ONE MODEL, TWO GAMES: CERTIFYING NASH EQUILIBRIA IN DISCOUNTED ZERO-SUM LOW-RANK MARKOV GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

An equilibrium of a learned game can admit profitable deviations in the true environment. We study how to certify approximate Nash equilibria in discounted zero-sum Markov games with unknown low-rank transitions. Our method solves two confidence games on one learned model. Under a policy-uniform two-sided value envelope and exact planning, pairing the lower game's maximizing policy with the upper game's minimizing policy minimizes the induced unilateral-deviation certificate. The reverse pairing controls its width and supplies sufficient exploration using the same solutions. A natural alternative combines a nominal equilibrium with a joint-action MDP maximizing bonus occupancy. We prove interaction upper bounds of the same order for both methods, while their fixed-envelope certificates can be strictly separated. Under finite-class realizability, known rewards in $[0, 1-\gamma]$, reset access, and certified oracles, $\widetilde{O}(d^2 J(J+d^2)/((1-\gamma)^4\epsilon^2))$ outer iterations suffice for an ϵ-Nash pair with high probability, where d is the feature dimension and J the joint-action count. Counting primitive transitions adds a factor $(1-\gamma)^{-1}$. The guarantee allows controlled estimation and planning errors despite changing learned features. Experiments distinguish certificate tightness from policy quality and exploration cost: crossing gives tighter envelope certificates on shared data, whereas bonus and uniform exploration are faster on delayed-conflict games. The sufficient confidence parameters remain conservative at these budgets.

1 INTRODUCTION

Computing an equilibrium of a learned game leaves a statistical question: can either player profit by deviating in the true environment? Each best response can direct play toward regions where the estimated transition model is inaccurate. A learner therefore needs a policy pair and a data-dependent upper bound on its true Nash gap.

We study discounted zero-sum Markov games with unknown low-rank transition factors. Existing results provide discounted low-rank MDP learning (Uehara et al., 2022) and finite-horizon game learning with upper/lower values and width-based output selection (Ni et al., 2023). We ask which pair minimizes the deviation bound induced by a two-sided value envelope, and whether the same planning solutions can guide exploration.

DREP-ZS solves two games on one learned model, with payoffs $r + b$ and $r - b$ and saddle policies (π^+, ν^+) and (π^-, ν^-). The conservative crossing (π^-, ν^+) combines the lower game's maximizing policy with the upper game's minimizing policy. Under an envelope valid for every stationary joint policy, exact solutions minimize the induced unilateral-deviation certificate. The reverse crossing (π^+, ν^-) bounds the confidence-game value width by its own bonus occupancy, providing an exploration rule from the same solutions (Figure 1).

A natural alternative computes a nominal (certainty-equivalent, CE) equilibrium and explores with a joint-action MDP maximizing bonus occupancy. Under the same fixed envelope and exact planning, the certificates satisfy

$$\underbrace{C_n(\pi_n^-, \nu_n^+)}_{\text{crossed minimum}} \le \underbrace{C_n(\pi_n^0, \nu_n^0)}_{\text{nominal-pair envelope}} \le \underbrace{2U_n + 2\zeta_n}_{\text{global bonus bound}} .$$

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.

Here C_n bounds unilateral deviations, U_n is the maximum discounted bonus return, and ζ_n is the envelope residual (Proposition 6). Both inequalities can be strict; they order certificates, not true gaps. We prove interaction upper bounds of the same order for the two methods. Their distinction lies in the certificate objective and the planning needed to obtain it.

Learning must make these bounds shrink despite changes in the fitted representation. Building on the occupancy and fixed-true-feature analysis of Uehara et al. (2022), we connect paired uniform-action probes to a selected-feature covariance bound. One confidence event covers all stationary joint policies and best responses. The argument applies to either method's adaptive data.

Our contributions are:

- **Certification and exploration from two plans.** We characterize the minimum of a fixed-envelope deviation certificate, prove strict separation from natural alternatives, and control its width through the reverse crossing.
- **Statistical realization and comparison.** We prove finite-sample guarantees with estimation and planning tolerances for both the crossings and nominal + bonus MDP, separating interactions from planning costs.
- **Evidence on tightness, quality, and cost.** Shared-data and matched-stopping studies distinguish tighter envelope certificates from better policies and faster exploration.

Closest related work. We build on REP-UCB's policy-uniform near-optimism and fixed true-feature potential (Uehara et al., 2022), and on width-based output selection in model-based low-rank games (Ni et al., 2023). Table 3 in Appendix A locates the stationary discounted reduction relative to these ingredients. Other nearby guarantees use contrastive low-rank representations (Qiu et al., 2022), known linear features (Xie et al., 2020; Chen et al., 2022; Yang et al., 2025), structured spectral embeddings (Zhou et al., 2024), or discounted tabular generative/offline access (Zhang et al., 2020; Yan et al., 2024). Their information models and representations differ from learning both factors through online roll-ins. Appendix A gives the full thematic comparison.

2 Problem and Learning Protocol

We consider a finite two-player zero-sum discounted Markov game $\mathcal{G} = (\mathcal{S}, \mathcal{A}, \mathcal{B}, P^\star, r, \gamma, d_0)$. Let $\gamma \in (0,1)$, $A = |\mathcal{A}|$, $B = |\mathcal{B}|$, $J = AB$, and $r(s,a,b) \in [0, 1-\gamma]$, so every unperturbed value lies in $[0,1]$. For stationary policies (π, ν), let $V^{\pi,\nu}_{P,r}(s)$ be the expected discounted return from state s, and write $v^{\pi,\nu}_{P,r} := \mathbb{E}_{s\sim d_0} V^{\pi,\nu}_{P,r}(s)$. We evaluate a pair by the two-sided deviation loss

Definition 1 (Nash duality gap)**.**

$$\mathrm{Gap}_P(\pi,\nu) := \max_{\pi'} v^{\pi',\nu}_{P,r} - \min_{\nu'} v^{\pi,\nu'}_{P,r}. \tag{1}$$

Thus $\mathrm{Gap}_P(\pi,\nu) \le \epsilon$ defines an ϵ-Nash pair. Stationary joint policies $\kappa : \mathcal{S} \to \Delta(\mathcal{A}\times\mathcal{B})$ include product policies and the unilateral deviations in equation 1. Writing $z = (a,b) \in \mathcal{Z} := \mathcal{A}\times\mathcal{B}$, their normalized discounted occupancy is

$$d^\kappa_{P,\gamma}(s,z) := (1-\gamma)\sum_{t\ge 0}\gamma^t \Pr_{P,\kappa}(s_t = s, z_t = z). \tag{2}$$

Stopping a trajectory independently with probability $1-\gamma$ returns the state marginal of this occupancy and uses $\gamma/(1-\gamma)$ transitions in expectation.

Assumption 2 (Unknown low-rank realizability)**.** *There are unknown functions $\phi^\star : \mathcal{S}\times\mathcal{A}\times\mathcal{B} \to \mathbb{R}^d$ and $\mu^\star : \mathcal{S} \to \mathbb{R}^d$ such that*

$$P^\star(s' \mid s,a,b) = \langle \mu^\star(s'), \phi^\star(s,a,b)\rangle. \tag{3}$$

They satisfy $\|\phi^\star(s,a,b)\|_2 \le 1$ and, for every $g : \mathcal{S} \to [0,1]$, $\|\sum_{s'} \mu^\star(s')g(s')\|_2 \le \sqrt{d}$. The learner knows a finite class $\mathcal{M} \subseteq \Psi \times \Phi$ containing $(\mu^\star, \phi^\star)$; every member induces a valid stochastic kernel and satisfies the same normalization. Here Ψ, Φ are the factor projections of $\mathcal{M}$, which need not contain all cross-pairings. We also write $P \in \mathcal{M}$ for a member's kernel.

Neither true factor is supplied to the learner. Planning uses the discounted Shapley operator

$$(\mathcal{T}_{P,c}V)(s) := \mathrm{val}(c(s,\cdot,\cdot) + \gamma PV(s,\cdot,\cdot)), \quad \mathrm{val}(Q) := \max_{x\in\Delta(\mathcal{A})}\min_{y\in\Delta(\mathcal{B})} x^\top Qy. \tag{4}$$

It is a γ-contraction (Shapley, 1953); hence approximate stationary saddle policies can be computed by value iteration. We use an η-accurate planning output $(\pi^\eta, \nu^\eta, v^\eta)$ satisfying

$$\max_\pi v_{P,c}^{\pi,\nu^\eta} \le v^\eta + \eta, \qquad \min_\nu v_{P,c}^{\pi^\eta,\nu} \ge v^\eta - \eta. \tag{5}$$

Assumption 3 (Learning protocol and computational access)**.** *The learner can reset to an independent $s_0 \sim d_0$; conditional on the past and queried context, each successor is drawn from $P^\star(\cdot \mid s,a,b)$. The reward, d_0, and γ are known. For prescribed deterministic tolerances $\tau_n \ge 0$, an estimation oracle returns a factorized model in $\mathcal{M}$ satisfying the cumulative likelihood condition equation 18. A discounted Shapley oracle satisfies equation 5 for every known model, bounded payoff, and $\eta \ge 0$. Planning tolerances may be fixed or chosen predictably from the past.*

This is reset-based online interaction: noninitial states are reached through roll-ins, rather than queried arbitrarily. The discounted simulation identities and geometric-roll-in proof used by the analysis are recorded in Appendix C.

3 Certification and Exploration from Two Games

Two policy roles arise from one pair of planning problems: one controls unilateral deviations, and the other directs data collection. We first state the confidence interface, then derive both roles.

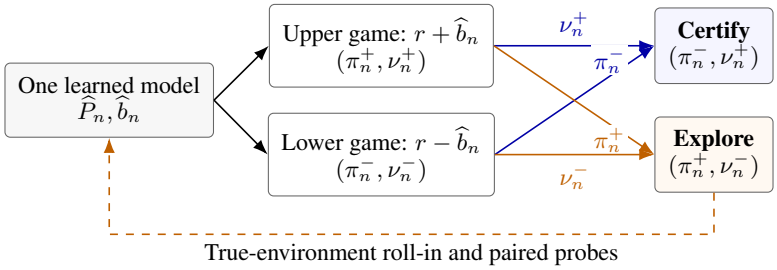


Figure 1: **One model, two games, two policy roles.** The conservative crossing minimizes the specified envelope certificate under exact planning. The reverse crossing controls its width and supplies the next round's data; both reuse the same game solutions.

3.1 Two-sided confidence certifies the conservative crossing

Definition 4 (Strategic-confidence envelope)**.** *For a true kernel $P^\star$, payoff c, and initial distribution d_0, a learned kernel $\widetilde{P}$, nonnegative bonus b, and residual $\zeta \ge 0$ form a strategic-confidence envelope if, simultaneously for every stationary joint policy κ,*

$$v_{\widetilde{P},c+b}^\kappa + \zeta \ge v_{P^\star,c}^\kappa \ge v_{\widetilde{P},c-b}^\kappa - \zeta. \tag{6}$$

This sufficient interface controls both sides of a Nash gap even after model-dependent best responses have been selected. Uniformity covers the deviations as well as the candidate being evaluated.

At iteration n, form two games from the same learned model:

$$\mathcal{G}_n^+ = (\widehat{P}_n, r + \widehat{b}_n, \gamma), \qquad \mathcal{G}_n^- = (\widehat{P}_n, r - \widehat{b}_n, \gamma). \tag{7}$$

Let $(\pi_n^+, \nu_n^+, \overline{v}_n)$ and $(\pi_n^-, \nu_n^-, \underline{v}_n)$ be their planning outputs with tolerance η_n. Define the candidate and its observable certificate by

$$\sigma_n^{\mathrm{cand}} := (\pi_n^-, \nu_n^+), \tag{8}$$

$$\widehat{\Delta}_n := \overline{v}_n - \underline{v}_n + 2\zeta_n + 2\eta_n. \tag{9}$$

Lemma 5 (Strategic-confidence reduction). *If $(\widehat{P}_n, \widehat{b}_n, \zeta_n)$ satisfies equation 6 for $c = r$, then*

$$\mathrm{Gap}_{P^\star}(\pi_n^-, \nu_n^+) \le \overline{v}_n - \underline{v}_n + 2\zeta_n + 2\eta_n = \widehat{\Delta}_n. \tag{10}$$

Proof. For every π, ν, the envelope and saddle inequalities give

$$v^{\pi,\nu_n^+}_{P^\star,r} \le v^{\pi,\nu_n^+}_{\widehat{P}_n, r+\widehat{b}_n} + \zeta_n \le \overline{v}_n + \eta_n + \zeta_n, \qquad v^{\pi_n^-,\nu}_{P^\star,r} \ge \underline{v}_n - \eta_n - \zeta_n.$$

Maximizing the first bound and minimizing the second proves the claim. □

Let (π_n^0, ν_n^0) be an exact equilibrium of the nominal game $(\widehat{P}_n, r)$, and define the global bonus value

$$U_n := \max_{\kappa} v^{\kappa}_{\widehat{P}_n, \widehat{b}_n}. \tag{11}$$

The maximum is over stationary joint policies, equivalently a discounted MDP with joint actions.

Proposition 6 (Minimum envelope certificate and comparison). *Under the same envelope, every stationary pair (π, ν) satisfies $\mathrm{Gap}_{P^\star}(\pi, \nu) \le C_n(\pi, \nu)$, where*

$$C_n(\pi, \nu) := \max_{\pi'} v^{\pi',\nu}_{\widehat{P}_n, r+\widehat{b}_n} - \min_{\nu'} v^{\pi,\nu'}_{\widehat{P}_n, r-\widehat{b}_n} + 2\zeta_n. \tag{12}$$

With exact planning, the conservative crossing minimizes this certificate:

$$\min_{\pi,\nu} C_n(\pi, \nu) = C_n(\pi_n^-, \nu_n^+) = \overline{v}_n - \underline{v}_n + 2\zeta_n. \tag{13}$$

Moreover, the same fixed envelope gives

$$C_n(\pi_n^-, \nu_n^+) \le C_n(\pi_n^0, \nu_n^0) \le 2U_n + 2\zeta_n. \tag{14}$$

Both inequalities can be strict; the ratio of the nominal envelope certificate to its crossed minimum can be arbitrarily large.

Proof. The envelope bounds each unilateral deviation in equation 12. Its two optimization terms depend separately on ν and π. Minimizing over ν gives the upper game's minimax value, and maximizing the subtracted term over π gives the lower game's minimax value. The corresponding saddle policies attain both values. For the nominal pair, saddle inequalities bound the upper response by $v_n^0 + U_n$ and the lower response by $v_n^0 - U_n$, proving equation 14. □

Exact planning is needed for this minimization identity; the observable certificate in Lemma 5 includes the planning error used by the implemented algorithm.

For the one-state construction in Appendix E, $t = 0.01$ and $\beta = 0.4$ give:

Candidate	True Nash gap	Envelope certificate
Nominal equilibrium	0	0.39
Conservative crossing	0.00975	0.00975

The global bonus bound is 0.80. Thus minimizing the envelope certificate can increase the true gap, even on the same model. The ordering applies to this fixed envelope; likelihood sets and different learning histories are separate confidence objects. Appendix E.3 realizes strict, arbitrarily large ratios with the learner's covariance bonus, including sufficient proof constants. That construction uses a redundant representation and rare histories, proving neither typical separation nor faster learning.

3.2 The reverse crossing controls the exploration width

Set the exploration pair to

$$\sigma_n^{\mathrm{exp}} := (\pi_n^+, \nu_n^-). \tag{15}$$

Lemma 7 (Exploration-width reduction). *For $\kappa_n = \kappa_{\pi_n^+, \nu_n^-}$,*

$$-2\eta_n \le \overline{v}_n - \underline{v}_n \le \frac{2}{1-\gamma} \mathbb{E}_{d^{\kappa_n}_{\widehat{P}_n,\gamma}}[\widehat{b}_n] + 2\eta_n. \tag{16}$$

The upper bound follows by evaluating the two saddle inequalities on this common pair: their payoffs differ by $2\widehat{b}_n$ along the same learned-model occupancy. The lower bound follows from payoff monotonicity and the planning tolerances. This provides the occupancy bound needed for learning, without requiring global maximization of bonus occupancy.

The reverse crossing need not itself have gap bounded by the value width (Theorem 14). Its role is to collect information sufficient for the conservative crossing's certificate.

Why both sides and the conservative crossing matter. A one-sided envelope can leave incompatible equilibria unresolved, and the upper-game or reverse-crossed pair can fail even under a valid two-sided envelope. Theorem 14 in Appendix C.2.1 formalizes both failures.

4 DREP-ZS: LEARN, CERTIFY, AND EXPLORE

DREP-ZS realizes the two-crossing reduction by learning one stationary model from paired samples. Algorithm 1 gives the procedure; Figure 1 shows how the next samples close the learning loop.

4.1 SAMPLING AND ESTIMATION

Under the preceding exploration pair $\sigma_{n-1}^{\exp}$, a geometric roll-in returns s from the state marginal of $d_{P^\star,\gamma}^{\sigma_{n-1}^{\exp}}$. The learner draws $z \sim \mathrm{Unif}(\mathcal{A} \times \mathcal{B})$, observes $s' \sim P^\star(\cdot \mid s, z)$, then draws another uniform z' and observes s''. It adds (s, z, s') and (s', z', s'') to $\mathcal{D}_n, \widetilde{\mathcal{D}}_n$, respectively:

$$\mathcal{D}_n = \mathcal{D}_{n-1} \cup \{(s, z, s')\}, \qquad \widetilde{\mathcal{D}}_n = \widetilde{\mathcal{D}}_{n-1} \cup \{(s', z', s'')\}. \tag{17}$$

Only these probe tuples enter the likelihood; all roll-in transitions still count toward interaction cost. The shifted tuple supplies the one-step training distribution needed by the confidence proof. Uniform local actions give $\kappa(z \mid s)/(1/J) \le J$ for every joint policy κ, while the roll-in determines the sampled state distribution. For $\mathcal{L}_n(P) := \sum_{(s,z,s') \in \mathcal{D}_n \cup \widetilde{\mathcal{D}}_n} \log P(s' \mid s, z)$, where the datasets are multisets and observations are counted with multiplicity, a certified τ_n-approximate MLE satisfies

$$\mathcal{L}_n(\widehat{P}_n) \ge \sup_{P \in \mathcal{M}} \mathcal{L}_n(P) - \tau_n, \qquad \widehat{P}_n(s' \mid s, z) := \langle \widehat{\mu}_n(s'), \widehat{\phi}_n(s, z) \rangle. \tag{18}$$

With

$$\widehat{\Sigma}_n := \sum_{(s,z,\cdot) \in \mathcal{D}_n} \widehat{\phi}_n(s, z) \widehat{\phi}_n(s, z)^\top + \lambda_n I_d, \qquad \widehat{b}_n(s, z) := \alpha_n \|\widehat{\phi}_n(s, z)\|_{\widehat{\Sigma}_n^{-1}}, \tag{19}$$

the bonus is occupancy-valid rather than a pointwise transition-error bound, because learned and true features need not share coordinates.

Algorithm 1 DREP-ZS

Require: $\mathcal{M}, N, \{\lambda_n, \alpha_n, \zeta_n, \eta_n, \tau_n\}_{n=1}^N$

1: Initialize $\sigma_0^{\exp}$ uniformly and $\mathcal{D}_0 = \widetilde{\mathcal{D}}_0 = \varnothing$
2: **for** $n = 1, \ldots, N$ **do**
3: Geometrically roll in under $\sigma_{n-1}^{\exp}$; collect the two uniform-action tuples in equation 17
4: Fit $\widehat{P}_n$ to likelihood tolerance τ_n via equation 18; form $\widehat{\Sigma}_n, \widehat{b}_n$
5: Solve $\mathcal{G}_n^+$ and $\mathcal{G}_n^-$ to tolerance η_n
6: Store $\sigma_n^{\mathrm{cand}}, \widehat{\Delta}_n$ and set $\sigma_n^{\exp} = (\pi_n^+, \nu_n^-)$
7: **end for**
8: **return** $\sigma_{n^\star}^{\mathrm{cand}}$, where $n^\star \in \arg\min_n \widehat{\Delta}_n$

Oracle budget. For N iterations, the reduction uses N approximate cumulative-MLE calls and $2N$ complete discounted zero-sum planning calls. Each planning call solves a known-model stochastic game to its stated tolerance, rather than a single matrix game. Feature evaluation, covariance construction, and the transitions in Corollary 10 are additional costs. Table 4 compares candidate and certificate interfaces without equating these different units of computation.

5 STATISTICAL GUARANTEES: SHRINKING THE CERTIFICATE

The analysis makes the envelope valid for every stationary joint policy and shows that exploration shrinks its width despite representation updates. We first control estimation errors on the two sampling distributions, transfer them to policy occupancies, and then sum uncertainty through a fixed true-feature potential. The argument supports both methods on their own data histories.

5.1 A SIMULTANEOUS CONFIDENCE EVENT

For sufficiently large universal constants, chosen in the order specified in Appendix D, set

$$L_n := \log \frac{c_0 |\mathcal{M}|(n+1)}{\delta}, \qquad \overline{\tau}_n := \max_{t \le n} \tau_t, \qquad \overline{L}_n := L_n + \overline{\tau}_n, \tag{20}$$

$$\lambda_n := c_\lambda \max\{1, d\overline{L}_n\}, \qquad \alpha_n := c_\alpha \sqrt{(J+d^2)\overline{L}_n}, \qquad \zeta_n := c_\zeta \sqrt{\frac{J\overline{L}_n}{n(1-\gamma)}}. \tag{21}$$

Appendix F.1 verifies the explicit sufficient choice $(c_0, c_\lambda, c_\alpha, c_\zeta) = (64, 32, 64, 12)$.

Theorem 8 (Low-rank realization of strategic confidence)**.** *Under Assumptions 2–3, with probability at least $1-\delta$ over the data, simultaneously for all $n \le N$ and stationary joint policies κ,*

$$v^{\kappa}_{\widehat{P}_n, r+\widehat{b}_n} \ge v^{\kappa}_{P^\star, r} - \zeta_n, \tag{22}$$

$$v^{\kappa}_{\widehat{P}_n, r-\widehat{b}_n} \le v^{\kappa}_{P^\star, r} + \zeta_n, \tag{23}$$

$$\left| v^{\kappa}_{\widehat{P}_n, r} - v^{\kappa}_{P^\star, r} \right| \le \frac{\mathbb{E}_{d^{\kappa}_{\widehat{P}_n,\gamma}}[\widehat{b}_n]}{1-\gamma} + \zeta_n. \tag{24}$$

The random event controls MLE errors under both training-context mixtures ρ_n, ρ'_n and covariance comparison uniformly over the finite feature class. Here ρ_n averages previous roll-in state occupancies with uniform actions, and ρ'_n follows one true transition and another uniform action. The latter is supplied by the shifted tuple. After this event is established, policy dependence enters only through deterministic occupancy inequalities, so no union bound over policies is needed. Complementing the normalized reward to $(1-\gamma) - r$ yields the lower envelope on the same event.

5.2 CONNECTING LEARNED UNCERTAINTY TO ONE FIXED POTENTIAL

Write $d_n = d^{\sigma_n^{\exp}}_{P^\star,\gamma}$, including the uniform initial pair at $n=0$, $f_n = \|\widehat{P}_n(\cdot \mid s,z) - P^\star(\cdot \mid s,z)\|_1$, and $B_n = \alpha_n/\sqrt{\lambda_n}$, so $0 \le \widehat{b}_n \le B_n$. The simulation identity transfers the learned occupancy in Lemma 7 to the true exploration occupancy:

$$\mathbb{E}_{d^{\sigma_n^{\exp}}_{\widehat{P}_n,\gamma}}[\widehat{b}_n] \le \mathbb{E}_{d_n}[\widehat{b}_n] + \frac{\gamma B_n}{1-\gamma} \mathbb{E}_{d_n}[f_n]. \tag{25}$$

On the same event, the training moments satisfy $n\mathbb{E}_{\rho_n}[\widehat{b}_n^2] \lesssim \alpha_n^2 d$ and $n\mathbb{E}_{\rho_n}[f_n^2] \lesssim \overline{L}_n$. The true-model one-step-back inequality converts these moments to an elliptical norm in the *fixed* true representation. With $\Sigma_{n,\star} = \lambda_n I + \sum_{i=0}^{n-1} \mathbb{E}_{d_i}[\phi^\star \phi^{\star\top}]$ and $\ell_N = 1 + \log(1 + N/(d\lambda_1))$,

$$\sum_{n=1}^{N} \mathbb{E}_{d_n}[\|\phi^\star(s,z)\|_{\Sigma_{n,\star}^{-1}}] \le \sqrt{2dN\ell_N}. \tag{26}$$

This potential accumulates information under the actual exploration policies even as $\widehat{\phi}_n$ changes. The fixed-feature device is inherited from low-rank MDP analysis (Uehara et al., 2022); here it sums the two terms in equation 25 to control a crossed-policy certificate. Lemmas 24–27 and Appendix D.7 give the transfer, training moments, and full summation. The algorithm never evaluates $\phi^\star$.

5.3 INTERACTION COMPLEXITY AND ORACLE TOLERANCES

Define

$$\mathfrak{K}_{d,J} := d\sqrt{J(J+d^2)} + (J+d^2),$$
$$\mathfrak{K}_{d,J,\gamma} := \frac{d\sqrt{J(J+d^2)}}{1-\gamma} + \frac{\gamma(J+d^2)}{(1-\gamma)^2}. \tag{27}$$

Theorem 9 (Cumulative certificate and returned-policy guarantee)**.** *Under Assumptions 2 and 3, with probability at least $1-\delta$, Algorithm 1 satisfies*

$$\mathrm{Gap}_{P^\star}(\widehat{\pi},\widehat{\nu}) \le \min_{n\le N} \widehat{\Delta}_n \le \frac{1}{N}\sum_{n=1}^{N} \widehat{\Delta}_n, \tag{28}$$

where, on the same event,

$$\frac{1}{N}\sum_{n=1}^{N} \widehat{\Delta}_n \le C\left[\mathfrak{K}_{d,J,\gamma}\sqrt{\frac{\overline{L}_N \ell_N}{N}} + \sqrt{\frac{J\overline{L}_N}{(1-\gamma)N}}\right] + \frac{4}{N}\sum_{n=1}^{N}\eta_n \tag{29}$$

for a universal C.

The two parts of $\mathfrak{K}_{d,J,\gamma}$ retain the distinct costs of bonus accumulation and model-to-true occupancy transfer. The remaining terms track the initial-distribution residual and planning error. Using $\mathfrak{K}_{d,J,\gamma} \le \mathfrak{K}_{d,J}/(1-\gamma)^2$ recovers the simpler rate below.

Corollary 10 (Tolerance-aware sample complexity)**.** *Under the theorem's assumptions, if $\overline{\tau}_n \le c_\tau L_n$ for every n and $N^{-1}\sum_n \eta_n \le \epsilon/8$, then*

$$N = \widetilde{O}\left(\frac{d^2 J(J+d^2)\log(|\mathcal{M}|/\delta)}{(1-\gamma)^4\epsilon^2}\right) \tag{30}$$

outer iterations suffice for an ϵ-Nash pair with probability at least $1-\delta$. The learner uses $2N$ MLE tuples; including geometric roll-ins, its high-probability primitive-transition bound is

$$\widetilde{O}\left(\frac{d^2 J(J+d^2)\log(|\mathcal{M}|/\delta)}{(1-\gamma)^5\epsilon^2}\right). \tag{31}$$

The hidden constant may depend on $1+c_\tau$; exact MLE has $c_\tau = 0$.

A formal nominal + bonus-MDP alternative. Let $\mathcal{R}_N$ denote the statistical term in equation 29.

Theorem 11 (Global-bonus certification and exploration)**.** *Replace the two confidence-game plans by a nominal η_n^0-saddle pair and a joint-action bonus-MDP plan $(\kappa_n^U, \overline{U}_n)$ satisfying*

$$U_n \le \overline{U}_n \le v^{\kappa_n^U}_{\widehat{P}_n,\widehat{b}_n} + \varepsilon_n^U. \tag{32}$$

Use κ_n^U for the next paired sample. Under the same statistical assumptions and parameter choices, with probability at least $1-\delta$, each nominal pair has gap at most $\Delta_n^U := 2\overline{U}_n + 2\zeta_n + 2\eta_n^0$, and

$$\frac{1}{N}\sum_{n=1}^{N}\Delta_n^U \le \mathcal{R}_N + \frac{2}{N}\sum_{n=1}^{N}(\varepsilon_n^U + \eta_n^0). \tag{33}$$

Selecting the smallest certificate gives interaction upper bounds of the same order as Corollary 10 when $N^{-1}\sum_n(\varepsilon_n^U + \eta_n^0) \le \epsilon/4$.

An optimal deterministic joint action is executable as two deterministic player actions. This comparator uses one zero-sum plan and one MDP plan; their computational costs need not match. Appendix E constructs the upper bound in equation 32 and proves the result under the comparator's own adaptive data sequence. The crossings instead minimize the envelope certificate and supply sufficient exploration from the same two game solutions, without maximizing global bonus occupancy.

The rates use $r \in [0, 1-\gamma]$. For $r_0 \in [0,1]$ and fixed unnormalized accuracy ϵ_0, substituting $\epsilon = (1-\gamma)\epsilon_0$ makes the displayed outer and primitive exponents six and seven. A fixed per-tuple likelihood error gives an accuracy floor: $\tau_n = 2n\varepsilon_{\mathrm{nll}}$ yields order $\widetilde{O}(\mathfrak{K}_{d,J}(1-\gamma)^{-2}\sqrt{\varepsilon_{\mathrm{nll}}})$.

Alternative confidence and output. Proposition 21 states the training-error and covariance obligations beyond finite classes. For an enumerable class, Corollary 28 permits selecting across envelope and likelihood-set certificates, with a split failure budget and unchanged exploration. Both extensions preserve their stated statistical conditions.

6 Experiments: Certificate Tightness and Exploration Cost

Our experiments ask two questions: with data and confidence objects fixed, how do candidate choice and certification objective affect the bound; with output and stopping fixed, how does exploration affect certification cost? The *proof* profile uses theoretically sufficient constants; the *calibrated* profile $(c_0, c_\lambda, c_\alpha, c_\zeta) = (8, 1, 0.025, 0.0125)$ uses independent calibration and carries no such guarantee. Appendix H gives protocols, seeds, and costs.

6.1 Comparing certificates on identical data

We evaluate both profiles at 170 stored checkpoints from 30 trajectories, holding each learned model and its data fixed. Table 1 compares crossed and nominal envelope certificates, the global bound $2\overline{U}_n + 2\zeta_n + 2\eta_n^0$, and the nominal likelihood-set (LR-set) certificate. Final calibrated crossing certificates improve on the nominal envelope by mean paired reductions of 4.4%, 2.4%, and 6.9% across the three settings; reductions from global bonus are 15.1%, 17.3%, and 30.8%. Pointwise paired intervals resolve the nominal-envelope improvement in every setting (Appendix H.4). None of the three calibrated envelope/global certificates reaches 0.05 on these trajectories. Checkpoints within a trajectory are not independent trials.

Table 1: **Fixed-data certificates.** Final-checkpoint medians over ten seeds per setting, calibrated profile. Zero means below 10^{-7}. Appendix H.4 reports both profiles, true gaps, paired effects, and matched cold-planning costs.

Setting	Cross	Nominal envelope	Global bonus	Nominal LR-set
Default random	0.653	0.678	0.774	0
Small random	0.408	0.421	0.503	0
Small hidden	0.492	0.529	0.712	0

The LR-set gives the tightest nominal certificate: all 30 final likelihood sets are singletons containing truth. Final nominal gaps are below 10^{-7}, whereas calibrated crossing gap medians are 0.0271, 0.0087, and 0.0577. All proof-profile envelope bounds exceed one. Thus certificate ordering need not improve policy quality or enable practical stopping.

