rows. Here $\mathcal{I}_+ = \{i \notin \mathcal{I}_K : N_i(t) > 0\}$ and Γ is the gamma function. Dirichlet integration gives, with $0 \log 0 = 0$,

$$\mathrm{pen}_t = \sum_{i \in \mathcal{I}_+} \left[\sum_{x \in X_i} N_{ix} \log \frac{N_{ix}}{N_i} - \log \Gamma(k_i/2) + \log \Gamma(N_i + k_i/2) \right.$$
$$\left. - \sum_{x \in X_i} \log \Gamma(N_{ix} + 1/2) + k_i \log \Gamma(1/2) \right]. \tag{34}$$

Unvisited rows contribute exactly zero. The penalty is nonnegative because $M_t \le L_t(\widehat{P}_t)$. Equation (34) is evaluated with log-gamma functions; it is not replaced by an interior asymptotic approximation at boundary counts.

Let pen_i denote row i's summand in (34). For a fixed alphabet and strictly positive limiting empirical proportions, Stirling's formula gives

$$\mathrm{pen}_i = \frac{k_i - 1}{2} \log \frac{N_i}{2\pi} + k_i \log \Gamma(1/2) - \log \Gamma(k_i/2) + o(1).$$

This explains the usual count-logarithmic sequential penalty. It is not a finite-sample upper bound: even the binary one-observation row has exact penalty $\log 2$. Boundary constants differ. Consequently the trajectory threshold is not obtained by inserting an effective sample size into (12). At modest counts either valid threshold can be numerically smaller; *the sampling model determines which one may be used.*

A simpler but different valid construction uses a uniform Dirichlet$(1, \dots, 1)$ prior. Its row mixture is $M_i^{(1)} = (k_i - 1)! \prod_x N_{ix}!/(N_i + k_i - 1)!$ and its saturated likelihood is $L_i(\widehat{P}) = \prod_x (N_{ix}/N_i)^{N_{ix}}$ for $N_i > 0$. Both are one at $N_i = 0$. Since the multinomial probability of the observed count vector under its own empirical distribution is at most one,

$$\frac{L_i(\widehat{P})}{M_i^{(1)}} = \binom{N_i + k_i - 1}{k_i - 1} \binom{N_i}{(N_{ix})_x} \prod_x (N_{ix}/N_i)^{N_{ix}} \le \binom{N_i + k_i - 1}{k_i - 1}.$$

Thus $\log(1/\beta) + \sum_i \log \binom{N_i + k_i - 1}{k_i - 1}$ is an anytime-valid budget for this *uniform-prior* construction. It is not asserted to upper-bound the half-prior penalty in (34). Evaluating the exact half-Dirichlet penalty avoids replacing it by a combinatorial upper bound; no ordering between the two mixture constructions is asserted.

G.4 Quantifying the difference between the two calibrations

Consider m covered unknown rows, each with alphabet size $k \ge 2$ and count N, and fixed β. Write b_{traj} for (13) evaluated at these counts. If the empirical proportions in each row converge to strictly positive limits, the preceding Stirling expansion gives

$$b_{\mathrm{traj}} = \frac{m(k-1)}{2} \log N + O_{m,k,\beta}(1), \qquad b_{\mathrm{iid}} = m(k-1) + \sqrt{2m(k-1)\log(1/\beta)} + \log(1/\beta).$$

Thus the trajectory construction's per-row radius is of order $\log N/N$, rather than $1/N$ for this fixed-sample construction. In the uncapped square-root terms of Theorem 15, substituting these budgets changes the bound by a factor of order $\sqrt{\log N}$ for fixed m, k, β. This is a comparison of two explicit sufficient thresholds, not a lower bound on the cost of anytime validity.

A finite-count comparison needs no asymptotic approximation. For binary rows with even N and exactly $N/2$ observations of each successor, the gamma duplication formula in (34) yields

$$b_{\mathrm{traj}} = \log(1/\beta) + m \log \left(\frac{\sqrt{\pi}\, \Gamma(N/2 + 1)}{\Gamma(N/2 + 1/2)} \right).$$

At $N = 2$, the per-row penalty is $\log 2$, whereas for large N it is $\frac{1}{2} \log(\pi N/2) + o(1)$. These identities explain why there is no uniform numerical ordering at small counts and why the sequential

penalty eventually grows. On exogenous fixed-sample data both constructions may be compared; on successor-dependent trajectories only the trajectory argument above justifies b_{traj}. A smaller observed number is not, by itself, permission to switch calibration models.

Equation (14) is the exact finite-count difference for these two constructions at the same covered-row counts. Its asymptotic interpretation keeps m, k, β fixed; boundary proportions require the exact gamma expression rather than the displayed interior expansion.

What transfers at a data-dependent stopping time. On the event $\{P^\star \in \mathcal{U}_t(b_t) \text{ for every } t\}$, the hull identity, fixed-policy repair/lower bounds, greedy optimality, and numerical-error bounds apply deterministically at every prefix. The comparator identity (15) and the explicit bound (35) therefore hold there simultaneously, including for data-selected policies, comparators, and finite stopping times. Use b_t, $N_i(t)$, and the corresponding covered/uncovered row sets, not fixed-design radii. The finite-coverage corollary still requires that the comparator visit no uncovered unknown row and uses its realized count-based C. No Markovian assumption is added on the behavior policy; the Markov conditional transition law and pre-transition measurability remain essential. A data-dependent feature class or an unsupported ellipsoid is not validated by this substitution.

H PESSIMISM AND COMPARATOR ANALYSIS

The comparison follows the standard pessimism argument: a lower value at the learned policy is compared with the same lower value at an arbitrary comparator, whose estimation error is then controlled (Jin et al., 2021; Xie et al., 2021; Uehara & Sun, 2023). We retain the comparator's coupled-to-rectangular gap before applying the transition simulation bound, keeping *approximation error separate from statistical and planning errors*.

Theorem 15 (Partial-coverage comparator bounds)**.** *On $P^\star \in \mathcal{U}_b$, let $\widehat{\pi}$ be ϵ_{alg}-optimal for $\mathcal{R}_b$, and let $\pi^\dagger \in \Pi$ be arbitrary. With $d_i = d^{\pi^\dagger}_{P^\star}(i)$,*

$$J^{\pi^\dagger}_{P^\star} - J^{\widehat{\pi}}_{P^\star} \le \epsilon_{\mathrm{alg}} + \frac{\gamma}{(1-\gamma)^2}\left[\sum_{i\in\mathcal{I}_+} d_i \min\left\{1, \sqrt{\frac{2b}{N_i}}\right\} + \sum_{i\in\mathcal{I}_0} d_i\right]. \tag{35}$$

Suppose $n \ge 1$, the comparator never visits an uncovered unknown row, and $C = \max_{i\in\mathcal{I}_+} d_i/\widehat{\rho}_i < \infty$. Then the non-algorithmic term in (35) *is at most*

$$\frac{\gamma}{(1-\gamma)^2}\sqrt{\frac{2bCm}{n}}. \tag{36}$$

For an exact maximizer π_{LR} of the coupled *objective, the same event and empirical C give*

$$J^{\pi^\dagger}_{P^\star} - J^{\pi_{\mathrm{LR}}}_{P^\star} \le \frac{\gamma}{(1-\gamma)^2}\sqrt{\frac{2bC}{n}}. \tag{37}$$

Proof of Theorem 15. On $P^\star \in \mathcal{U}_b \subseteq \mathcal{R}_b$, every policy obeys $J^\pi_{\mathrm{R}} \le J^\pi_{\mathrm{LR}} \le J^\pi_{P^\star}$. This is a simultaneous deterministic consequence of the event, so data-dependent policy selection is permitted. By ϵ_{alg}-optimality of $\widehat{\pi}$,

$$J^{\pi^\dagger}_{P^\star} - J^{\widehat{\pi}}_{P^\star} \le J^{\pi^\dagger}_{P^\star} - J^{\widehat{\pi}}_{\mathrm{R}} \le J^{\pi^\dagger}_{P^\star} - J^{\pi^\dagger}_{\mathrm{R}} + \epsilon_{\mathrm{alg}}.$$

Let Q be a rectangular worst kernel for $\pi^\dagger$. Both $P^\star_i$ and Q_i lie in $\mathcal{B}_i(b)$ on covered rows. Pinsker's inequality (Cover & Thomas, 2006) and the triangle inequality give

$$\mathrm{TV}(P^\star_i, Q_i) \le \sqrt{\tfrac{1}{2}D_{\mathrm{KL}}(\widehat{p}_i\|P^\star_i)} + \sqrt{\tfrac{1}{2}D_{\mathrm{KL}}(\widehat{p}_i\|Q_i)} \le \sqrt{2b/N_i}.$$

TV is at most one; it is zero on known rows and may be one on uncovered unknown rows. Applying (21) with reference occupancy $d^{\pi^\dagger}_{P^\star}$ proves (35).

If no uncovered unknown row is visited, weighted Cauchy–Schwarz and $\sum_{i\in\mathcal{I}_+} d_i \le 1$ yield

$$\sum_{i\in\mathcal{I}_+} d_i\sqrt{2b/N_i} \le \sqrt{\sum_{i\in\mathcal{I}_+} d_i}\sqrt{\sum_{i\in\mathcal{I}_+}\frac{2bd_i}{N_i}} \le \sqrt{\frac{2b}{n}\sum_{i\in\mathcal{I}_+}\frac{d_i}{\widehat{\rho}_i}} \le \sqrt{\frac{2bCm}{n}}.$$

The coverage coefficient is inside the square root. No population coefficient is substituted for this empirical quantity.

For the coupled policy, select a coupled worst kernel P^- for $\pi^\dagger$. The same policy-comparison argument gives $J^{\pi^\dagger}_{P^\star} - J^{\pi_{\mathrm{LR}}}_{P^\star} \le J^{\pi^\dagger}_{P^\star} - J^{\pi^\dagger}_{P^-}$. For $i \in \mathcal{I}_+$, Pinsker's inequality and $(\sqrt{x}+\sqrt{y})^2 \le 2(x+y)$ for $x, y \ge 0$ give

$$\mathrm{TV}(P_i^\star, P_i^-)^2 \le D_{\mathrm{KL}}(\widehat{p}_i \| P_i^\star) + D_{\mathrm{KL}}(\widehat{p}_i \| P_i^-).$$

Both kernels obey the same global budget, so $\sum_{i\in\mathcal{I}_+} \widehat{\rho}_i \,\mathrm{TV}(P_i^\star, P_i^-)^2 \le 2b/n$. Consequently

$$\begin{aligned}\sum_{i\in\mathcal{I}_+} d_i \,\mathrm{TV}(P_i^\star, P_i^-) &\le \sqrt{\sum_{i\in\mathcal{I}_+} d_i}\sqrt{\sum_{i\in\mathcal{I}_+} d_i \,\mathrm{TV}(P_i^\star, P_i^-)^2} \\ &\le \sqrt{C \sum_{i\in\mathcal{I}_+} \widehat{\rho}_i \,\mathrm{TV}(P_i^\star, P_i^-)^2} \le \sqrt{2bC/n}.\end{aligned}$$

Known rows contribute zero and uncovered unknown rows have zero comparator occupancy. Applying (21) completes the proof. □

What is and is not being compared. The coupled inequality is a tabular matched-assumption comparison, not a restatement of every CPPO-LR guarantee or a claim to improve its general-model rates (Uehara & Sun, 2023). The rectangular bound retains the effect of independently usable row budgets. *Neither proof assumes the learned policy is covered*; pessimism transfers the comparator comparison to it. If the comparator reaches an uncovered unknown row, its contribution in (35) remains and the finite-C corollaries cannot be invoked. When $n = 0$, no expression involving $1/n$ or $\widehat{\rho}$ is used: all unknown rows are unconstrained and the direct uncovered-row bound remains meaningful.