Native crossing uses two zero-sum plans for certification and exploration. Nominal + global bonus uses one zero-sum plan and one joint-action MDP; certifying a nominal pair through the envelope adds two best responses and requires an exploration rule. Nominal + bonus plans faster in all three settings: matched final median times are 2.2541/1.0673, 0.0774/0.0473, and 0.0768/0.0468 seconds (crossing/nominal + bonus), respectively. These serial cold-planning measurements exclude interaction, model fitting, and true-gap evaluation. We report call types and time separately; the fixed-data comparison measures certification and computation, not learning under alternative roll-ins.

6.2 Comparing exploration under a common stopping rule

Checking whether interaction is necessary. The deep-branch family permits zero-sample certification: searching its 32 nominal candidates against all 32 models yields minimum certificate 0.00820045 on each of ten instances, using 32 zero-sum plans, 2,048 best responses, and mean 6.80 seconds. This uses no true-model access or robust equilibrium solver. Its fixed-candidate roll-in comparison illustrates a computation–interaction tradeoff (Appendix H.2).

The delayed-conflict family instead places incompatible games at the ends of two paths. For $c = a\gamma^{L+1}$ and initial first-action probabilities p, q averaged over aliases, the endpoint gaps satisfy

$$\mathrm{Gap}_+(\pi, \nu) = c(2 - p - q), \qquad \mathrm{Gap}_-(\pi, \nu) = c(p + q).$$

Their sum is $2c$, and uniform initial actions attain worst-case gap c. Hence the initial minimax gap over stationary pairs equals c (Proposition 39). With $\gamma = 0.9$, $L \in \{2, 4, 6\}$, and $a \in \{0.25, 0.4\}$, every initial minimax gap exceeds 0.1195, ruling out common certification at 0.05 without data. Two probes from the initial state cannot reach the terminal disagreement.

For each setting, ten paired seeds under each true endpoint compare five roll-ins (600 trajectories): reverse, upper same-side (π^+, ν^+), bonus MDP, nominal, and uniform. All share nominal output, the anytime LR-set certificate ($\delta_{\text{set}} = 0.05$), and a 12,000-transition budget including roll-ins. The common output and stopping rule isolate the effect of roll-in choice; this experiment evaluates calibrated exploration, not the native proof-profile stopping rule.

Table 2: **Delayed conflict with common nominal output and LR-set stopping.** Mean primitive transitions to certificate 0.05, capped at 12,000. Every non-nominal cell has 20/20 threshold hits; nominal has 0/20. All three inaccurate certified outputs are retained.

L	a	Reverse	Upper	Bonus MDP	Nominal	Uniform
2	0.25	3,111.8	1,762.2	629.0	12,000.0	834.2
2	0.4	2,601.8	2,479.3	188.7	12,000.0	265.4
4	0.25	3,119.6	1,562.2	735.2	12,000.0	1,444.0
4	0.4	2,410.5	1,649.5	165.3	12,000.0	176.8
6	0.25	2,315.6	1,439.8	682.1	12,000.0	1,101.3
6	0.4	2,199.8	1,545.2	315.3	12,000.0	494.6

Bonus MDP and uniform exploration use fewer transitions in all six settings (Table 2), with paired 95% intervals excluding zero. Upper same-side has lower means too; its difference is unresolved at $(L, a) = (2, 0.4)$. Three runs exclude truth and certify an inaccurate pair: two reverse and one uniform. Accurate certified outputs total 118/120 for reverse, 120/120 each for upper and bonus, 119/120 for uniform, and 0/120 for nominal. Intervals resample ten seed blocks, retaining both truths and all methods. Observable stopping includes inaccurate certifications; true-gap diagnostics never affect it. Appendix H.3 gives curves, endpoint strata, and computation. Path actions are irrelevant, and uniform exploration enters a path with probability one half, so this family tests delayed disagreement rather than action-sequence discovery.

Reverse crossing is also slower in wall time: its two games are additional to the common output computation. Sufficient exploration need not be the fastest exploration on these instances.

7 Scope and Extensions

The interaction guarantees require finite-class realizability, known normalized rewards, resets, and certified oracles. Uniform paired probes are required by the stated confidence analysis. Fitting and planning can depend on the state count; the action and discount rates are not known to be minimax optimal. The experiments expose conservative constants and faster alternative exploration, while the delayed-conflict sweep couples depth and rank. Appendix B details these boundaries and the restricted-budget, three-seed 4,096-state tabular comparison. The finite-class experiments do not test large implicit model classes.

Oracle-conditional stationary CCE. With exact MLE, evaluation and best responses, shared randomness, and a stationary oracle, candidate/deviation sampling extends the interface to coarse correlated equilibria (CCE; Theorem 38). The oracle is PPAD-hard (Daskalakis et al., 2023); our experiments supply it on a team-potential family (Appendix G).

8 Conclusion

Two confidence-game solutions minimize the specified deviation certificate of a fixed envelope under exact planning and supply sufficient exploration. The learner's interaction upper bound has the same order as nominal output with bonus-MDP exploration. Experiments distinguish envelope tightness from policy quality, planning cost, and exploration speed. Tighter practical confidence and selective exploration remain open directions.

AI USE STATEMENT

Generative-AI tools assisted with literature organization, drafting, code implementation, debugging, and consistency checks. The authors take responsibility for the final text, theoretical claims, code, experiments, and supplementary artifacts.

REPRODUCIBILITY STATEMENT

The problem assumptions, learning protocol, oracle tolerances, algorithms, and main guarantees are stated in Sections 2–5 and Section 7. Appendix D provides the complete zero-sum proofs, while Appendix G provides the stationary-CCE derivations. Appendix F gives explicit constants, the likelihood-set certificate, output selection, and numerical error corrections. Appendix E proves the CE + bonus-MDP guarantee and strict certificate separation. The fixed-data replay includes all 170 saved checkpoints, both parameter profiles, global-bonus upper bounds, and host-matched planning times. The independent-exploration artifact includes every training tuple, enabling replay of likelihoods, certificates, and stopping. It also includes complete initial candidate-search matrices, the proved delayed-conflict construction, both endpoint strata, planner/response counts, and all observed confidence failures. Appendix H documents the experimental families, finite-model implementation, calibration protocol, evaluation policies, seed counts, stopping rules, uncertainty reporting, and compute accounting. The accompanying supplementary artifact archives the per-task configurations, raw traces, machine-readable summaries, and plots used to reproduce the reported tables and figures.

REFERENCES

Alekh Agarwal, Sham M. Kakade, Akshay Krishnamurthy, and Wen Sun. FLAMBE: Structural complexity and representation learning of low rank MDPs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20095–20107, 2020.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 551–560, 2020.

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2159–2170, 2020.

Zixiang Chen, Dongruo Zhou, and Quanquan Gu. Almost optimal algorithms for two-player zero-sum linear mixture markov games. In *Proceedings of the 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pp. 227–261, 2022.

Qiwen Cui and Simon S. Du. When are offline two-player zero-sum markov games solvable? In *Advances in Neural Information Processing Systems*, volume 35, pp. 25779–25791, 2022.

Qiwen Cui, Kaiqing Zhang, and Simon S. Du. Breaking the curse of multiagents in a large state space: RL in markov games with independent linear function approximation. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2651–2652, 2023.

Constantinos Daskalakis, Noah Golowich, and Kaiqing Zhang. The complexity of markov equilibrium in stochastic games. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 4180–4234, 2023.

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

Liad Erez, Tal Lancewicki, Uri Sherman, Tomer Koren, and Yishay Mansour. Regret minimization and convergence to equilibria in general-sum markov games. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9343–9373, 2023.

Junyi Fan, Yuxuan Han, Jialin Zeng, Jian-Feng Cai, Yang Wang, Yang Xiang, and Jiheng Zhang. RL in markov games with independent function approximation: Improved sample complexity bound under the local access model. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2035–2043, 2024.

Songtao Feng, Ming Yin, Yu-Xiang Wang, Jing Yang, and Yingbin Liang. Improving sample efficiency of model-free algorithms for zero-sum markov games. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 13387–13422, 2024.

Arlington M. Fink. Equilibrium in a stochastic n-person game. *Journal of Science of the Hiroshima University, Series A-I (Mathematics)*, 28(1):89–93, 1964. doi: 10.32917/hmj/1206139508.

Amy Greenwald and Keith Hall. Correlated q-learning. In *Proceedings of the Twentieth International Conference on Machine Learning*, pp. 242–249, 2003.

Junling Hu and Michael P. Wellman. Nash q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, 4:1039–1069, 2003.

Audrey Huang, Jinglin Chen, and Nan Jiang. Reinforcement learning in low-rank MDPs with density features. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13710–13752, 2023.

Baihe Huang, Jason D. Lee, Zhaoran Wang, and Zhuoran Yang. Towards general function approximation in zero-sum markov games. In *International Conference on Learning Representations*, 2022.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1704–1713, 2017.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2137–2143, 2020.

Chi Jin, Qinghua Liu, and Tiancheng Yu. The power of exploiter: Provable multi-agent RL in large state spaces. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10251–10279, 2022.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning: A simple, efficient, decentralized algorithm for multiagent reinforcement learning. *Mathematics of Operations Research*, 49(4): 2295–2322, 2024. doi: 10.1287/moor.2021.0317.

Yujia Jin, Vidya Muthukumar, and Aaron Sidford. The complexity of infinite-horizon general-sum stochastic games. In *14th Innovations in Theoretical Computer Science Conference*, volume 251 of *Leibniz International Proceedings in Informatics*, pp. 76:1–76:20, 2023. doi: 10.4230/LIPIcs.ITCS.2023.76.

Chris Junchi Li, Dongruo Zhou, Quanquan Gu, and Michael I. Jordan. Learning two-player markov games: Neural function approximation and correlated equilibrium. In *Advances in Neural Information Processing Systems*, volume 35, pp. 33262–33274, 2022.

Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 157–163, 1994. doi: 10.1016/B978-1-55860-335-6.50027-1.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7001–7010, 2021.

Weichao Mao, Lin Yang, Kaiqing Zhang, and Tamer Başar. On improving model-free algorithms for decentralized multi-agent reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15007–15049, 2022.

Zakaria Mhammedi, Adam Block, Dylan J. Foster, and Alexander Rakhlin. Efficient model-free exploration in low-rank MDPs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66782–66817, 2023.

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

Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free representation learning and exploration in low-rank MDPs. *Journal of Machine Learning Research*, 25(6):1–76, 2024.

Chengzhuo Ni, Yuda Song, Xuezhou Zhang, Zihan Ding, Chi Jin, and Mengdi Wang. Representation learning for general-sum low-rank markov games. In *International Conference on Learning Representations*, 2023.

Julien Perolat, Florian Strub, Bilal Piot, and Olivier Pietquin. Learning nash equilibrium for general-sum markov games from batch data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 232–241, 2017.

Shuang Qiu, Xiaohan Wei, Jieping Ye, Zhaoran Wang, and Zhuoran Yang. Provably efficient fictitious play policy optimization for zero-sum markov games with structured transitions. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8715–8725, 2021.

Shuang Qiu, Lingxiao Wang, Chenjia Bai, Zhuoran Yang, and Zhaoran Wang. Contrastive UCB: Provably efficient contrastive self-supervised learning in online reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18168–18210, 2022.

Shuang Qiu, Ziyu Dai, Han Zhong, Zhaoran Wang, Zhuoran Yang, and Tong Zhang. Posterior sampling for competitive RL: Function approximation and partial observation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 26585–26637, 2023.

Tongzheng Ren, Tianjun Zhang, Csaba Szepesvári, and Bo Dai. A free lunch from the noise: Provable and practical exploration for representation learning. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1686–1696, 2022.

Lloyd S. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10): 1095–1100, 1953. doi: 10.1073/pnas.39.10.1095.

Aaron Sidford, Mengdi Wang, Lin Yang, and Yinyu Ye. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2992–3002, 2020.

Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum markov games with a large number of players sample-efficiently? In *International Conference on Learning Representations*, 2022.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 2898–2933, 2019.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline RL in low-rank MDPs. In *International Conference on Learning Representations*, 2022.

Yuanhao Wang, Qinghua Liu, Yu Bai, and Chi Jin. Breaking the curse of multiagency: Provably efficient decentralized multi-agent RL with function approximation. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2793–2848, 2023.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4259–4299, 2021.

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3674–3682, 2020.

Yuling Yan, Gen Li, Yuxin Chen, and Jianqing Fan. Model-based reinforcement learning for offline zero-sum markov games. *Operations Research*, 72(6):2430–2445, 2024. doi: 10.1287/opre.2022.0342.

Lin F. Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6995–7004, 2019.

Tong Yang, Bo Dai, Lin Xiao, and Yuejie Chi. Incentivize without bonus: Provably efficient model-based online multi-agent RL for markov games. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70704–70744, 2025.

Asrin Efe Yorulmaz and Tamer Başar. Near optimal convergence to coarse correlated equilibrium in general-sum markov games. In *Proceedings of the 8th Annual Learning for Dynamics and Control Conference*, volume 331 of *Proceedings of Machine Learning Research*, pp. 316–345, 2026.

Kaiqing Zhang, Sham M. Kakade, Tamer Başar, and Lin F. Yang. Model-based multi-agent RL in zero-sum markov games with near-optimal sample complexity. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1166–1178, 2020.

Xuezhou Zhang, Yuda Song, Masatoshi Uehara, Mengdi Wang, Alekh Agarwal, and Wen Sun. Efficient reinforcement learning in block MDPs: A model-free representation learning approach. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26517–26547, 2022.

Yuheng Zhang, Yu Bai, and Nan Jiang. Offline learning in markov games with general function approximation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 40804–40829, 2023.

Canzhe Zhao, Ruofeng Yang, Baoxiang Wang, Xuezhou Zhang, and Shuai Li. Learning adversarial low-rank markov decision processes with unknown transition and full-information feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 59107–59123, 2023.

Yulai Zhao, Yuandong Tian, Jason D. Lee, and Simon Du. Provably efficient policy optimization for two-player zero-sum markov games. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 2736–2761, 2022.

Chenhao Zhou, Zebang Shen, Chao Zhang, Hanbin Zhao, and Hui Qian. Solving zero-sum markov games with continuous state via spectral dynamic embedding. In *Advances in Neural Information Processing Systems*, volume 37, pp. 68622–68657, 2024. doi: 10.52202/079017-2192.

A Extended Related Work

Discounted stochastic games and equilibrium computation. Discounted stochastic games originate with the fixed-point formulation of Shapley (1953); stationary equilibria for discounted general-sum games were established by Fink (1964), and Littman (1994) brought the model into multi-agent reinforcement learning. Early value-based algorithms include Nash Q-learning (Hu & Wellman, 2003) and correlated Q-learning (Greenwald & Hall, 2003). Under distinct information structures, later work develops batch general-sum learning (Perolat et al., 2017), near-optimal generative-model solvers for discounted turn-based zero-sum games (Sidford et al., 2020), near-minimax model-based guarantees for simultaneous-move zero-sum games with generative access (Zhang et al., 2020), and decentralized last-iterate convergence from online interaction (Wei et al., 2021). For offline data, unilateral coverage is necessary in finite-horizon games (Cui & Du, 2022), while Yan et al. (2024) give a minimax-optimal pessimistic model-based analysis for discounted tabular games with a leading $S(A+B)$ action–state factor under unilateral clipped concentrability. This sharper action dependence belongs to a fixed-data tabular setting. Our online protocol instead generates its own coverage while learning both factors of an unknown low-rank transition, so the statistical comparison is not obtained by substituting $J = AB$ into the offline bound.

Function approximation and representation learning in MDPs. State-space-independent RL guarantees have been developed through Bellman rank and witness rank (Jiang et al., 2017; Sun et al., 2019), latent state decoding and abstraction (Du et al., 2019; Misra et al., 2020), and known linear transition or value features (Yang & Wang, 2019; Jin et al., 2020). Low-rank MDPs make the state–action factor of the transition kernel unknown. Model-based methods learn it using a realizable transition class (Agarwal et al., 2020; Uehara et al., 2022); model-free alternatives use minimax representation objectives, latent block structure, spectral noise assumptions, or barycentric spanners (Modi et al., 2024; Zhang et al., 2022; Ren et al., 2022; Mhammedi et al., 2023). Related variants learn representations through a contrastive objective (Qiu et al., 2022), exploit next-state density features (Huang et al., 2023), or study adversarial losses with an unknown low-rank transition (Zhao et al., 2023). Our statistical layer is closest to the discounted occupancy, finite-class MLE, and fixed-ground-truth potential analysis of Uehara et al. (2022). REP-UCB provides policy-uniform near-optimism and handles changing representations through a fixed true-feature potential. We use this methodology to realize a two-sided envelope and sum the certificate induced by the two policy crossings.

Learning two-player zero-sum Markov games. For finite-horizon tabular games, provable self-play progressed from upper/lower confidence value iteration (Bai & Jin, 2020) to near-optimal model-free and model-based rates (Bai et al., 2020; Liu et al., 2021); the discounted generative-model counterpart was studied by Zhang et al. (2020). With function approximation, existing guarantees cover known linear features or linear-mixture transitions (Xie et al., 2020; Chen et al., 2022), structured-transition policy optimization (Qiu et al., 2021; Zhao et al., 2022), neural or kernel approximation (Li et al., 2022), general complexity measures such as minimax Eluder/Bellman–Eluder dimensions and posterior-sampling coefficients (Huang et al., 2022; Jin et al., 2022; Qiu et al., 2023), and improved finite-horizon model-free rates (Feng et al., 2024). Qiu et al. (2022) also treat low-rank Markov games using contrastive representation recovery. For discounted continuous-state games, Zhou et al. (2024) construct spectral features from structured Gaussian transition dynamics and prove state-space-independent last-iterate convergence for natural policy gradient. Their result concerns spectral policy evaluation and a one-sided equilibrium criterion, whereas we learn both factors of an unknown low-rank transition and certify the two-sided Nash gap from online data. These works either assume tabular or known/structured features, impose a different representation-identification condition, or analyze a finite horizon. Our distinction is the conjunction of an unknown model-based low-rank factorization, a single stationary representation, discounted online occupancies, and a computable Nash-gap certificate.

The planning structure gives a concrete distinction from directly extending the all-player optimistic-payoff construction in Algorithm 1 of Ni et al. (2023). In a zero-sum game, adding a positive bonus to each player's payoff gives $(r+b, -r+b)$, whose sum $2b$ generally depends on the state and action. That auxiliary problem need not remain zero-sum. Our two auxiliary games each retain a zero-sum Shapley operator; their crossed saddle policies supply both a stationary candidate and its exploration rule. A separate nominal game and joint-action bonus MDP is another valid route: Theorem 11

establishes its certificate and interaction upper bounds of the same order. Proposition 6 distinguishes the resulting fixed-envelope bounds. This comparison concerns those planning constructions, not a lower bound on every extension of prior algorithms or a claim of state-independent computation.

General-sum learning, CCE, and computational boundaries. In episodic tabular games, decentralized V-learning and related no-regret procedures learn correlated and coarse correlated equilibria without enumerating the full joint action space (Song et al., 2022; Jin et al., 2024; Mao et al., 2022). Under function approximation, Wang et al. (2023) study decentralized linear approximation, Cui et al. (2023) develop independent linear Markov games, Ni et al. (2023) learn unknown low-rank representations for finite-horizon Nash, correlated, and coarse correlated equilibria, and Fan et al. (2024) analyze independent linear approximation with local simulator access. Offline interval estimation under general function approximation is developed by Zhang et al. (2023), while online model-based learning with known linear features is studied by Yang et al. (2025). Convergence of no-regret dynamics in known general-sum Markov games is analyzed by Erez et al. (2023) and, with sharper CCE rates, by Yorulmaz & Başar (2026). The stationary discounted objective changes the computational picture: Jin et al. (2023) establish PPAD complexity for stationary Nash equilibria, and Daskalakis et al. (2023) prove PPAD-hardness even for approximate stationary Markov CCE while giving a tractable nonstationary alternative. Accordingly, our general-sum theorem assumes a stationary-CCE oracle and isolates its error from representation-estimation and interaction costs.

Table 3: **Closest theorem-level comparisons.** "Unknown" means that both transition factors are learned. Width-based selection and fixed-feature potentials have precedents; the last row combines a stationary discounted certificate with its exploration rule. Rates require aligned reward/gap units and access protocols.

Work	Setting / representation	Confidence and output	Computation / analysis
Uehara et al. (2022)	Discounted MDP; unknown	All-policy near-optimism; near-optimal control	MLE and planning; fixed true-feature potential
Ni et al. (2023)	Finite-horizon MG; unknown	Upper/lower width; Nash, correlated, and coarse correlated candidates	MLE and equilibrium oracles; stagewise analysis
Yang et al. (2025)	Discounted extension; known features	Value-incentivized model objective; NE/CCE regret	Optimization oracles; known-feature structure
This work	Discounted zero-sum MG; unknown	Conservative crossing certifies; reverse crossing explores	Two zero-sum Shapley operators; approximate MLE; stationary occupancies

B Detailed Scope and Limitations

Statistical assumptions. The main theorem assumes a finite realizable transition class, certified approximate MLE, resets to d_0, and uniform local action sampling. Proposition 21 replaces finiteness by two training-error and selected-feature covariance conditions, but gives no concrete infinite-class instantiation; likelihood covering alone does not verify all of these obligations. Known rewards are relaxed only through structured pointwise confidence in Appendix C.4. Extending the proof to predictable nonuniform actions would replace J by a density-ratio term and require adapted covariances; unconstrained policy-dependent sampling lacks the all-policy coverage used here.

Computational scope. The guarantees concern interactions: generic MLE and Shapley computation still scale with the observed state space, and the general-sum result assumes a PPAD-hard stationary-CCE oracle. The discount and joint-action dependences are not established as minimax-optimal. Here $J = AB$ comes from joint-policy importance weighting; $A + B$ requires separate control of unilateral deviations, while existing tabular results use fixed-data coverage (Yan et al., 2024). Known-feature lower bounds also use generative access and a uniform control criterion, so Proposition 15 applies only after aligning protocols. The one-sided separation rules out shared-envelope certificates, not methods that exploit raw data or additional structure.

The CE + bonus-MDP comparator has an interaction upper bound of the same order and can be cheaper to plan. Our minimum-certificate statement is restricted to the best-response certificate of a

fixed envelope. Its attainable unbounded-ratio construction uses a redundant representation and rare histories; it is not a typical-case or information-complexity separation. The saved-data comparison tests actual learned confidence without establishing a benefit from changing the exploration policy.

Candidate quality and certification. The experiments probe specific mechanisms in controlled finite classes. On common data, CE has a numerical-zero final gap and a modestly larger calibrated-envelope certificate than the conservative crossing. Likelihood sets identify the true model and certify CE much more tightly, whereas explicit sufficient proof constants give raw envelope bounds above one at every tested checkpoint. The proof-profile output-selection corollary preserves the theoretical guarantee, but is not the algorithm used in the calibrated exploration studies. Enumeration makes these comparisons feasible; it does not establish scalability to large implicit model classes.

Exploration comparisons. Under fixed current-MLE/CE output and set certification, reverse crossing uses fewer interactions on the deep branch. That class already admits a nominal pair certifiable at 0.05 before interaction, so the finding concerns computation versus interaction under the chosen candidate rule. Random games often certify without data, and the hidden family does not resolve a reverse-versus-CE difference. The delayed-conflict family excludes a common accurate pair analytically. Across six depth/separation settings and balanced true endpoints, bonus MDP and uniform exploration certify faster than calibrated reverse crossing. Upper same-side also has lower means, with one unresolved paired interval. Three of 600 runs lose true-model coverage and falsely certify (two reverse, one uniform); all remain in the results. Depth also changes rank, and actions along the paths are irrelevant, so this construction tests delayed disagreement rather than isolated depth scaling or action-sequence discovery. The two-planner reduction supplies a certified exploration rule, but these comparisons do not establish its empirical superiority.

Coverage and additional boundaries. The proof-profile guarantee does not apply to the calibrated bonus. The finite policy audit covers 299/300 seeds, with a 98.16% lower endpoint of a two-sided 95% confidence interval; policy-uniform validity comes from the theorem. The held-out misspecification study is empirical and establishes neither a misspecified-class guarantee nor monotone degradation. The $\gamma = 0.95$ results are right-censored, the 4,096-state tabular comparison uses three seeds and a restricted budget, and the CCE oracle is supplied only on a team-potential family. These studies do not test broad benchmark performance or establish tight constants or worst-case exponents.

C Supplementary Zero-Sum Statements

C.1 Discounted identities

Proposition 12 (Geometric roll-in)**.** *Starting from $s_0 \sim d_0$, stop at the current state with probability $1-\gamma$, and otherwise take one transition under a stationary joint policy κ. The returned state has marginal $d^{\kappa}_{P,\gamma}(s)$, and the expected number of transitions is $\gamma/(1-\gamma)$.*

Proof. The procedure returns s_t with probability $(1-\gamma)\gamma^t$; summing over t gives equation 2. The transition count is geometric. □

Lemma 13 (Discounted simulation identity)**.** *For transition kernels P, P', bounded payoffs c, c', and a stationary joint policy κ,*

$$v^{\kappa}_{P',c'} - v^{\kappa}_{P,c} = \frac{1}{1-\gamma}\mathbb{E}_{d^{\kappa}_{P',\gamma}}\left[(c'-c) + \gamma(P'-P)V^{\kappa}_{P,c}\right], \tag{34}$$

$$= \frac{1}{1-\gamma}\mathbb{E}_{d^{\kappa}_{P,\gamma}}\left[(c'-c) + \gamma(P'-P)V^{\kappa}_{P',c'}\right]. \tag{35}$$

Its proof appears in Appendix D.2.

C.2 STRATEGIC REDUCTIONS AND A FAILURE EXAMPLE

Proof of Lemma 5. Fix n. For every maximizing policy π, upper strategic confidence and the upper game's approximate-saddle guarantee imply

$$v^{\pi,\nu_n^+}_{P^\star,r} \le v^{\pi,\nu_n^+}_{\widehat{P}_n,r+\widehat{b}_n} + \zeta_n \le \overline{v}_n + \eta_n + \zeta_n.$$

Likewise, for every minimizing policy ν, lower strategic confidence and the lower game's approximate-saddle guarantee give

$$v^{\pi_n^-,\nu}_{P^\star,r} \ge v^{\pi_n^-,\nu}_{\widehat{P}_n,r-\widehat{b}_n} - \zeta_n \ge \underline{v}_n - \eta_n - \zeta_n.$$

Taking the maximum in the first display and the minimum in the second, then subtracting, proves equation 10. □

Proof of Lemma 7. Let V_n^+ and V_n^- denote the exact saddle values of the upper and lower confidence games. The two conditions in equation 5 imply $|\overline{v}_n - V_n^+| \le \eta_n$ and $|\underline{v}_n - V_n^-| \le \eta_n$. Payoff monotonicity gives $V_n^+ \ge V_n^-$, proving the lower bound.

For the upper bound, apply the maximizing-policy side of equation 5 to π_n^+ in the upper game and the minimizing-policy side to ν_n^- in the lower game:

$$\overline{v}_n \le v^{\pi_n^+,\nu_n^-}_{\widehat{P}_n,r+\widehat{b}_n} + \eta_n,$$
$$\underline{v}_n \ge v^{\pi_n^+,\nu_n^-}_{\widehat{P}_n,r-\widehat{b}_n} - \eta_n.$$

Their difference is $2(1-\gamma)^{-1}\mathbb{E}_{d^{\kappa_n}_{\widehat{P}_n,\gamma}}[\widehat{b}_n] + 2\eta_n$, which proves equation 16. □

C.2.1 ONE-SIDED SEPARATION AND FAILURE OF INCORRECT CROSSING

Theorem 14 (One-sided strategic-confidence separation)**.** *For every $\gamma \in (0,1)$, two rank-two discounted zero-sum games can share the same known reward and a common learned kernel with $b = \zeta = 0$ whose values upper-bound both games for every stationary joint policy, while*

$$\inf_{\pi,\nu} \max_{i\in\{1,2\}} \mathrm{Gap}_{P_i}(\pi,\nu) \ge \frac{\gamma}{2}. \tag{36}$$

The analogous lower-only statement holds. Moreover, a zero-residual two-sided envelope can have upper-game and reverse-crossed candidates of gap 1, while the conservative crossing has gap 0 and the value width is $7/10$.

The construction uses two conflicting 2×2 transition matrices and a common learned kernel that always reaches the rewarding absorbing state. Their gaps sum to γ for every policy pair. This restriction concerns the supplied envelope; it does not exclude using the underlying data or a different confidence object. The following proof gives both constructions and verifies the low-rank normalization.

Proof of Theorem 14. One-sided separation. Fix $\gamma \in (0,1)$. Let the state space be $\{s_0, g, \ell\}$, let both players have two actions at s_0, and let d_0 put unit mass on s_0. The states g and ℓ are absorbing, their actions are immaterial, and the common known reward is

$$r(s_0,\cdot,\cdot) = 0, \qquad r(g,\cdot,\cdot) = 1-\gamma, \qquad r(\ell,\cdot,\cdot) = 0.$$

Define

$$M_1 = \begin{pmatrix} 1/2 & 1 \\ 0 & 1/2 \end{pmatrix}, \qquad M_2 = \begin{pmatrix} 1/2 & 0 \\ 1 & 1/2 \end{pmatrix}.$$

For $i \in \{1,2\}$, kernel P_i moves from s_0 to g with probability $M_i(a,b)$ and to ℓ otherwise. Let the common learned kernel $\widetilde{P}$ move from s_0 to g with probability one; all three kernels use the same absorbing transitions at g, ℓ. Their transition rows lie in $\mathrm{span}\{\mathbf{e}_g, \mathbf{e}_\ell\}$, so every kernel has rank at most two. Explicitly, take $\mu(g) = e_1, \mu(\ell) = e_2, \mu(s_0) = 0$ and $\phi(s,a,b) = (P(g \mid s,a,b), P(\ell \mid$

$s, a, b)$). Then $\|\phi\|_2 \le 1$, while $\|\sum_{s'} \mu(s')h(s')\|_2 \le \sqrt{2}$ for every $h : \mathcal{S} \to [0,1]$, verifying the normalization in Assumption 2 with $d = 2$.

For every stationary joint policy κ,

$$v^{\kappa}_{\widetilde{P},r} = \gamma, \qquad v^{\kappa}_{P_i,r} = \gamma\, \mathbb{E}_{(a,b)\sim\kappa(\cdot|s_0)}[M_i(a,b)] \le \gamma.$$

Hence the upper side of the strategic-confidence interface holds simultaneously for both true kernels with $b = 0$ and $\zeta = 0$. If instead the common learned kernel moves from s_0 to ℓ with probability one, its value is zero under every policy and the lower side holds simultaneously for the same two true kernels.

Now let $p = \pi(1 \mid s_0)$ and $q = \nu(1 \mid s_0)$. Direct best-response calculation in the two matrix games gives

$$\mathrm{Gap}_{P_1}(\pi,\nu) = \gamma\left(1 - \frac{p+q}{2}\right), \qquad \mathrm{Gap}_{P_2}(\pi,\nu) = \gamma\frac{p+q}{2}.$$

Their sum is γ, so their maximum is at least $\gamma/2$ for every output pair. Because the reward, learned kernel, bonus, discount, and initial distribution supplied by the one-sided interface are identical in the two instances, no certificate derived only from that interface can be uniformly smaller.