Equation (15) follows directly from the first policy-comparison inequality in the proof by writing $J^{\pi^\dagger}_{P^\star} - J^{\pi^\dagger}_{\mathrm{R}} = (J^{\pi^\dagger}_{P^\star} - J^{\pi^\dagger}_{\mathrm{LR}}) + G_{\pi^\dagger}$. The proof of (37) in fact bounds the first summand by $\gamma(1-\gamma)^{-2}\sqrt{2bC/n}$. Combining these inequalities under the finite-C assumptions of Theorem 15 yields (16). A valid upper bound on the comparator's gap can replace $G_{\pi^\dagger}$ in (16). This can improve the generic rectangular bound, but a small gap for $\widehat{\pi}$ alone cannot be substituted. Neither C nor the true comparator occupancy is assumed known to the learner.

Comparing (36) and (37), the factor $\sqrt{m}$ is an upper-bound comparison, not a matching lower bound. The analytical examples demonstrate actual loss of robust value and policy ranking, but they do not prove that every statistically calibrated problem incurs that factor. Independent rowwise calibration is a separate baseline: condition on the exogenous row labels and select $\beta_i > 0$ with $\sum_i \beta_i \le \beta$, using those labels but not successors. Applying (33) per covered row with budget $b_i = (k_i - 1) + \sqrt{2(k_i-1)\log(1/\beta_i)} + \log(1/\beta_i)$ and taking a union bound gives a product confidence set with level $1-\beta$. Outcome-dependent failure allocation would require a separate validity argument. Likewise, selecting the smallest of several individually valid budgets or changing the mixture measure after observing successors does not preserve the stated level without an additional simultaneous-validity argument.

I VALID LINEAR-MIXTURE EXTENSIONS

I.1 THE CONSTRAINED LIKELIHOOD REGION AND AN ELLIPSOIDAL OUTER SET

The matrices A_i and valid parameter domain Θ are defined in Section 5. For the ellipsoidal results let $\Sigma \in \mathbb{R}^{d\times d}$ be positive definite, $R \ge 0$, and

$$\mathcal{C} = \Theta \cap \{\theta : \|\Sigma^{1/2}(\theta - \widehat{\theta})\|_2 \le R\}.$$

All references to $\mathcal{C}$ in the ellipsoidal statements below use this definition; it is not identified with the exact likelihood set Θ_b^{LR}.

A finite linear-mixture model is $P_{\theta,i} = A_i\theta$, with all transitions valid exactly when $\theta \in \Theta$. Normalization and any known-row equalities are included in $H\theta = h$; all nonnegativity constraints are

retained, including probabilities of unobserved successors. Additional fixed linear inequalities can be retained in the constrained optimization domains. A claim that an entire ellipsoid satisfies such extra constraints also requires their corresponding support-function inequalities. Let

$$F(\theta) = -\sum_{i\in\mathcal{I}_+,x:N_{ix}>0} N_{ix}\log(\varphi_{ix}^\top\theta), \qquad \widehat{\theta}\in\arg\min_{\theta\in\Theta} F(\theta),$$

and assume this constrained optimum is attained and finite. Its likelihood region is $\Theta_b^{\mathrm{LR}} = \{\theta\in\Theta : F(\theta) - F(\widehat{\theta}) \le b\}$. For a finite likelihood point,

$$F(\theta) - F(\widehat{\theta}) = \sum_{i\in\mathcal{I}_+} N_i D_{\mathrm{KL}}(\widehat{p}_i\|A_i\theta) - \sum_{i\in\mathcal{I}_+} N_i D_{\mathrm{KL}}(\widehat{p}_i\|A_i\widehat{\theta}).$$

The second term cannot be dropped unless the model class attains the saturated empirical kernel. This is an important difference from (1).

Attainment is a statement about kernels. Stack the matrices A_i into A. The image of the polyhedron Θ under the linear map A is a closed polyhedron contained in a product of finite probability simplexes, hence compact. Intersecting this image with the lower-semicontinuous likelihood sublevel preserves compactness. Thus every row objective in (18), every fixed-policy coupled value, and every feasible-repair minimum attains its optimum. Unbounded directions in $\ker A$ have no effect on these statements. This uses linear images of polyhedra, not the false assertion that all linear images of closed sets are closed. The ellipsoidal set $\mathcal{C}$ is itself compact because $\Sigma \succ 0$.

Proposition 16 (A sufficient outer-ellipsoid condition)**.** *Let $\Sigma \succ 0$ be specified. For every $\theta\in\Theta_b^{\mathrm{LR}}$, set $u = \theta - \widehat{\theta}$ and assume that*

$$u^\top\nabla^2 F(\widehat{\theta}+tu)u \ge u^\top\Sigma u \qquad \textit{for all } t\in[0,1].$$

Then

$$\Theta_b^{\mathrm{LR}} \subseteq \Theta\cap\{\theta : \|\Sigma^{1/2}(\theta-\widehat{\theta})\|_2 \le \sqrt{2b}\}.$$

Suppose the curvature condition holds for every realized dataset on which this construction is used, and $P_{\theta^\star}$ is the true model for a fixed $\theta^\star\in\Theta$. A saturated budget from either regime of Theorem 7 then gives the same coverage for this intersection. At a trajectory stopping time, the condition must hold for that prefix; a construction valid at every prefix retains the simultaneous event. If the curvature condition is not verified, retain the exact LR region. The resulting rule has unconditional coverage; this does not assert coverage conditional on selecting an ellipsoid.

Proof. Let $u = \theta - \widehat{\theta}$. The constrained first-order condition gives $\nabla F(\widehat{\theta})^\top u \ge 0$. Finite observed probabilities at both endpoints stay positive on the segment, so the integral Taylor formula applies:

$$F(\theta) - F(\widehat{\theta}) = \nabla F(\widehat{\theta})^\top u + \int_0^1 (1-t)u^\top\nabla^2 F(\widehat{\theta}+tu)u\,dt \ge \tfrac{1}{2}\|\Sigma^{1/2}u\|_2^2.$$

The likelihood sublevel condition proves inclusion. For coverage, let F_{sat} be the minimum negative log likelihood over the product of all valid row simplexes. Since the structured class is a subset, $F(\widehat{\theta}) \ge F_{\mathrm{sat}}$. Thus

$$F(\theta^\star) - F(\widehat{\theta}) \le F(\theta^\star) - F_{\mathrm{sat}} = K(P^\star).$$

The saturated confidence event implies membership in the structured LR region, and hence in the displayed outer set. □

A concrete conservative curvature bound is available from the observed features:

$$\nabla^2 F(\theta) = \sum_{i\in\mathcal{I}_+,x:N_{ix}>0} \frac{N_{ix}\varphi_{ix}\varphi_{ix}^\top}{(\varphi_{ix}^\top\theta)^2} \succeq G := \sum_{i\in\mathcal{I}_+,x\in X_i} N_{ix}\varphi_{ix}\varphi_{ix}^\top,$$

since valid probabilities are at most one. If G is positive definite on the relevant free coordinates, it can supply the required curvature. If G is singular in a direction affecting feasible transitions, it gives no finite likelihood-derived radius in that direction. Adding a ridge to G does not preserve this lower

Hessian inequality and cannot be justified without a parameter bound or another calibrated argument. The exact convex LR oracle remains available even when an ellipsoidal shortcut is unsupported.

Equality elimination must be done explicitly. Writing $\theta = \theta_0 + Bz$ with columns of B spanning $\ker H$ makes the transition representation affine in z and replaces curvature by $B^\top GB$. Unidentifiable directions can be removed only when they leave all relevant transition predictions unchanged, or when independently justified constraints control them. *A small parameter count by itself is not a proof of a dimension-adaptive confidence rate*; the saturated calibration above may still pay the full tabular dimension.

Accounting for two different relaxations. Under the hypotheses of Proposition 16, take $R = \sqrt{2b}$ with the stated Σ, so that $\Theta_b^{\mathrm{LR}} \subseteq \mathcal{C}$. Define the exact structured likelihood image $\mathcal{P}_{\mathrm{LR}} = \{P_\theta : \theta \in \Theta_b^{\mathrm{LR}}\}$, the outer-set image $\mathcal{P}_{\mathcal{C}} = \{P_\theta : \theta \in \mathcal{C}\}$, and its hull $\mathcal{R}_{\mathcal{C}} = \mathrm{Rect}_{sa}(\mathcal{P}_{\mathcal{C}})$. For fixed π, let $J^\pi_{\mathcal{D}} = \min_{P \in \mathcal{D}} J^\pi_P$ for these compact kernel sets. Their inclusions imply

$$J^\pi_{\mathcal{P}_{\mathrm{LR}}} - J^\pi_{\mathcal{R}_{\mathcal{C}}} = \underbrace{J^\pi_{\mathcal{P}_{\mathrm{LR}}} - J^\pi_{\mathcal{P}_{\mathcal{C}}}}_{\text{outer-set approximation}} + \underbrace{J^\pi_{\mathcal{P}_{\mathcal{C}}} - J^\pi_{\mathcal{R}_{\mathcal{C}}}}_{\text{rectangularization}},$$

with both terms nonnegative. Proposition 20 controls the second term. In particular, a repair feasible *only in* $\mathcal{C}$ gives an upper bound on $J^\pi_{\mathcal{P}_{\mathcal{C}}}$, not necessarily on $J^\pi_{\mathcal{P}_{\mathrm{LR}}}$.

To bound the total change, let Q be a worst kernel for $\mathcal{R}_{\mathcal{C}}$ and repair into the *exact* likelihood set:

$$\min_{\theta \in \Theta_b^{\mathrm{LR}}} \frac{1}{2} \sum_i d^\pi_Q(i) \|A_i\theta - Q_i\|_1.$$

This is convex because the negative log likelihood and the objective are convex and the validity constraints are linear. Every feasible θ obeys

$$0 \le J^\pi_{\mathcal{P}_{\mathrm{LR}}} - J^\pi_Q \le J^\pi_{P_\theta} - J^\pi_Q \le \frac{\gamma}{(1-\gamma)^2} \sum_i d^\pi_Q(i)\, \mathrm{TV}(A_i\theta, Q_i).$$

The last inequality is Lemma 8; no return-convexity assumption is used. This exact-likelihood repair retains convex log-likelihood constraints (Boyd & Vandenberghe, 2004, Chapters 3–4); unlike the ellipsoidal repair, no SOCP representation is asserted. Alternatively, row projection and repair can use the exact likelihood set throughout, eliminating the outer-set approximation term. The two-solve binary-mixture construction in Appendix I.4 takes this latter route.

I.2 Valid row projections and conditional norm formulas

Theorem 17 (Valid-kernel projection and oracle)**.** *The smallest (s,a)-rectangular superset of $\{P_\theta : \theta \in \mathcal{C}\}$ has row sets $\{A_i\theta : \theta \in \mathcal{C}\}$. For $g_i(v) = A_i^\top v$, its row objective is the second-order cone program (SOCP)*

$$\psi_{\mathcal{C}}(g) := \min_{\theta \in \mathcal{C}} \theta^\top g, \qquad g = g_i(v). \tag{38}$$

The unconstrained formula $\widehat{\theta}^\top g - R\|\Sigma^{-1/2} g\|_2$ equals (38) *if and only if an unconstrained minimizer lies in Θ. For $R > 0$ and $L \in \mathbb{R}^{d \times q}$, the affine ellipsoid $E_L = \{\widehat{\theta} + Lz : z \in \mathbb{R}^q,\ \|z\|_2 \le R\}$ consists entirely of valid parameters if and only if*

$$H\widehat{\theta} = h, \qquad HL = 0, \qquad \varphi_{ix}^\top \widehat{\theta} \ge R\|L^\top \varphi_{ix}\|_2 \quad \text{for all } (i,x). \tag{39}$$

For such an ellipsoid the exact row formula is $\widehat{\theta}^\top g - R\|L^\top g\|_2$. In either valid construction, contraction and greedy optimality follow as in Section 3.

Proof of Theorem 17. The valid intersection $\mathcal{C}$ is nonempty, convex, and compact. Its image under $\theta \mapsto P_\theta$ is therefore a compact convex set of stochastic kernels. The ith row projection is exactly $A_i\mathcal{C}$ by the definition of projection. Their product is the unique smallest (s,a)-rectangular superset by the same product argument as Proposition 1.

For a fixed row and value vector, $P_{\theta,i}^\top v = \theta^\top A_i^\top v$. Minimizing this linear objective subject to a norm inequality, linear equalities, and linear nonnegativity inequalities is the SOCP in (38). The unconstrained ellipsoid has minimum

$$\widehat{\theta}^\top g - R\|\Sigma^{-1/2} g\|_2.$$

For $R > 0$ and $g \neq 0$ its unique minimizer is $\theta_{\text{ell}} = \widehat{\theta} - R\Sigma^{-1}g/\|\Sigma^{-1/2}g\|_2$. Since $\mathcal{C}$ is a subset of the ellipsoid, equality with the constrained minimum holds exactly when this optimizer is feasible. If $g = 0$ or $R = 0$, a feasible unconstrained minimizer exists automatically and the claim is interpreted accordingly. Merely finding the unconstrained expression is not enough to define a valid robust MDP backup.

For $E_L = \{\widehat{\theta} + Lz : \|z\|_2 \leq R\}$ with $R > 0$, all equality constraints hold throughout the ball if and only if $H\widehat{\theta} = h$ and $HL = 0$, by considering $z = 0$ and opposite directions. Each nonnegativity constraint holds throughout if and only if

$$\min_{\|z\|_2 \leq R} \varphi_{ix}^\top(\widehat{\theta} + Lz) = \varphi_{ix}^\top\widehat{\theta} - R\|L^\top\varphi_{ix}\|_2 \geq 0.$$

This proves (39) and both its necessity and sufficiency. Linear minimization over the wholly valid ball gives $\widehat{\theta}^\top g - R\|L^\top g\|_2$. For $R = 0$, validity reduces only to $\widehat{\theta} \in \Theta$; $HL = 0$ is then unnecessary. Finally, all projected rows are probability distributions, so the contraction, evaluation, and greedy arguments of Appendix A apply. □

An equality-aware norm shortcut. Suppose $H\widehat{\theta} = h$. Let B have orthonormal columns spanning $\ker(H\Sigma^{-1/2})$ and set $L = \Sigma^{-1/2}B$. Then $HL = 0$, $L^\top\Sigma L = I$, and

$$\{\theta : H\theta = h,\ \|\Sigma^{1/2}(\theta - \widehat{\theta})\|_2 \leq R\} = \{\widehat{\theta} + Lz : \|z\|_2 \leq R\}.$$

The equality-constrained ellipsoid therefore has objective $\widehat{\theta}^\top g - R\|L^\top g\|_2$. If $L^\top g \neq 0$, its minimizer is $\widehat{\theta} - RLL^\top g/\|L^\top g\|_2$. One must still verify all nonnegativity inequalities for this point, or verify (39) for the whole ellipsoid. This separates normalization from positivity rather than treating either as automatic.

Lemma 18 (Primal–dual bounds with transition validity)**.** *Stack the nonnegativity rows into A. For any $\lambda \geq 0$ and any equality multiplier y, set $q = g + H^\top y - A^\top\lambda$. Then*

$$D(\lambda, y) := \widehat{\theta}^\top q - h^\top y - R\|\Sigma^{-1/2}q\|_2 \ \leq\ \psi_{\mathcal{C}}(g) \ \leq\ g^\top\theta_f \quad \textit{for every } \theta_f \in \mathcal{C}. \tag{40}$$

Thus an admissible dual pair and a feasible model give a row-objective enclosure even without strong duality or successful solver termination. If, after eliminating affine equalities and inequality faces that hold identically, a point has norm strictly below R and strictly satisfies every remaining inequality, conic strong duality applies. This condition is not needed for the displayed enclosure.

Proof. Add $y^\top(H\theta - h) - \lambda^\top A\theta$ to the objective and minimize over E. At a feasible θ this added term is nonpositive, so the resulting infimum is a lower bound on the constrained minimum. The support function of the ellipsoid gives exactly $D(\lambda, y)$; any feasible primal point gives the upper bound. Relative Slater conditions are used only for the optional equality of optimal primal and dual values, as in Boyd & Vandenberghe (2004, Section 5.2.3). Weak duality and (40) need no such condition. Rounding, multiplier admissibility, and primal feasibility must be accounted for before treating numerical values as endpoints. □

Corollary 19 (When a likelihood ellipsoid needs no validity truncation)**.** *Work in an identifiable parameterization after removing directions that leave all transitions unchanged. Let B span its equality-free directions* $\ker H$*. Define*

$$G = \sum_{i \in \mathcal{I}_+, x \in X_i} N_{ix}\varphi_{ix}\varphi_{ix}^\top, \qquad W = B^\top GB.$$

Assume $W \succ 0$ and let $L = BW^{-1/2}$. The equality-constrained ellipsoid $E_L = \{\widehat{\theta} + Lz : \|z\|_2 \leq \sqrt{2b}\}$ contains Θ_b^{LR} in this parameterization (equivalently, contains every feasible kernel image before the removal). Put

$$R_{\text{valid}} = \min_{(i,x):\|L^\top\varphi_{ix}\|_2 > 0} \frac{\varphi_{ix}^\top\widehat{\theta}}{\|L^\top\varphi_{ix}\|_2}, \qquad \min\varnothing = +\infty. \tag{41}$$

For $b > 0$, the entire E_L is valid exactly when $\sqrt{2b} \leq R_{\text{valid}}$. In that case the likelihood-calibrated outer model admits the exact row formula $\widehat{\theta}^\top g - \sqrt{2b}\|L^\top g\|_2$. Otherwise retain $E_L \cap \Theta$ and solve the constrained row problem. For $b = 0$, the singleton $\{\widehat{\theta}\}$ is valid. If no free coordinates remain, it is a singleton for every radius.

Proof. Along every valid likelihood segment, $\nabla^2 F \succeq G$ since each observed transition probability is at most one. In the coordinates $\theta = \widehat{\theta} + Bz$, the Taylor and constrained first-order argument of Proposition 16 gives $z^\top W z \le 2b$. This proves containment. The identities $HL = 0$ and $H\widehat{\theta} = h$, together with Theorem 17, make nonnegativity equivalent to every inequality in (41). The formula is then the support function on E_L. The zero-radius and no-free-coordinate cases follow directly. □

Validity and coverage are different requirements. Corollary 19 gives a computable sufficient route from likelihood calibration to a wholly valid ellipsoid, not an automatic property of a calibrated radius. If the test fails, reducing $\sqrt{2b}$ to R_{valid} can exclude likelihood-feasible models and the truth. Intersecting with Θ cannot exclude the true parameter on the coverage event because that parameter is already valid. If W is singular, use the exact LR set or an independently justified bound in unidentified directions; adding a ridge is not a likelihood-calibration argument. Known rows and declared structural zeros remain exact constraints in either construction; a positive-radius ambient ball cannot relax them. An empirical boundary estimate with $\varphi_{ix}^\top \widehat{\theta} = 0$ and $L^\top \varphi_{ix} \neq 0$ forces $R_{\text{valid}} = 0$.

Two elementary domain diagnostics. For $A_i = I_2$, $H = (1, 1)$, $\widehat{\theta} = (1/2, 1/2)$, and $\Sigma = I_2$, a bare positive-radius ball gives $\min \theta^\top \mathbf{1} = 1 - R\sqrt{2}$, whereas every stochastic row must give one. Hence even preservation of a constant continuation vector can fail without normalization. Normalization alone is insufficient: the affine interval $\theta = (.9 + t, .1 - t)$, $|t| \le .2$, has minimum second-coordinate prediction $-.1$; imposing nonnegativity gives zero. These are analytic counterexamples to dropping constraints, not empirical baselines. When the whole affine ellipsoid passes (41), the feature-norm expression and constrained backup agree exactly. Existing forest experiments use exact valid likelihood intervals, not an unchecked ellipsoid. Deterministic checks of the constrained oracle, including non-target-row validity, appear in Appendix K.5.