Failure of alternative candidates for the value-width certificate. Consider one state with a deterministic self-loop and two actions per player. Let e_1, e_2 denote the pure actions and define

$$r = (1-\gamma)R, \qquad b = (1-\gamma)B, \qquad R = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} 2 & 3/2 \\ 1/5 & 0 \end{pmatrix}.$$

Set $\widehat{P} = P^\star$ and $\zeta = \eta = 0$. Because $b \ge 0$, Definition 4 holds exactly for every stationary joint policy. The corresponding normalized upper and lower return matrices are

$$R + B = \begin{pmatrix} 2 & 3/2 \\ 6/5 & 1 \end{pmatrix}, \qquad R - B = \begin{pmatrix} -2 & -3/2 \\ 4/5 & 1 \end{pmatrix}.$$

The first matrix has the unique saddle pair (e_1, e_2), while the second has the unique saddle pair (e_2, e_1). Consequently, the upper-game pair is (e_1, e_2), the reverse crossing is (e_1, e_1), and the conservative crossing is (e_2, e_2). Under the true matrix R, their Nash gaps are respectively $1, 1, 0$. The upper and lower game values are $\overline{v} = 3/2$ and $\underline{v} = 4/5$, so their value-width certificate is $\overline{v} - \underline{v} = 7/10$. Both alternative candidates have gap $1 > 7/10$ and therefore violate this certificate, whereas the conservative crossing satisfies it. This example concerns the value-width certificate, rather than uniqueness of the conservative crossing among all possible certification rules. □

C.3 Inherited lower-bound baseline

Proposition 15 (Discounted MDP lower bounds transfer)**.** *Let $\mathfrak{M}$ be any class of discounted low-rank MDPs under a fixed interaction protocol. Embed every MDP as a zero-sum Markov game by giving the minimizing player a single action and keeping the same states, maximizing actions, transition kernel, reward, discount, and initial distribution. Any algorithm that returns an ϵ-Nash pair for every game in the embedded class also returns an ϵ-optimal policy for every MDP in $\mathfrak{M}$, with exactly the same interaction count. Consequently, every minimax sample-complexity lower bound for $\mathfrak{M}$ is also a lower bound for the corresponding game class.*

Proof. The minimizing policy is unique. For the embedded game, Definition 1 therefore reduces to

$$\mathrm{Gap}_P(\pi,\nu) = \max_{\pi'} v^{\pi',\nu}_{P,r} - v^{\pi,\nu}_{P,r},$$

which is exactly the MDP suboptimality of π at d_0. The game and MDP generate identical observations under every maximizing policy, so the reduction neither adds information nor changes the number of environment transitions. □

C.4 Structured unknown rewards

The main theorem treats the normalized payoff as known. The following assumption gives a modular extension through a separate reward confidence sequence.

Assumption 16 (Structured reward confidence)**.** *The unknown payoff $r^\star : \mathcal{S} \times \mathcal{Z} \to [0, 1-\gamma]$ is observed through a reward-estimation protocol that, for any $\delta_{\mathrm{r}} \in (0,1)$, returns history-measurable functions $\widehat{r}_n$ and $b_n^{\mathrm{r}} \geq 0$ such that, with probability at least $1-\delta_{\mathrm{r}}$, simultaneously for every $n \leq N$ and every (s,z),*

$$|\widehat{r}_n(s,z) - r^\star(s,z)| \leq b_n^{\mathrm{r}}(s,z). \tag{37}$$

The estimates and radii are bounded, and the model-based planning oracle can optimize the resulting bounded confidence payoffs.

This condition can be instantiated, for example, by a finite realizable reward class with a uniform likelihood guarantee or by a known-feature linear reward model with a valid regression radius. Such structure supplies pointwise confidence; useful sample complexity additionally requires the radius to be small on the exploration occupancies, which is exposed explicitly below.

Proposition 17 (Unknown-reward composition)**.** *Suppose Assumptions 2 and 16 hold, together with Assumption 3 after replacing its known-reward clause by the reward feedback in Assumption 16. Run the transition learner with confidence level $\delta/2$, write $\widehat{b}_n^{\mathrm{P}} := \widehat{b}_n$, and in iteration n solve the two confidence games with base payoff $\widehat{r}_n$ and total bonus $\widehat{b}_n^{\mathrm{P}} + b_n^{\mathrm{r}}$. Let κ_n be their reverse-crossed exploration policy and define*

$$\mathfrak{R}_N^{\mathrm{r}} := \frac{2}{N(1-\gamma)} \sum_{n=1}^{N} \mathbb{E}_{d^{\kappa_n}_{\widehat{P}_n,\gamma}}[b_n^{\mathrm{r}}]. \tag{38}$$

Using reward confidence level $\delta/2$, the best certified crossed candidate satisfies, with probability at least $1-\delta$,

$$\mathrm{Gap}_{P^\star,r^\star}(\widehat{\pi},\widehat{\nu}) \leq C\left[\frac{\mathfrak{K}_{d,J}}{(1-\gamma)^2}\sqrt{\frac{\overline{L}_N \ell_N}{N}} + \sqrt{\frac{J\overline{L}_N}{(1-\gamma)N}}\right] + \mathfrak{R}_N^{\mathrm{r}} + \frac{4}{N}\sum_{n=1}^{N}\eta_n, \tag{39}$$

where $\overline{L}_N$ is evaluated at confidence level $\delta/2$. Consequently, the known-reward rate is retained whenever the structured reward estimator guarantees $\mathfrak{R}_N^{\mathrm{r}} = O(\epsilon)$.

Proof. Intersect the transition-confidence event at level $\delta/2$ with the event in equation 37 at level $\delta/2$. The union bound gives probability at least $1-\delta$. For every stationary joint policy κ, pointwise reward confidence, payoff monotonicity, and Theorem 8 applied to the true payoff $r^\star$ give

$$\begin{aligned} v^{\kappa}_{\widehat{P}_n,\widehat{r}_n+\widehat{b}_n^{\mathrm{P}}+b_n^{\mathrm{r}}} &\geq v^{\kappa}_{\widehat{P}_n,r^\star+\widehat{b}_n^{\mathrm{P}}} \geq v^{\kappa}_{P^\star,r^\star} - \zeta_n, \\ v^{\kappa}_{\widehat{P}_n,\widehat{r}_n-\widehat{b}_n^{\mathrm{P}}-b_n^{\mathrm{r}}} &\leq v^{\kappa}_{\widehat{P}_n,r^\star-\widehat{b}_n^{\mathrm{P}}} \leq v^{\kappa}_{P^\star,r^\star} + \zeta_n. \end{aligned}$$

Thus the modified games provide the same policy-uniform two-sided envelope for the true payoff. The proof of Lemma 5 therefore certifies the conservative crossing without change.

Applying the proof of Lemma 7 with total bonus $\widehat{b}_n^{\mathrm{P}} + b_n^{\mathrm{r}}$ yields

$$\overline{v}_n - \underline{v}_n \leq \frac{2}{1-\gamma}\mathbb{E}_{d^{\kappa_n}_{\widehat{P}_n,\gamma}}[\widehat{b}_n^{\mathrm{P}} + b_n^{\mathrm{r}}] + 2\eta_n.$$

All transition concentration and fixed-$\phi^\star$ potential arguments are uniform over the predictable exploration policies, so the proof of Theorem 9 bounds the $\widehat{b}_n^{\mathrm{P}}$ part exactly as before. After averaging, the remaining reward-radius contribution is precisely $\mathfrak{R}_N^{\mathrm{r}}$ from equation 38. The best-iterate rule and the preceding candidate certificate complete equation 39. □

Discount accounting. Bonus accumulation contributes $(1-\gamma)^{-1}$, and occupancy transfer contributes $\gamma(1-\gamma)^{-2}$, as retained in $\mathfrak{K}_{d,J,\gamma}$. Bounding both by $(1-\gamma)^{-2}$ and squaring gives the displayed $(1-\gamma)^{-4}$ outer-iteration bound. Geometric roll-in adds one effective-horizon factor to the primitive transition count.

D COMPLETE PROOFS

This appendix proves Theorems 8 and 9. The proof has three layers. First, adaptive MLE and covariance concentration define a single good event. Second, one-step-back inequalities convert errors measured by the training distributions into errors under arbitrary discounted occupancies. Third, a fixed-feature log-determinant argument sums the exploration widths.

D.1 NOTATION, FILTRATION, AND TRAINING DISTRIBUTIONS

Write $\mathcal{Z} = \mathcal{A} \times \mathcal{B}$, so $|\mathcal{Z}| = J$. A stationary joint-action policy is a map $\kappa : \mathcal{S} \to \Delta(\mathcal{Z})$. Every product pair (π, ν) induces $\kappa_{\pi,\nu}(a, b \mid s) = \pi(a \mid s)\nu(b \mid s)$. All statistical statements below hold for arbitrary κ, and hence for every product policy and every unilateral best response used in the Nash gap.

Symbol	Meaning
$J = AB,\ \mathcal{Z} = \mathcal{A} \times \mathcal{B}$	Joint-action count and space.
$d^{\kappa}_{P,\gamma}$	Normalized discounted state–joint-action occupancy.
ρ_n, ρ'_n	Context laws for the first and shifted MLE tuples.
$\widehat{P}_n, f_n$	Learned kernel and its pointwise ℓ_1 error.
$\widehat{\Sigma}_n, \Sigma_{n,\star}$	Learned-feature empirical covariance and fixed-$\phi^\star$ population covariance.
$\widehat{b}_n$	Self-normalized bonus built from $\widehat{\phi}_n$.
$L_n, \overline{L}_n, \mathfrak{G}_n$	Finite-class log complexity, tolerance-inflated complexity, and generic statistical-interface complexity.
$\lambda_n, \alpha_n, \zeta_n, \eta_n$	Ridge level, bonus scale, statistical residual, and planning error.
$\mathfrak{K}_{d,J}$	Dimension–action coefficient in the certificate bound.
$\widehat{\Delta}_n$	Observable crossed-policy Nash certificate.

The proof dependency is Theorem 8 $\to$ Lemmas 5 and 7 $\to$ Lemma 24, the training-moment bounds, and Lemma 22 $\to$ Lemma 27 $\to$ Theorem 9. Corollary 10 then substitutes the certified MLE and planning tolerances into the PAC bound.

Let

$$\kappa_n := \kappa_{\pi_n^+, \nu_n^-}, \qquad d_n := d^{\kappa_n}_{P^\star,\gamma}.$$

The initial policy κ_0 is uniform. Before fitting the n-th model, the first transition contexts have mixture distribution

$$\rho_n(s, z) := \left(\frac{1}{n}\sum_{i=0}^{n-1} d_i(s)\right)\frac{1}{J}, \tag{40}$$

$$\rho'_n(s', z') := \left[\sum_{s,z} \rho_n(s, z)P^\star(s' \mid s, z)\right]\frac{1}{J}. \tag{41}$$

The distribution ρ'_n is the context distribution of the second tuple. For the potential argument, define the mixture of the actual exploration occupancies

$$\overline{d}_n := \frac{1}{n}\sum_{i=0}^{n-1} d_i \quad \text{and} \quad \Sigma_{n,\star} := n\mathbb{E}_{\overline{d}_n}[\phi^\star\phi^{\star\top}] + \lambda_n I_d. \tag{42}$$

For a candidate feature $\phi \in \Phi$, also define

$$\Sigma^{\mathrm{tr}}_{n,\phi} := n\mathbb{E}_{\rho_n}[\phi\phi^\top] + \lambda_n I_d, \qquad \widehat{\Sigma}_{n,\phi} := \sum_{(s,z,\cdot)\in\mathcal{D}_n} \phi(s, z)\phi(s, z)^\top + \lambda_n I_d. \tag{43}$$

Thus $\widehat{\Sigma}_n = \widehat{\Sigma}_{n,\widehat{\phi}_n}$. Finally, let

$$f_n(s, z) := \|\widehat{P}_n(\cdot \mid s, z) - P^\star(\cdot \mid s, z)\|_1.$$

Let $\mathcal{F}_{n,0}$ be the sigma-field immediately before the n-th geometric roll-in. Within an iteration, enlarge the filtration after the roll-in, after the first transition, and after the second transition. Conditional on the corresponding context, each successor state is sampled from $P^\star$. The policies and all model estimates are measurable with respect to the past. This is the only adaptivity used below.

D.2 Proof of the simulation identity

Proof of Lemma 13. The policy-evaluation equations give

$$V^{\kappa}_{P',c'} - V^{\kappa}_{P,c} = (I - \gamma P'_{\kappa})^{-1} \left[c'_{\kappa} - c_{\kappa} + \gamma (P'_{\kappa} - P_{\kappa}) V^{\kappa}_{P,c} \right].$$

Multiplying by d_0 and expanding the resolvent $(I - \gamma P'_{\kappa})^{-1}$ yields the normalized occupancy in equation 34. Expanding the same difference with $(I - \gamma P_{\kappa})^{-1}$ proves equation 35. □

D.3 The simultaneous statistical good event

Choose c_0 so that $L_n \geq 1$, then fix sufficiently large c_{mle} and c_λ for the two concentration lemmas below. We take $c_{\mathrm{cov}} = 2$ and choose c_α, c_ζ afterward, as specified in the confidence proof. All these constants are universal. Set

$$\xi_n := c_{\mathrm{mle}} \frac{\overline{L}_n}{n}.$$

We use the event $\mathcal{E} = \mathcal{E}_{\mathrm{mle}} \cap \mathcal{E}_{\mathrm{cov}}$, where

$$\mathcal{E}_{\mathrm{mle}} := \left\{ \mathbb{E}_{\rho_n}[f_n^2] \leq \xi_n, \quad \mathbb{E}_{\rho'_n}[f_n^2] \leq \xi_n \text{ for every } n \leq N \right\}, \tag{44}$$

$$\mathcal{E}_{\mathrm{cov}} := \left\{ c_{\mathrm{cov}}^{-1} \Sigma^{\mathrm{tr}}_{n,\phi} \preceq \widehat{\Sigma}_{n,\phi} \preceq c_{\mathrm{cov}} \Sigma^{\mathrm{tr}}_{n,\phi} \text{ for every } n \leq N,\ \phi \in \Phi \right\}. \tag{45}$$

Lemma 18 (Reverse bounded-process inequality)**.** *Let $(X_i, \mathcal{F}_i)_{i \geq 1}$ be an adapted process with $0 \leq X_i \leq B$ for some $B > 0$, and write $\mu_i := \mathbb{E}[X_i \mid \mathcal{F}_{i-1}]$. For every $\varepsilon \in (0,1)$, with probability at least $1 - \varepsilon$, simultaneously for all $n \geq 1$,*

$$\sum_{i=1}^{n} \mu_i \leq 2 \sum_{i=1}^{n} X_i + 2B \log \frac{1}{\varepsilon}.$$

Proof. Let $c := 1 - e^{-1}$. For $x \in [0,1]$, convexity gives $e^{-x} \leq 1 - cx$. Consequently,

$$Z_n := \exp\left(\frac{c}{B} \sum_{i=1}^{n} \mu_i - \frac{1}{B} \sum_{i=1}^{n} X_i \right)$$

is a nonnegative supermartingale, because

$$\mathbb{E}[e^{-X_i/B} \mid \mathcal{F}_{i-1}] \leq 1 - \frac{c\mu_i}{B} \leq e^{-c\mu_i/B}.$$

Ville's inequality gives $\sup_{n \geq 1} Z_n \leq 1/\varepsilon$ with probability at least $1 - \varepsilon$. Rearranging on this event and using $c^{-1} < 2$ proves the claim. □

Lemma 19 (Adaptive finite-class MLE)**.** *For a sufficiently large c_{mle}, $\Pr(\mathcal{E}_{\mathrm{mle}}) \geq 1 - \delta/4$.*

Proof. Order the two successor-state observations in each iteration chronologically and denote the resulting context–outcome pairs by $(X_t, Y_t)_{t=1}^{2n}$. For a fixed $P \in \mathcal{M}$, define the conditional squared Hellinger distance

$$h_t^2(P) := \frac{1}{2} \sum_{y \in \mathcal{S}} \left(\sqrt{P(y \mid X_t)} - \sqrt{P^\star(y \mid X_t)} \right)^2.$$

The process

$$M_m(P) := \prod_{t=1}^{m} \sqrt{\frac{P(Y_t \mid X_t)}{P^\star(Y_t \mid X_t)}} \exp\left(\sum_{t=1}^{m} h_t^2(P) \right)$$

is a nonnegative supermartingale. Indeed, conditional on the context and the past,

$$\mathbb{E}\left[\sqrt{\frac{P(Y_t \mid X_t)}{P^\star(Y_t \mid X_t)}} \,\middle|\, X_t, \mathcal{F}_{t-1} \right] = \sum_{y} \sqrt{P(y \mid X_t) P^\star(y \mid X_t)} = 1 - h_t^2(P) \leq e^{-h_t^2(P)}.$$

Ville's inequality and a union bound over $P \in \mathcal{M}$ imply that, for each fixed n, with probability at least $1 - \delta/(32(n+1)^2)$, every $P \in \mathcal{M}$ satisfies

$$\sum_{t=1}^{2n} h_t^2(P) \le 2\log\left(\frac{32|\mathcal{M}|(n+1)^2}{\delta}\right) + \frac{1}{2}[\mathcal{L}_n(P^\star) - \mathcal{L}_n(P)]\,. \tag{46}$$

Indeed, this follows by taking logarithms of $M_{2n}(P)$ and rearranging; the factor 2 only enlarges the confidence term. Realizability and equation 18 give $\mathcal{L}_n(P^\star) - \mathcal{L}_n(\widehat{P}_n) \le \tau_n$. Hence the right-hand side for $\widehat{P}_n$ is $O(L_n + \tau_n)$.

It remains to replace the realized contexts by their conditional distributions. For iteration i, group the two context losses as

$$H_i(P) := h_{i,1}^2(P) + h_{i,2}^2(P) \in [0, 2],$$

where the subscripts identify the first and second tuples. Conditional on $\mathcal{F}_{i,0}$, the expectation of $H_i(P)$ is the sum of the Hellinger errors under that iteration's first-context distribution and its one-step-shifted second-context distribution. Thus, after summing over $i \le n$, these conditional expectations equal

$$n\mathbb{E}_{\rho_n}[h^2(P^\star,P)] + n\mathbb{E}_{\rho'_n}[h^2(P^\star,P)].$$

Apply Lemma 18 with $B = 2$, and take a union bound over $P \in \mathcal{M}$. With the corresponding share of the failure probability, simultaneously over $P \in \mathcal{M}$ and $n \ge 1$,

$$\sum_{i=1}^{n} \mathbb{E}[H_i(P) \mid \mathcal{F}_{i,0}] \le 2\sum_{i=1}^{n} H_i(P) + c\log\left(\frac{32|\mathcal{M}|(n+1)^2}{\delta}\right). \tag{47}$$

Moreover, $\|p - q\|_1^2 \le 8h^2(p, q)$. Combining equations 46 and 47, dividing by n, and enlarging the universal constant gives

$$\mathbb{E}_{\rho_n}[f_n^2] + \mathbb{E}_{\rho'_n}[f_n^2] \le c_{\mathrm{mle}}\frac{\overline{L}_n}{n}.$$

A union bound using $\sum_{n\ge1}(n+1)^{-2} < 1$ proves the lemma. □

Lemma 20 (Uniform covariance comparison)**.** *For sufficiently large c_λ, the choice of λ_n in equation 21 satisfies* $\Pr(\mathcal{E}_{\mathrm{cov}}) \ge 1 - \delta/4$.

Proof. Fix $\phi \in \Phi$. Let $x_i = \phi(s_i, z_i)$ be the first context feature in roll-in i, and write

$$G_{n,\phi} = \sum_{i=1}^{n} \mathbb{E}[x_i x_i^\top \mid \mathcal{F}_{i,0}], \qquad \widehat{G}_{n,\phi} = \sum_{i=1}^{n} x_i x_i^\top.$$

The conditional first-context law gives $G_{n,\phi} = n\mathbb{E}_{\rho_n}[\phi\phi^\top]$. In particular, $\|x_i\|_2 \le 1$, and both matrices lie between 0 and nI_d in the Loewner order.

We first establish a scalar bound that allows adaptive contexts. For a fixed unit vector u, let

$$Y_i = (u^\top x_i)^2, \quad \mu_i = \mathbb{E}[Y_i \mid \mathcal{F}_{i,0}], \quad S_n = \sum_{i=1}^{n} Y_i, \quad T_n = \sum_{i=1}^{n} \mu_i.$$

Here $0 \le Y_i \le 1$. For $\theta = 1/4$, convexity on $[0, 1]$ yields

$$\mathbb{E}[e^{\theta Y_i} \mid \mathcal{F}_{i,0}] \le 1 + (e^\theta - 1)\mu_i \le e^{(e^\theta-1)\mu_i},$$
$$\mathbb{E}[e^{-\theta Y_i} \mid \mathcal{F}_{i,0}] \le 1 - (1 - e^{-\theta})\mu_i \le e^{-(1-e^{-\theta})\mu_i}.$$

Thus, under the filtration after each iteration's tuple observations, both

$$Z_n^+ = \exp\{\theta S_n - (e^\theta - 1)T_n\}, \qquad Z_n^- = \exp\{(1 - e^{-\theta})T_n - \theta S_n\}$$

are nonnegative supermartingales starting at one. Ville's inequality, together with

$$\frac{e^{1/4} - 1}{1/4} < \frac{5}{4}, \qquad \frac{1 - e^{-1/4}}{1/4} > \frac{3}{4},$$

implies that, for any $t > 0$, except on an event of probability at most $2e^{-t}$,

$$|S_n - T_n| \le \frac{1}{4} T_n + 4t. \tag{48}$$

For each $n \ge 1$, choose a deterministic Euclidean ε_n-net $\mathcal{N}_n$ of the unit sphere, where

$$\varepsilon_n = \frac{1}{16(n+1)}, \qquad |\mathcal{N}_n| \le [1 + 32(n+1)]^d.$$

Apply the scalar bound with

$$t_n = \log \frac{8|\Phi|(n+1)^2 |\mathcal{N}_n|}{\delta}.$$

A union bound over n, ϕ, and $u \in \mathcal{N}_n$ has total failure probability at most

$$\sum_{n \ge 1} |\Phi|\, |\mathcal{N}_n|\, 2e^{-t_n} = \frac{\delta}{4} \sum_{n \ge 1} (n+1)^{-2} < \frac{\delta}{4}.$$

On the complementary event, writing $D_{n,\phi} = \widehat{G}_{n,\phi} - G_{n,\phi}$, we have

$$|u^\top D_{n,\phi} u| \le \frac{1}{4} u^\top G_{n,\phi} u + 4t_n \qquad (u \in \mathcal{N}_n).$$

To extend this bound to every unit vector v, choose $u \in \mathcal{N}_n$ with $\|u - v\|_2 \le \varepsilon_n$. For every symmetric matrix A,

$$|v^\top A v - u^\top A u| \le 2\varepsilon_n \|A\|_{\mathrm{op}}.$$

Since $\|G_{n,\phi}\|_{\mathrm{op}} \le n$ and $\|D_{n,\phi}\|_{\mathrm{op}} \le n$, it follows that

$$\begin{aligned} |v^\top D_{n,\phi} v| &\le \frac{1}{4} v^\top G_{n,\phi} v + 4t_n + \frac{5}{2} \varepsilon_n n \\ &\le \frac{1}{4} v^\top G_{n,\phi} v + a_n, \end{aligned}$$

where the deterministic quantity

$$a_n := 4 \log \frac{8|\Phi|(n+1)^2 [1 + 32(n+1)]^d}{\delta} + 1$$

is an upper bound on the additive error. Consequently,

$$\frac{3}{4} G_{n,\phi} - a_n I_d \preceq \widehat{G}_{n,\phi} \preceq \frac{5}{4} G_{n,\phi} + a_n I_d.$$

Because $|\Phi| \le |\mathcal{M}|$, $d \ge 1$, and $\overline{L}_n \ge L_n$, we have

$$a_n \le C \left[d \log(n+1) + \log \frac{|\Phi|(n+1)^2}{\delta} \right] \le C' d L_n \le C' d \overline{L}_n.$$

Choose the universal c_λ in equation 21 sufficiently large that $\lambda_n \ge 2a_n$ for all n. Adding $\lambda_n I_d$ then gives, simultaneously over n, ϕ,

$$\frac{1}{2}(G_{n,\phi} + \lambda_n I_d) \preceq \widehat{G}_{n,\phi} + \lambda_n I_d \preceq 2(G_{n,\phi} + \lambda_n I_d).$$

This proves $\Pr(\mathcal{E}_{\mathrm{cov}}) \ge 1 - \delta/4$, with $c_{\mathrm{cov}} = 2$. □

Lemmas 19 and 20 show that

$$\Pr(\mathcal{E}) \ge 1 - \delta/2. \tag{49}$$

All remaining statistical arguments are deterministic on $\mathcal{E}$.

D.4 ESTIMATOR-AGNOSTIC STATISTICAL INTERFACE

Proposition 21 (Statistical-interface generalization)**.** *Retain the true-kernel factorization and normalization in Assumption 2, and the sampling and planning access in Assumption 3. The model class may now be arbitrary. Replace finite-class MLE by an estimator that returns a valid rank-d factorization $(\widehat{P}_n, \widehat{\mu}_n, \widehat{\phi}_n)$ satisfying the same normalization. Suppose there is a deterministic nondecreasing sequence $\mathfrak{G}_n \geq 1$ and a universal c_{int} such that, with probability at least $1-\delta$, simultaneously for every $n \leq N$,*

$$\max\{\mathbb{E}_{\rho_n}[f_n^2],\, \mathbb{E}_{\rho'_n}[f_n^2]\} \leq \frac{\mathfrak{G}_n}{n}, \tag{50}$$

$$c_{\mathrm{int}}^{-1}\Sigma^{\mathrm{tr}}_{n,\widehat{\phi}_n} \preceq \widehat{\Sigma}_n \preceq c_{\mathrm{int}}\Sigma^{\mathrm{tr}}_{n,\widehat{\phi}_n}. \tag{51}$$

For sufficiently large constants depending only on c_{int}, run DREP-ZS *with*

$$\lambda_n = c_\lambda \max\{1, d\mathfrak{G}_n\}, \qquad \alpha_n = c_\alpha\sqrt{(J+d^2)\mathfrak{G}_n}, \qquad \zeta_n = c_\zeta\sqrt{\frac{J\mathfrak{G}_n}{n(1-\gamma)}}. \tag{52}$$

Then, on the same event,

$$\mathrm{Gap}_{P^\star}(\widehat{\pi}, \widehat{\nu}) \leq C\left[\frac{\mathfrak{K}_{d,J}}{(1-\gamma)^2}\sqrt{\frac{\mathfrak{G}_N\ell_N}{N}} + \sqrt{\frac{J\mathfrak{G}_N}{(1-\gamma)N}}\right] + \frac{4}{N}\sum_{n=1}^{N}\eta_n. \tag{53}$$

Proof. The proof of Theorem 8 uses randomness only through the two training-error bounds in equation 44 and the selected-feature instance of equation 45. Conditions (50) and (51) provide exactly these two statements with $\xi_n = \mathfrak{G}_n/n$ and $c_{\mathrm{cov}} = c_{\mathrm{int}}$. Substituting this ξ_n into equation 58, the choices in equation 52 dominate its self-normalized and initial-distribution terms. The complement-payoff argument is deterministic, so the two-sided strategic-confidence conclusion remains simultaneous over all stationary joint policies.

For the cumulative argument, equation 50 gives $n\mathbb{E}_{\rho_n}[f_n^2] \leq \mathfrak{G}_n$, while equation 51 gives $n\mathbb{E}_{\rho_n}[\widehat{b}_n^2] \leq c_{\mathrm{int}}\alpha_n^2 d$. These are the only statistical inputs to Lemma 25. The occupancy-transfer lemma and the fixed-$\phi^\star$ log-determinant potential are unchanged. Repeating the summation in the proof of Theorem 9, with $\overline{L}_n$ replaced by $\mathfrak{G}_n$, yields the right-hand side of equation 53. Finally, Lemma 5 and the best-iterate rule convert that average certificate bound into the displayed Nash gap. □

For the finite realizable class, Lemmas 19 and 20 instantiate the interface with $\mathfrak{G}_n = O(\overline{L}_n)$. For an infinite class, equations 50–51 are explicit proof obligations: a covering-number statement for likelihood alone is insufficient unless it also controls approximation error, likelihood tails or support, and the data-dependent selected feature.

D.5 ONE-STEP-BACK INEQUALITIES

The next two lemmas are the discounted replacement for finite-horizon backward induction. They hold for any nonnegative $g : \mathcal{S}\times\mathcal{Z} \to [0, G]$.