I.3 Valid-kernel comparator and sharing bounds

Proposition 20 (Sensitivity bounds and a structured repair)**.** *Let $P^\star = P_{\theta^\star}$ with $\theta^\star \in \mathcal{C}$, and define $D_i = \|A_i \Sigma^{-1/2}\|_{2\to 1}$, the induced norm from ℓ_2 to ℓ_1. For an ϵ_{alg}-optimal policy $\widehat{\pi}$ in the rectangular hull of $\{P_\theta : \theta \in \mathcal{C}\}$ and any comparator $\pi^\dagger$,*

$$J_{P^\star}^{\pi^\dagger} - J_{P^\star}^{\widehat{\pi}} \le \epsilon_{\text{alg}} + \frac{\gamma}{(1-\gamma)^2} \sum_i d_{P^\star}^{\pi^\dagger}(i) \min\{1, RD_i\}.$$

For fixed π and a rectangular worst kernel Q, the sharing gap obeys

$$\begin{aligned} 0 &\le \min_{\theta \in \mathcal{C}} J_{P_\theta}^\pi - J_Q^\pi \\ &\le \frac{\gamma}{(1-\gamma)^2} \min_{\theta \in \mathcal{C}} \sum_i d_Q^\pi(i)\, \mathrm{TV}(A_i\theta, Q_i) \le \frac{\gamma}{(1-\gamma)^2} \sum_i d_Q^\pi(i) \min\{1, RD_i/2\}. \end{aligned}$$

The middle minimization is (29)*; every feasible parameter also gives an evaluable upper endpoint for the shared robust value.*

Proof. For $\theta, \theta' \in \mathcal{C}$,

$$\mathrm{TV}(A_i\theta, A_i\theta') \le \tfrac{1}{2} D_i \|\Sigma^{1/2}(\theta - \theta')\|_2 \le RD_i.$$

Together with TV at most one, (21), and the pessimistic comparator argument of Appendix H, this proves the first result. For the sharing gap use Proposition 10. The feasible center $\widehat{\theta}$ satisfies $\mathrm{TV}(A_i\widehat{\theta}, Q_i) \le RD_i/2$ because $Q_i = A_i\theta_i$ for some $\theta_i \in \mathcal{C}$. Substituting this center proves the final bound. The comparison uses transition distances; no linearity of return in the parameter is assumed. □

The $2 \to 1$ operator norm is a mathematical sensitivity quantity; computing it exactly can itself be nontrivial in high dimension. The readily computable bound $D_i \le \sqrt{|X_i|}\|A_i\Sigma^{-1/2}\|_{2\to 2}$ follows from $\|z\|_1 \le \sqrt{|X_i|}\|z\|_2$. Alternatively, evaluating a feasible solution of (29) avoids this norm and returns an explicit kernel bound.

I.4 A SHARED BINARY-MIXTURE SPECIALIZATION

A one-parameter model makes the computational reduction explicit without an ellipsoidal approximation. For each unknown row i, fix two distinct successors x_i, y_i in advance and let

$$P_{p,i} = (1-p)e_{x_i} + pe_{y_i}, \qquad p \in [0,1],$$

where e_x is the point mass at x; known rows are independent of p. This is a valid linear mixture with parameter $(1-p, p)$ on a one-dimensional simplex. The same binary event, observing y_i rather than x_i, has probability p in every unknown row. Let $n_1 = \sum_{i\in\mathcal{I}_+} N_{iy_i}$, $n = \sum_{i\in\mathcal{I}_+} N_i \geq 1$, and $\bar{p} = n_1/n$. Define

$$\mathrm{kl}(a\|p) := a\log(a/p) + (1-a)\log((1-a)/(1-p))$$

with the usual extended-real conventions, and let $F(p)$ be the negative transition log likelihood in this shared model.

Proposition 21 (Pooled likelihood and endpoint planning)**.** *The constrained likelihood region is the interval*

$$I_b = \{p \in [0,1] : F(p) - F(\bar{p}) \leq b\} = \{p : n\,\mathrm{kl}(\bar{p}\|p) \leq b\} = [p_-, p_+].$$

Its endpoints can be computed by at most two monotone binary-KL solves. The row minimum in the state–action rectangular hull of $\mathcal{U}_b^{\mathrm{B}} = \{P_p : p \in I_b\}$ is

$$\min_{p\in I_b} P_{p,i}^\top v = \min\{(1-p_-)v(x_i) + p_- v(y_i),\ (1-p_+)v(x_i) + p_+ v(y_i)\}.$$

Thus no row-specific nonlinear solve is needed during value iteration after the endpoints are computed. Under an exogenous fixed row design with independent successors, set

$$b_{\mathrm{B}} = 1 + \sqrt{2\log(1/\beta)} + \log(1/\beta).$$

Then $I_{b_{\mathrm{B}}}$ contains each fixed true $p^\star$ with probability at least $1-\beta$, conditionally on the row labels.

Fix a stationary policy π and let v be its rectangular robust value for this interval model. If all nonzero slopes $v(y_i) - v(x_i)$ on unknown rows with $\pi(a \mid s) > 0$ have the same sign, the coupled and rectangular policy values agree for every μ. If $p_- < p_+$, μ has full support, and both slope signs occur on such rows, the coupled value is strictly larger.

Proof. The likelihood is $p^{n_1}(1-p)^{n-n_1}$, so its maximizer is $\bar{p}$ and subtraction gives $n\,\mathrm{kl}(\bar{p}\|p)$. Convexity and monotonicity on either side of $\bar{p}$ give the interval and its endpoint solves. A row expectation is affine in p, so one endpoint attains its minimum. Under the stated design, the binary indicators are independent Bernoulli$(p^\star)$ even though their row labels can differ. Applying Agrawal (2020, Theorem I.3) with alphabet size two and the Chernoff calculation in Appendix G proves the threshold.

For the final claim, all slopes of one sign have a common minimizing endpoint; zero slopes impose no restriction. The kernel at that endpoint makes every policy-relevant row residual vanish and is coupled-feasible. The simulation identity therefore gives exactness. Conversely, two opposite nonzero slopes require different endpoints. No shared parameter can minimize both rows. Under full-support initialization both rows have positive occupancy for every feasible kernel. The nonnegative residual identity then excludes a zero gap at any shared parameter; continuity on the compact interval makes the minimum gap strictly positive. □

At $\bar{p} = 0$, $I_b = [0, 1 - e^{-b/n}]$; at $\bar{p} = 1$, $I_b = [e^{-b/n}, 1]$. With no observations the interval is $[0,1]$ and no expression involving $1/n$ is used. The binary threshold is justified by *this model's pooled sufficient statistic*, not by substituting a parameter dimension into a general tabular threshold. At adaptive stopping, a single binary Dirichlet mixture on the pooled counts gives the corresponding anytime budget by the same predictable-sampling proof as Appendix G; the fixed-sample budget cannot simply be reused. This specialization illustrates the bridge using elementary likelihood and interval geometry, rather than claiming a new scalar-optimization principle.

I.5 ALIGNMENT IS A FROZEN-SURROGATE STATEMENT

The proof uses concavity of a pointwise infimum of linear functions and the support function of an ellipsoid (Boyd & Vandenberghe, 2004, Chapter 3). Its role is to delimit what a fixed-surrogate alignment calculation implies; these convex-analytic identities are not new on their own.

Proposition 22 (Constrained and unconstrained alignment)**.** *Fix a continuation vector v and $\nu \in \Delta(\mathcal{I})$, and write $\mathbb{E}_\nu f_i = \sum_i \nu_i f_i$, $g_i = A_i^\top v$, and $\bar{g} = \mathbb{E}_\nu g_i$. For the validity-constrained ellipsoid $\mathcal{C}$,*

$$0 \le \psi_{\mathcal{C}}(\bar{g}) - \mathbb{E}_\nu \psi_{\mathcal{C}}(g_i) \le R\mathbb{E}_\nu \|\Sigma^{-1/2}(g_i - \bar{g})\|_2 \le R\sqrt{\mathbb{E}_\nu \|\Sigma^{-1/2}(g_i - \bar{g})\|_2^2}. \tag{42}$$

Replacing $\mathcal{C}$ by a wholly valid E_L, the difference equals $R(\mathbb{E}_\nu \|L^\top g_i\|_2 - \|L^\top \bar{g}\|_2)$.

Proof. The function $\psi_{\mathcal{C}}(g) = \min_{\theta \in \mathcal{C}} \theta^\top g$ is concave, so Jensen's inequality gives $\psi_{\mathcal{C}}(\bar{g}) \ge \mathbb{E}_\nu \psi_{\mathcal{C}}(g_i)$. To bound the difference, center it at $\widehat{\theta}$: $f(g) = \psi_{\mathcal{C}}(g) - \widehat{\theta}^\top g = \min_{\theta \in \mathcal{C}} (\theta - \widehat{\theta})^\top g$. For any g, g', Cauchy–Schwarz gives

$$|f(g) - f(g')| \le R\|\Sigma^{-1/2}(g - g')\|_2.$$

The linear center terms cancel in the Jensen gap, whence

$$\psi_{\mathcal{C}}(\bar{g}) - \mathbb{E}_\nu \psi_{\mathcal{C}}(g_i) = \mathbb{E}_\nu[f(\bar{g}) - f(g_i)] \le R\mathbb{E}_\nu \|\Sigma^{-1/2}(g_i - \bar{g})\|_2.$$

Cauchy–Schwarz proves the second upper bound in (42). For a wholly valid E_L, insert the exact support formula from Theorem 17; the linear terms cancel and the displayed identity follows. □

For a wholly valid ellipsoid with $R > 0$, shared and independent minima agree exactly when the nonzero vectors $L^\top g_i$ with $\nu_i > 0$ are nonnegative multiples of one vector; all-zero vectors also give equality. This is the equality condition in the Euclidean triangle inequality. At $R = 0$, the two minima always agree. With validity constraints active, the exact norm identity may fail but the centered Lipschitz inequality remains valid. This conclusion concerns $\min_\theta \mathbb{E}_\nu \theta^\top g_i$ versus $\mathbb{E}_\nu \min_\theta \theta^\top g_i$ *for frozen inputs*. In an MDP, the actual return is a rational function of transitions and the occupancy depends on the model. The main paper's return bounds therefore use Lemma 8 instead.

Relation to an adversarial model-fitting penalty. For fixed $\lambda \ge 0$, consider the Lagrangian value

$$\inf_{\theta \in \Theta:\, F(\theta) < \infty} \{J^\pi_{P_\theta} + \lambda(F(\theta) - F(\widehat{\theta}) - b)\}.$$

It lower-bounds $\min_{\theta \in \Theta_b^{\mathrm{LR}}} J^\pi_{P_\theta}$: the penalty is nonpositive at every likelihood-feasible parameter. This resembles the data-fit-regularized return minimization in RAMBO (Rigter et al., 2022). Strong duality is not automatic because $J^\pi_{P_\theta}$ is generally nonconvex in θ. If return is replaced by the frozen linear surrogate and a suitable relative Slater condition holds, convex strong duality can apply. The paper makes neither a global strong-duality claim for nonlinear return nor a claim that practical alternating updates compute the exact coupled optimum.

J COMPUTATIONAL DETAILS AND INTERPRETATION OF THE GUARANTEES

J.1 A COMPLETE TABULAR WORKFLOW

The computational task should be specified before choosing a numerical method. Rectangular planning solves a different problem from evaluating or optimizing a policy against one coupled kernel. Algorithm 1 returns a rectangular policy and bounds on its coupled robust value; it does not optimize the coupled return.