Lemma 22 (True-model one-step-back inequality)**.** *For every stationary joint policy κ,*

$$\mathbb{E}_{d^\kappa_{P^\star,\gamma}}[g] \leq \sqrt{\gamma}\,\mathbb{E}_{d^\kappa_{P^\star,\gamma}}\left[\|\phi^\star(s,z)\|_{\Sigma^{-1}_{n,\star}}\right]\sqrt{nJ\mathbb{E}_{\rho_n}[g^2] + \lambda_n dG^2} + \sqrt{(1-\gamma)J\mathbb{E}_{\rho_n}[g^2]}. \tag{54}$$

Proof. Let $\upsilon = d^\kappa_{P^\star,\gamma}$, and define

$$u_g := \sum_{s\in\mathcal{S}}\mu^\star(s)\sum_{z\in\mathcal{Z}}\kappa(z\mid s)g(s,z).$$

The discounted occupancy recursion and the low-rank factorization give

$$\mathbb{E}_\upsilon[g] = (1-\gamma)\mathbb{E}_{d_0,\kappa}[g] + \gamma\,\mathbb{E}_{(\widetilde{s},\widetilde{z})\sim\upsilon}[\phi^\star(\widetilde{s},\widetilde{z})^\top u_g]. \tag{55}$$

The second term is at most

$$\gamma \mathbb{E}_{\upsilon}[\|\phi^\star\|_{\Sigma_{n,\star}^{-1}}]\|u_g\|_{\Sigma_{n,\star}}.$$

Because $0 \le g \le G$, Assumption 2 gives $\|u_g\|_2 \le G\sqrt{d}$. Jensen's inequality and the definition of $\overline{d}_n$ give

$$\begin{aligned}\|u_g\|_{\Sigma_{n,\star}}^2 &= n\mathbb{E}_{(\widetilde{s},\widetilde{z})\sim\overline{d}_n}[(\phi^\star(\widetilde{s},\widetilde{z})^\top u_g)^2] + \lambda_n\|u_g\|_2^2 \\ &\le n\mathbb{E}_{(\widetilde{s},\widetilde{z})\sim\overline{d}_n,\, s\sim P^\star(\cdot|\widetilde{s},\widetilde{z}),\, z\sim\kappa(\cdot|s)}[g^2(s,z)] + \lambda_n d G^2.\end{aligned}$$

The state marginal in equation 40 obeys

$$\rho_n(s) = (1-\gamma)d_0(s) + \gamma\mathbb{E}_{(\widetilde{s},\widetilde{z})\sim\overline{d}_n} P^\star(s \mid \widetilde{s},\widetilde{z}).$$

Importance sampling from $\kappa(\cdot \mid s)$ to the uniform distribution on $\mathcal{Z}$ therefore implies

$$\gamma\mathbb{E}_{\overline{d}_n,P^\star,\kappa}[g^2] \le J\mathbb{E}_{\rho_n}[g^2].$$

Consequently,

$$\gamma\|u_g\|_{\Sigma_{n,\star}} \le \sqrt{\gamma}\sqrt{nJ\mathbb{E}_{\rho_n}[g^2] + \lambda_n d G^2}.$$

Finally, $\rho_n(s) \ge (1-\gamma)d_0(s)$, so Cauchy–Schwarz and the same importance sampling argument yield

$$(1-\gamma)\mathbb{E}_{d_0,\kappa}[g] \le \sqrt{(1-\gamma)J\mathbb{E}_{\rho_n}[g^2]}.$$

Substitution into equation 55 proves the lemma. □

Lemma 23 (Learned-model one-step-back inequality)**.** *On $\mathcal{E}_{\mathrm{mle}}$, for every stationary joint policy κ,*

$$\begin{aligned}\mathbb{E}_{d^\kappa_{\widehat{P}_n,\gamma}}[g] \le \sqrt{\gamma}\,\mathbb{E}_{d^\kappa_{\widehat{P}_n,\gamma}}\left[\|\widehat{\phi}_n(s,z)\|_{(\Sigma^{\mathrm{tr}}_{n,\widehat{\phi}_n})^{-1}}\right] \\ \times \sqrt{2nJ\mathbb{E}_{\rho'_n}[g^2] + 2nG^2\xi_n + \lambda_n d G^2} \\ + \sqrt{(1-\gamma)J\mathbb{E}_{\rho_n}[g^2]}.\end{aligned} \tag{56}$$

Proof. Let $\widehat{d} = d^\kappa_{\widehat{P}_n,\gamma}$, and define

$$\widehat{u}_g := \sum_s \widehat{\mu}_n(s)\sum_z \kappa(z \mid s)g(s,z).$$

The learned-model occupancy recursion gives

$$\mathbb{E}_{\widehat{d}}[g] = (1-\gamma)\mathbb{E}_{d_0,\kappa}[g] + \gamma\mathbb{E}_{\widehat{d}}[\widehat{\phi}_n^\top\widehat{u}_g].$$

Cauchy–Schwarz in the $\Sigma^{\mathrm{tr}}_{n,\widehat{\phi}_n}$-geometry bounds the second term by

$$\gamma\mathbb{E}_{\widehat{d}}\left[\|\widehat{\phi}_n\|_{(\Sigma^{\mathrm{tr}}_{n,\widehat{\phi}_n})^{-1}}\right]\|\widehat{u}_g\|_{\Sigma^{\mathrm{tr}}_{n,\widehat{\phi}_n}}.$$

As before, $\|\widehat{u}_g\|_2 \le G\sqrt{d}$. For a training context $x = (s,z)$, let

$$m_{\widehat{P}}(x) = \mathbb{E}_{s'\sim\widehat{P}_n(\cdot|x),\, z'\sim\kappa(\cdot|s')}g(s',z'), \quad m_\star(x) = \mathbb{E}_{s'\sim P^\star(\cdot|x),\, z'\sim\kappa(\cdot|s')}g(s',z').$$

Then $|m_{\widehat{P}}(x) - m_\star(x)| \le G f_n(x)$. Therefore,

$$\begin{aligned}\|\widehat{u}_g\|^2_{\Sigma^{\mathrm{tr}}_{n,\widehat{\phi}_n}} &= n\mathbb{E}_{\rho_n}[m^2_{\widehat{P}}] + \lambda_n\|\widehat{u}_g\|_2^2 \\ &\le 2n\mathbb{E}_{\rho_n}[m^2_\star] + 2nG^2\mathbb{E}_{\rho_n}[f_n^2] + \lambda_n d G^2 \\ &\le 2nJ\mathbb{E}_{\rho'_n}[g^2] + 2nG^2\xi_n + \lambda_n d G^2.\end{aligned}$$

The last inequality uses Jensen, the definition of ρ'_n, and importance sampling over the joint action. The initial-distribution term is bounded exactly as in the proof of Lemma 22. □

D.6 PROOF OF TWO-SIDED CONFIDENCE

Proof of Theorem 8. Fix n and a stationary joint policy κ, and work on $\mathcal{E}$. The simulation identity in Lemma 13 gives

$$(1-\gamma)\left(v^{\kappa}_{\widehat{P}_n, r+\widehat{b}_n} - v^{\kappa}_{P^\star, r}\right) = \mathbb{E}_{d^{\kappa}_{\widehat{P}_n,\gamma}}\left[\widehat{b}_n + \gamma(\widehat{P}_n - P^\star)V^{\kappa}_{P^\star,r}\right]. \tag{57}$$

Because $r \in [0, 1-\gamma]$, $\|V^{\kappa}_{P^\star,r}\|_\infty \le 1$. Hence the right-hand side of equation 57 is at least

$$\mathbb{E}_{d^{\kappa}_{\widehat{P}_n,\gamma}}[\widehat{b}_n - f_n].$$

Apply Lemma 23 with $g = f_n$ and $G = 2$. On $\mathcal{E}_{\mathrm{mle}}$, both training second moments are bounded by ξ_n, so

$$\mathbb{E}_{d^{\kappa}_{\widehat{P}_n,\gamma}}[f_n] \le q^{\mathrm{tr}}_{n,\kappa}\sqrt{2nJ\xi_n + 8n\xi_n + 4\lambda_n d} + \sqrt{(1-\gamma)J\xi_n}, \tag{58}$$

where

$$q^{\mathrm{tr}}_{n,\kappa} := \mathbb{E}_{d^{\kappa}_{\widehat{P}_n,\gamma}}\left[\|\widehat{\phi}_n\|_{(\Sigma^{\mathrm{tr}}_{n,\widehat{\phi}_n})^{-1}}\right].$$

The covariance event implies

$$\|\widehat{\phi}_n(s,z)\|_{\widehat{\Sigma}_n^{-1}} \ge c_{\mathrm{cov}}^{-1/2}\|\widehat{\phi}_n(s,z)\|_{(\Sigma^{\mathrm{tr}}_{n,\widehat{\phi}_n})^{-1}}.$$

Since $n\xi_n = c_{\mathrm{mle}}\overline{L}_n$, $\lambda_n d = c_\lambda d^2 \overline{L}_n$, and $c_{\mathrm{cov}} = 2$, the choices

$$c_\alpha^2 \ge 20c_{\mathrm{mle}} + 8c_\lambda, \qquad c_\zeta^2 \ge c_{\mathrm{mle}}$$

give

$$c_{\mathrm{cov}}\big(2nJ\xi_n + 8n\xi_n + 4\lambda_n d\big) \le \alpha_n^2.$$

Thus the empirical bonus satisfies

$$\mathbb{E}_{d^{\kappa}_{\widehat{P}_n,\gamma}}[\widehat{b}_n] \ge q^{\mathrm{tr}}_{n,\kappa}\sqrt{2nJ\xi_n + 8n\xi_n + 4\lambda_n d}.$$

Combining the last three displays yields

$$v^{\kappa}_{\widehat{P}_n, r+\widehat{b}_n} - v^{\kappa}_{P^\star,r} \ge -\sqrt{\frac{J\xi_n}{1-\gamma}} \ge -\zeta_n,$$

where the final inequality follows from the choice of c_ζ.

For the other sign, define $\bar{r}(s,z) = (1-\gamma) - r(s,z)$. This payoff also lies in $[0, 1-\gamma]$, and for every P, κ,

$$v^{\kappa}_{P,\bar{r}} = 1 - v^{\kappa}_{P,r}.$$

Applying the upper-confidence result to $\bar{r}$ gives

$$v^{\kappa}_{\widehat{P}_n, \bar{r}+\widehat{b}_n} \ge v^{\kappa}_{P^\star,\bar{r}} - \zeta_n.$$

Substituting $v^{\kappa}_{\widehat{P}_n,\bar{r}+\widehat{b}_n} = 1 - v^{\kappa}_{\widehat{P}_n, r-\widehat{b}_n}$ proves equation 23. Moreover, discounted occupancy gives

$$v^{\kappa}_{\widehat{P}_n, r\pm\widehat{b}_n} = v^{\kappa}_{\widehat{P}_n,r} \pm \frac{1}{1-\gamma}\mathbb{E}_{d^{\kappa}_{\widehat{P}_n,\gamma}}[\widehat{b}_n].$$

Rearranging equation 22 bounds $v^{\kappa}_{P^\star,r} - v^{\kappa}_{\widehat{P}_n,r}$, and rearranging equation 23 bounds the opposite difference. Taking their maximum proves equation 24. The same event $\mathcal{E}$ covers all stationary joint policies, including data-dependent choices of κ, so the inequalities are simultaneous. □

D.7 CHANGING REPRESENTATIONS AND CUMULATIVE WIDTH

Lemma 24 (Transfer from learned to true occupancy)**.** *For any stationary joint policy κ,*

$$\mathbb{E}_{d^{\kappa}_{\widehat{P}_n,\gamma}}[\widehat{b}_n] \le \mathbb{E}_{d^{\kappa}_{P^\star,\gamma}}[\widehat{b}_n] + \frac{\gamma B_n}{1-\gamma}\mathbb{E}_{d^{\kappa}_{P^\star,\gamma}}[f_n], \tag{59}$$

where $B_n = \alpha_n/\sqrt{\lambda_n}$.

Proof. Apply equation 35 to the two models with payoff $\widehat{b}_n$:

$$v^{\kappa}_{\widehat{P}_n,\widehat{b}_n} - v^{\kappa}_{P^\star,\widehat{b}_n} = \frac{\gamma}{1-\gamma}\mathbb{E}_{d^{\kappa}_{P^\star,\gamma}}\left[(\widehat{P}_n - P^\star)V^{\kappa}_{\widehat{P}_n,\widehat{b}_n}\right].$$

Since $0 \le \widehat{b}_n \le B_n$, $\|V^{\kappa}_{\widehat{P}_n,\widehat{b}_n}\|_\infty \le B_n/(1-\gamma)$. Multiplying the preceding identity by $1-\gamma$ and using the definition of f_n proves the lemma. □

Lemma 25 (Training moments)**.** *On $\mathcal{E}$, for every $n \le N$,*

$$n\mathbb{E}_{\rho_n}[\widehat{b}_n^2] \le c_{\mathrm{cov}}\alpha_n^2 d, \tag{60}$$

$$n\mathbb{E}_{\rho_n}[f_n^2] \le c_{\mathrm{mle}}\overline{L}_n. \tag{61}$$

Proof. The second inequality is the MLE event. For the first, the covariance comparison gives

$$\widehat{\Sigma}_n^{-1} \preceq c_{\mathrm{cov}}(\Sigma^{\mathrm{tr}}_{n,\widehat{\phi}_n})^{-1}.$$

Therefore,

$$\begin{aligned} n\mathbb{E}_{\rho_n}[\widehat{b}_n^2] &= n\alpha_n^2\mathbb{E}_{\rho_n}[\widehat{\phi}_n^\top\widehat{\Sigma}_n^{-1}\widehat{\phi}_n] \\ &\le c_{\mathrm{cov}}\alpha_n^2\,\mathrm{tr}\Big(n\mathbb{E}_{\rho_n}[\widehat{\phi}_n\widehat{\phi}_n^\top](\Sigma^{\mathrm{tr}}_{n,\widehat{\phi}_n})^{-1}\Big) \le c_{\mathrm{cov}}\alpha_n^2 d. \end{aligned}$$

□

Lemma 26 (Elliptical-potential inequality)**.** *Let $A_0 \succ 0$, let $X_1, \dots, X_T \succeq 0$, and define $A_t = A_{t-1} + X_t$. If $X_t \preceq A_{t-1}$ for every t, then*

$$\sum_{t=1}^{T}\mathrm{tr}(X_tA_{t-1}^{-1}) \le 2\log\frac{\det(A_T)}{\det(A_0)}.$$

Proof. Let $Y_t = A_{t-1}^{-1/2}X_tA_{t-1}^{-1/2}$. Its eigenvalues lie in $[0,1]$, and $x \le 2\log(1+x)$ on this interval. Hence

$$\begin{aligned} \mathrm{tr}(X_tA_{t-1}^{-1}) &= \mathrm{tr}(Y_t) \\ &\le 2\log\det(I_d + Y_t) = 2\log\frac{\det(A_t)}{\det(A_{t-1})}. \end{aligned}$$

Summing over t telescopes. □

Lemma 27 (Fixed-feature potential)**.** *Let*

$$q_n := \mathbb{E}_{d_n}[\|\phi^\star(s,z)\|_{\Sigma_{n,\star}^{-1}}].$$

Then

$$\sum_{n=1}^{N} q_n \le \sqrt{2dN\ell_N}. \tag{62}$$

Proof. Let $G_n = \mathbb{E}_{d_n}[\phi^\star \phi^{\star\top}]$. Jensen and Cauchy–Schwarz give

$$\sum_{n=1}^{N} q_n \le \sqrt{N \sum_{n=1}^{N} \operatorname{tr}(G_n \Sigma_{n,\star}^{-1})}.$$

Because $\lambda_n \ge \lambda_1$,

$$\Sigma_{n,\star} \succeq A_{n-1} := \lambda_1 I_d + \sum_{i=0}^{n-1} G_i.$$

Thus $\operatorname{tr}(G_n \Sigma_{n,\star}^{-1}) \le \operatorname{tr}(G_n A_{n-1}^{-1})$. Moreover, $G_n \preceq I_d \preceq A_{n-1}$ because $\|\phi^\star\|_2 \le 1$ and $\lambda_1 \ge 1$. Apply Lemma 26 with initial matrix $A_0 = \lambda_1 I_d + G_0$. Since $\det(\lambda_1 I_d + G_0) \ge \det(\lambda_1 I_d)$, the resulting bound can be relaxed to

$$\sum_{n=1}^{N} \operatorname{tr}(G_n \Sigma_{n,\star}^{-1}) \le 2 \log \frac{\det(\lambda_1 I_d + \sum_{i=0}^{N} G_i)}{\det(\lambda_1 I_d)}.$$

Since $\operatorname{tr}(G_i) \le 1$, the arithmetic–geometric mean inequality upper-bounds the log determinant by $d \log(1 + (N+1)/(d\lambda_1)) \le d\ell_N$. Combining the displays proves the result. □

Cumulative-certificate part of Theorem 9. Work on $\mathcal{E}$. Lemma 7 and Lemma 24, with $\kappa = \kappa_n$, give

$$\overline{v}_n - \underline{v}_n \le \frac{2}{1-\gamma} \mathbb{E}_{d_n}[\widehat{b}_n] + \frac{2\gamma B_n}{(1-\gamma)^2} \mathbb{E}_{d_n}[f_n] + 2\eta_n. \tag{63}$$

Apply Lemma 22 first with $g = \widehat{b}_n$, $G = B_n$, and then with $g = f_n$, $G = 2$. Using Lemma 25 yields

$$\mathbb{E}_{d_n}[\widehat{b}_n] \le q_n \sqrt{cJ\alpha_n^2 d + \lambda_n d B_n^2} + \alpha_n \sqrt{\frac{c(1-\gamma)Jd}{n}}, \tag{64}$$

$$\mathbb{E}_{d_n}[f_n] \le q_n \sqrt{cJ\overline{L}_n + 4\lambda_n d} + \sqrt{\frac{c(1-\gamma)J\overline{L}_n}{n}}. \tag{65}$$

The parameter choices imply, uniformly over $n \le N$,

$$\alpha_n \le c\sqrt{(J+d^2)\overline{L}_N}, \qquad \lambda_n \le cd\overline{L}_N,$$

$$B_n \le c\sqrt{\frac{J+d^2}{d}}, \qquad \sum_{n=1}^{N} n^{-1/2} \le 2\sqrt{N}.$$

For the first bound below, the q_n-term is at most

$$\sqrt{dN\ell_N}\,\sqrt{cdJ(J+d^2)\overline{L}_N} = cd\sqrt{J(J+d^2)}\sqrt{N\overline{L}_N\ell_N},$$

where $J + 1 \le 2J$ absorbs the $\lambda_n d B_n^2$ contribution. Its $n^{-1/2}$-term is smaller by at least a factor $d^{1/2}$. For the second bound, multiplying the q_n-term by $B_n \le c\sqrt{(J+d^2)/d}$ gives

$$c(J+d^2)\sqrt{N\overline{L}_N\ell_N};$$

the corresponding $n^{-1/2}$-term is no larger. Combining these facts with Lemma 27 gives

$$\sum_{n=1}^{N} \mathbb{E}_{d_n}[\widehat{b}_n] \le cd\sqrt{J(J+d^2)}\sqrt{N\overline{L}_N\ell_N},$$

$$\sum_{n=1}^{N} B_n \mathbb{E}_{d_n}[f_n] \le c(J+d^2)\sqrt{N\overline{L}_N\ell_N}.$$

Summing equation 63 therefore gives

$$\sum_{n=1}^{N} (\overline{v}_n - \underline{v}_n) \le c\left(\frac{d\sqrt{J(J+d^2)}}{1-\gamma} + \frac{\gamma(J+d^2)}{(1-\gamma)^2}\right)\sqrt{N\overline{L}_N\ell_N} + 2\sum_{n=1}^{N}\eta_n.$$

Finally,

$$\sum_{n=1}^{N} \zeta_n \le 2c_\zeta \sqrt{\frac{JN\overline{L}_N}{1-\gamma}}.$$

Adding the $2\eta_n$ already present in each stored certificate, dividing by N, enlarging the universal constant, and using equation 49 prove equation 29. This event has probability at least $1-\delta/2$, which is stronger than the $1-\delta$ probability asserted by the theorem. □

D.8 Main theorem and interaction count

Returned-policy part of Theorem 9. Work on the common event $\mathcal{E}$ in equation 49, on which Theorem 8 and the cumulative bound equation 29 hold. Lemma 5 gives, for every n,

$$\mathrm{Gap}_{P^\star}(\pi_n^-, \nu_n^+) \le \widehat{\Delta}_n.$$

Since $n^\star \in \arg\min_{n \le N} \widehat{\Delta}_n$,

$$\mathrm{Gap}_{P^\star}(\widehat{\pi}, \widehat{\nu}) \le \widehat{\Delta}_{n^\star} \le \frac{1}{N}\sum_{n=1}^{N} \widehat{\Delta}_n.$$

This proves equation 28; combining it with equation 29 gives the claimed numerical bound. □

Proof of Corollary 10. The tolerance condition gives $\overline{L}_N \le (1+c_\tau)L_N$. Using $\mathfrak{K}_{d,J,\gamma} \le \mathfrak{K}_{d,J}/(1-\gamma)^2$, the resulting first term dominates the second up to universal constants for $d, J \ge 1$ and $\gamma \in [0,1)$. The planning contribution is at most $\epsilon/2$ whenever $N^{-1}\sum_{n=1}^{N} \eta_n \le \epsilon/8$. A sufficient implicit condition for the statistical contribution to be at most $\epsilon/2$ is

$$N \ge \frac{C(1+c_\tau)d^2 J(J+d^2)L_N \ell_N}{(1-\gamma)^4 \epsilon^2}.$$

Here we used

$$\mathfrak{K}_{d,J}^2 \le 2d^2J(J+d^2) + 2(J+d^2)^2 \le 6d^2J(J+d^2),$$

where the last inequality follows from $J + d^2 \le 2d^2J$ for $d, J \ge 1$. Standard inversion of the logarithms yields equation 30.

It remains to count primitive environment transitions. Each outer iteration uses two transitions after the roll-in. The number T_n of transitions in the roll-in has, for integer $t \ge 0$, the geometric tail

$$\Pr(T_n \ge t \mid \mathcal{F}_{n,0}) = \gamma^t \le e^{-(1-\gamma)t}.$$

Taking $t_\star = \lceil (1-\gamma)^{-1}\log(2N/\delta) \rceil$ and applying a union bound gives, with probability at least $1-\delta/2$,

$$\sum_{n=1}^{N} T_n \le N t_\star.$$

Adding the $2N$ tuple transitions and substituting equation 30 proves equation 31. The statistical event fails with probability at most $\delta/2$, and the geometric-tail event fails with probability at most $\delta/2$. Their intersection therefore has probability at least $1-\delta$, jointly proving the gap and interaction bounds. □

D.9 Approximate planning

We record why the coefficient of the planning error in equation 29 is four. For the upper game, equation 5 gives

$$\max_{\pi} v^{\pi,\nu_n^+}_{\widehat{P}_n, r+\widehat{b}_n} \le \overline{v}_n + \eta_n, \qquad v^{\pi_n^+,\nu}_{\widehat{P}_n, r+\widehat{b}_n} \ge \overline{v}_n - \eta_n$$

for every ν. The analogous lower-game inequalities hold with $(\pi_n^-, \nu_n^-, \underline{v}_n)$. The upper game's maximizing-deviation bound and the lower game's minimizing-deviation bound yield the $2\eta_n$ term in

equation 10. The upper game's guarantee for π_n^+ and the lower game's guarantee for ν_n^-, evaluated at the crossed exploration pair (π_n^+, ν_n^-), yield

$$\overline{v}_n - \underline{v}_n \le \frac{2}{1-\gamma}\mathbb{E}_{d^{\kappa_n}_{\widehat{P}_n,\gamma}}[\widehat{b}_n] + 2\eta_n.$$

Thus the stored valid certificate already contains $2\eta_n$, and bounding its width contributes a second $2\eta_n$. Averaging gives exactly $4N^{-1}\sum_{n=1}^{N}\eta_n$. No numerical error is hidden in the statistical term.

E Global Bonus Exploration and Certificate Separation

E.1 A complete nominal + bonus-MDP interface

Write $h = 1-\gamma$. We abbreviate the nominal (certainty-equivalent) pair as CE. At round n, fit the same model and bonus as in Algorithm 1. Solve the nominal zero-sum game for $(\pi_n^0, \nu_n^0, v_n^0)$ with the saddle tolerance η_n^0 from Assumption 3. Separately solve the discounted MDP with kernel $\widehat{P}_n$, reward $\widehat{b}_n$, and action space $\mathcal{A}\times\mathcal{B}$. It returns a stationary joint policy κ_n^U and an upper value $\overline{U}_n$ satisfying equation 32. Store the nominal pair and Δ_n^U, and collect the next paired sample under κ_n^U. Initialize with the same uniform pair and return the nominal pair with the smallest stored certificate. No true-model evaluation enters the procedure. The two planning interfaces are one zero-sum game and one joint-action MDP per round, not two calls of interchangeable cost.

Execution and a computable approximate oracle. For a finite discounted MDP there is an optimal deterministic policy. An action $z(s) = (a(s), b(s))$ is implemented by the product of the two Dirac policies, so this comparator needs no correlated execution device. For an approximate value vector w, let

$$(T_b w)(s) = \max_z\{b(s,z) + \gamma\widehat{P}w(s,z)\}, \qquad e = \|T_b w - w\|_\infty,$$

and take a deterministic policy κ greedy with respect to w. Contraction gives $\|V_b^* - w\|_\infty \le e/h$. Because $T_b^\kappa w = T_b w$, the same bound holds for $\|V_b^\kappa - w\|_\infty$. Therefore

$$U \le \overline{U} := d_0^\top w + e/h \le v^\kappa_{\widehat{P},b} + 2e/h.$$

An upper allowance for numerical residual error is included in e in our implementation. Taking $e \le h\varepsilon^U/2$ realizes the oracle. A near-optimal policy's evaluated return alone is a lower bound on U; it cannot be substituted for $\overline{U}$ in a certificate.

Proof of Theorem 11. First fix a round on the strategic-confidence event. Linearity in the reward and the nominal saddle inequalities imply

$$\max_\pi v^{\pi,\nu_n^0}_{P^\star,r} \le \max_\pi\{v^{\pi,\nu_n^0}_{\widehat{P}_n,r} + v^{\pi,\nu_n^0}_{\widehat{P}_n,\widehat{b}_n}\} + \zeta_n \le v_n^0 + \eta_n^0 + U_n + \zeta_n,$$
$$\min_\nu v^{\pi_n^0,\nu}_{P^\star,r} \ge \min_\nu\{v^{\pi_n^0,\nu}_{\widehat{P}_n,r} - v^{\pi_n^0,\nu}_{\widehat{P}_n,\widehat{b}_n}\} - \zeta_n \ge v_n^0 - \eta_n^0 - U_n - \zeta_n.$$

Subtracting and using $U_n \le \overline{U}_n$ proves validity.

For learning, all distributions must now refer to the comparator's own history. Set $d_0^{\exp}$ to the true occupancy of the initial uniform pair and, for $n \ge 1$, set $d_n^U = d^{\kappa_n^U}_{P^\star,\gamma}$. Define ρ_n^U as the average state marginal of $d_0^{\exp}, d_1^U, \ldots, d_{n-1}^U$ with uniform joint actions, and $\rho_n^{U\prime}$ by one true transition and a second uniform action. These are the predictable contexts of the two stored tuples under the paired protocol. The MLE and covariance proofs in Appendix D apply to any such adapted query sequence. Reapplying them establishes an event $\mathcal{E}^U$ with the same probability bound, on which the confidence envelope, both training moments, and selected-feature covariance comparison hold for this history. This step does not presume that the reverse-crossing data and bonus-MDP data agree.

With $B_n = \alpha_n/\sqrt{\lambda_n}$, the oracle and Lemma 24 give

$$\Delta_n^U \le \frac{2}{h}\mathbb{E}_{d_n^U}[\widehat{b}_n] + \frac{2\gamma B_n}{h^2}\mathbb{E}_{d_n^U}[f_n] + 2\zeta_n + 2(\varepsilon_n^U + \eta_n^0). \tag{66}$$

Define

$$\Sigma^U_{n,\star} = \lambda_n I + \mathbb{E}_{d_0^{\exp}}[\phi^\star \phi^{\star\top}] + \sum_{i=1}^{n-1} \mathbb{E}_{d_i^U}[\phi^\star \phi^{\star\top}].$$

The one-step-back inequalities and training moments, followed by the fixed-feature potential lemma, give exactly

$$\sum_{n=1}^{N} \mathbb{E}_{d_n^U}[\widehat{b}_n] \le Cd\sqrt{J(J+d^2)}\sqrt{N\overline{L}_N \ell_N},$$

$$\sum_{n=1}^{N} B_n \mathbb{E}_{d_n^U}[f_n] \le C(J+d^2)\sqrt{N\overline{L}_N \ell_N}.$$

Indeed these are the two sums proved after equation 64–equation 65; their proof uses the exploration sequence only through its predictable moments and $\Sigma^U_{n,\star}$. Summing equation 66, using $\sum_n \zeta_n \le 2c_\zeta\sqrt{JN\overline{L}_N/h}$, and dividing by N proves equation 33. Selecting the minimum certificate bounds the returned gap by this average. Under the stated likelihood and planning tolerances, the statistical term is at most $\epsilon/2$ at the same order of N as Corollary 10. The geometric roll-in count is independent of the choice of policies and contributes the same extra h^{-1} factor, with the same failure-budget allocation. □

What is and is not compared. At an exact fixed snapshot, Proposition 6 also follows from $v_n^+ - v_n^- \le 2v^{\pi_n^+,\nu_n^-}_{\widehat{P}_n,b_n} \le 2U_n$. The intermediate CE envelope certificate is stronger than the global bonus bound because it evaluates only unilateral deviations against CE. The crossed pair further minimizes that fixed-envelope certificate over both candidate policies. For approximate confidence-game plans with tolerance η_n, the computed crossed certificate is at most the exact minimum plus $4\eta_n$. For an approximate nominal pair the second inequality in equation 14 acquires $2\eta_n^0$. Numerical interval endpoints require their respective error allowances; raw floating-point values need not preserve the exact ordering. These comparisons concern raw certificates. Capping by the known gap bound of one preserves validity and weak ordering, but can erase strict differences. They do not order true gaps or likelihood-set certificates. Evaluating U_n on a saved reverse-crossing trajectory tests the fixed-snapshot comparison only, not the interaction guarantee for the different data sequence in Theorem 11.

E.2 Strict separation within a fixed envelope

For the example in Section 3, take a true one-state self-loop with $r = (1-\gamma)R$, $b = (1-\gamma)B$, and $\zeta = 0$, where

$$R = \begin{pmatrix} 0 & t \\ t & 2t \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 0 \\ 0 & \beta \end{pmatrix}, \qquad 0 < t < \beta, \quad 2t \le 1.$$

The true and learned kernels coincide, so every nonnegative b with $\zeta = 0$ is a valid envelope. Scalar factors $\phi = \mu = 1$ verify rank-one normalization. The nominal game R has the unique saddle (e_2, e_1) and value t. The upper game $R + B$ has the same saddle and value t. In the lower game $R - B$, both players put probability t/β on action 2; its value is t^2/β. Thus the crossing uses row mixture $(1 - t/\beta, t/\beta)$ against column e_1. Its exact certificate and true gap are both $t - t^2/\beta$. For nominal CE, the upper best response is t and the lower best response is $2t - \beta$, so its envelope certificate is $\beta - t$, despite true gap zero. The bonus MDP chooses $(2, 2)$ forever and has value β. Consequently

$$C_{\text{cross}} = t - t^2/\beta < C_{\text{CE}} = \beta - t < 2U = 2\beta, \qquad \frac{C_{\text{CE}}}{C_{\text{cross}}} = \frac{\beta}{t}.$$

For example $t = 0.01, \beta = 0.4$ gives 0.00975, 0.39, and 0.80. The separation is in certificate tightness, not policy quality. A uniform positive bonus floor x adds $2x$ to all three certificates; an additional residual ζ adds 2ζ to each.

E.3 A COVARIANCE-BONUS REALIZATION AND ITS SCOPE

The zero entries in the preceding example need not be attainable by a finite-sample covariance bonus. The following construction connects the separation to the bonus actually used by the learner. It is an existence result, not a claim about typical learning histories.

Represent a one-state self-loop with two coordinates: set $\phi(s,z) = e_1$ for $z \neq (2,2)$ and $\phi(s,2,2) = e_2$, and $\mu(s) = (1,1)^\top$. Then $P(s \mid s,z) = 1$, $\|\phi\|_2 = 1$, and $\|\sum_s \mu(s)g(s)\|_2 \leq \sqrt{2}$ for $|g| \leq 1$. Use the singleton factor class, exact MLE, and the sufficient constants in equation 67 with $d = 2, J = 4$. This is a valid, nonminimal low-rank factorization, even though its kernel is known. With k of the first n first-tuple uniform actions equal to $(2,2)$, the empirical covariance and normalized bonus matrix are

$$\widehat{\Sigma}_n = \operatorname{diag}(\lambda_n + n - k, \lambda_n + k), \qquad \frac{\widehat{b}_n}{h} = x_n \mathbf{1}\mathbf{1}^\top + \beta_n e_2 e_2^\top,$$

where

$$x_n = \frac{\alpha_n}{h\sqrt{\lambda_n + n - k}}, \qquad \beta_n = \frac{\alpha_n}{h\sqrt{\lambda_n + k}} - x_n.$$

For $k < n/2$, choose the fixed reward hR with $0 < t < \min\{\beta_n, 1/2\}$. On that history, the three certificates are

$$t - t^2/\beta_n + 2x_n + 2\zeta_n, \qquad \beta_n - t + 2x_n + 2\zeta_n, \qquad 2\beta_n + 2x_n + 2\zeta_n,$$

with strict inequalities. Thus strict separation is compatible with the learner's actual covariance and residual formulas, including proof constants.

For an arbitrarily large ratio, take histories with $k = 0$. Their probability is $(3/4)^n > 0$, because each first probe uses an independent uniform joint action. Here $x_n \to 0$, $\zeta_n \to 0$, and $\beta_n \to \beta_\infty = 64\sqrt{8}/(8h) > 0$. First choose a sufficiently small fixed $t > 0$, then a sufficiently large finite n. The ratio of CE to crossed certificates approaches β_∞/t and hence has no universal finite bound. The reward is fixed before sampling. All envelopes are valid in this example because the fitted kernel is exactly true. However, these histories become rare and the representation is redundant; the construction proves neither a typical ratio nor an advantage in information complexity. Section H.4 instead measures certificate differences on all saved checkpoints of the stored nontrivial finite-model learning runs.