For $n = 0$, all unknown rows are uncovered; skip formulas involving N_i^{-1}, n^{-1}, or empirical coverage ratios. The uncertainty set is already a product and the rectangularization gap is zero, although pessimism can be extreme and the true-policy comparator guarantee can be vacuous. Known rows and one-successor alphabets should not be treated as uncertain numerical optimization problems.

Algorithm 1 Likelihood-based planning with approximation bounds

Require: Counts, predeclared alphabets, known rows, r, γ, μ, level β, and numerical tolerances.

1: Calibrate b for the actual sampling law (Theorem 7); form covered-row radii b/N_i for $i \in \mathcal{I}_+$, full uncovered-row simplexes, and fixed known rows.
2: Plan using the reverse-KL oracle and Proposition 13; extract a greedy policy π.
3: Evaluate π robustly. Retain an approximate value v, a feasible row kernel $Q \in \mathcal{R}_b$, and a residual-corrected bound $L_\pi \le J^\pi_{\mathrm{R}}$ (Lemma 14).
4: Compute d^π_Q by solving its discounted occupancy equation. Repair Q using (7); verify the final $K(P) \le b$ before evaluating $U_\pi = J^\pi_P$.
5: Construct $\underline{d}$, solve the dual of (9), and correct for the residual of v to obtain $\underline{G}_\pi \le G_\pi$ (Proposition 11).
6: Return $[L_\pi + \underline{G}_\pi, U_\pi]$ for J^π_{LR}. Enclose linear-system errors separately when evaluation is approximate.
7: Optionally compute the policy-selection bound (11), using Corollary 12 for approximate values.

J.2 Distinguishing approximation bounds from statistical guarantees

Table 2: Quantities that must not be interchanged. "Deterministic" means for each fixed dataset and valid numerical enclosure; it is not a conditional coverage probability.

Quantity	Guaranteed statement	What it does not establish
Corrected rectangular value L_π	Deterministically $L_\pi \le J^\pi_{\mathrm{R}} \le J^\pi_{\mathrm{LR}}$; on model containment, also $L_\pi \le J^\pi_{P^\star}$.	A tight estimate of the coupled value or of true return.
Feasible repaired value J^π_P	Deterministically $J^\pi_{\mathrm{LR}} \le J^\pi_P$.	A pessimistic lower bound for true return.
Lower residual dual	A lower bound on G_π, after value-residual correction.	Exactness when the bound is zero; zero may be inconclusive.
Bellman policy bound	An upper bound on $J^\star_{\mathrm{LR}} - J^{\pi_{\mathrm{R}}}_{\mathrm{LR}}$.	Statistical comparison to $P^\star$ without coverage, or necessity when nonzero.
Frozen alignment gap	A shared-versus-row-wise difference for one fixed linear surrogate.	The nonlinear adversarial return gap or convergence of model-fitting updates.

Interpreting the bounds. Table 2 separates deterministic and statistical statements. Let $\underline{G}_\pi \ge 0$ lower-bound G_π and let $P \in \mathcal{U}_b$. A *small upper bound on* G_π shows that rectangularization has little effect on that policy's robust value. By contrast, a narrow interval $[J^\pi_{\mathrm{R}} + \underline{G}_\pi, J^\pi_P]$ can locate the coupled value accurately even when the gap is large. Neither statement establishes policy optimality: Appendix E.3 has exact evaluation but a coupled-suboptimal selected policy. A positive lower gap bound establishes a change in the robust objective, not poor true return. Statistical performance additionally requires the model-containment event and the relevant comparator assumptions.

J.3 Complexity and arithmetic requirements

A dense policy-evaluation or occupancy solve costs $O(|\mathcal{S}|^3)$ by standard elimination. Sparse iterative evaluation is also possible, but its residual must be converted into a value enclosure. If each scalar row oracle uses at most B evaluations and $E = \sum_i |X_i|$ is the number of admissible transitions, one Bellman sweep costs $O(EB)$. This is an implementation count, not a uniform bit-complexity theorem: B depends on the objective tolerance, numerical scale, empirical probabilities, and support branch. For exact row objectives and initial error $e_0 = \|v_0 - V^\star_{\mathrm{R}}\|_\infty > 0$, a target error $\varepsilon > 0$ is achieved after

$$k \ge \max\left\{0, \left\lceil \frac{\log(e_0/\varepsilon)}{-\log\gamma} \right\rceil \right\}$$

sweeps; no sweep is needed if $e_0 = 0$. Appendix F includes row-error accumulation.

Mixture repair uses one shared multiplier and separable scalar row solves; the lower residual program uses a shared multiplier and scalar normalization duals. General valid-kernel repair, linear-mixture row optimization, and the policy-selection program can be passed to a suitable convex solver. An exponential-cone representation handles the reverse-KL epigraph; the validity-constrained ellipsoidal row oracle is an SOCP, whereas an exact likelihood region generally retains exponential-cone constraints. These are distinct from minimizing the nonlinear discounted return over a coupled model class. Standard convex duality is sufficient for the convex programs (Boyd & Vandenberghe, 2004); it is not invoked for the nonlinear return problem.

Every reported bound needs the correct numerical direction. Feasible coupled kernels give upper values. Lower-program dual points give lower objectives. Bellman residuals enclose fixed-point error. Binary-root lower endpoints give lower occupancy entries. For a machine-arithmetic guarantee, use directed rounding or interval arithmetic throughout logarithms, roots, sums, and linear solves. Ordinary floating-point residuals require separate rounding-error bounds to imply the same guarantee. Extremely large likelihood budgets can create probabilities below floating-point range; the objective shortcuts in Appendix F avoid silently setting observed probabilities to zero.

J.4 Proof dependencies and the scope of the contribution

The results combine established statistical and optimization tools with a *likelihood-specific approximation analysis*. The following distinctions identify what is inherited and what the present proofs establish.

Likelihood geometry and Bellman planning. CPPO-LR supplies the coupled offline objective (Uehara & Sun, 2023); likelihood-based model regions and rectangular robust dynamic programming are established (Wiesemann et al., 2013; Iyengar, 2005; Nilim & El Ghaoui, 2005). Proposition 1 identifies the exact row projections of one global saturated-likelihood region, including covered, uncovered, and known rows. The fixed-point and greedy-policy proofs instantiate classical theory for that identified set. They do not establish tractability of the unrelaxed coupled objective.

Evaluation and policy approximation. Lemma 8 is a standard simulation identity, and the primal–dual arguments use standard convex optimization (Uehara & Sun, 2023; Boyd & Vandenberghe, 2004). The approximation results are the occupancy-weighted feasible repair in Theorem 4, the complementary lower program in Proposition 5, and their relation to the reverse-KL exactness criterion in Theorem 2. The criterion and its initialization-aware extension (Theorem 3) specialize compatibility to a fixed likelihood region, distinct from general simultaneous-solvability theory (Grand-Clément et al., 2026). The separate Bellman program in Proposition 6 addresses policy loss, which a fixed-policy gap alone does not control. The analytical constructions show when these distinctions affect the answer.

Scalar optimization and numerical error. The positive-reference Burg-entropy root is already explicit in Adam & Mácha (2022, Theorem 3.3, arXiv version); efficient divergence-robust planning is studied by Ho et al. (2022). Appendix F supplies the empirical-zero cases and objective enclosures required by this likelihood model, then uses contraction arguments to propagate their accuracy. *Efficient scalar optimization alone is not the claimed theoretical advance.*

Statistical calibration. The fixed-sample proof applies the multinomial exponential-moment bound of Agrawal (2020, Theorem I.3). The trajectory proof applies Dirichlet mixture integration and Ville's inequality (Krichevsky & Trofimov, 1981; Lindon & Malek, 2022; Howard et al., 2020). These are established concentration tools; the statistical consequence here is their combination with the explicit decomposition of coupled uncertainty, rectangularization error, and numerical error under the stated sampling and comparator assumptions.

Linear-mixture geometry. Robust RL with structured transitions is established: Liu & Xu (2025) study linear mixtures, whereas Tang et al. (2025) study linearly structured f-divergence regularization. The extension identifies the validity constraints needed for row projections and support-function formulas, and transfers the occupancy-sensitive approximation analysis to one shared parameter set. Ellipsoid support calculations and Jensen inequalities are standard convex analysis (Boyd &

Vandenberghe, 2004). Their frozen-surrogate interpretation does not imply a nonlinear-return identity or convergence of RAMBO-style model fitting (Rigter et al., 2022).

J.5 Detailed related work and statistical context

Table 3 and the discussion below distinguish *optimization objectives*, representations, and sampling assumptions in offline and robust RL.

Table 3: **Which uncertainty is optimized?** Our comparison fixes a coupled likelihood objective and quantifies the error of its tractable rectangular relaxation.

Formulation	Core object	Relation to this work
CPPO-LR (Uehara & Sun, 2023)	One likelihood-constrained model	Supplies the coupled reference objective and its partial-coverage analysis.
Bellman-consistent pessimism (Xie et al., 2021)	Consistent value functions	Controls value-space uncertainty; we quantify relaxation of a transition-likelihood set.
ATAC / ARMOR (Cheng et al., 2022; Bhardwaj et al., 2023)	Worst-case relative performance	Relative pessimism supports reference-policy improvement; our objective is absolute worst-case return.
Divergence-robust planning (Ho et al., 2022)	Rectangular uncertainty	Supplies efficient backups; shared likelihood constraints need not be rectangular.
This work	A coupled set, its hull, and feasible repairs	Two-sided evaluation bounds, likelihood-specific exactness, and a separate policy-loss bound.

Likelihood confidence is not the same as likelihood estimation risk. Our statistical object is a set containing each fixed true transition kernel with prescribed frequentist probability. Bilodeau et al. (2023) instead characterize minimax conditional-density estimation under population KL risk through empirical Hellinger entropy, and identify settings in which maximum likelihood is suboptimal for that risk. Their results do not make a calibrated saturated-likelihood region invalid, nor does our coverage theorem establish *optimal KL risk* for the empirical model. The direction and role of divergence matter: $K(P) = \sum_i N_i D_{\mathrm{KL}}(\widehat{p}_i \| P_i)$ is an observed likelihood contrast, whereas population prediction risk averages $D_{\mathrm{KL}}(P_i^\star \| \widehat{P}_i)$ under a sampling law. The finite-alphabet proof uses a multinomial exponential-moment bound; no general nonparametric likelihood rate is inferred from it.

Representation assumptions determine what coverage can buy. Chen & Jiang (2019) analyze the interaction between coverage and function approximation. Foster et al. (2022) show that standard concentrability and value-function realizability can still require sample complexity polynomial in the state-space size. This is a barrier for the stated value-function representation, not a claim that every tabular or realizable transition-model problem is intractable. Positive results change relevant assumptions: Xie & Jiang (2021) use stronger exploratory coverage, while Zhan et al. (2022) combine value and density-ratio realizability with single-policy concentrability. Here transitions are represented directly by tabular probabilities or a specified valid linear-mixture model. These are additional structural assumptions, and our linear-mixture sensitivity inequality alone is not a dimension-dependent statistical rate. The required confidence radius and identifiable directions must be justified separately.

Computational pessimism already has several forms. Jin et al. (2021) analyze pessimistic value iteration; Xie et al. (2021) impose pessimism on Bellman-consistent value functions and develop a policy-iteration approach. ATAC uses a relative-pessimism actor–critic game (Cheng et al., 2022), and ARMOR adversarially optimizes a model-based relative-performance objective (Bhardwaj et al., 2023). Their computational results are not contradicted by the coupling in our likelihood region.

Even within model-based methods, the objectives differ: for a reference policy π_0, generally

$$\min_{P\in\mathcal{U}_b}(J_P^\pi - J_P^{\pi_0}) \neq \min_{P\in\mathcal{U}_b} J_P^\pi - \min_{P\in\mathcal{U}_b} J_P^{\pi_0}.$$

For example, if two admissible models give value pairs $(J_P^\pi, J_P^{\pi_0}) = (0,0)$ and $(1,2)$, the two expressions are -1 and 0. Thus an approximation guarantee for absolute worst-case return does not automatically inherit a robust reference-policy-improvement guarantee. Our comparison fixes the *absolute CPPO-LR objective* (Uehara & Sun, 2023); the new step is to bound the change caused by row-wise relaxation of that same set.

Robust planning and coupled uncertainty. Classical theory supplies rectangular Bellman recursion (Iyengar, 2005; Nilim & El Ghaoui, 2005; Wiesemann et al., 2013). Simultaneous solvability and non-rectangular policy gradients address broader tractability questions (Grand-Clément et al., 2026; Li et al., 2026; Satheesh et al., 2026); distributionally robust offline and linear-mixture methods study related uncertainty objectives (Shi & Chi, 2024; Panaganti et al., 2025; Liu & Xu, 2025; Tang et al., 2025). We instead fix one likelihood region and analyze its row-product relaxation. Its minimality is *among rectangular supersets of that region*, not all tractable methods. Budget saturation and reachability identify exactness; occupancy quantifies failure.

Offline occupancy is not online coverability. Xie et al. (2023) study how the existence of a suitable covering distribution enables online exploration, and Amortila et al. (2024) develop online learning through density-ratio representations, including a role for offline learning in their analysis. Our empirical coverage ratio instead refers to the *given dataset* and one comparator's occupancy. We neither choose an exploration distribution nor guarantee that the observed counts realize a favorable cover. Predictable-trajectory calibration makes likelihood inference valid under adaptive collection; it is *not an exploration or regret theorem.*

Linear transition models are not linear value approximation. Amortila et al. (2026) analyze linear off-policy evaluation and show why feature dynamics, rather than feature covariance alone, enter the relevant coverage condition. Section 5 assumes a specified transition representation $P_{\theta,i} = A_i\theta$ and minimizes linear continuation objectives over valid parameter sets. It does not derive an off-policy-evaluation guarantee from value realizability or a covariance condition alone. This is why the curvature requirement, normalization constraints, and nonnegativity constraints are explicit, and why the exact binary-mixture specialization is stated separately from a general ellipsoidal approximation.

J.6 Assumptions that must be checked in an application

Model specification and support. Finite-state calibration presumes a fixed, correctly specified conditional transition law. An empirical zero is not evidence of impossibility. An aggregated representation of a continuous environment need not be Markov, even when the original environment is Markov; applying the theorem after arbitrary discretization requires a separate justification. Unmodeled reward uncertainty is not included in the present confidence set.

Stationary policy class. The coupled optimum is over stationary randomized policies. History-dependent policies that infer a latent model while interacting can define a different robust decision problem. The rectangular optimum is attained by a deterministic stationary policy, but no such restriction is imposed on the coupled optimizer without proof. Likewise, a componentwise worst-case value vector for a coupled set need not be realized by one kernel for all initial states. *The scalar objective fixes μ throughout.*

Confidence construction. The fixed-sample theorem conditions on exogenous row labels and uses independent conditional multinomial samples; the row design may be fixed in advance or randomized independently of successor outcomes. Stopping a trajectory when selected counts reach a target does not, by itself, justify that i.i.d. proof. The trajectory mixture theorem instead works with predictable row selection and can be stopped adaptively. A known transition row can be omitted from both the likelihood ratio and the mixture. The mixture is fixed in advance; validity holds for each fixed true kernel, not only under a prior average. Choosing a smaller empirical alphabet, changing this mixture after observing data, or selecting among thresholds using successor outcomes requires an additional coverage argument.

Structured confidence. The same geometry applies to any compact convex set of valid shared kernels: take row projections, compute a rectangular policy, and repair against the original coupled set. An ellipsoid is an optional outer approximation, not a license to drop normalization or nonnegativity. If curvature is unavailable in an important direction, retain the exact constrained likelihood set or explicitly account for additional regularization; an unsupported radius can invalidate every downstream statistical statement.

J.7 Structured repair and frozen-surrogate alignment

What an alignment statement can establish. Robust Adversarial Model-Based Offline RL (RAMBO) fits a shared adversarial dynamics model using return and data fit (Rigter et al., 2022). Proposition 22 isolates parameter sharing for fixed continuation values v and row weights ν: inequality (42) compares one shared linear minimization with independently minimized row objectives. Its norm identity applies to a wholly valid affine ellipsoid. These formulas concern fixed ν and v, not nonlinear return or convergence of RAMBO updates. Changing a model changes its occupancy and continuation values; return comparisons therefore use simulation identities, not Jensen's inequality on kernels.

Repairing into the valid shared model. The tabular row-mixture repair need not preserve a shared feature parameter. Instead, use the convex program

$$\min_{\theta\in\mathcal{C}} \frac{1}{2}\sum_i d_Q^\pi(i)\|A_i\theta - Q_i\|_1. \tag{43}$$

Every feasible parameter gives an upper endpoint $J^\pi_{P_\theta}$ for the shared robust value, and its objective times $\gamma/(1-\gamma)^2$ bounds the sharing gap. An ℓ_1 epigraph makes (43) an SOCP. It repairs into the *valid shared set*; it does not assume that averaging parameters averages likelihoods or returns. The same residual lower program works with $P \in \{P_\theta : \theta \in \mathcal{C}\}$ and any uniform occupancy lower envelope (Appendices C and D).

J.8 Notation reference

Counts and confidence budgets are unnormalized unless explicitly divided by n.

Symbol	Meaning
$i=(s,a)$; X_i	A transition row; its predeclared successor alphabet.
$\mathcal{I}_+, \mathcal{I}_0, \mathcal{I}_K$	Covered unknown, uncovered unknown, and known rows.
$N_i, \widehat{p}_i, \widehat{\rho}_i$; n, m	Row count, empirical law, row proportion; total unknown-row observations and number of covered rows.
$b, K(P)$	Global LR budget and unnormalized empirical reverse-KL cost.
$\mathcal{U}_b, \mathcal{R}_b$	Coupled LR set and its smallest (s,a)-rectangular superset.
$\phi_i(v), T^\pi_{\mathrm{R}}, T_{\mathrm{R}}$	Row minimum, robust policy operator, robust optimality operator.
$J^\pi_{\mathrm{LR}}, J^\pi_{\mathrm{R}}, G_\pi$	Coupled value, rectangular value, and their nonnegative difference.
J, G_π	Normalized return $(1-\gamma)J$ and normalized gap $(1-\gamma)G_\pi$.
$\mathcal{C}_\pi(v), Z_j^\infty$	Critical selected rows and the greatest closed minimizing-row set for candidate j (Theorem 3).
$Q, P(\alpha)$	An assembled rectangular adversary and a feasible repaired adversary.
$d^\pi_P, \underline{d}$	Normalized discounted occupancy and its uniform lower envelope.
$B_\pi(Q), A_{\underline{d}}(v)$	Convex repair cost and convex residual lower bound.
Δ_B	Minimum Bellman residual; $\Delta_B/(1-\gamma)$ bounds coupled policy loss.
$\Theta, \Theta_b^{\mathrm{LR}}$	Valid parameter polyhedron and exact centered likelihood region.
$\mathcal{C}, E_L$	Validity-constrained ellipsoid and affine ellipsoid (valid under (39)).
$\psi_\mathcal{C}, \mathcal{P}_\mathcal{C}, \mathcal{R}_\mathcal{C}$	Linear minimum over $\mathcal{C}$, its shared-kernel image, and the rectangular hull of that image.
$\beta, b_t, \mathcal{U}_t(b_t)$	Failure level, time-dependent likelihood budget, and its region using counts through time t.
$\delta, \xi, \epsilon_{\mathrm{alg}}$	Row-objective error bound, computed Bellman residual bound, and tolerance on rectangular policy loss.

K EMPIRICAL PROTOCOL AND COMPLETE RESULTS

The empirical scope is the three components in Section 6: matched-objective rectangular planning, same-region approximation bounds, and common-parameter compatibility. Every sampled family retains *all specified conditions and all 20 datasets per condition*; the separate deterministic implementation check retains all 36 prescribed cases. This appendix gives the data and accuracy contracts; Appendix L derives the coupled reference and the occupancy diagnostics.

K.1 SAMPLING, SUPPORT, AND IMPLEMENTATION

Unknown rows retain their predeclared alphabets, including unobserved successors; singleton and absorbing rows are known. The learner receives counts, support, rewards, and model structure, but not the true kernel. Samples are independent multinomials at exogenous row counts, with $\gamma = .9$ and $\beta = .05$. Timing uses the saturated tabular budget from Theorem 7; the shared binary study uses Proposition 21. Reported returns, gaps, and planning-loss tolerances use the normalization $1 - \gamma$. Likelihood budgets and optimization-objective tolerances retain their stated units.

The slippery FrozenLake maps are from Gymnasium (Towers et al., 2025), with continuing absorbing-goal reward one and hole reward zero; their values are not episodic success rates. The forest equations follow MDP Toolbox (Python MDP Toolbox contributors, n.d.): waiting resets age to zero with probability p and otherwise increments it, capped at the oldest state; cutting resets age deterministically. Forest rewards use $r_1 = 4, r_2 = 2$, divided by four, $p = .1$, and initial state zero.

The recorded platform is an Apple M2 Pro, macOS 26.5, Python 3.13.13, NumPy 2.3.5, SciPy 1.17.0, and Gymnasium 1.3.0. The conic baseline uses CVXPY 1.9.2 (Diamond & Boyd, 2016; Agrawal et al., 2018) and Clarabel 0.11.1 (Goulart & Chen, 2024). Both methods use one worker and one BLAS thread; the dependency lock records the execution environment. These are implementation-specific timings, not a comparison with the best specialized robust-MDP solver (Ho et al., 2026).