F EXPLICIT CONSTANTS AND ALTERNATIVE CERTIFICATES

F.1 ONE EXECUTABLE CHOICE OF SUFFICIENT CONSTANTS

The finite-class confidence result admits the following conservative, unoptimized choice:

$$c_0 = 64, \quad c_{\mathrm{mle}} = 128, \quad c_\lambda = 32, \quad c_\alpha = 64, \quad c_\zeta = 12. \tag{67}$$

Here c_{mle} is an analysis constant; the other four specify the algorithm through equation 20–equation 21. We verify every sufficient condition used by the confidence proof.

Write $A_n = 32|\mathcal{M}|(n+1)^2/\delta$, so $\log A_n \leq 2L_n$. In Lemma 19, use the empirical bound $\sum_i H_i(\widehat{P}_n) \leq 2\log A_n + \tau_n/2$. Lemma 18 with $B = 2$, and failure allocation $\delta/(32|\mathcal{M}|(n+1)^2)$, bounds the predictable sum by twice the empirical sum plus $4\log A_n$, uniformly over the finite model class. Consequently,

$$n\big(\mathbb{E}_{\rho_n}[f_n^2] + \mathbb{E}_{\rho'_n}[f_n^2]\big) \leq 64\log A_n + 8\tau_n \leq 128\overline{L}_n.$$

The time and model union bounds have total failure probability less than $\delta/4$, as in the original proof.

For the covariance comparison, its deterministic error satisfies

$$\begin{aligned} a_n &= 4\log\frac{8|\Phi|(n+1)^2[1+32(n+1)]^d}{\delta} + 1 \\ &\leq 4(d+2)L_n + 1 \leq 13dL_n. \end{aligned}$$

Thus $\lambda_n = 32d\overline{L}_n \geq 2a_n$, yielding $c_{\mathrm{cov}} = 2$. Finally,

$$64^2 = 4096 \geq 20 \cdot 128 + 8 \cdot 32 = 2816, \qquad 12^2 = 144 \geq 128,$$

which verifies the bonus and residual conditions in the confidence proof. The remaining arguments use the resulting statistical event and do not require a larger algorithmic constant.

These constants establish executability, rather than tight finite-budget certification. With exact MLE, $J = 4$, $|\mathcal{M}| = 32$, $\gamma = 0.9$, $\delta = 0.05$, and $n = 10{,}000$, they give $2\zeta_n \simeq 6.7594$. Lemma 7 implies $\widehat{\Delta}_n \geq 2\zeta_n$, whereas normalized Nash gaps are at most one. Clipping an upper certificate to one preserves validity but does not improve on that trivial bound in this regime. The calibrated experiments use a different parameter profile and measure its empirical behavior.

F.2 An anytime finite-model likelihood certificate

Let $\mathcal{L}_n(P)$ denote the sum of exact conditional log likelihoods of the observed training tuples, in chronological order. An impossible observed successor has log likelihood $-\infty$. Define

$$\mathcal{M}_n^{\mathrm{LR}} := \left\{ P \in \mathcal{M} : \max_{Q \in \mathcal{M}} \mathcal{L}_n(Q) - \mathcal{L}_n(P) \leq \log \frac{|\mathcal{M}|}{\delta_{\mathrm{set}}} \right\}. \tag{68}$$

With probability at least $1 - \delta_{\mathrm{set}}$, the true kernel belongs to this set simultaneously at every iteration. To see this, for fixed Q, the conditional likelihood ratio

$$R_t(Q) = \prod_{j=1}^{t} \frac{Q(Y_j \mid X_j)}{P^\star(Y_j \mid X_j)}$$

is a nonnegative supermartingale under $P^\star$: each likelihood-ratio increment has conditional expectation at most one, including when the next context depends on the preceding successor. Ratios are defined on $P^\star$-almost-sure observations, whose denominators are positive. Ville's inequality and a union bound over $Q \in \mathcal{M}$ give $\sup_t R_t(Q) \leq |\mathcal{M}|/\delta_{\mathrm{set}}$ for all Q, on one event of the stated probability. This implies equation 68 contains $P^\star$. No independence between the two tuples is required.

On this same event, every stationary pair, including a data-selected pair, has the alternative certificate

$$\mathrm{Gap}_{P^\star}(\pi, \nu) \leq C_n^{\mathrm{set}}(\pi, \nu) := \max_{P \in \mathcal{M}_n^{\mathrm{LR}}} \mathrm{Gap}_P(\pi, \nu). \tag{69}$$

This remains valid at a data-dependent stopping time. It requires two best-response MDP computations for each retained model and candidate; it does not require optimizing a robust equilibrium over the set. It is an alternative finite-class diagnostic, not a claim that Proposition 6 minimizes every certificate.

F.3 Oracle budgets and certificate selection

Table 4: **Planning and certification interfaces at one fitted model.** $K_n = |\mathcal{M}_n^{\mathrm{LR}}|$. A ZS call solves a full known-model discounted zero-sum game; a BR call solves a discounted MDP with one opponent fixed; a joint-MDP call maximizes bonus over both actions. The table excludes shared estimation, feature/covariance construction and environment interaction. It counts calls, not equal-cost operations.

Candidate and certificate	ZS	BR	Joint MDP	Models used
Crossing and native width	2	0	0	$\widehat{P}_n$
CE and global bonus	1	0	1	$\widehat{P}_n$
Given pair and envelope	0	2	0	$\widehat{P}_n$
CE and envelope	1	2	0	$\widehat{P}_n$
Given pair and set certificate	0	$2K_n$	0	$\mathcal{M}_n^{\mathrm{LR}}$
CE and set certificate	1	$2K_n$	0	$\widehat{P}_n, \mathcal{M}_n^{\mathrm{LR}}$
Both outputs and selection	3	$2K_n$	0	Both interfaces

The first two rows also supply exploration under their respective cumulative guarantees. CE-envelope and CE-set rows account for output only; an exploration policy is additional. In particular, CE-envelope certification combined with bonus-MDP exploration uses one ZS, two BR, and one joint-MDP call. Its smaller output bound remains controlled by Theorem 11 on that exploration history.

The last row augments the original two-call exploration algorithm with a nominal plan and its set certificate. No additional environment samples are required, but the added planning and certification work is real. Our finite-class implementation also updates likelihoods and feature covariances by enumeration; the logarithmic statistical dependence on $|\mathcal{M}|$ does not describe this computational cost. In particular, the set certificate optimizes best responses to a supplied pair, not a robust equilibrium over the entire set.

Corollary 28 (Output selection on a common confidence event)**.** *Fix positive $\delta_{\mathrm{env}}, \delta_{\mathrm{set}}$ with $\delta_{\mathrm{env}} + \delta_{\mathrm{set}} \le \delta$. Run Algorithm 1 under Theorem 9 with failure budget δ_{env}, retaining its original exploration rule and parameter schedule. In round n, also construct any stationary pair σ_n^{CE} from the observed data (for example, an approximate nominal equilibrium). Associate $\widehat{\Delta}_n$ with σ_n^{cand} and $C_n^{\mathrm{set}}(\sigma_n^{\mathrm{CE}})$ with σ_n^{CE}, where the set uses δ_{set}. Let $\widehat{\sigma}_N^{\mathrm{sel}}$ be a candidate attaining*

$$B_N := \min_{n \le N} \{\widehat{\Delta}_n,\ C_n^{\mathrm{set}}(\sigma_n^{\mathrm{CE}})\}. \tag{70}$$

With probability at least $1 - \delta$, simultaneously for the considered rounds,

$$\mathrm{Gap}_{P^\star}(\widehat{\sigma}_N^{\mathrm{sel}}) \le B_N \le \min_{n \le N} \widehat{\Delta}_n. \tag{71}$$

Consequently the returned-gap and PAC upper bounds of Theorem 9 and Corollary 10 apply with δ_{env} in their logarithmic terms. Choosing $\delta_{\mathrm{env}} = \delta_{\mathrm{set}} = \delta/2$ preserves their stated order.

Proof. Intersect the envelope event at level $1 - \delta_{\mathrm{env}}$ with the anytime truth-containment event at level $1 - \delta_{\mathrm{set}}$. Their intersection has probability at least $1 - \delta$. On it, every stored crossing is bounded by its native certificate and every stored CE pair is bounded by its set certificate. Selecting among these inequalities proves equation 71; no union bound over candidates or rounds is needed because both events already cover their data-dependent choices. The unchanged exploration process has the original cumulative-width bound with the revised confidence budget, proving the final assertion. □

Numerical upper bounds on C_n^{set} can replace its exact value in this argument. Each bound entering the minimum must be valid: a calibrated envelope is not made valid by taking its minimum with a valid set certificate. Likewise, selecting CE as the next exploration policy would require a new cumulative-width analysis. The empirical exploration comparison below uses a common valid set certificate and does not claim to run the proof-parameter hybrid in this corollary.

F.4 Numerical errors and common-data comparisons

For a fixed opponent and payoff, let T_{BR} be its discounted best-response Bellman operator. An approximate vector w with residual $e = \|T_{\mathrm{BR}} w - w\|_\infty$ satisfies $\|w - V_{\mathrm{BR}}\|_\infty \le e/(1-\gamma)$. If the upper and lower response estimates have errors bounded by $e_+/(1-\gamma)$ and $e_-/(1-\gamma)$, the implemented envelope certificate adds $(e_+ + e_-)/(1-\gamma)$ to their difference plus $2\zeta_n$. The finite-model certificate applies this correction before maximizing over models. The implementation also adds a small floating-point allowance and rejects a solver that does not attain its requested accuracy.

Common-data comparisons hold the data, learned kernel, and parameter profile fixed, and evaluate both the conservative crossing and the certainty-equivalent equilibrium. Recomputing the proof constants at a checkpoint tests the confidence interface on that data; it is not a new trajectory following the proof-profile exploration rule. Only the proof-profile envelope and the likelihood-set certificate carry the stated confidence guarantees; calibrated-envelope coverage is empirical. For the two valid certificate families, a joint $1 - \delta$ guarantee can allocate $\delta/2$ to each. Two separate $1 - \delta$ guarantees do not imply that their intersection has the same confidence level.

G Proofs for Oracle-Conditional Stationary CCE

G.1 Formal setting and oracle boundary

Consider an m-player discounted general-sum Markov game with finite joint-action space $\mathcal{Z} = \prod_{i=1}^{m} \mathcal{A}_i$, $J = |\mathcal{Z}|$, shared transition kernel $P^\star$, and known rewards $r_i : \mathcal{S} \times \mathcal{Z} \to [0, 1-\gamma]$. The kernel satisfies Assumption 2, with $\mathcal{A} \times \mathcal{B}$ replaced by $\mathcal{Z}$. A stationary correlation device is a policy $\sigma : \mathcal{S} \to \Delta(\mathcal{Z})$. Writing σ_{-i} for its marginal over the other players, define a coarse deviation by

$$(\pi_i \diamond \sigma_{-i})(z \mid s) := \pi_i(a_i \mid s)\sigma_{-i}(z_{-i} \mid s) \tag{72}$$

and the initial-distribution stationary CCE gap by

$$\mathrm{Gap}_{\mathrm{CCE,P}}(\sigma) := \max\left\{0, \max_{i\in[m]} \left[\sup_{\pi_i} v_{P,r_i}^{\pi_i \diamond \sigma_{-i}} - v_{P,r_i}^{\sigma}\right]\right\}. \tag{73}$$

The deviator does not condition on the current correlation signal, as required by the coarse-correlated information structure. Because σ_{-i} is stationary Markov, the supremum is attained by a deterministic stationary best response in the resulting discounted MDP.

Assumption 29 (Stationary-CCE computational access)**.** *The learner has the reset, conditional-independence, known-d_0, and known-γ access of Assumption 3, uses exact MLE ($\tau_n = 0$), and can execute stationary correlation devices using shared randomness. For every $P \in \mathcal{M}$, reward vector, and $\eta \geq 0$, $\mathrm{CCEOracle}(P, \{r_i\}, \eta)$ returns a stationary σ with $\mathrm{Gap}_{\mathrm{CCE,P}}(\sigma) \leq \eta$. Exact policy evaluation and stationary best responses are available for a known model.*

Finite discounted games admit stationary Nash equilibria, and hence stationary CCEs (Fink, 1964); the assumption concerns their computation. Since approximate stationary Markov CCE is PPAD-hard in general (Daskalakis et al., 2023), the theorem exposes stationary-CCE computation through an explicit oracle.

G.2 DRep-CCE and the strategic bridge

Let $q = m + 1$. At the end of iteration n, maintain the candidate and its m optimistic unilateral deviations,

$$\kappa_{n,0} := \sigma_n, \qquad \kappa_{n,i} := \pi_{i,n}^{+} \diamond \sigma_{n,-i}, \qquad i \in [m]. \tag{74}$$

At iteration $n+1$, every $\kappa_{n,k}$ independently generates a geometric roll-in and the two uniform-action tuples used in equation 17. The pooled $2qn$ tuples are fitted by exact MLE, as required by Assumption 29. Define

$$\widehat{\Sigma}_n^{\mathrm{cce}} := \sum_{t=1}^{n}\sum_{k=0}^{m} \widehat{\phi}_n(s_{t,k}, z_{t,k})\widehat{\phi}_n(s_{t,k}, z_{t,k})^\top + \lambda_n I_d, \tag{75}$$

$$\widehat{b}_n(s,z) := \alpha_n \|\widehat{\phi}_n(s,z)\|_{(\widehat{\Sigma}_n^{\mathrm{cce}})^{-1}}. \tag{76}$$

The oracle returns σ_n. For every player, compute

$$\pi_{i,n}^{+} \in \arg\max_{\pi_i} v_{\widehat{P}_n, r_i + \widehat{b}_n}^{\pi_i \diamond \sigma_{n,-i}}, \tag{77}$$

$$U_{i,n} := v_{\widehat{P}_n, r_i + \widehat{b}_n}^{\pi_{i,n}^{+} \diamond \sigma_{n,-i}}, \qquad L_{i,n} := v_{\widehat{P}_n, r_i - \widehat{b}_n}^{\sigma_n}. \tag{78}$$

Store

$$\widehat{\Delta}_n^{\mathrm{cce}} := \max\{0, \max_i (U_{i,n} - L_{i,n}) + 2\zeta_n\}, \tag{79}$$

$$\widehat{\sigma} := \sigma_{n^\star}, \qquad n^\star \in \arg\min_{n \leq N} \widehat{\Delta}_n^{\mathrm{cce}}. \tag{80}$$

Lemma 30 (Stationary-CCE certificate)**.** *On the event of Lemma 34,*

$$\mathrm{Gap}_{\mathrm{CCE,P}^\star}(\sigma_n) \leq \widehat{\Delta}_n^{\mathrm{cce}} \qquad \textit{for every } n. \tag{81}$$

Lemma 31 (Oracle-residual width decomposition)**.** *For every n,*

$$\max_{i\in[m]} (U_{i,n} - L_{i,n}) \leq \eta_n + \frac{1}{1-\gamma} \sum_{k=0}^{m} \mathbb{E}_{d_{\widehat{P}_n,\gamma}^{\kappa_{n,k}}}[\widehat{b}_n]. \tag{82}$$

The first lemma is the general-sum analogue of Lemma 5; the second shows why the algorithm samples the candidate and every unilateral optimistic deviation.

Algorithm 2 DREP-CCE with a stationary-CCE oracle

Require: $\mathcal{M}, N, \{\eta_n, \lambda_n, \alpha_n, \zeta_n\}_{n=1}^N$
Initialize $\kappa_{0,0}, \dots, \kappa_{0,m}$ uniformly
for $n = 1, \dots, N$ **do**
  for $k = 0, \dots, m$ **do**
    Roll in under $\kappa_{n-1,k}$ and collect two uniform-action transition tuples
  end for
  Fit $\widehat{P}_n$ by exact MLE and form $\widehat{\Sigma}_n^{\mathrm{cce}}, \widehat{b}_n$
  $\sigma_n \leftarrow \mathrm{CCEOracle}(\widehat{P}_n, \{r_i\}_{i=1}^m, \eta_n)$
  for $i = 1, \dots, m$ **do**
    Compute $\pi_{i,n}^+, U_{i,n}, L_{i,n}$ from equation 77–equation 78
    Set $\kappa_{n,i} = \pi_{i,n}^+ \diamond \sigma_{n,-i}$
  end for
  Set $\kappa_{n,0} = \sigma_n$ and store $\widehat{\Delta}_n^{\mathrm{cce}}$
end for
return the best iterate in equation 80

G.3 PARAMETERS AND TRAINING DISTRIBUTIONS

Let

$$K_n := qn, \qquad L_n^{\mathrm{cce}} := \log\left(\frac{c_0 mq|\mathcal{M}|(n+1)^2}{\delta}\right), \tag{83}$$

$$\lambda_n := c_\lambda \left(q + dL_n^{\mathrm{cce}}\right), \qquad \alpha_n := c_\alpha \sqrt{(J+d^2)L_n^{\mathrm{cce}} + qd}, \tag{84}$$

$$\zeta_n := c_\zeta \sqrt{\frac{JL_n^{\mathrm{cce}}}{qn(1-\gamma)}}. \tag{85}$$

Choose c_0 so that $L_n^{\mathrm{cce}} \geq 1$, then fix sufficiently large universal c_{mle} and c_λ for the statistical event. Take $c_{\mathrm{cov}} = 2$ and choose c_α, c_ζ afterward to satisfy the inequalities in the player-wise confidence proof. The q term in λ_n controls the q-policy batch added in one outer iteration.

For $d_{t,k} := d_{P^\star,\gamma}^{\kappa_{t,k}}$, define

$$\overline{d}_n^{\mathrm{cce}} := \frac{1}{qn} \sum_{t=0}^{n-1} \sum_{k=0}^{m} d_{t,k}, \tag{86}$$

$$\rho_n^{\mathrm{cce}}(s,z) := \overline{d}_n^{\mathrm{cce}}(s) \frac{1}{J}, \tag{87}$$

$$\rho_n^{\mathrm{cce},\prime}(s',z') := \left[\sum_{s,z} \rho_n^{\mathrm{cce}}(s,z) P^\star(s' \mid s,z)\right] \frac{1}{J}. \tag{88}$$

The true covariance used only in the analysis is

$$\Sigma_{n,\star}^{\mathrm{cce}} := K_n \mathbb{E}_{\overline{d}_n^{\mathrm{cce}}}[\phi^\star \phi^{\star\top}] + \lambda_n I_d. \tag{89}$$

For a candidate feature $\phi \in \Phi$, set

$$\Sigma_{n,\phi}^{\mathrm{cce,tr}} := K_n \mathbb{E}_{\rho_n^{\mathrm{cce}}}[\phi\phi^\top] + \lambda_n I_d, \tag{90}$$

$$\widehat{\Sigma}_{n,\phi}^{\mathrm{cce}} := \sum_{t=1}^{n} \sum_{k=0}^{m} \phi(s_{t,k}, z_{t,k}) \phi(s_{t,k}, z_{t,k})^\top + \lambda_n I_d. \tag{91}$$

G.4 STATISTICAL EVENT AND ONE-STEP-BACK CONSEQUENCES

Let

$$f_n(s,z) := \|\widehat{P}_n(\cdot \mid s,z) - P^\star(\cdot \mid s,z)\|_1, \qquad \xi_n^{\mathrm{cce}} := c_{\mathrm{mle}} \frac{L_n^{\mathrm{cce}}}{K_n}.$$

Lemma 32 (Block statistical event)**.** *Under Assumptions 2 and 29, with probability at least* $1-\delta/2$*, simultaneously for every* $n \le N$*,*

$$\mathbb{E}_{\rho_n^{\mathrm{cce}}}[f_n^2] \le \xi_n^{\mathrm{cce}}, \qquad \mathbb{E}_{\rho_n^{\mathrm{cce},\prime}}[f_n^2] \le \xi_n^{\mathrm{cce}}, \tag{92}$$

$$c_{\mathrm{cov}}^{-1}\Sigma_{n,\phi}^{\mathrm{cce,tr}} \preceq \widehat{\Sigma}_{n,\phi}^{\mathrm{cce}} \preceq c_{\mathrm{cov}}\Sigma_{n,\phi}^{\mathrm{cce,tr}} \qquad \textit{for every } \phi \in \Phi. \tag{93}$$

Proof. Order the $2qn$ context–successor observations chronologically. Every exploration policy used to generate a context is measurable before its roll-in, and every successor is conditionally sampled from $P^\star$. The Hellinger supermartingale in the proof of Lemma 19 therefore applies without change. Unioning over the finite model class and over n, the likelihood comparison for the MLE gives a total predictable squared Hellinger error of order L_n^{cce} over the qn first contexts and the qn second contexts. The two sums are nonnegative, so each is at most the total. The inequality $\|p-q\|_1^2 \le 8h^2(p,q)$ and division by $K_n = qn$ prove equation 92 with failure probability at most $\delta/4$.

Order the $K_n = qn$ first contexts chronologically and condition before each roll-in. For every fixed ϕ, their conditional second moments sum to $K_n\mathbb{E}_{\rho_n^{\mathrm{cce}}}[\phi\phi^\top]$. The scalar supermartingale argument in the proof of Lemma 20 applies to this ordered sequence. For outer iteration n, use a deterministic sphere net with

$$\varepsilon_n^{\mathrm{cce}} = \frac{1}{16(qn+1)}, \qquad |\mathcal{N}_n^{\mathrm{cce}}| \le [1+32(qn+1)]^d.$$

Allocate the scalar tail probabilities over outer iterations, features, and net directions exactly as in that proof. The two unregularized Gram matrices have operator norm at most qn, so the extension error is bounded by a universal constant. The resulting additive error is at most

$$a_n^{\mathrm{cce}} := 4\log\frac{8|\Phi|(n+1)^2[1+32(qn+1)]^d}{\delta} + 1 \le CdL_n^{\mathrm{cce}}.$$

The last inequality uses $qn+1 \le q(n+1)$ and the definition of L_n^{cce}, which includes $\log q$. Choosing c_λ sufficiently large in $\lambda_n = c_\lambda(q + dL_n^{\mathrm{cce}})$ ensures $\lambda_n \ge 2a_n^{\mathrm{cce}}$. Thus equation 93 holds with $c_{\mathrm{cov}} = 2$ and failure probability at most $\delta/4$. Together with the MLE event, this proves the stated block statistical event. □

The next result records the two consequences of the one-step-back arguments that will be used below.

Lemma 33 (Block one-step-back inequalities)**.** *On the event of Lemma 32, for every stationary joint policy* κ *and every* $g : \mathcal{S}\times\mathcal{Z} \to [0,G]$*,*

$$\mathbb{E}_{d^\kappa_{P^\star,\gamma}}[g] \le \sqrt{\gamma}\,\mathbb{E}_{d^\kappa_{P^\star,\gamma}}\left[\|\phi^\star\|_{(\Sigma_{n,\star}^{\mathrm{cce}})^{-1}}\right]\sqrt{K_nJ\mathbb{E}_{\rho_n^{\mathrm{cce}}}[g^2] + \lambda_n dG^2} + \sqrt{(1-\gamma)J\mathbb{E}_{\rho_n^{\mathrm{cce}}}[g^2]}, \tag{94}$$

$$\mathbb{E}_{d^\kappa_{\widehat{P}_n,\gamma}}[g] \le \sqrt{\gamma}\,\mathbb{E}_{d^\kappa_{\widehat{P}_n,\gamma}}\left[\|\widehat{\phi}_n\|_{(\Sigma_{n,\widehat{\phi}_n}^{\mathrm{cce,tr}})^{-1}}\right] \times \sqrt{2K_nJ\mathbb{E}_{\rho_n^{\mathrm{cce},\prime}}[g^2] + 2K_nG^2\xi_n^{\mathrm{cce}} + \lambda_n dG^2} + \sqrt{(1-\gamma)J\mathbb{E}_{\rho_n^{\mathrm{cce}}}[g^2]}. \tag{95}$$

Proof. Replace n, ρ_n, ρ_n', and $\Sigma_{n,\star}$ in Lemmas 22 and 23 by $K_n, \rho_n^{\mathrm{cce}}, \rho_n^{\mathrm{cce},\prime}$, and $\Sigma_{n,\star}^{\mathrm{cce}}$, respectively. The occupancy recursion is unchanged. The only property of the training state distribution used in those proofs is

$$\rho_n^{\mathrm{cce}}(s) = (1-\gamma)d_0(s) + \gamma\mathbb{E}_{(\widetilde{s},\widetilde{z})\sim\overline{d}_n^{\mathrm{cce}}}P^\star(s \mid \widetilde{s},\widetilde{z}),$$

which follows by averaging the discounted occupancy recursion over all (t,k). Uniform joint actions again contribute the importance ratio J. □

G.5 UNIFORM CONFIDENCE AND CCE CERTIFICATION

Lemma 34 (Player-wise two-sided confidence)**.** *On the event of Lemma 32, simultaneously for every $n \le N$, every $i \in [m]$, and every stationary joint policy κ,*

$$v^{\kappa}_{\widehat{P}_n, r_i + \widehat{b}_n} \ge v^{\kappa}_{P^\star, r_i} - \zeta_n, \tag{96}$$

$$v^{\kappa}_{\widehat{P}_n, r_i - \widehat{b}_n} \le v^{\kappa}_{P^\star, r_i} + \zeta_n. \tag{97}$$

Proof. Fix n, i, κ. The simulation identity and $\|V^{\kappa}_{P^\star, r_i}\|_\infty \le 1$ give

$$(1-\gamma)\left(v^{\kappa}_{\widehat{P}_n, r_i + \widehat{b}_n} - v^{\kappa}_{P^\star, r_i}\right) \ge \mathbb{E}_{d^{\kappa}_{\widehat{P}_n, \gamma}}[\widehat{b}_n - f_n].$$

Apply equation 95 with $g = f_n$, $G = 2$. Equations 92 and 93 imply that the feature-norm term is at most

$$c\mathbb{E}_{d^{\kappa}_{\widehat{P}_n, \gamma}}\left[\|\widehat{\phi}_n\|_{(\widehat{\Sigma}^{\mathrm{cce}}_n)^{-1}}\right] \sqrt{JL^{\mathrm{cce}}_n + \lambda_n d}.$$

More explicitly, the same calculation as in the zero-sum confidence proof requires

$$2\big(2K_n J\xi^{\mathrm{cce}}_n + 8K_n\xi^{\mathrm{cce}}_n + 4\lambda_n d\big) \le \alpha_n^2.$$

Since $K_n\xi^{\mathrm{cce}}_n = c_{\mathrm{mle}} L^{\mathrm{cce}}_n$ and $\lambda_n d = c_\lambda(qd + d^2 L^{\mathrm{cce}}_n)$, it suffices to choose $c_\alpha^2 \ge 20c_{\mathrm{mle}} + 8c_\lambda$. This makes the feature-norm contribution no larger than $\mathbb{E}_{d^{\kappa}_{\widehat{P}_n, \gamma}}[\widehat{b}_n]$. The remaining initial-distribution term is

$$\sqrt{(1-\gamma)J\xi^{\mathrm{cce}}_n}.$$

With $c_\zeta^2 \ge c_{\mathrm{mle}}$, dividing this term by $1-\gamma$ gives at most ζ_n, proving equation 96.

For the lower inequality, apply the already proved upper inequality to $\bar{r}_i = (1-\gamma) - r_i$. Since $v^{\kappa}_{P, \bar{r}_i} = 1 - v^{\kappa}_{P, r_i}$ for every P, κ, the result is exactly equation 97. The same statistical event covers every player i and stationary joint policy κ, including data-dependent choices, so no policy-class union bound is needed. □

Proof of Lemma 30. For every player i, uniform upper confidence gives

$$\begin{aligned}\sup_{\pi_i} v^{\pi_i \diamond \sigma_{n,-i}}_{P^\star, r_i} &\le \sup_{\pi_i} v^{\pi_i \diamond \sigma_{n,-i}}_{\widehat{P}_n, r_i + \widehat{b}_n} + \zeta_n \\ &= U_{i,n} + \zeta_n.\end{aligned}$$

Uniform lower confidence gives $v^{\sigma_n}_{P^\star, r_i} \ge L_{i,n} - \zeta_n$. Subtracting, maximizing over i, and taking the positive part proves equation 81. □

Proof of Lemma 31. Fix i and write $\kappa_{n,i} = \pi^+_{i,n} \diamond \sigma_{n,-i}$. Because all terms use the same learned transition model,

$$\begin{aligned}U_{i,n} - L_{i,n} = &\left(v^{\kappa_{n,i}}_{\widehat{P}_n, r_i + \widehat{b}_n} - v^{\kappa_{n,i}}_{\widehat{P}_n, r_i}\right) \\ &+ \left(v^{\kappa_{n,i}}_{\widehat{P}_n, r_i} - v^{\sigma_n}_{\widehat{P}_n, r_i}\right) \\ &+ \left(v^{\sigma_n}_{\widehat{P}_n, r_i} - v^{\sigma_n}_{\widehat{P}_n, r_i - \widehat{b}_n}\right).\end{aligned}$$

The first and third differences equal, respectively,

$$\frac{1}{1-\gamma}\mathbb{E}_{d^{\kappa_{n,i}}_{\widehat{P}_n, \gamma}}[\widehat{b}_n], \qquad \frac{1}{1-\gamma}\mathbb{E}_{d^{\kappa_{n,0}}_{\widehat{P}_n, \gamma}}[\widehat{b}_n].$$

The middle difference is at most $\mathrm{Gap}_{\mathrm{CCE}, \widehat{\mathrm{P}}_n}(\sigma_n) \le \eta_n$ by the oracle guarantee. Maximizing over i and upper-bounding the maximum by the sum of all q nonnegative bonus expectations proves equation 82. □

G.6 CUMULATIVE STATISTICAL BOUND

The following two lemmas adapt the occupancy-transfer and potential arguments to a batch of q exploration policies.

Lemma 35 (Training moments and occupancy transfer)**.** *On the event of Lemma 32, for every n,*

$$K_n \mathbb{E}_{\rho_n^{\mathrm{cce}}}[\widehat{b}_n^2] \le c\alpha_n^2 d, \qquad K_n \mathbb{E}_{\rho_n^{\mathrm{cce}}}[f_n^2] \le cL_n^{\mathrm{cce}}. \tag{98}$$

Moreover, for every stationary joint policy κ,

$$\mathbb{E}_{d^{\kappa}_{\widehat{P}_n,\gamma}}[\widehat{b}_n] \le \mathbb{E}_{d^{\kappa}_{P^\star,\gamma}}[\widehat{b}_n] + \frac{\gamma B_n}{1-\gamma}\mathbb{E}_{d^{\kappa}_{P^\star,\gamma}}[f_n], \qquad B_n := \frac{\alpha_n}{\sqrt{\lambda_n}}. \tag{99}$$

Proof. Equation 98 follows from the covariance and MLE events exactly as in Lemma 25, with n replaced by K_n. Equation 99 is Lemma 24; it holds for arbitrary stationary joint policies and does not use the zero-sum structure. □

For $G_{n,k} := \mathbb{E}_{d_{n,k}}[\phi^\star \phi^{\star\top}]$, define

$$q^\star_{n,k} := \mathbb{E}_{d_{n,k}}\left[\|\phi^\star\|_{(\Sigma^{\mathrm{cce}}_{n,\star})^{-1}}\right]. \tag{100}$$

Let

$$\ell_N^{\mathrm{cce}} := 1 + \log\left(1 + \frac{q(N+1) + d\lambda_{N+1}}{d\lambda_1}\right). \tag{101}$$

Lemma 36 (Batched fixed-feature potential)**.** *For the q-policy exploration family,*

$$\sum_{n=1}^{N}\sum_{k=0}^{m} q^\star_{n,k} \le \sqrt{2qdN\ell_N^{\mathrm{cce}}}. \tag{102}$$

Proof. Let $H_n := \sum_{k=0}^m G_{n,k}$. Jensen and Cauchy–Schwarz give

$$\sum_{n=1}^{N}\sum_{k=0}^{m} q^\star_{n,k} \le \sqrt{qN \sum_{n=1}^{N} \mathrm{tr}\left(H_n(\Sigma^{\mathrm{cce}}_{n,\star})^{-1}\right)}.$$

Each $G_{n,k} \preceq I_d$, hence $H_n \preceq qI_d$. The parameter choice gives $\Sigma^{\mathrm{cce}}_{n,\star} \succeq \lambda_n I_d \succeq qI_d$, so every eigenvalue of $(\Sigma^{\mathrm{cce}}_{n,\star})^{-1/2} H_n (\Sigma^{\mathrm{cce}}_{n,\star})^{-1/2}$ is at most one. Therefore,

$$\mathrm{tr}\left(H_n(\Sigma^{\mathrm{cce}}_{n,\star})^{-1}\right) \le 2\log\frac{\det(\Sigma^{\mathrm{cce}}_{n,\star} + H_n)}{\det(\Sigma^{\mathrm{cce}}_{n,\star})}.$$

Because λ_n is nondecreasing, $\Sigma^{\mathrm{cce}}_{n+1,\star} \succeq \Sigma^{\mathrm{cce}}_{n,\star} + H_n$. The log determinants telescope. Using $\sum_{t=0}^N \mathrm{tr}(H_t) \le q(N+1)$, $\Sigma^{\mathrm{cce}}_{1,\star} \succeq \lambda_1 I_d$, and the arithmetic–geometric mean inequality yields

$$\sum_{n=1}^{N} \mathrm{tr}\left(H_n(\Sigma^{\mathrm{cce}}_{n,\star})^{-1}\right) \le 2d\ell_N^{\mathrm{cce}}.$$

Substitution proves equation 102. □

Define

$$\overline{B}_N := \max_{n\le N} B_n.$$

For a universal constant $C > 0$, let

$$\begin{aligned}\mathfrak{R}_N^{\mathrm{cce}} := \frac{C}{\sqrt{N}}\Bigg[&\frac{\alpha_N d\sqrt{qJ\ell_N^{\mathrm{cce}}}}{1-\gamma} + \frac{\alpha_N\sqrt{q(1-\gamma)Jd}}{1-\gamma} \\ &+ \frac{\gamma\overline{B}_N\sqrt{qd\ell_N^{\mathrm{cce}}(JL_N^{\mathrm{cce}} + \lambda_N d)}}{(1-\gamma)^2} \\ &+ \frac{\gamma\overline{B}_N\sqrt{q(1-\gamma)JL_N^{\mathrm{cce}}}}{(1-\gamma)^2} + \sqrt{\frac{JL_N^{\mathrm{cce}}}{q(1-\gamma)}}\Bigg].\end{aligned} \tag{103}$$

Theorem 37 (Cumulative stationary-CCE certificate)**.** *Under Assumptions 2 and 29, with the parameters in equations 83–85, with probability at least* $1-\delta$*,*

$$\frac{1}{N}\sum_{n=1}^{N}\widehat{\Delta}_n^{\mathrm{cce}} \le \mathfrak{R}_N^{\mathrm{cce}} + \frac{1}{N}\sum_{n=1}^{N}\eta_n. \tag{104}$$

Proof. Work on the event of Lemma 32. Combine Lemma 31 with equation 99 for all q policies:

$$\max_i(U_{i,n}-L_{i,n}) \le \eta_n + \frac{1}{1-\gamma}\sum_{k=0}^{m}\mathbb{E}_{d_{n,k}}[\widehat{b}_n] + \frac{\gamma B_n}{(1-\gamma)^2}\sum_{k=0}^{m}\mathbb{E}_{d_{n,k}}[f_n]. \tag{105}$$

Apply equation 94 to $g=\widehat{b}_n$, $G=B_n$, and use equation 98. For every k,

$$\mathbb{E}_{d_{n,k}}[\widehat{b}_n] \le C\alpha_n\sqrt{Jd}\,q^\star_{n,k} + C\alpha_n\sqrt{\frac{(1-\gamma)Jd}{qn}}. \tag{106}$$

Apply the same inequality to $g=f_n$, $G=2$, to obtain

$$\mathbb{E}_{d_{n,k}}[f_n] \le C\sqrt{JL_n^{\mathrm{cce}}+\lambda_n d}\,q^\star_{n,k} + C\sqrt{\frac{(1-\gamma)JL_n^{\mathrm{cce}}}{qn}}. \tag{107}$$

Sum equations 106 and 107 over n,k. Use Lemma 36, monotonicity of $L_n^{\mathrm{cce}},\lambda_n,\alpha_n$, the definition of $\overline{B}_N$, and $\sum_{n=1}^{N} n^{-1/2} \le 2\sqrt{N}$. Because the right-hand side of equation 105 is nonnegative, taking the positive part in equation 79 preserves the same upper bound after adding $2\zeta_n$. Substitution into equation 105 gives the first four terms in equation 103, plus the average oracle residual. Finally,

$$\frac{2}{N}\sum_{n=1}^{N}\zeta_n \le 4c_\zeta\sqrt{\frac{JL_N^{\mathrm{cce}}}{q(1-\gamma)N}},$$

which is the last term of $\mathfrak{R}_N^{\mathrm{cce}}$ after enlarging C. The common statistical event has probability at least $1-\delta/2$, which is stronger than the theorem's stated probability. □

G.7 Stationary-CCE PAC guarantee

Let

$$D_q := J + d^2 + qd, \qquad \mathfrak{C}_q := d\sqrt{qJD_q} + \sqrt{d}\,D_q. \tag{108}$$

Theorem 38 (Oracle-conditional PAC guarantee for stationary CCE)**.** *Suppose Assumptions 2 and 29 hold. Run Algorithm 2 with equations 83–85. Then, with probability at least* $1-\delta$*,*

$$\mathrm{Gap}_{\mathrm{CCE},\mathrm{P}^\star}(\widehat{\sigma}) \le \mathfrak{R}_N^{\mathrm{cce}} + \frac{1}{N}\sum_{n=1}^{N}\eta_n. \tag{109}$$

Moreover,

$$\mathfrak{R}_N^{\mathrm{cce}} \le \frac{C\mathfrak{C}_q L_N^{\mathrm{cce}}\sqrt{\ell_N^{\mathrm{cce}}}}{(1-\gamma)^2\sqrt{N}}. \tag{110}$$

Consequently, if $N^{-1}\sum_{n=1}^{N}\eta_n \le \epsilon/2$*, then, up to logarithmic factors,*

$$N = \widetilde{O}\left(\frac{d^2qJD_q + dD_q^2}{(1-\gamma)^4\epsilon^2}\right) \tag{111}$$

outer iterations suffice for $\mathrm{Gap}_{\mathrm{CCE},\mathrm{P}^\star}(\widehat{\sigma}) \le \epsilon$*. The algorithm uses* $2qN$ *MLE transition tuples and, with the same probability after adjusting constants, at most*

$$\widetilde{O}\left(\frac{q\left(d^2qJD_q + dD_q^2\right)}{(1-\gamma)^5\epsilon^2}\right) \tag{112}$$

primitive environment transition steps.

Proof. Lemma 30 and the best-iterate rule give

$$\mathrm{Gap}_{\mathrm{CCE},\mathrm{P}^\star}(\widehat{\sigma}) \le \min_{n\le N} \widehat{\Delta}_n^{\mathrm{cce}} \le \frac{1}{N}\sum_{n=1}^{N} \widehat{\Delta}_n^{\mathrm{cce}}.$$

Theorem 37 therefore gives

$$\mathrm{Gap}_{\mathrm{CCE},\mathrm{P}^\star}(\widehat{\sigma}) \le \mathfrak{R}_N^{\mathrm{cce}} + \frac{1}{N}\sum_{n=1}^{N} \eta_n.$$

It remains to simplify the rate. Equations 84 and 108 imply

$$\alpha_N \le C\sqrt{D_q L_N^{\mathrm{cce}}}, \qquad \overline{B}_N \le C\sqrt{\frac{D_q L_N^{\mathrm{cce}}}{q}},$$
$$JL_N^{\mathrm{cce}} + \lambda_N d \le C D_q L_N^{\mathrm{cce}}.$$

Substituting these inequalities into equation 103, using $d, J, q, L_N^{\mathrm{cce}}, \ell_N^{\mathrm{cce}} \ge 1$, and enlarging C proves equation 110. Solving the statistical term in equation 110 for N, and using

$$\mathfrak{C}_q^2 \le 2d^2 qJD_q + 2dD_q^2,$$

gives equation 111.

Each outer iteration performs q independent geometric roll-ins and collects two transition tuples after each roll-in. Every roll-in length has tail $\Pr(T \ge t) \le e^{-(1-\gamma)t}$. Set $t_\star = \lceil (1-\gamma)^{-1}\log(2qN/\delta)\rceil$. A union bound over the qN roll-ins shows that their total length is at most $qNt_\star = \widetilde{O}(qN/(1-\gamma))$, except on an event of probability at most $\delta/2$. Adding $2qN$ tuple transitions and substituting equation 111 proves equation 112. Intersecting with the statistical event of probability at least $1-\delta/2$ gives joint success probability at least $1-\delta$. □

H Extended Experimental Details and Results

Throughout the experimental tables, figures, and archived method names, CE denotes the nominal (certainty-equivalent) candidate or roll-in. The main text uses "nominal" for the same method.

We use controlled finite-model experiments to examine three distinct tasks: finding a candidate, collecting informative data, and certifying a candidate. The matched exploration studies hold output and certification fixed, while the common-data study holds the training trajectory fixed. The baseline ablations examine the full learning procedure, envelope coverage, and observed-state scaling at fixed rank. Additional studies probe estimation and planning tolerances, misspecification, and the stationary-CCE reduction.

H.1 Setup

Games and model classes. Every candidate is a valid low-rank transition model. Except in the explicit misspecification study below, the true model belongs to the finite candidate class. We construct $\mathcal{S} = [d] \times [R]$, where the second coordinate is an observational alias, and use

$$P_{\mu,\phi}(s' \mid s, z) = \sum_{k=1}^{d} \phi_k(s,z)\mu_k(s'). \tag{113}$$

In the random families, action-dependent bounded logits generate $\phi(s,z) \in \Delta([d])$; the random-game families use smoothed emissions $\mu_k \in \Delta(\mathcal{S})$ with finite likelihoods. The baseline deep branch and delayed-conflict family use disjoint emission supports and exact-support likelihoods (Sections H.2 and H.3). Decoys perturb latent assignments, logits, or rare-action branches while sharing the known reward. Exact enumeration implements the MLE. Besides random instances, we use a hidden-branch game in which a profitable unilateral deviation reaches model disagreement that the nominal occupancy rarely visits, and a duality trap with a nearly correct scalar value but potential

unilateral deviations. A two-step shifted-branch family makes all candidate models agree on the first branch transition and disagree only at its successor, isolating the role of the second MLE tuple.

Unless varied, the baseline zero-sum configuration is

$$|\mathcal{S}| = 256, \quad d = 4, \quad A = B = 3, \quad |\mathcal{M}| = 32, \quad \gamma = 0.9, \quad \delta = 0.05.$$

The baseline default comparison uses 30 paired seeds. State, rank, action, discount, and model-class sweeps use 10 seeds per configuration; the planning tolerance experiment uses 20. The controlled general-sum study uses 10 seeds for each of 48 configurations. The additional approximate-MLE study uses 30 paired seeds for each of two families, the exhaustive micro-game audit uses 300 seeds, and the shifted-tuple study uses 30 paired seeds. The long-horizon certificate study uses 20 seeds for each of four configurations; the fixed-precision study uses 10 seeds for each of six configurations. The held-out misspecification confirmation uses 100 new seeds for every micro-game mode–level–condition cell and 30 new paired seeds for every default-scale family–level–method cell, for 2,320 additional tasks.

Table 5: **Settings and outcomes beyond the default comparison.** Each row uses its own configuration; counts are seeds per configuration.

Study	$\lvert\mathcal{S}\rvert$	d	J	$\lvert\mathcal{M}\rvert$	Seeds	Outcome
Envelope micro-game	4	2	4	8	300	299/300 seeds cover; $\gamma = 0.8$
Long-run certificate	64	4	4	32	20	All 80 runs cover at checkpoints
Fixed precision	64	4	4–16	16	10	Target 0.25; no hits at $\gamma = 0.95$
Shifted second tuple	64	4	4	32	30	Removing tuple: envelope 100% to 80%

Methods and evaluation. The baseline ablation suite compares DREP-ZS with a known-representation control, certainty-equivalent MLE planning without bonuses, one-sided optimism, uncrossed output, and uniform rather than targeted exploration. The known-representation control fixes the true $\phi^\star$ for every candidate and learns only the next-state factor μ. Hidden-branch experiments additionally test bonus clipping and removal of the shifted second tuple. These variants can change both output and exploration. In contrast, the matched exploration studies keep CE output and set certification fixed: their upper same-side roll-in is distinct from the one-sided-output ablation.

For each method's output $(\pi_n^{\mathrm{out}}, \nu_n^{\mathrm{out}})$, ground-truth discounted best responses give the true zero-sum gap

$$G_n^{\mathrm{zs}} = \max_{\pi} v_{P^\star,r}^{\pi,\nu_n^{\mathrm{out}}} - \min_{\nu} v_{P^\star,r}^{\pi_n^{\mathrm{out}},\nu}.$$

The full algorithm uses $(\pi_n^{\mathrm{out}}, \nu_n^{\mathrm{out}}) = (\pi_n^-, \nu_n^+)$. We report the final gap and area under the current-gap curve (AUC) against primitive environment transitions. The baseline suite uses ground-truth policy-evaluation residual at most 10^{-10}; the matched certification studies use the corrected best-response bounds specified in Appendix F.4. We also audit the envelope

$$\left| v_{\widehat{P}_n,r}^{\kappa} - v_{P^\star,r}^{\kappa} \right| \le \frac{1}{1-\gamma} \mathbb{E}_{d_{\widehat{P}_n,\gamma}^{\kappa}} [\widehat{b}_n] + \zeta_n$$

on candidate, exploration, true best-response, random product, and random correlated policies. Coverage is counted per seed only when the inequality holds at every checkpoint; hence it matches the simultaneous form of the theorem more closely than pointwise coverage. The exhaustive micro-game audit strengthens this finite test set to all 256 deterministic stationary joint policies, 1,000 random product or correlated policies, and four algorithm-dependent policies at every checkpoint. We additionally optimize the upper and lower signed envelope violations over stationary stochastic policies. A softmax parameterization, analytic discounted-value gradients, 32 random restarts, and at most 100 L-BFGS-B iterations produce two adversarial policies per checkpoint. This yields 1,262 audited policies per checkpoint. The adversarial-search random stream is separated from the learning random stream, so strengthening the audit does not alter the trajectory being evaluated.

Finite-class implementation and cost accounting. We update each candidate's cumulative likelihood from the two new tuples and select the largest score, so exact MLE requires one scan of $|\mathcal{M}|$ stored scores per outer iteration. A Shapley Bellman sweep computes the low-rank continuation as

Table 6: **Fixed-candidate exploration in the deep-branch family.** Each cell gives certifications out of ten at target 0.05 and mean primitive transitions capped at 12,000. All seeds enter the mean. The deep branch admits zero-sample full-class certification by a different nominal candidate; these results are conditional on the current-MLE candidate rule.

Configuration $(\|\mathcal{S}\|, d, J)$	Reverse crossing	CE roll-in	Uniform roll-in
Random $(256, 4, 9)$	10/10; 211	10/10; 201	10/10; 238
Random $(64, 4, 4)$	10/10; 171	10/10; 230	10/10; 189
Hidden $(64, 4, 4)$	10/10; 922	10/10; 904	10/10; 1,322
Deep branch $(64, 8, 4)$	10/10; 2,354	0/10; 12,000	2/10; 10,213

μV followed by $\phi^\top(\mu V)$, which costs $O(|\mathcal{S}|d + |\mathcal{S}|Jd)$ before solving $|\mathcal{S}|$ statewise matrix games. The implementation uses a closed form for 2×2 games, support enumeration through 4×4, and a HiGHS linear-program fallback. Planning targets a Bellman residual at most $(1-\gamma)\eta$. With exact statewise matrix-game solutions, this controls both value error and the saddle inequalities in equation 5; numerical response corrections for the matched certification studies are given in Appendix F.4. The code logs MLE, planning, and total time separately. Low-rank likelihood updates and planning operate on the factors, without a dense $|\mathcal{S}| \times J \times |\mathcal{S}|$ transition tensor.

Calibration and reporting. On an independent validation set comprising random, hidden, and duality families, we tested six pairs of universal multipliers. We first required at least 95% simultaneous coverage for both the Nash certificate and the finite envelope audit, then minimized median certificate AUC. This selects

$$c_\alpha = 0.025, \qquad c_\zeta = 0.0125,$$

which are retained in the calibrated test profiles, including both matched exploration studies. The common-data comparison also evaluates the explicit proof profile separately. Both validation coverages are 100%; the next smaller pair $0.01/0.005$ retains 100% certificate coverage but only 33.3% envelope coverage. Baseline learning curves show medians with 95% bootstrap intervals; Figure 6 shows medians only. Paired comparisons resample seed pairs 10,000 times. Baseline paired hitting-time comparisons retain only pairs finite in both arms; the matched exploration studies in Sections H.2 and H.3 report all-seed censoring and capped costs. The baseline realizable-class campaign contains 2,773 successful task executions. Adding 30 common-data trajectories, 120 deep-branch matched-exploration runs, and 600 delayed-conflict runs gives 3,523 realizable-class tasks, excluding technical pilots. These task executions have no runtime failures. The policy-uniformity audit uses 300 seed settings, including the 100 settings in the baseline diagnostic. The separate held-out misspecification confirmation contains 2,320 successful tasks and no runtime failures. Its levels, constants, horizons, and seeds were frozen before inspection of the confirmation outcomes. The baseline and matched exploration studies use Python 3.11; the common-data candidate comparison uses the documented Python 3.12 environment. The added global-bonus replay uses Python 3.11.15 on one host, retiming every comparator together; these times are not pooled across hosts. It reuses the 170 snapshots and adds no learning trajectories.

H.2 Deep-branch exploration and zero-sample candidate search

Controlled comparison. Three roll-in rules—reverse crossing, certainty-equivalent (CE), and uniform—collect their own trajectories on four configurations, with ten paired seeds each. The first roll-in is uniform for every method. All subsequent differences arise from the roll-in policy and its computational requirements. The two uniform-action training tuples, exact-support MLE, nominal candidate, likelihood-set certificate, confidence level, and numerical tolerances are common. Random and hidden settings match the common-data comparison's game families, with fresh seeds. The deep branch has $d = 8$; comparisons are paired within each family, not across ranks.

The primary target is $C_n^{\mathrm{set}} \le 0.05$ with a budget of 12,000 primitive transitions. We also record first passage below 0.1 on the same trajectory. Certificates are evaluated at $n = 0$ and after every complete tuple pair, so these are actual online stopping times rather than first-logged crossings on a sparse grid. Roll-in transitions and both local probes count toward the budget. An incomplete final pair consumes remaining transitions without an MLE update. A run stops at the primary target or budget exhaustion. All ten seeds enter each summary, with failed runs right censored at 12,000; the empirical

restricted mean is the mean of these capped costs. Intervals use 10,000 paired seed resamples and are descriptive, without a multiple-comparison correction.

A multistep information branch. There are five path components $x_0, \dots, x_4$ and three absorbing components s, h, ℓ, each with eight observed aliases. Every candidate has the same transitions at x_k for $k < 4$:

$$(0,0) \to s, \quad (0,1) \to h, \quad (1,0) \to x_{k+1}, \quad (1,1) \to \ell.$$

At x_4, only $(1,0)$ depends on the model: it reaches h with probability p and ℓ otherwise. Truth has $p = 0.95$; the 31 decoys have $p_k = 0.05 + 0.30k/31$. Path and safe rewards are $0.5(1-\gamma)$; high and low rewards are $0.95(1-\gamma)$ and $0.05(1-\gamma)$. The initial distribution is uniform over aliases of x_0. Emissions are random within disjoint component supports; model order and feature coordinates are randomly permuted. Mixture weights lie in the simplex and emissions are distributions, giving a realizable factorization with the required norm bounds. Exact zero likelihoods are handled as $-\infty$, without probability flooring.

Each decoy's nominal pair chooses $(0,0)$ on the path. In truth, the maximizer can deviate to action 1 along the path against that minimizing policy, obtaining gap $0.405\gamma^5 = 0.23914845$. Thus the multistep information affects both reported accuracy targets. Both players' actions influence access to it. Two local probes from x_0 cannot reach the informative context; the roll-in must advance into the path. In contrast, the hidden family's disagreement is at the initial component, and its first probe queries that context with probability at least $(1-\gamma)/J = 0.025$ under every roll-in rule.

Pilot, confirmation, and numerical checks. Technical pilots use one independent seed per family and a 1,200-transition budget. The first deep-branch draft sent every nonadvancing action to the safe component. Its entrance ties allowed valid zero-sample certificates for some tie-broken nominal pairs, weakening its exploration test. The high/low exits remove this degeneracy; alias-invariant decoy gaps were checked before confirmation. Both pilot versions are retained; neither enters the ten-seed summaries. Confirmation fixes all settings and seeds in its manifest before execution. No parameters were tuned on confirmation results and no seeds were replaced.

Planning requests $\eta = 10^{-7}$ and best-response values use tolerance 10^{-8} with the residual corrections of Appendix F.4. All methods reuse nominal plans indexed by the selected model and fixed-model response bounds indexed by the candidate/model pair. Reuse is exact for this fixed finite class; the current likelihood set is recomputed every round and may grow again. No monotone elimination assumption is used. Raw training tuples permit replay of every likelihood, candidate choice, and certificate. All 120 trajectories retain truth and satisfy numerical certificate coverage at all 26,353 evaluated rounds, including the initial evaluations.

Interaction results and interpretation. Table 6 and Figure 2 use the primary target 0.05. The following paired differences compare mean capped primitive costs to reverse crossing (positive favors reverse; 95% bootstrap intervals):

Family	CE minus reverse	Uniform minus reverse
Default random	-10.0 $[-113.1, 72.9]$	26.2 $[-30.2, 91.1]$
Small random	59.3 $[0.0, 177.9]$	18.3 $[-144.3, 199.2]$
Small hidden	-17.9 $[-72.0, 18.3]$	399.5 $[238.6, 573.3]$
Deep branch	9645.7 $[8644.7, 10409.3]$	7858.8 $[5322.1, 9977.1]$

Default and small random games have six and seven zero-transition successes, respectively, common to all roll-ins. At the secondary 0.1 target, every random instance already certifies at $n = 0$; hidden and deep first-passage costs equal those at 0.05. These controls make no general case for exploration superiority. On the deep branch, reverse crossing has median 1,893.5 transitions and mean 2,354.3, versus capped means 12,000 for CE and 10,213.1 for uniform. Reverse crossing sees a median of one informative training context, compared with zero for each control. Its median set size remains 32, while the median selected CE certificate falls to 0.00820. Certification can improve through selecting a pair valid across the retained models, without identifying a singleton model. The uninformative controls retain median true gap 0.239. In all ten successful reverse runs the set still contains all 32 models. The resulting certificate therefore also witnesses a suitable pair for the initial full class. A learner that enumerates all nominal candidates and certifies each before interaction could find such a

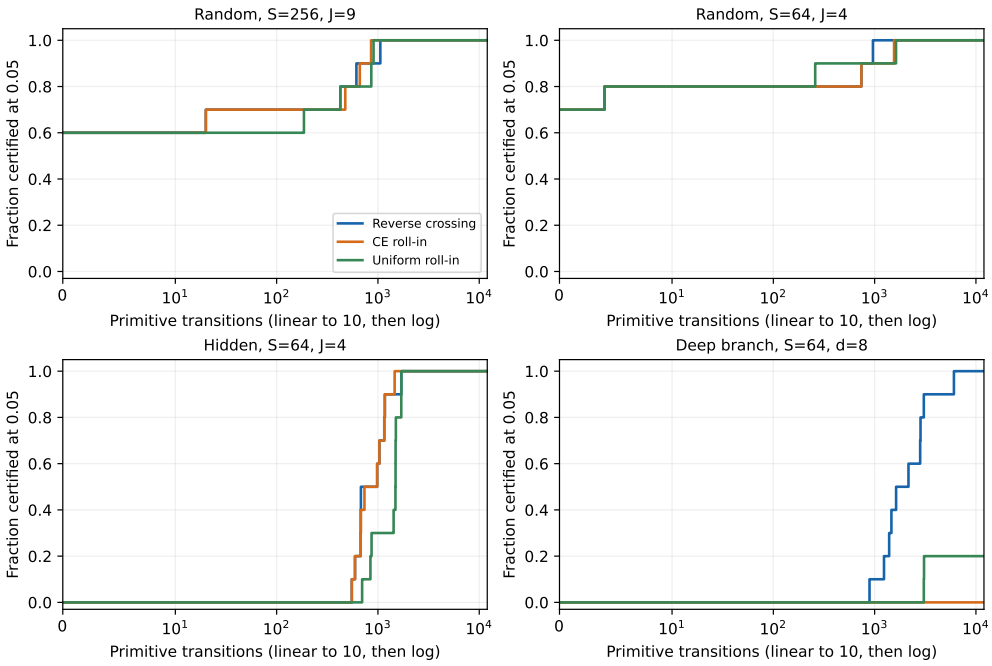


Figure 2: **Independent data collection, common certification.** Empirical fraction reaching $C_n^{\mathrm{set}} \leq 0.05$ versus primitive transitions, ten paired seeds per family and method. Horizontal scaling is linear through ten transitions and logarithmic thereafter, retaining zero-sample successes. Curves include all seeds; unfinished deep-branch CE/uniform runs remain censored at 12,000.

pair without data. This experiment isolates exploration under the fixed current-MLE/CE output rule; it is not an information lower bound against broader candidate search.

End-to-end computation. The confirmation uses Python 3.11.15, NumPy 2.4.6, SciPy 1.17.1 and pandas 3.0.5, with twelve workers and one BLAS thread per worker. The table gives mean elapsed seconds through stopping or censoring, excluding separately timed true-model diagnostics but including sampling, likelihood updates, covariances, planning, certification, and loop overhead. These are parallel learning runs, distinct from the earlier serial cold-checkpoint timings.

Family	Reverse crossing	CE roll-in	Uniform roll-in
Default random	33.53	6.80	5.68
Small random	1.60	0.70	0.59
Small hidden	4.75	0.69	0.71
Deep branch	8.20	0.97	0.93

Reverse crossing pays for two additional confidence-game plans between checks. In particular, its deep-branch interaction advantage comes with higher computation; the faster controls mostly fail to certify there. These results compare exploration under the calibrated profile and a valid set certificate. They neither instantiate the proof-profile hybrid of Corollary 28 nor establish its worst-case sample-complexity rate empirically.

Exhaustive initial nominal-candidate search. For each of the same ten deep-branch classes, we solve every model's nominal game and evaluate its pair against every initial model. The selection uses only the known class, rewards, discount, and initial distribution, with no true-model argument and zero interactions. It minimizes

$$F_0 = \min_{Q \in \mathcal{M}} \max_{P \in \mathcal{M}} \mathrm{Gap}_P(\pi_Q, \nu_Q)$$

over the computed nominal pairs, adding best-response residuals to obtain an upper certificate. This is finite candidate search, not a global robust equilibrium optimization; it is also a legal initial-MLE tie-breaking rule because every initial log likelihood is zero.

We deliberately measure a simple complete implementation: cold nominal plans, cold response solves, no candidate deduplication, no early stopping, and no cache across candidates. Each class costs exactly 32 complete zero-sum planning calls and $2 \cdot 32^2 = 2048$ fixed-opponent best-response calls. All ten minimum upper certificates lie in $[0.0082004453426234, 0.0082004453426281]$. The mean wall time is 6.803 seconds (range 6.771–6.823), comprising 0.461 seconds of nominal planning and 6.342 seconds of certification. Timings are serial with one BLAS thread on an Intel Core i9-9980XE CPU, using the numerical software versions above. Unlike the parallel learning times, this is a serial cold-search measurement; it does not establish a precise end-to-end speed ranking. More efficient search may use fewer calls. This measured feasible baseline cannot be excluded by its quadratic raw call count. It shows that this branch tests how candidate computation can substitute for interaction, rather than unavoidable information demand. In particular $\Gamma_0 \le F_0 \le 0.008201$; we do not claim this upper witness is the exact minimax gap over arbitrary stationary pairs.

H.3 Delayed conflict: an information requirement before experimentation

The original deep branch admits a full-class accurate candidate. We therefore construct a different family with provably incompatible equilibria, adapting the conflict-matrix mechanism of Theorem 14. This is an identifiability diagnostic, not a sample-complexity lower bound or a claim that reverse crossing must outperform other explorers.

Construction. There are an initial component x_0, two paths $u_1, \dots, u_L$ and $v_1, \dots, v_L$, and absorbing components safe, high, and low. At x_0, joint actions $(0,0), (1,1)$ lead to safe, $(0,1)$ to u_1, and $(1,0)$ to v_1. Each path advances deterministically, independently of actions. In model P_θ, the last component u_L transitions to high with probability θ and to low otherwise; v_L interchanges these probabilities. Rewards are $(1-\gamma)/2$ at the initial, path, and safe components, $1-\gamma$ at high, and zero at low. The known class contains $\theta = 1/2 \pm a$, with $0 < a < 1/2$, and can contain additional parameters between them. All models share the reward and every preterminal transition.

Each component emits into its own disjoint set of observed aliases, with an arbitrary fixed probability distribution. Mixture weights over next components give a factor ϕ, and the emission distributions give μ, of dimension $d = 2L + 4$. Their norms satisfy $\|\phi\|_2 \le 1$ and $\|\sum_s \mu(s) f(s)\|_2 \le \sqrt{d}\|f\|_\infty$. The initial distribution is supported on aliases of x_0.

Proposition 39 (No zero-sample common approximate equilibrium)**.** *For this family, let $c = a\gamma^{L+1}$. Over all stationary product-policy pairs, including policies depending on observed aliases,*

$$\Gamma_0 := \inf_{\pi,\nu} \max_{P \in \mathcal{M}} \mathrm{Gap}_P(\pi, \nu) = c.$$

Thus no pair has a valid full-class gap certificate below c. For $L \ge 2$, the two consecutive local probes starting at x_0 observe no model-dependent transition.

Proof. The transition into high or low occurs at time $L+1$. The normalized discounted return for each initial joint action is therefore given by

$$A_+ = \begin{pmatrix} 1/2 & 1/2 + c \\ 1/2 - c & 1/2 \end{pmatrix}, \qquad A_- = \begin{pmatrix} 1/2 & 1/2 - c \\ 1/2 + c & 1/2 \end{pmatrix}.$$

Every action after the initial state has the same transition and reward. Let $p = \mathbb{E}_{s \sim d_0} \pi(0 \mid s)$ and $q = \mathbb{E}_{s \sim d_0} \nu(0 \mid s)$. The dominant actions in each endpoint matrix are independent of the alias. Computing both best-response values and taking their difference gives

$$\mathrm{Gap}_+(\pi, \nu) = c(2 - p - q), \qquad \mathrm{Gap}_-(\pi, \nu) = c(p + q).$$

These gaps sum to $2c$, so their maximum is at least c for every pair. Choosing uniform initial actions gives gap c at both endpoints and at most c for every intermediate parameter, establishing equality. The first two transitions from the initial component can reach only the first two path components or safe. Model dependence occurs when leaving u_L or v_L, after at least three transitions when $L \ge 2$. □

For $(\gamma, L, a) = (0.9, 4, 0.4)$ this gives $\Gamma_0 = 0.236196 > 0.05$. For our entire grid $L \in \{2, 4, 6\}$ and $a \in \{0.25, 0.4\}$, the smallest value is 0.119574225, above both recorded targets 0.05 and 0.1. This excludes *all* stationary common pairs, a stronger statement than failure of an enumerated nominal-candidate pool. Data can still select a pair with high-probability validity under the retained likelihood class; the proposition only precludes uniform certification on the initial class.