Balanced designs allocate N observations to every unknown row. Lake imbalance gives left-action rows weight 10 and others weight 1; forest imbalance gives wait row s weight $2^{-s/3}$. We rescale to the same total sample count and apply largest remainders, breaking ties by row index. Timing uses $N = 250$ and seeds 12000–12019 in each of eight conditions; forest uses $N \in \{25, 250, 2500\}$ and seeds 0–19 in all twelve conditions. The recurrent study uses seeds 11000–11019. Timing/forest pilot seed 11999 and recurrent pilot seed 10999 are excluded. Study selection followed exploratory work; pilot tuning did not use the new timing or repair cohorts. The forest accuracy rerun is not independent replication.

K.2 MATCHED PLANNING: SEPARATING COMPUTATION FROM ACCURACY

Both row solvers use the same Bellman loop and handling of known, uncovered, and constant-value rows. Task and paired-method order are randomized with fixed seeds. One cold construction/solve precedes three warm solves, each restarting the value at zero. Clarabel reuses solver structure, not a previously optimal value iterate. A row is accepted only when reconstructed feasible-primal/admissible-dual objectives meet the common accuracy target. Every tolerance refinement and retry is timed. Table 1 reports construction-plus-first-solve totals separately from warm times; startup, independent verification, data generation, and gap-bound computation are excluded.

Independent checks use normalized accuracy 10^{-7}. If v has backup residual ξ and row width δ, the normalized optimal value is at most $(1 - \gamma)\mu^\top v + \xi + \gamma\delta$, plus the recorded arithmetic allowance. Subtracting a separately verified policy lower value gives a control-loss upper bound. All 1,280 solves over 160 datasets meet the requested 10^{-3} target; the maximum checked control-loss upper bound is $2.79 \cdot 10^{-6}$. The speed comparison therefore holds *the rectangular objective and its accepted policy accuracy fixed*.

K.3 RECURRENT LIKELIHOOD BOUNDS

The recurrent model and its three starts are defined in Section 6. Counts are (N, N) or $(N/5, 9N/5)$, for $N \in \{25, 250, 2500\}$, giving 120 independent datasets and 360 dependent evaluations. Planning targets 10^{-8} normalized accuracy; repair and lower-program tolerances are 10^{-10}. Likelihood

feasibility is checked at 70 digits, with any required inward correction made before recomputing the repaired value. All 1,080 saved primal witnesses pass this check. Appendix L.1 gives the covering-partition argument for all coupled reference lower endpoints.

Table 4 reports every condition, averaging starts within datasets before aggregation. The minimum of two feasible upper endpoints is a valid upper endpoint; its nonnegative improvement over either constituent is by construction. Its *magnitude* and remaining distance from the independently enclosed gap are empirical. Both constituents are retained: occupancy-only repair has a larger direct upper value in 70/360 evaluations, as is possible because the optimized objective is a TV surrogate. No evaluation is removed or relabeled by its outcome.

Counts / N	Lower	Reference	Uniform upper	Occupancy upper	Minimum upper	Upper slack
B / 25	0.02896	0.07039	0.09144	0.08917	0.08355	0.01316
B / 250	0.02653	0.03633	0.04122	0.03648	0.03648	0.00016
B / 2500	0.00990	0.01095	0.01240	0.01099	0.01097	0.00002
I / 25	0.00897	0.02979	0.11422	0.07161	0.06904	0.03925
I / 250	0.01973	0.03931	0.07130	0.04108	0.04108	0.00177
I / 2500	0.00996	0.01262	0.01966	0.01262	0.01262	0.00001

Table 4: All six recurrent conditions, with all 20 datasets per condition. Means first average the three dependent starts within a dataset. Both constituent upper bounds, their minimum, and minimum-upper slack are shown. Unrounded values, numerical enclosures, and paired-dataset sampling intervals are available in the accompanying code.

K.4 SHARED-MODEL COMPATIBILITY

All 240 forest datasets use normalized planning accuracy 10^{-7}. Pooling the observed binary fire outcomes gives one likelihood interval, used consistently for planning and verification. The evaluated policy is the output of the shared planner. A feasible common endpoint supplies an unnormalized ordinary value vector v. Let ϵ_P and ϵ_B bound its raw ordinary and fixed-policy rectangular Bellman residuals. Enclose the exact likelihood endpoints by $p_{\mathrm{lo,out}} \le p_- \le p_{\mathrm{lo,in}}$ and $p_{\mathrm{hi,in}} \le p_+ \le p_{\mathrm{hi,out}}$, and define

$$\epsilon_{\mathrm{end}} = \mathrm{span}(v) \max\{p_{\mathrm{lo,in}} - p_{\mathrm{lo,out}},\ p_{\mathrm{hi,out}} - p_{\mathrm{hi,in}}\}.$$

The normalized sharing gap is at most $\epsilon_P + \epsilon_B + \gamma\epsilon_{\mathrm{end}}$, plus the recorded rounding allowance. Endpoints and final parameter feasibility are recomputed at 70 digits and converted outward to binary arithmetic. Table 5 reports every condition. Its maximum upper bound, $4.91 \cdot 10^{-12}$, supports common-endpoint compatibility for the evaluated policies. The claim is *same-region evaluation exactness to numerical accuracy*, not return improvement over nominal planning or an independent-row model.

States	Counts	N	Datasets	Mean parameter width	Maximum gap bound
10	B	25	20	0.13355	$1.686 \cdot 10^{-12}$
10	B	250	20	0.04291	$1.626 \cdot 10^{-12}$
10	B	2500	20	0.01361	$1.694 \cdot 10^{-12}$
10	I	25	20	0.13560	$1.777 \cdot 10^{-12}$
10	I	250	20	0.04309	$1.672 \cdot 10^{-12}$
10	I	2500	20	0.01364	$1.687 \cdot 10^{-12}$
30	B	25	20	0.07838	$4.663 \cdot 10^{-12}$
30	B	250	20	0.02480	$4.863 \cdot 10^{-12}$
30	B	2500	20	0.00787	$4.908 \cdot 10^{-12}$
30	I	25	20	0.07913	$4.691 \cdot 10^{-12}$
30	I	250	20	0.02489	$4.841 \cdot 10^{-12}$
30	I	2500	20	0.00786	$4.909 \cdot 10^{-12}$

Table 5: All twelve shared-forest conditions. B/I denote balanced/imbalanced counts. Parameter widths are means over 20 datasets; the last column is the maximum normalized sharing-gap upper bound. Each policy and feasible endpoint are evaluated within the same pooled likelihood set.

K.5 CONSTRAINED-BACKUP IMPLEMENTATION CHECKS

The model domain is part of the backup. The checks instantiate Lemma 18 on five analytic settings: normalization of a constant continuation, nonnegativity, validity of a non-target transition row, agreement with a licensed equality-aware norm formula, and exact known rows. Four zero-radius or zero-objective variants give nine base fixtures. Each also receives the same three continuation transforms: shifts by $\pm 3/8$ and perturbation $(1/10, -1/5)$, giving 36 cases in total. These are deterministic raw row-objective checks, not additional datasets or normalized-return experiments.

The non-target fixture makes the distinction concrete. Write $\theta = (1, t)$ with $|t| \leq 1$. Its two transition rows are $(.5 + .1t, .5 - .1t)$ and $(.5 + t, .5 - t)$. Minimizing the second probability of the first row while ignoring the second row admits $t = 1$ and value $.4$. Enforcing *both* stochastic rows requires $|t| \leq .5$ and yields $.45$. An unrestricted or target-row-only expression is an algebraic domain diagnostic here, not a competing valid planner.

Fixture	Analytic optimum	Objective-width upper
Constant continuation	1	$3.01 \cdot 10^{-11}$
Nonnegativity	0	$1.06 \cdot 10^{-11}$
Non-target row	.45	$5.39 \cdot 10^{-11}$
Licensed norm formula	.4	$5.98 \cdot 10^{-11}$
Known rows	$-.2$	$4.65 \cdot 10^{-11}$
Constant, $R = 0$	1	$2.00 \cdot 10^{-12}$
Constant, $g = 0$	0	$3.01 \cdot 10^{-11}$
Known rows, $R = 0$	$-.2$	$2.00 \cdot 10^{-12}$
Known rows, $g = 0$	0	$4.65 \cdot 10^{-11}$

Table 6: The nine base constrained-backup fixtures. Widths are rounded upward; all quantities use raw row-objective units. The three prescribed transforms of each fixture are also retained in the accompanying code. A negative optimum for a known row is allowed because the continuation vector may have negative components.

All 28 nonzero-radius conic solves terminated optimally; eight zero-radius cases used a direct singleton branch. There were no recorded solver warnings or unsuccessful terminations. Every raw objective agrees with the independent analytic optimum within $1.69 \cdot 10^{-11}$. *Acceptance does not rely on that solver status*: an exactly feasible rational parameter supplies the upper endpoint, and the multipliers supply a lower endpoint through (40), evaluated at 80 digits with a 10^{-12} allowance. The largest objective interval width is $8.32 \cdot 10^{-11}$, below the prescribed 10^{-7} target; the prescribed objective-agreement target is 10^{-8}. The raw optimizer parameters are retained even when their small constraint residuals prevent exact feasibility. They are not substituted for the verified upper witnesses.

For the fixed target row, write $\phi(v) = \min_{\theta \in \mathcal{C}} v^\top A_i \theta$. Stochasticity implies $\phi(v + c\mathbf{1}) = \phi(v) + c$ and $|\phi(v + \Delta v) - \phi(v)| \leq \|\Delta v\|_\infty$. All 18 shift comparisons pass, with maximum raw discrepancy $1.84 \cdot 10^{-16}$; all nine perturbation comparisons satisfy the .2 sup-norm bound, with largest enclosed change below .141. The checks use objectives formed from the target transition features while enforcing all model rows. A separate singular-curvature test rejects an unsupported radius without adding a ridge.

The recorded environment uses CVXPY 1.9.2 and Clarabel 0.11.1 with one worker, one numerical-library thread, tolerances 10^{-10}, at most 200 solver iterations, and a 30-second per-case cap. The accompanying code retains all inputs, attempted outputs, raw multipliers, and verification records. These checks support the implementation of the validity-constrained backup; they do not validate the statistical calibration of a likelihood ellipsoid, establish general conic efficiency, or constitute formally directed arithmetic.

K.6 UNCERTAINTY AND SCOPE

Sampling intervals use 10,000 percentile-bootstrap draws over independent paired datasets within each condition. *Timing repetitions and initializations are clustered, not extra replications.* These approximate intervals describe dataset sampling variation, not simultaneous coverage or numerical error. Recorded numerical allowances and high-precision checks are not formally directed transcendental

interval proofs. The code retains complete counts, policies, witnesses, and optimization statuses; incomplete cohorts do not pass the publication builder.