Frozen collection and controls. The confirmed class has only the two endpoints, with four aliases per component, giving $(d, |\mathcal{S}|) = (8, 32), (12, 48), (16, 64)$ as depth increases. Each of the six (L, a) settings uses ten independent seeds for emissions, model ordering, and feature permutations. Each seed is evaluated under *both* true endpoints and all five methods. The resulting 600 trajectories use independent confirmation seeds; a 60-trajectory technical pilot at budget 1,200 is excluded from confirmation summaries. No bonus parameter was retuned. Tests of stochasticity, normalization, low-rank norms, alias-dependent gap identities, and probe reachability preceded the pilot. The two nominal endpoint candidates have full-class gap $2c$, whereas arbitrary common pairs have optimum c.

All methods start with one uniform roll-in and then differ only in their roll-in policy. Exact MLE, paired uniform probes, first-tuple covariance, calibrated bonus, CE output, and LR-set certification are common. Only the probe tuples train the learner, matching the original access protocol; every roll-in transition still consumes budget. Any action at either terminal path component is informative. The CE policy enters safe at the root; after the initial uniform roll-in, its two probes cannot reach a model-dependent transition. Uniform exploration selects one of the two paths with probability one half at the root, and subsequent path actions are irrelevant. Thus delayed information here does not by itself make uniform exploration exponentially unlikely in the path length.

The bonus baseline treats the joint action as a single MDP control and maximizes discounted bonus under the same current fitted kernel. A deterministic joint-action policy factors into two deterministic policies for the common roll-in interface. This differs from solving a zero-sum game with bonus payoff. Upper same-side uses the upper game's full saddle pair. Nominal and response caches, online checks, residual tolerances, and handling of incomplete final tuple pairs match Section H.2. Warm starts do not remove planner calls from the accounting.

Interaction results, uncertainty, and observed failures. Table 2 includes all seeds and counts observable threshold hits, including three incorrect certifications. The 0.1 and 0.05 first-passage times coincide in these two-model games. No configuration has a zero-sample hit. Numerical replay verifies every likelihood, selected MLE, certificate, informative-tuple count, and stopping point at 191,965 checkpoints. Three trajectories exclude truth and then stop on an inaccurate pair: reverse at $(L, a) = (6, 0.25)$ on two seed/endpoint combinations, and uniform at $(2, 0.25)$ on one. At confidence level 0.95, true-model exclusion is possible; it is not discarded as a numerical failure. The pilot also has two such events, retained separately. Truth enters only environment sampling and separately timed diagnostics, never fitting or stopping.

Figure 3 gives complete threshold curves. Intervals in Table 7 use 10,000 paired bootstrap resamples of the ten independent seed blocks, retaining both endpoints and all methods together. They are descriptive and not adjusted for multiple comparisons. Bonus and uniform minus reverse intervals exclude zero in all six settings; upper minus reverse is unresolved at $(2, 0.4)$. Table 8 reports both truth strata, preventing an apparent benefit from depending on a single privileged endpoint.

Table 7: **Paired capped-cost differences: comparator minus reverse.** Negative favors the comparator. Each entry is mean [95% paired bootstrap interval], in primitive transitions; both true endpoints stay together when resampling seeds.

L	a	Upper minus reverse	Bonus minus reverse	Uniform minus reverse
2	0.25	-1350 $[-1898, -877]$	-2483 $[-3086, -1789]$	-2278 $[-2848, -1696]$
2	0.4	-122 $[-378, 158]$	-2413 $[-2650, -2183]$	-2336 $[-2595, -2081]$
4	0.25	-1557 $[-2080, -1042]$	-2384 $[-2845, -1947]$	-1676 $[-2554, -750]$
4	0.4	-761 $[-1062, -487]$	-2245 $[-2589, -1907]$	-2234 $[-2574, -1906]$
6	0.25	-876 $[-1425, -367]$	-1633 $[-2127, -1141]$	-1214 $[-1561, -859]$
6	0.4	-655 $[-968, -366]$	-1884 $[-2241, -1494]$	-1705 $[-2052, -1336]$

Table 8: **Balanced true-model strata.** Mean capped primitive transitions at 0.05, ten seeds in each row. Every non-CE row has ten threshold hits; CE has none. The three inaccurate outputs remain included.

L	a	Truth	Reverse	Upper	Bonus MDP	CE	Uniform
2	0.25	$-$	3,250.5	1,747.5	672.6	12,000.0	547.2
2	0.25	$+$	2,973.1	1,776.8	585.4	12,000.0	1,121.3
2	0.4	$-$	2,376.1	2,494.2	148.0	12,000.0	244.9
2	0.4	$+$	2,827.5	2,464.4	229.4	12,000.0	286.0
4	0.25	$-$	3,046.4	1,352.1	826.3	12,000.0	1,177.1
4	0.25	$+$	3,192.7	1,772.4	644.1	12,000.0	1,710.8
4	0.4	$-$	2,515.8	1,666.6	165.3	12,000.0	181.7
4	0.4	$+$	2,305.2	1,632.3	165.4	12,000.0	171.8
6	0.25	$-$	2,065.5	1,425.9	542.3	12,000.0	779.7
6	0.25	$+$	2,565.7	1,453.6	822.0	12,000.0	1,422.9
6	0.4	$-$	2,471.5	1,504.4	349.0	12,000.0	471.7
6	0.4	$+$	1,928.0	1,586.0	281.6	12,000.0	517.6

Table 9: **End-to-end learner wall time in seconds.** Means over twenty confirmation runs per cell, through stopping or censoring; true-model diagnostics are excluded.

L	a	Reverse	Upper	Bonus MDP	CE	Uniform
2	0.25	13.99	2.75	0.82	0.71	0.11
2	0.4	6.11	2.75	0.28	0.68	0.06
4	0.25	14.36	2.60	0.96	0.81	0.16
4	0.4	8.29	2.33	0.28	0.87	0.08
6	0.25	10.81	2.70	0.99	0.79	0.14
6	0.4	8.25	2.49	0.43	0.83	0.11

Table 10: **Actual solver calls per trajectory.** Means over all 120 confirmation runs of each method. CE and exploration ZS columns count complete zero-sum plans; BR counts individual fixed-opponent best responses, and MDP counts joint-action control plans. All methods use the same exact nominal/response cache.

Roll-in	CE ZS	BR	Exploration ZS	Bonus MDP
Reverse	1.60	6.40	476.02	0.00
Upper	1.57	6.27	157.25	0.00
Bonus MDP	1.57	6.27	0.00	40.76
CE	1.03	4.13	0.00	0.00
Uniform	1.58	6.33	0.00	0.00

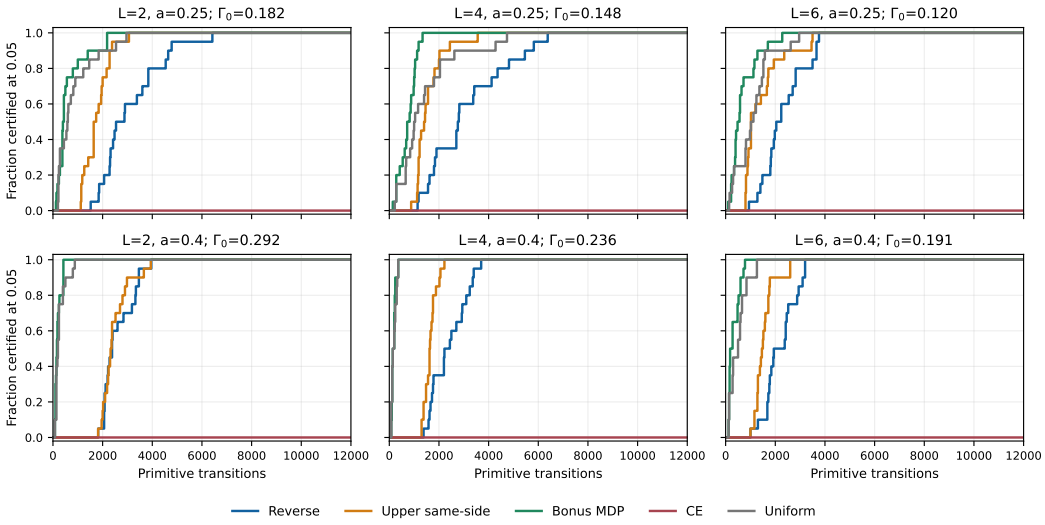


Figure 3: **A proved information requirement does not ensure superior reverse exploration.** Each panel averages twenty runs per method over ten paired seeds and both true endpoints. Curves record observable certificate hits at 0.05, including three inaccurate outputs (two reverse at $L = 6, a = 0.25$; one uniform at $L = 2, a = 0.25$). CE runs are censored at 12,000; all other runs reach the threshold.

Post hoc accuracy sensitivity. As an evaluation-only sensitivity check, charging an inaccurate certified output the full 12,000-transition budget changes the uniform $(2, 0.25)$ mean to 1,405.10 and the reverse $(6, 0.25)$ mean to 3,201.05. Other cells are unchanged; bonus and uniform still have lower means than reverse in every configuration. This diagnostic is not an implementable stopping rule because it uses truth after the run. No run is removed.

Computational interpretation. Reverse crossing requires two complete zero-sum exploration plans per update in this CE/set implementation. Upper same-side needs one, and bonus occupancy needs one control-MDP plan. Table 10 counts these separately from nominal CE plans and fixed-opponent best responses. It includes the final planning call when a roll-in exhausts the budget without producing another tuple pair. Timings use the same CPU and software as the initial search, twelve worker processes, and one BLAS thread per worker. Elapsed learner time includes sampling, updates, planning, certification, and overhead, but excludes true-model diagnostics. These measured parallel-run times are descriptive, not hardware-independent complexity bounds. Reverse has higher mean wall time than upper, bonus, and uniform in every configuration.

In the native envelope algorithm, the two confidence-game solutions are already needed for certification; reverse crossing reuses them for the width-occupancy argument without a third exploration optimization. This oracle reuse is the justified reason to use the reduction when that certification interface is desired. It does not imply dominance when another output/certification interface, such as finite-class CE/set, is available. These results distinguish a sufficient theoretical mechanism from an empirically best exploration rule; they give no evidence that the calibrated reverse crossing is the latter.

H.4 Global-bonus comparison on all saved checkpoints

Immutable data and matched confidence. This analysis reuses the 30 trajectories and all 170 checkpoints described in Section H.5; it collects zero additional transitions. At each checkpoint we restore the selected MLE, exact-support likelihood scores, and per-model first-tuple covariances. Both proof and calibrated bonuses use these same sufficient statistics. For each of the 340 checkpoint/profile combinations, we solve the nominal and two confidence games, compute both candidates' envelope and LR-set certificates, and solve the joint-action bonus MDP. The latter returns an interval for U_n, an executable deterministic policy, and the upper bound required by Theorem 11. Neither the true model nor true-gap evaluation is used in any of these operations.

Table 11: **Full fixed-checkpoint certificate comparison.** Final-checkpoint medians over ten seeds per setting; all quantities are in normalized-return units. CE true-gap medians are below 10^{-7} in every row. Zero denotes numerical values below that threshold. Raw proof bounds are shown; capped values and every checkpoint remain in the artifact.

Profile	Setting	Cross env.	CE env.	Global	Cross LR	CE LR	Cross gap
Calibrated	Default random	0.653	0.678	0.774	0.027	0	0.027
Calibrated	Small random	0.408	0.421	0.503	0.009	0	0.009
Calibrated	Small hidden	0.492	0.529	0.712	0.058	0	0.058
Proof	Default random	411.873	439.104	466.254	0.204	0	0.204
Proof	Small random	329.450	345.079	368.437	0.152	0	0.152
Proof	Small hidden	478.299	483.548	504.875	0.112	0	0.112

The recorded global certificate is $2\overline{U}_n + 2\zeta_n + 2\eta_n^0$. The bonus solver stops when its residual, with a scale-aware numerical allowance, implies value error at most 10^{-8}. The nominal and confidence-game policies are additionally checked against their statewise one-sided Bellman inequalities; their achieved saddle errors enter all native certificates. Envelope and LR-set evaluations use the same best-response residual corrections as the original comparison. Task and iteration identifiers link each row to its saved snapshot. The analysis also checks every parameter, candidate-specific envelope and LR-set value, and true gap against the original 680 comparison rows. Numerical differences and the largest bonus-oracle slack are saved in `integrity.json`.

Resolved differences and their magnitude. Every checkpoint and parameter matches the original comparison; the largest reproduced envelope discrepancy is 5.1×10^{-12}. The bonus-oracle slack is at most 2.0×10^{-8}. The certificate ordering in Proposition 6 holds in all 340 rows. Numerical intervals resolve strict crossing-versus-CE separation in 324 rows and strict CE-versus-global separation in 320 rows. These are descriptive checkpoint counts, not 340 independent statistical samples.

At final checkpoints, mean paired percentage reductions in the calibrated crossing certificate relative to global bonus are $15.1\%\,[12.8, 17.2]$, $17.3\%\,[14.5, 20.1]$, and $30.8\%\,[30.2, 31.3]$ for default random, small random, and small hidden games, respectively. Relative to CE's envelope, they are $4.4\%\,[3.6, 5.2]$, $2.4\%\,[1.2, 3.7]$, and $6.9\%\,[6.7, 7.1]$. Brackets are pointwise 95% percentile bootstrap intervals from 10,000 resamples of the ten paired seeds within each setting; no independence across checkpoints is assumed. Both profiles' paired effects are archived. Table 11 shows that these certificate differences coexist with numerical-zero final CE gaps and much tighter LR-set certificates. All 30 final likelihood sets are singletons containing truth. None of the three calibrated envelope/global certificates reaches 0.05 on these trajectories; all proof-profile raw bounds exceed one, so their capped versions are uninformative.

The joint-action MDP is inexpensive here: its calibrated final median time is 0.0111, 0.0082, and 0.0083 seconds. Including the nominal plan, CE + bonus takes 1.0673, 0.0473, and 0.0468 seconds, versus 2.2541, 0.0774, and 0.0768 seconds for native crossing. These host-matched measurements support a tighter-envelope versus cheaper-planning tradeoff, not a claim that avoiding an additional MDP necessarily saves runtime.

Planning counts and serial timing. We recompute all candidates and certificates on one host, using Python 3.11.15, NumPy 2.4.6, SciPy 1.17.1, pandas 3.0.5, one worker, and one BLAS thread. Every plan starts from zero; no policy, value, or response cache is shared across checkpoints or methods. Reported times are one serial cold measurement per checkpoint/profile and method, not repeated microbenchmark averages. Bonus construction is logged separately because it is common to the envelope methods; model fitting, interaction, and true-gap evaluation are excluded.

Native crossing uses two full zero-sum plans and no BR or joint-MDP call. CE-envelope uses one zero-sum plan and two BR calls for output; CE + global bonus uses one zero-sum and one joint-MDP call for output and exploration. CE + LR-set uses one zero-sum plan and $2K_n$ BR calls, where K_n is the active class size. CE-envelope and CE + LR-set require an additional exploration policy for learning. The diagnostic BR calls used to check the crossing's envelope value are excluded from its native cost. LR-set certification is also recorded for the crossed candidate.

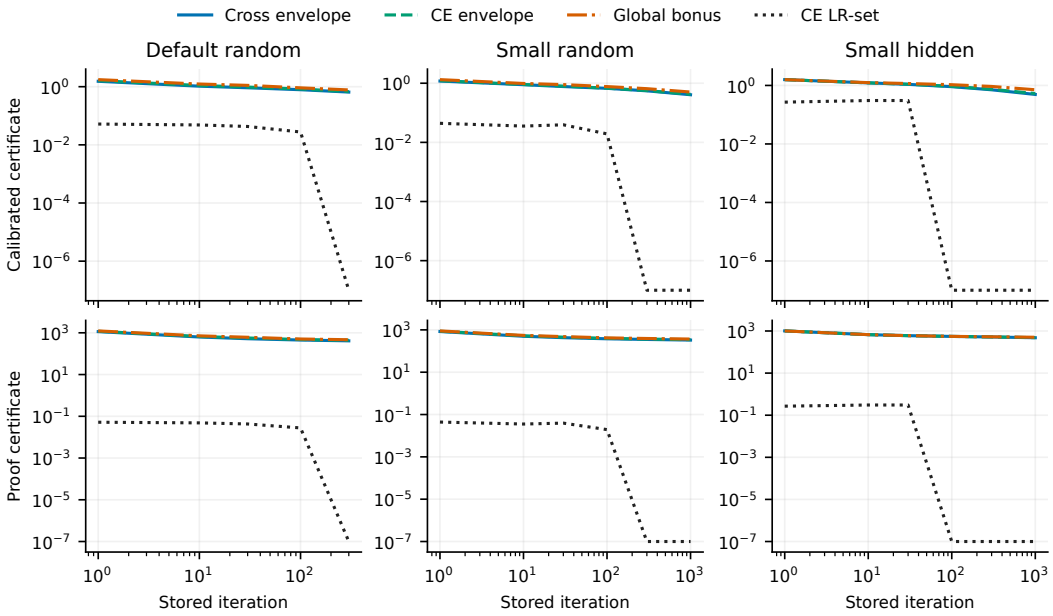


Figure 4: **Global and deviation certificates on identical learned models.** Medians over ten seeds per panel, with a display floor of 10^{-7}. The two rows use the calibrated and sufficient proof parameters on the same stored histories. Curves do not represent new runs of either a proof-profile learner or a bonus-exploration learner.

Table 12: **Matched cold-planning and certification times (seconds).** Final-checkpoint medians over ten seeds. The first four numerical columns include candidate planning and the indicated certificate; the last column isolates the bonus-MDP call and is included in CE + bonus. Exploration is supplied natively only by crossing and CE + bonus.

Profile	Setting	Native cross	CE envelope	CE + bonus	CE + LR	Bonus MDP only
Calibrated	Default random	2.2541	1.0699	1.0673	1.0702	0.0111
Calibrated	Small random	0.0774	0.0458	0.0473	0.0459	0.0082
Calibrated	Small hidden	0.0768	0.0452	0.0468	0.0452	0.0083
Proof	Default random	2.9028	1.0737	1.0706	1.0692	0.0151
Proof	Small random	0.1053	0.0476	0.0503	0.0456	0.0114
Proof	Small hidden	0.1067	0.0473	0.0500	0.0452	0.0116

This comparison quantifies a fixed-confidence tradeoff. It neither establishes faster reverse exploration nor tests avoidance of an irrelevant unknown branch. The strict analytic example and attainable covariance construction delimit possible certificate separation; the saved runs measure its magnitude on ordinary learning histories. The delayed-conflict study remains the evidence about interaction efficiency, including its unfavorable results for reverse crossing.

H.5 Candidate and certificate comparison on shared data

Protocol and scope. Thirty fresh data trajectories use three configurations and ten seeds per configuration. Default random games have $|\mathcal{S}| = 256, J = 9, N = 300$; small random and small hidden games have $|\mathcal{S}| = 64, J = 4, N = 1000$. All use $d = 4, |\mathcal{M}| = 32, \gamma = 0.9, \delta = 0.05$. The saved checkpoints are $1, 10, 30, 100, 300$, with 1,000 added in the two small settings. The manifest fixes configurations and seeds before collection. A technical pilot reruns the first prespecified task to check execution; these summaries count each prespecified task only once. Each data trajectory follows calibrated DREP-ZS; the alternative candidates never alter its exploration or random stream.

The NPZ checkpoints preserve exact-support log likelihoods, first-context feature covariances for every candidate model, and the selected MLE. The comparator recomputes the calibrated and proof-

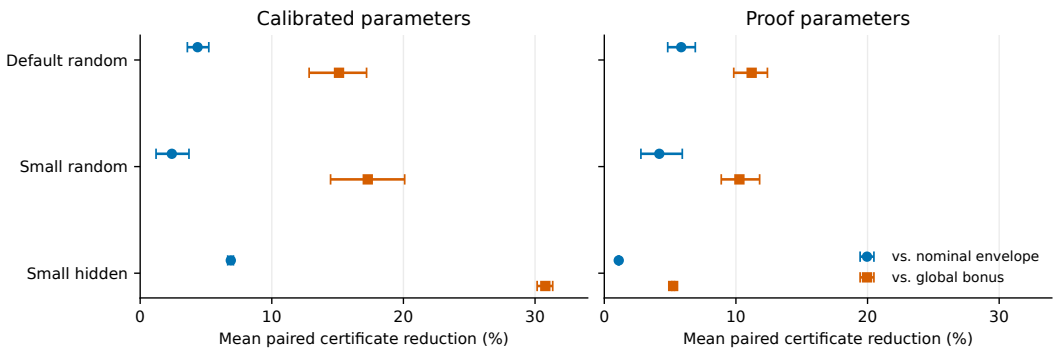


Figure 5: **Certificate differences on a relative scale.** Mean final-checkpoint reduction $100(1 - C_{\text{cross}}/C_{\text{comparator}})$ across ten paired seeds per setting, with pointwise 95% percentile bootstrap intervals from 10,000 resamples. Both profiles use the same stored data. These are raw certificate comparisons; proof-profile bounds exceed one. The full-range comparison, including the tighter LR-set certificate, is retained in Figure 4.

Table 13: **Common-data certification at the final checkpoint.** Medians over ten paired seeds per configuration under the calibrated profile. Each paired cell reports crossing / CE; numerical zero denotes values below 10^{-7}.

Family	N	Gap: cross / CE	C_n: cross / CE	C_n^{set}: cross / CE	$\lvert\mathcal{M}_N^{\text{LR}}\rvert$
Default random	300	0.0271 / $\simeq 0$	0.6530 / 0.6777	0.0271 / $\simeq 0$	1
Small random	1000	0.0087 / $\simeq 0$	0.4079 / 0.4211	0.0087 / $\simeq 0$	1
Small hidden	1000	0.0577 / $\simeq 0$	0.4919 / 0.5289	0.0577 / $\simeq 0$	1

profile bonuses from the same statistics, solves their upper/lower games, and solves the nominal game for CE. All four candidate/profile rows at a checkpoint use the same saved snapshot. In particular, the proof-profile test concerns confidence on these data, not the sample complexity of a new proof-profile trajectory. All selected checkpoint models maximize the exact-support likelihood as well as the legacy implementation's floored likelihood.

Certificates and numerical precision. The envelope certificate is evaluated for both candidates by two best-response MDPs. The set certificate uses the threshold $\log(|\mathcal{M}|/0.05)$; each retained model supplies two best responses for the fixed candidate. The set certificate has failure level 0.05. The proof-profile envelope also has a stated confidence guarantee, whereas calibrated-envelope coverage is empirical. Two valid 95% marginal guarantees would not imply joint 95% coverage without splitting the budget. Candidate planning requests $\eta = 10^{-7}$; response solvers request value error 10^{-8} and include the residual corrections in Appendix F.4. The largest observed envelope response correction is 2.3×10^{-8}. Coverage is checked against the upper endpoint of the numerical true-gap interval, capped by the known gap bound of one, with 10^{-10} comparison tolerance. All 680 envelope and all 680 set comparisons cover at the logged checkpoints. Every likelihood set retains the true model. These finite audits do not replace either theorem.

Paired effects. The final likelihood sets are singletons in all 30 trajectories, and all CE true-gap upper endpoints are below 10^{-7}. For calibrated certificates, paired median CE-minus-crossing differences (95% bootstrap intervals from 10,000 seed-pair resamples) are:

Family	True gap difference	Envelope certificate difference
Default random	$-0.0271\ [-0.0441, -0.0203]$	$0.0276\ [0.0251, 0.0376]$
Small random	$-0.0087\ [-0.0162, -0.0027]$	$0.0076\ [0.0021, 0.0203]$
Small hidden	$-0.0577\ [-0.0596, -0.0539]$	$0.0369\ [0.0351, 0.0378]$

Raw numerical gaps remain in the artifact; displayed numerical zeros do not carry evidence of ranking below the solver's precision. The crossing order agrees with Proposition 6 up to numerical

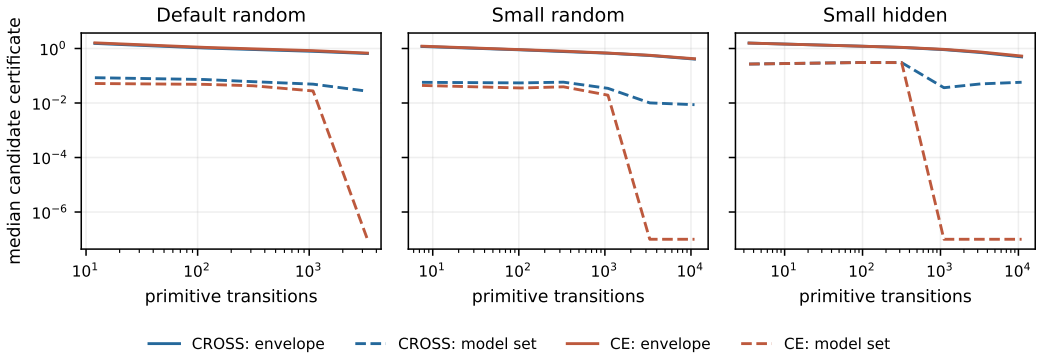


Figure 6: **Certificates on identical data (ten seeds per configuration).** Solid lines: common calibrated-envelope certificates; dashed lines: likelihood-set certificates. Curves are medians, with display floor 10^{-7}. Candidate formation and certification consume no additional environment transitions.

tolerances at every checkpoint. The set certificate can favor CE even though crossing minimizes the common-envelope certificate. The artifact also reports success rates at 0.1 and 0.05, per-seed censoring, and transition costs capped at each trajectory's budget. These costs include every seed; they are not conditional averages over successful runs. Hitting times are first observed on the prespecified checkpoint grid and do not claim exact online stopping times between checkpoints. At threshold 0.05, calibrated-envelope certificates reach the target in 0/10 seeds for either candidate in each family. Likelihood-set certificates reach it in 10/10 CE seeds in every family, versus 8/10, 10/10, and 8/10 crossing seeds in default random, small random, and small hidden games. For CE-set, the corresponding median first-logged-hit costs are 133.5, 19, and 1139.5 primitive transitions. All seeds enter these summaries; crossing results retain two censored seeds in each of the first and third families. These comparisons share a data-collection trajectory and do not measure CE-driven exploration.

Proof-profile conservatism. Median raw proof-profile certificates (crossing / CE) are 411.9/439.1, 329.4/345.1, 478.3/483.5 for the three respective configurations. All exceed one, even before considering the practically useful thresholds 0.1 or 0.05. Reporting the capped value one would preserve validity while hiding the scale of the conservative profile, so raw values are retained here. The constants demonstrate an executable sufficient schedule, not a practical tuning recommendation.

Original computation records. The following table and Figure 7 retain the Python 3.12 timings. The matched comparison including global bonus is reported separately in Section H.4. The baseline checkpoint comparison runs serially, with one BLAS thread and no concurrent trajectory collection. Candidate planners start without a warm-start vector. The following final-checkpoint times are medians in seconds. Crossing's width uses its two Shapley plans without additional best-response calls; CE-envelope uses one nominal plan and two responses; CE-set uses one nominal plan and two responses per active model. The diagnostic responses used to evaluate crossing's C_n are excluded from its native width cost. Ground-truth evaluation is timed separately and excluded from every column.

Family	Crossing width	CE-envelope	CE-set
Default random	7.169	4.105	4.105
Small random	0.173	0.104	0.102
Small hidden	0.172	0.108	0.114

These are checkpoint planning/certification costs, not end-to-end learning benchmarks. Class likelihood scores have already been accumulated for MLE. The full active-class sizes, separate planning/response times, and bootstrap intervals are supplied in the artifact. In this identifiable 32-model family, set-based certification becomes inexpensive once only one model remains; the comparison supplies no general runtime dominance claim for the single-model method.

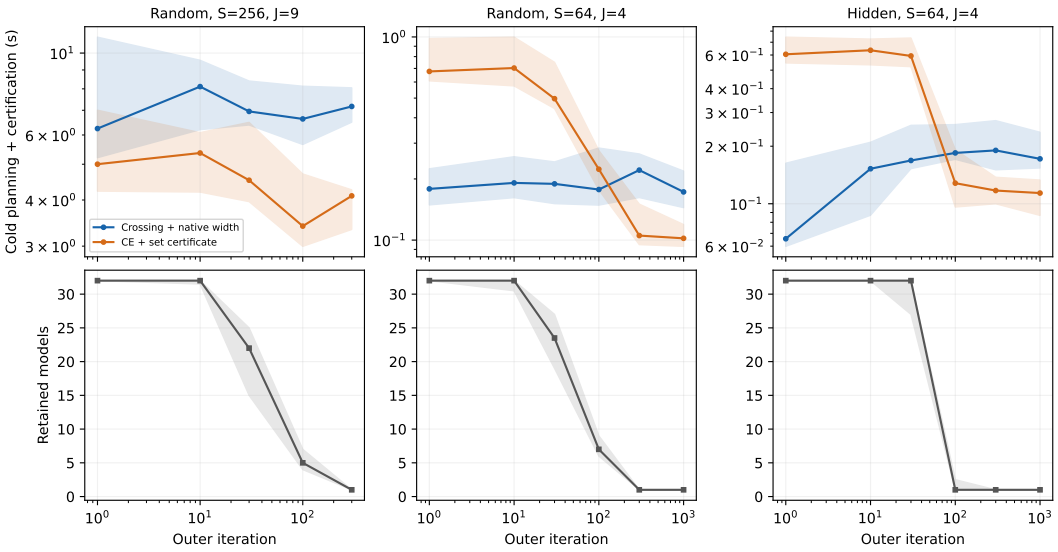


Figure 7: **Checkpoint computation as the likelihood set contracts.** All saved checkpoints from the common-data comparison, with ten paired seeds per configuration. Top: cold-start candidate planning plus native width or set certification. Bottom: retained-model counts. Bands are pointwise 95% bootstrap intervals for medians (10,000 seed resamples). These costs exclude trajectory collection and fitting.

Figure 7 shows a stage-dependent tradeoff. At the first checkpoint, both small configurations retain a median of 32 models: crossing versus CE-set costs are 0.179 versus 0.676 seconds on random games and 0.066 versus 0.605 seconds on hidden games. At the final checkpoint the sets are singletons and the ordering reverses, as above. Default random games favor CE-set even at the first checkpoint (5.001 versus 6.245 seconds). Counting two Shapley calls versus $2K_n$ response calls alone therefore cannot determine runtime; both call type and the evolving active class matter.

The common-data study uses Python 3.12.14, with dependency versions pinned in the accompanying artifact; the baseline campaign used Python 3.11. The 30 common-data trajectories are additional to the baseline 2,773 realizable-class runs. The 680 comparisons are repeated evaluations on 170 checkpoints, not independent learning runs.

H.6 Zero-sum certification and strategic ablations

Figure 8 shows the default random-family learning curves and the certificate audit. Across all 30 seeds and all checkpoints, DREP-ZS attains 100% simultaneous Nash-certificate coverage and 100% finite-envelope coverage, with no audited violations. The stored certificate covers the true gap with a substantial margin.

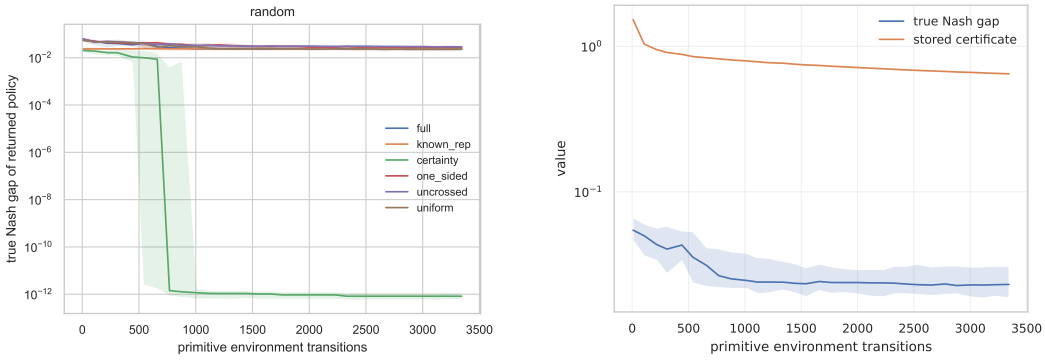


Figure 8: **Default zero-sum experiment (30 paired seeds).** Left: returned-policy true Nash gap versus primitive transitions; DREP-ZS and known representation use minimum-certificate selection, and other methods use the current candidate. Shading is a 95% bootstrap interval for the median. Certainty-equivalent MLE becomes exact after identifying the finite true model, whereas confidence-based policies retain a conservative gap. Right: DREP-ZS's stored certificate covers the true gap at every checkpoint but is loose.

Table 14 quantifies the comparison. Paired bootstrap differences (baseline minus DREP-ZS) show smaller gap AUC for the full method than for one-sided optimism by 0.00394 $[0.00155, 0.00668]$, and than for uncrossed output by 0.00370 $[0.00075, 0.00535]$. Uniform and targeted exploration are not distinguishable on random instances: their paired AUC difference is 0.00003 with interval $[-0.00160, 0.00154]$. Known representation reduces AUC by 0.00365 $[0.00252, 0.00494]$. Its paired final-gap difference (known representation minus DREP-ZS) is -0.00004 $[-0.00038, 0.00012]$, localizing the measurable representation overhead to early learning rather than the final checkpoint.

Table 14: **Default random zero-sum games.** Entries are medians across 30 paired seeds; brackets are 95% bootstrap intervals for the median. Certainty-equivalent MLE has the smallest numerical gap after exact model identification, but the online nominal baseline reports no certificate; the common-data comparison in Section H.5 supplies one.

Method	Final true Nash gap	Current-gap AUC
DREP-ZS	0.0231 $[0.0192, 0.0308]$	0.0285 $[0.0231, 0.0365]$
Known representation	0.0239 $[0.0183, 0.0307]$	0.0239 $[0.0190, 0.0293]$
Certainty-equivalent MLE	Numerical zero	0.00375 $[0.00316, 0.00495]$
One-sided optimism	0.0282 $[0.0223, 0.0318]$	0.0332 $[0.0281, 0.0399]$
Uncrossed output	0.0289 $[0.0225, 0.0319]$	0.0335 $[0.0272, 0.0396]$
Uniform exploration	0.0232 $[0.0185, 0.0319]$	0.0290 $[0.0258, 0.0349]$

Hidden unilateral deviations. The random family does not reliably expose the strategic ablations, but the hidden-branch stress test does. DREP-ZS has median final gap 0.05323. Relative to it, the paired median excess gap is 0.00819 $[0.00611, 0.00872]$ for one-sided optimism, 0.01140 $[0.01062, 0.01216]$ for uncrossed output, and 0.01300 $[0.01252, 0.01547]$ for uniform exploration. Removing the second tuple produces no clear final-gap difference, and clipping is identical on this instance because the realized bonus does not enter the clipped regime. All methods solve the duality trap by the final checkpoint, so that construction does not separate their long-run behavior. These are ablations of the full procedure, so their differences can combine output quality and data collection. The matched exploration studies isolate the roll-in choice under a common output and certificate.

Finite-model identification. The true model is selected by the finite-class MLE after roughly 600–900 primitive transitions in the default family. Its nominal policy then reaches numerical zero gap, while DREP-ZS keeps paying its confidence bonus. The baseline online baseline reports no certificate. Section H.5 therefore supplies both the best-response certificate from Proposition 6 and a likelihood-set certificate for CE on common data.

H.7 State size, computation, and numerical planning

Table 15 separates statistical and computational scaling. Increasing only the number of observational aliases from 64 to 4096 keeps the median primitive-interaction count between 3174 and 3307, and the final gap between 0.0084 and 0.0130. Simultaneous certificate and envelope coverage remain 100% at every state size. In contrast, DREP-ZS runtime grows from 7.38 to 52.83 seconds, and a dense tabular transition tensor grows from 0.063 to 256.1 MiB. This is consistent with state-independent statistical control and state-dependent computation.

Table 15: **Observed-state scaling at fixed $d = 4$.** Interactions, gap, and runtime are medians for DREP-ZS over 10 seeds. The tabular memory column is the dense transition-model footprint. The 4096-state tabular point uses only 3 seeds and 50 iterations.

$\lvert\mathcal{S}\rvert$	Steps	Gap	Time (s)	Tabular MiB
64	3307	0.0097	7.38	0.063
256	3174	0.0123	10.76	1.004
1024	3226	0.0130	20.16	16.016
4096	3296	0.0084	52.83	256.063

Figure 9 shows that the tabular control also requires 1.40 seconds per iteration at 4096 states, versus 0.041 at 64 states. The strongest additional computational trend is the joint-action sweep: increasing J from 4 to 16 raises median runtime from 8.34 to 261.35 seconds and the final gap from 0.0107 to 0.0433. Increasing γ from 0.5 to 0.95 raises runtime from 1.58 to 17.57 seconds and the gap from 0.00149 to 0.02069.

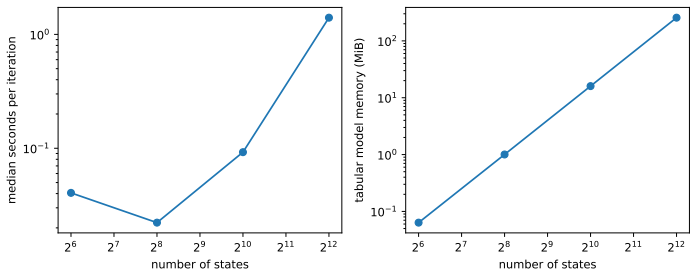


Figure 9: **Computational state-space diagnostic.** Tabular time per iteration and dense-model memory grow with the number of observed states, despite nearly stable interaction counts.

Approximate Shapley planning. Requested tolerances $\eta \in \{10^{-8}, 10^{-4}, 10^{-3}, 10^{-2}, 5 \times 10^{-2}\}$ all yield a median final gap near 0.0524 on the margin-separated hidden-branch games. Runtime nevertheless falls from 40.24 seconds at 10^{-8} to about 8.2 seconds at 10^{-2} and 5×10^{-2}. Across this margin-separated suite, coarser planning reduces runtime without changing the selected policy.

H.8 Long-horizon certificates and fixed-precision scaling

Long-horizon decay. The preceding learning curves end before the conservative certificate has stabilized. We therefore run 10,000 outer iterations with $|\mathcal{S}| = 64, d = 4, J = 4, |\mathcal{M}| = 32$, evaluating every 100 iterations. The random family uses $\gamma \in \{0.5, 0.75, 0.9\}$, and the hidden-branch family uses $\gamma = 0.9$; each configuration has 20 seeds. Table 16 reports the best stored certificate, its corresponding true gap, a descriptive log–log slope fitted to the latter half of the certificate trace, and the first logged iteration at which each threshold is reached. All logged checkpoints in all 80 runs satisfy the stored-certificate inequality.

Table 16: **Long-horizon certificate behavior.** Entries are medians over 20 seeds. Threshold times are observed on a 100-iteration evaluation grid; "–" denotes no hit by iteration 10,000. The fitted slopes summarize the finite experimental range and are not estimates of a theorem exponent.

Family	γ	Certificate	True gap	Slope	$N_{.25}$	$N_{.10}/N_{.05}$
Random	0.50	0.0368	0.0010	-0.388	100	700 / 4850
Random	0.75	0.0733	0.0026	-0.385	300	4750 / –
Random	0.90	0.1774	0.0033	-0.385	4450	– / –
Hidden	0.90	0.1933	0.0317	-0.417	5600	– / –

Figure 10 makes two conclusions visible. First, the certificate continues to shrink throughout the tested horizon, with no visible optimization floor. Second, increasing γ slows this shrinkage much more than it changes the realized final true gap, which is consistent with the discount dependence of the confidence radius.

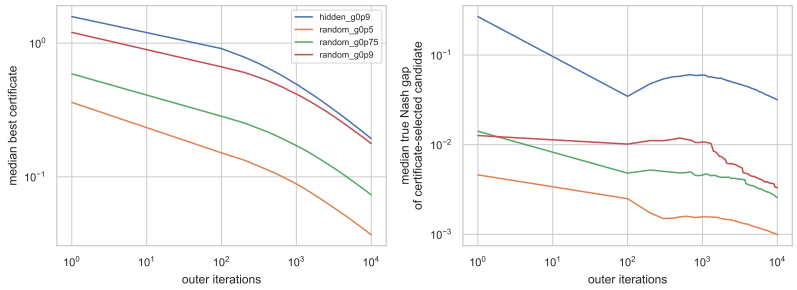


Figure 10: **Long-horizon certificates (20 seeds per configuration).** Median best stored certificate (left) and the true gap of its selected candidate (right), aggregated by the median, versus outer iteration. Every plotted certificate covers its corresponding true gap.

Fixed-precision scaling. We next stop at the first iteration for which the best certificate is at most 0.25, subject to a 10,000-iteration budget. This study uses $|\mathcal{S}| = 64, d = 4, |\mathcal{M}| = 16$ and 10 seeds per configuration. The discount sweep fixes $J = 4$; the joint-action sweep fixes $\gamma = 0.9$. As shown in Table 17, every seed reaches the target in the tested discount configurations through $\gamma = 0.9$ and the tested action configurations through $J = 16$. At $\gamma = 0.95$, no seed reaches the target within the budget, so the reported iteration and transition counts are right-censored rather than successful hitting times; the median terminal certificate is 0.344. All 60 runs retain simultaneous certificate coverage at every logged checkpoint.

Table 17: **Fixed-precision scaling at target certificate 0.25.** Medians use 10 seeds. Primitive transitions include geometric roll-ins and both MLE tuples. The $\gamma = 0.95$ row is censored at the common iteration budget.

Sweep	γ	J	Hit rate	Iterations	Transitions
Discount	0.50	4	100%	4.0	13.0
Discount	0.75	4	100%	240.5	1,226.0
Discount	0.90	4	100%	4,472.5	48,907.5
Discount	0.95	4	0%	10,000	210,462.5
Actions	0.90	9	100%	6,356.5	70,798.5
Actions	0.90	16	100%	8,153.5	89,692.5

H.9 Approximation, policy uniformity, and proof diagnostics

Certified approximate MLE. We tested the likelihood condition in equation 18 on random and hidden families with $|\mathcal{S}| = 64, d = 4, A = B = 2, |\mathcal{M}| = 32, \gamma = 0.9$, 300 iterations, and 30 paired seeds per configuration. Besides exact MLE, we used $\tau_n = c_\tau L_n$ for $c_\tau \in \{0.25, 1, 4\}$,

and a per-tuple negative-log-likelihood error $\varepsilon_{\mathrm{nll}} \in \{10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}\}$, so that $\tau_n = N_n\varepsilon_{\mathrm{nll}}$ for N_n accumulated tuples. To expose the allowed slack without using the true model, the approximate oracle deterministically returns the lowest-likelihood admissible candidate. The bonus uses $L_n + \max_{t\le n}\tau_t$, exactly as in the theorem.

Table 18 and Figure 11 show the predicted distinction between a small per-tuple negative-log-likelihood error and cumulative likelihood slack of order L_n. All 540 runs retain simultaneous certificate and audited-envelope coverage. At $\varepsilon_{\mathrm{nll}} = 10^{-2}$, gap-AUC remains close to exact MLE: 0.0121 versus 0.0106 on random games and 0.0374 versus 0.0409 on hidden games. In contrast, $c_\tau = 4$ raises gap-AUC to 0.0321 and 0.2420, respectively, and reduces final model-selection accuracy to 0% and 76.7%. Thus larger certified slack delays identification as anticipated; realized final Nash gaps are nonmonotone in the tolerance. Paired intervals sharpen this comparison. Relative to exact MLE, $c_\tau = 1$ increases gap-AUC by $0.00970\,[0.00584, 0.01280]$ on random games and $0.02933\,[0.02337, 0.03553]$ on hidden games; at $c_\tau = 4$, the increases are $0.01993\,[0.01334, 0.02878]$ and $0.20061\,[0.18464, 0.21052]$. In contrast, the $\varepsilon_{\mathrm{nll}} = 10^{-2}$ AUC difference is small and positive on random games, $0.00121\,[0.00029, 0.00181]$, but negative on hidden games, $-0.00354\,[-0.00381, -0.00321]$. The theorem permits this realized nonmonotonicity because the tolerance controls admissibility and the certificate, not a monotone sample-path ordering of policies.

Table 18: **Certified approximate-MLE stress test.** Medians and final MLE accuracy use 30 seeds. Every row has 100% simultaneous certificate and finite-audit coverage.

Family	MLE tolerance	Gap-AUC	Final MLE accuracy
Random	exact	0.0106	100.0%
Random	$\varepsilon_{\mathrm{nll}} = 10^{-2}$	0.0121	96.7%
Random	$c_\tau = 1$	0.0228	53.3%
Random	$c_\tau = 4$	0.0321	0.0%
Hidden	exact	0.0409	100.0%
Hidden	$\varepsilon_{\mathrm{nll}} = 10^{-2}$	0.0374	100.0%
Hidden	$c_\tau = 1$	0.0699	100.0%
Hidden	$c_\tau = 4$	0.2420	76.7%

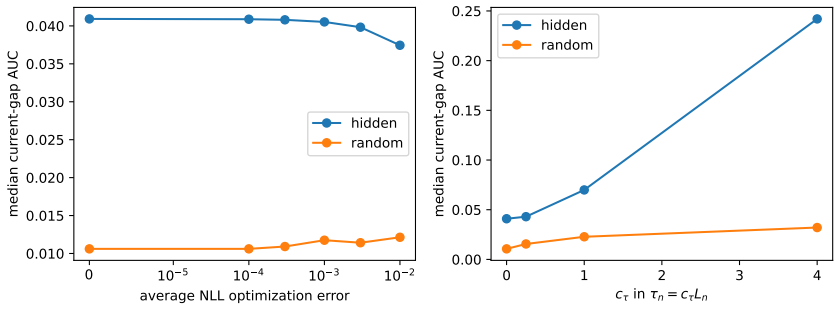


Figure 11: **Approximate-MLE tolerance (30 paired seeds).** Median current-gap AUC changes only modestly under the tested small average NLL errors (left), but increases with logarithmic likelihood slack c_τ (right), especially on hidden-branch games.

Exhaustive policy-uniform audit. To strengthen the policy-uniform audit, we use $|\mathcal{S}| = 4, d = 2, J = 4, |\mathcal{M}| = 8, \gamma = 0.8$, enumerate all 256 deterministic stationary joint policies, and add 1,004 random or algorithm-dependent policies plus the two adversarially optimized policies described above. Figure 12 visualizes the resulting maximum ratios together with the shifted-tuple learning curves. Across 300 seeds, seven checkpoints, and 1,262 policies per checkpoint, this gives 2,650,200 policy–checkpoint evaluations. Of the 300 seeds, 299 have simultaneous coverage. The sole failing environment is the previously identified seed 80089 at iteration 120: the lower envelope is violated by 0.00643 and the maximum error-to-radius ratio is 1.049. Adversarial optimization recovers the same

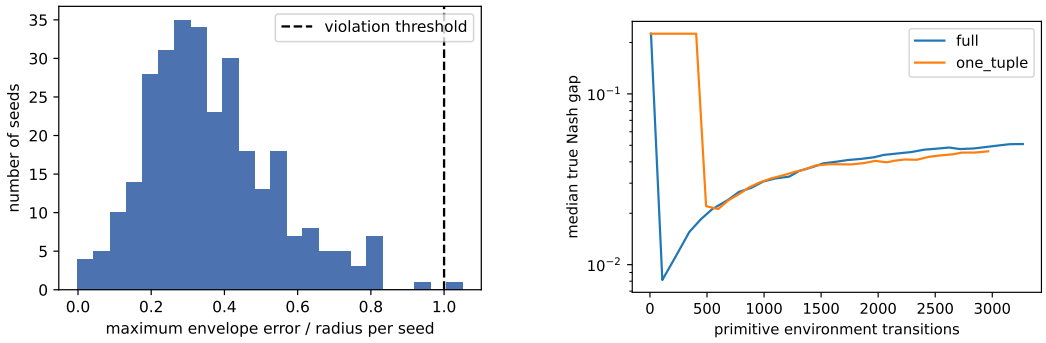


Figure 12: **Policy-uniform and shifted-tuple stress tests.** Left: maximum error-to-radius ratios for the 300-seed exhaustive and adversarial micro-game audit; the dashed line marks a violation. Right: the second tuple improves early learning on a branch where disagreement is shifted one transition beyond the first context.

lower-side failure but finds no new failing environment; there are no upper-envelope violations and every optimization call reports success. Together with the ground-truth policy-evaluation residual below 10^{-10}, these diagnostics classify the small, early lower-side breach as a finite-sample coverage failure rather than a numerical solver failure. The empirical seed-level coverage is 99.67%, with a two-sided 95% Clopper–Pearson lower bound of 98.16%. This strengthened finite stress test complements the theorem's policy-uniform event.

Why the shifted tuple matters. In the two-step branch, candidate models agree on the root-to-intermediate transition and disagree only at the intermediate state. The full method has median gap-AUC 0.0411, compared with 0.0603 after deleting the second tuple. It selects the true model permanently after a median 108 rather than 438 transitions, and reaches a sustained gap at most 0.05 after 254 rather than 582 transitions. Removing the tuple also reduces simultaneous audited-envelope coverage from 100% to 80%, although stored certificate coverage remains 100% in both variants. Paired bootstrap differences (one tuple minus full) are $0.01949\,[0.01133, 0.03332]$ for gap-AUC and $304\,[180, 542]$ transitions for permanent model identification. The sustained-gap comparison uses only 11 pairs finite in both arms and has interval $230\,[-17, 1079]$; the final-gap difference is $-0.00464\,[-0.00863, 0.00037]$. Thus the evidence supports a finite-sample AUC and identification benefit, while neither the sustained threshold time nor final gap is statistically resolved by this suite.

Near-margin planning, separation, and roll-in accounting. A two-state rank-two near-margin construction varies seven action margins and eight requested Shapley tolerances. Fourteen of 56 configurations switch the root action and incur positive gap. The largest observed ratio of Nash gap to achieved oracle residual is 0.064, so every gap lies well below the displayed 2η reference, supporting the one-call additive planning-error control. Separately, an exact grid evaluation of the rank-two hard pair recovers one-sided minimax gaps $\gamma/2 \in \{0.25, 0.45, 0.495\}$ for $\gamma \in \{0.5, 0.9, 0.99\}$. In the distinct one-state example of Theorem 14, the conservative crossing has zero gap, while the upper-game and reverse-crossed candidates have gap one. The latter comparison concerns the candidate, not its use for exploration.

Finally, the observed median primitive transitions per outer iteration are 2.995, 5.043, 11.005, and 20.648 for $\gamma = 0.5, 0.75, 0.9, 0.95$. These values are within 1.7% of $2 + \gamma/(1-\gamma) = 3, 5, 11, 21$, respectively, checking the transition accounting behind the additional geometric-roll-in factor in Corollary 10. Figure 13 compares both accounting diagnostics.

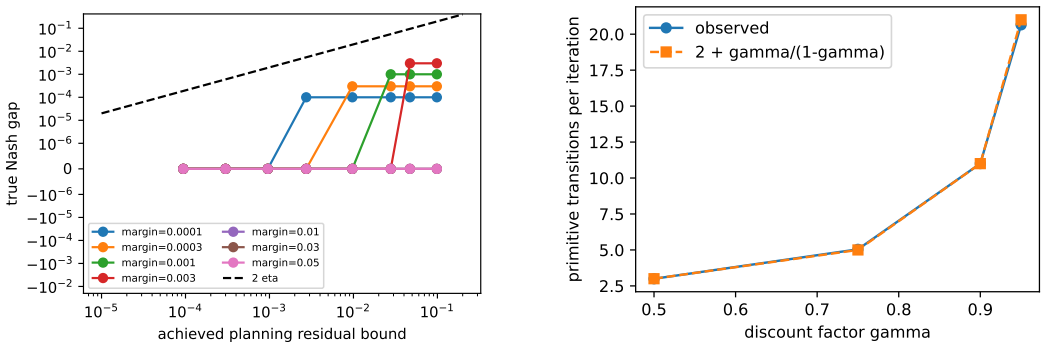


Figure 13: **Planning and transition-accounting diagnostics.** Left: the true gap remains below the displayed 2η reference on near-margin instances. Right: observed primitive transitions per outer iteration track $2 + \gamma/(1-\gamma)$.

H.10 Transition-class misspecification stress test

The certificate analysis assumes exact transition-class realizability. We use a separate controlled study to distinguish how policy performance changes when $P^\star \notin \mathcal{M}$ from when the realizability-based certificate and policy-uniform envelope first fail empirically. We keep $\mathcal{M}$, all algorithmic constants, and every training and evaluation rule fixed. Starting from the realizable model $P_0^\star = P_{\mu^\star,\phi^\star}$, we form either

$$\mu_\eta = (1-\eta)\mu^\star + \eta\widetilde{\mu}, \qquad P_\eta^\star = P_{\mu_\eta,\phi^\star}, \tag{114}$$

$$\phi_\eta = (1-\eta)\phi^\star + \eta\widetilde{\phi}, \qquad P_\eta^\star = P_{\mu^\star,\phi_\eta}, \tag{115}$$

where the alternatives are independently sampled simplex-valued factors. Both constructions preserve a valid rank-d kernel and isolate finite-class misspecification. Because the same η induces different errors across instances and modes, we report the measured projection

$$\varepsilon_{\mathrm{mis}} = \min_{P\in\mathcal{M}} \max_{s,z} \mathrm{TV}\big(P_\eta^\star(\cdot \mid s,z), P(\cdot \mid s,z)\big).$$

For each environment, M1 runs DREP-ZS with the original misspecified class. M2 replaces the original true candidate by $P_\eta^\star$, keeping the class size and all decoys fixed, and is therefore a matched-realizable control. M3 performs certainty-equivalent MLE planning in the original misspecified class without bonuses or a certificate. We selected the displayed levels on a discovery set, froze the protocol, and evaluated disjoint confirmation seeds; all results below use only the confirmation stage.

Held-out micro-game confirmation. We use $|\mathcal{S}| = 4, d = 2, A = B = 2, |\mathcal{M}| = 8, \gamma = 0.8$, $\eta \in \{0, 0.1, 0.2, 0.5\}$, and 100 new seeds per mode–level–condition cell. M1 and M2 run for 10,000 iterations and are audited at $n \in \{120, 300, 1000, 3000, 10000\}$. Each checkpoint enumerates all 256 deterministic stationary joint policies, adds 1,000 random and four algorithm-dependent policies, and performs the two signed adversarial searches. Coverage in Table 19 treats the seed as the statistical unit: one failing checkpoint makes the entire seed fail.

Table 19: **Held-out transition-class micro audit.** Coverage entries are successes among 100 new seeds with two-sided 95% Clopper–Pearson intervals. ΔAUC is the median of 100 paired seedwise M1-minus-M2 current-gap AUC differences, with a 95% bootstrap interval. "Adaptive" is the final rate at which M1 selects the candidate minimizing KL under its realized context distribution.

Mode	η	Median ε_{mis}	Stored certificate	Envelope	Adaptive	ΔAUC
Emission	0	0	100 [.964, 1]	100 [.964, 1]	100%	0 [0, 0]
Emission	.1	.0557	100 [.964, 1]	39 [.294, .493]	99%	.00147 [.00086, .00238]
Emission	.2	.1082	100 [.964, 1]	5 [.016, .113]	97%	.00526 [.00376, .00738]
Emission	.5	.2529	90 [.824, .951]	0 [0, .036]	92%	.00930 [.00687, .01666]
Feature	0	0	100 [.964, 1]	100 [.964, 1]	100%	0 [0, 0]
Feature	.1	.0740	100 [.964, 1]	40 [.303, .503]	98%	.00259 [.00144, .00385]
Feature	.2	.1460	98 [.930, .998]	5 [.016, .113]	95%	.00822 [.00634, .01179]
Feature	.5	.3357	80 [.708, .873]	0 [0, .036]	96%	.02195 [.01651, .02971]

The held-out results reproduce the discovery-stage ordering. At the mild level, every stored Nash certificate still covers, yet simultaneous envelope coverage is only 39%–40%. At the two strongest measured errors, envelope coverage falls to at most 5%; stored-certificate failures appear later, reaching 10/100 emission seeds and 20/100 feature seeds at $\eta = 0.5$. The paired AUC intervals are already positive at $\eta = 0.1$ but increase gradually with measured error. Thus the policy-uniform event is more sensitive than the observed policy performance or stored certificate. Across the nonzero cells, the final MLE selects the candidate minimizing KL under its realized context distribution in 92%–99% of seeds.

Every matched-realizable run retains stored-certificate coverage. These controls attain 100/100 envelope coverage in seven cells and 99/100 in the remaining feature-shift cell. In that seed, the lower envelope is violated by 0.01584 at $n = 120$; all four later checkpoints cover.

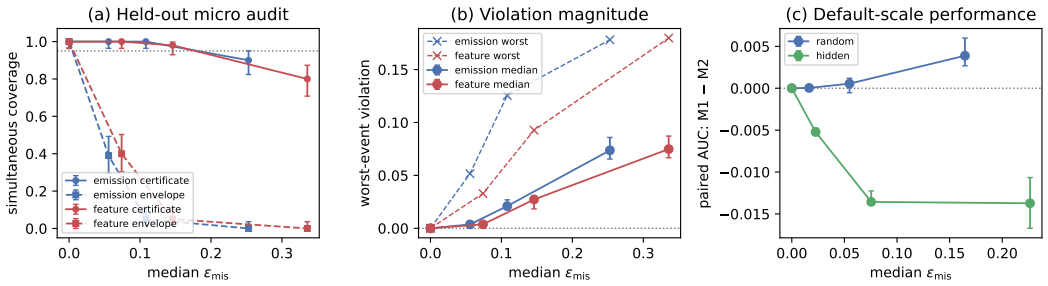


Figure 14: **Frozen held-out misspecification confirmation.** Left: simultaneous M1 stored-certificate and policy-envelope coverage; bars are Clopper–Pearson intervals and the dotted line marks 95%. Middle: for each seed, we maximize current/best certificate and signed envelope violations over checkpoints; solid curves show the median with bootstrap intervals and dashed curves the worst seed. Right: paired M1-minus-M2 gap-AUC differences on the default-scale families; positive values mean misspecification hurts M1 relative to its matched control. The matched controls, omitted for legibility, have 100% stored-certificate coverage and at least 99% envelope coverage in every micro cell.

Default-scale policy performance. The performance suite applies emission-factor shifts to the random 3×3 and strategic hidden 2×2 families, with $\eta \in \{0, 0.03, 0.1, 0.3\}$, 3,000 iterations, and 30 new paired seeds per family–level–method cell. At $\eta = 0$, the paired M1–M2 AUC difference is numerically zero for both families. At nonzero levels, Table 20 shows distinct paired effects across the two families.

Table 20: **Held-out default-scale misspecification.** AUC entries are medians over 30 paired seeds. The last column is the median of the 30 seedwise M1-minus-M2 AUC differences, with a 95% bootstrap interval. M3 is certainty-equivalent and has no strategic confidence certificate.

Family	η	Median $\varepsilon_{\mathrm{mis}}$	M1	M2	M3	M1–M2
Random	.03	.0165	.0263	.0252	.00174	$.00004\ [-.00019, .00024]$
Random	.10	.0548	.0261	.0239	.00462	$.00055\ [-.00054, .00119]$
Random	.30	.1645	.0312	.0236	.01436	$.00388\ [.00267, .00599]$
Hidden	.03	.0226	.0495	.0545	.01225	$-.00522\ [-.00539, -.00476]$
Hidden	.10	.0755	.0385	.0508	.02241	$-.01356\ [-.01383, -.01225]$
Hidden	.30	.2264	.0339	.0476	.02821	$-.01373\ [-.01672, -.01065]$

On random games, only the largest level resolves a performance loss relative to M2. On hidden-branch games, all nonzero paired intervals are negative, so M1 attains lower AUC than its matched control on this designed family. M3 has lower AUC than M1 in every tested cell but supplies no certificate. The micro audit therefore shows that the policy-envelope event fails before stored-certificate coverage, while the default-scale study reveals gradual, family-dependent policy effects.

H.11 Controlled general-sum CCE study

We finally test DRep-CCE on team-potential hidden-branch games with $m \in \{2, 3, 4\}$. For each finite model we enumerate a small cache of stationary correlation devices and verify its oracle residual by exact policy evaluation and one discounted best-response MDP per player. We compare all-player optimistic-deviation sampling with sampling only the candidate, the currently worst deviation, or a uniform policy. This study audits the statistical reduction with verified oracle residuals.

With a numerically exact oracle, all-player sampling has gap AUC 0.00633, 0.00676, and 0.00660 for two, three, and four players, respectively. At four players, this improves over candidate-only sampling (0.02265) and uniform sampling (0.02916). All 48 combinations of player count, tolerance, and sampling mode have 100% simultaneous empirical certificate coverage.

Table 21: **Exact-oracle CCE gap AUC.** Medians over 10 hidden-branch seeds. "All" samples the candidate and every optimistic unilateral deviation.

Players	All	Worst	Candidate	Uniform
2	0.00633	0.00653	0.00683	0.00683
3	0.00676	0.00738	0.00859	0.00714
4	0.00660	0.00740	0.02265	0.02916

Figure 15 gives the complementary oracle tolerance result. For two players, requested tolerances $10^{-8}, 10^{-2}, 5 \times 10^{-2}, 10^{-1}$ produce median final gaps $0, 0.0060, 0.0498, 0.0964$, matching the verified residuals after the statistical error vanishes. For four players, the achieved oracle residual saturates at 0.0410 for the two largest requested tolerances, and the true gap saturates with it. This clean floor supports the statistical-plus-oracle decomposition in Theorem 38.

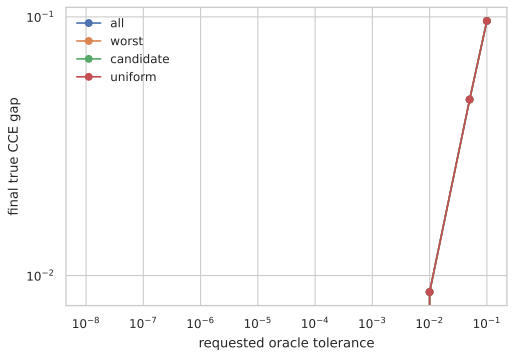


Figure 15: **Supplementary stationary-CCE oracle diagnostic.** The final true CCE gap tracks the verified oracle residual after finite-model identification.

Reproducibility. The artifact archives per-task configurations, raw traces, summaries, and plots; interaction counts include roll-ins and both MLE tuples, while runtimes and oracle residuals are logged separately.