The experiments quantify a fixed likelihood relaxation and its compatibility mechanisms. They do not empirically establish a new minimax rate, exact general coupled control, or an end-to-end assessment speed advantage. The companion records retain the wider exploratory context separately; the present comparisons and their mathematical targets are specified above.

L Reference construction and occupancy mechanisms

This appendix supplies the independent reference and interprets the complete recurrent cohort. The occupancy-envelope and initialization analyses reuse those 120 datasets; they were specified after inspection and are *post hoc* diagnostics, not fresh confirmation. All three starts, both count designs, and all sample levels are retained in the numerical records.

L.1 Enclosing the coupled reference globally

Let n_X, n_Y be the row counts and k_X, k_Y the counts reaching the reward-one and reward-zero absorbers from X and Y, respectively. Set $\hat{p} = k_X/n_X$, $\hat{q} = k_Y/n_Y$. Using the Bernoulli divergence kl from Appendix I.4, define the adverse endpoints

$$p_-(u) = \min\{p : n_X \,\mathrm{kl}(\hat{p}\|p) \le u\}, \qquad q_+(u) = \max\{q : n_Y \,\mathrm{kl}(\hat{q}\|q) \le u\}.$$

For μ supported on $\{X, Y\}$, let $\mu_X = \mu(X)$ and $\mu_Y = \mu(Y)$. Eliminating the transient Bellman equations gives the normalized return

$$F_\mu(p, q) = \frac{\gamma p[\mu_X + \gamma(1-q)\mu_Y]}{1 - \gamma^2(1-p)(1-q)}.$$

Its denominator is positive; F_μ is nondecreasing in p and nonincreasing in q. A feasible pair can therefore be replaced by adverse endpoints using its row costs and any remaining budget without increasing value. The exact coupled minimum is

$$\min_{0 \le u \le b} F_\mu\big(p_-(u), q_+(b-u)\big). \tag{44}$$

For mixed initialization, (44) minimizes the average once under one kernel, rather than averaging separately minimized returns.

For any allocation cell $[a, c] \subseteq [0, b]$, monotonicity implies

$$F_\mu\big(p_-(c), q_+(b-a)\big) \le \inf_{u \in [a,c]} F_\mu\big(p_-(u), q_+(b-u)\big). \tag{45}$$

The lower-bounding pair may spend $c + b - a > b$; it is a *relaxation of the whole cell*, not a feasible upper witness. Upper witnesses use one allocation and its complement and pass the original likelihood constraint.

The independent audit applies (45), beginning at $[0, b]$ and bisecting unresolved intervals. A cell is terminal only when its outward lower value exceeds the saved reference lower minus 10^{-10}. *Terminal cells must cover the entire allocation interval*; their minimum lower value is the audited global lower endpoint. A cap or nonrepresentable midpoint gives an unresolved audit, not acceptance. All 360 covers complete, comprising 784,033 terminal intervals. Audited gap brackets include this acceptance allowance and rectangular-value uncertainty and have width below $1.01 \cdot 10^{-7}$. Partition digests are recorded for the original arithmetic environment; cross-platform comparison also checks the cover, terminal count, and lower endpoint rather than interpreting float-bearing digest equality as mathematical validity.

Binary roots only propose endpoints. Acceptance uses integer-count likelihoods at 70 digits: an outer lower-success endpoint lies below the exact empirical probability and has cost at least its allocated budget; an outer upper-failure endpoint satisfies the converse directional test. Budgets are converted from their exact binary representations before taking complements. Zero budget uses the rational empirical probability; empirical zeros/ones use the appropriate one-sided formulas. A 10^{-55} arithmetic margin precedes outward conversion. Feasible upper witnesses are checked separately. These tests validate numerical directions with allowances, not a formal proof of transcendental interval evaluation.

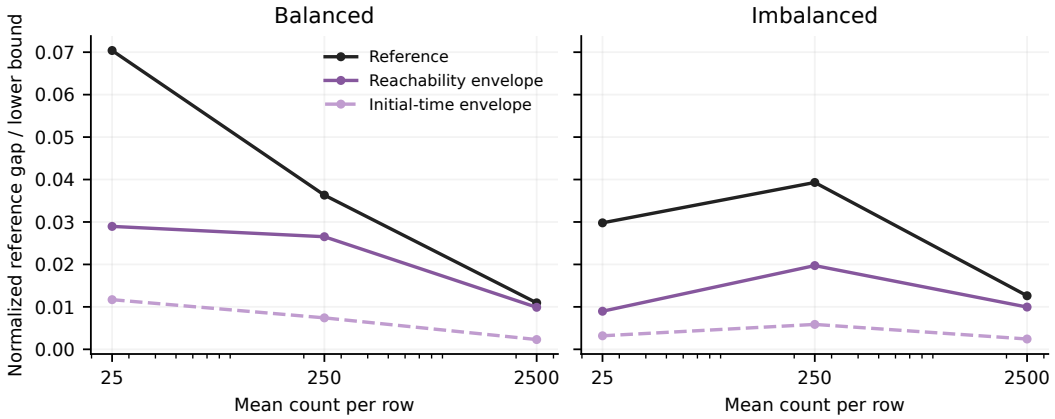


Figure 3: **Recurrence exposes incompatibility without changing the region.** This post hoc comparison changes only the occupancy lower envelope. The initial-time envelope uses $t = 0$; the reachability envelope includes future-visit terms in (46). Each point averages the three starts within all 20 datasets. The reference is used only for comparison.

L.2 Why recurrent visits sharpen the lower bound

The initial-time envelope is $\underline{d}^{(0)}(s,a) = (1-\gamma)\mu(s)\pi(a \mid s)$. Let $\underline{P}_{sa,x}$ lower-bound transition probability over the row projection and put $S_\pi(s,x) = \sum_a \pi(a \mid s)\underline{P}_{sa,x}$. Every feasible P satisfies $P_\pi \geq S_\pi \geq 0$, hence $P_\pi^t \geq S_\pi^t$ by induction. For every finite T,

$$\underline{d}^{(T)}(s,a) = (1-\gamma)\left[\sum_{t=0}^{T} \gamma^t \mu^\top S_\pi^t\right]_s \pi(a \mid s) \leq d_P^\pi(s,a). \tag{46}$$

Thus (46) can replace the time-zero envelope in Proposition 5 without changing the likelihood set. Known rows are exact, and unknown rows retain their declared support. The scalar row intervals supply the coordinate lower bounds in the recurrent study.

From a pure initial state, only one uncertain row has nonzero time-zero weight. Assigning its full budget and leaving the other row empirical makes that initial-envelope objective zero. Positive recurrence terms expose the cost of later visits to the other row. This accounts for the .00579–.01911 increase in mean lower values in Figure 3. Both variants use the same rectangular value and 10^{-10} optimization target; all 360 initial-envelope dual evaluations meet that target. Complete initialization-specific means and intervals are available in the accompanying code.

L.3 Interpreting the remaining interval width

Let L be a normalized residual lower bound on $\bar{G}_\pi$. For a repaired feasible kernel P_m, put $U_m = (1-\gamma)(J_{P_m}^\pi - J_{\mathrm{R}}^\pi)$ and $S_m = \gamma(1-\gamma)^{-1}\sum_i d_Q^\pi(i)\,\mathrm{TV}(Q_i, (P_m)_i)$. In exact arithmetic,

$$L \leq \bar{G}_\pi \leq U_m \leq S_m, \qquad S_m - L = (S_m - U_m) + (U_m - \bar{G}_\pi) + (\bar{G}_\pi - L).$$

The three terms are TV-inequality slack, repaired-adversary suboptimality, and residual-lower-bound slack. The last term need not be caused solely by the occupancy envelope: minimizing the relaxed residual can select a different kernel from the coupled-value minimizer. A repair primal–dual width concerns its TV objective, not any of these value differences. At $N = 2500$, the mean TV-minus-direct upper-bound differences are .00584 and .00678 under balance and imbalance, while occupancy-repair upper slack is $3.79 \cdot 10^{-5}$ and $6.61 \cdot 10^{-6}$. Ordinary evaluation therefore extracts substantially tighter upper information than the TV estimate in those conditions.

For approximate Q, the implemented TV endpoint additionally bounds $(1-\gamma)J_Q^\pi - \bar{J}_{\mathrm{R}}{}^\pi$ and the occupancy/evaluation roundoff. A difference of two padded upper endpoints is labeled a *difference of bounds*, not exact TV slack. Let $[j_m^-, j_m^+]$ and $[j_C^-, j_C^+]$ enclose the normalized repaired and coupled values. Their difference encloses repaired-adversary suboptimality as $[j_m^- - j_C^+, j_m^+ - j_C^-] \cap [0, \infty)$;

the rectangular value cancels. These conventions prevent numerical error from being attributed to the approximation mechanism.

L.4 One adversary for a mixed initialization

Proposition 9 gives the nonnegative premium

$$\bar{G}_\pi((\delta_X + \delta_Y)/2) - \tfrac{1}{2}\{\bar{G}_\pi(\delta_X) + \bar{G}_\pi(\delta_Y)\}.$$

It is the additional value of requiring one coupled model for both starts, since the rectangular value is linear in initialization. Table 7 reports numerical mean enclosures for every condition. The lower endpoint is positive in 119/120 datasets; the remaining sign is unresolved. These reused-data intervals are *not population sampling intervals* or a prespecified confirmation test.

Counts	N	Mean premium enclosure	Positive lower / 20
B	25	[.01235887, .01235908]	20
B	250	[.00680267, .00680288]	20
B	2500	[.00216221, .00216242]	20
I	25	[.00065924, .00065944]	19
I	250	[.00177036, .00177057]	20
I	2500	[.00108922, .00108943]	20

Table 7: Post hoc initialization premium on all six 20-dataset conditions. Endpoints are rounded outward; intervals concern numerical uncertainty in the sample mean, not sampling confidence.

L.5 Analytical controls

The 66 controls instantiate proved mechanisms, not sampled performance estimates. Appendices E.2–E.3 give 28 symmetric-branching and 28 unreachable-row cases on $m \in \{1, 2, 4, 8, 16, 32, 64\}$, $\zeta \in \{.005, .02, .1, .25\}$, plus six safe-action cases with $m > 1$, $\zeta = .1$. Repaired values and lower bounds match branching references within $6.44 \cdot 10^{-9}$ and $7.59 \cdot 10^{-9}$; exact safe-action evaluation coexists with coupled policy losses .03487–.36579. In four shared-parameter controls, a known router selects equally between success probabilities p and $1 - p$, with $p \in [\ell, u]$. The normalized gap $\gamma^2(u - \ell)/2$ is $.405, .243, .081, 0$ for $[0, 1]$, $[.2, .8]$, $[.4, .6]$, and $\{.5\}$, respectively. All controls are available in the accompanying code; no sampling confidence interval applies.