CALIBRATING POLICY COVERS FOR NASH EQUILIBRIUM LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

A policy cover can reach every important state while its uniform mixture provides poor data for equilibrium learning. We study how to calibrate a finite cover in zero-sum Block games without identifying the latent-state decoder. Under common emissions and a finite evaluable decoder class containing the true decoder, constraints on all candidate-decoder events are equivalent at the population level to latent coverage constraints. Incorrect candidates therefore do not require a more conservative population mixture. Every library admits a coverage factor $R \leq \min\{M, S\}$, where M is its size and S the latent state count. We estimate the event constraints from shared rollouts and implement them by linear programming. Under finite-class fitting conditions, the resulting Nash guarantee replaces M by R in the sufficient fitting budget, while charging cover construction and calibration. A hidden-coordinate class gives a polynomial-time implementation. In a three-stage family with an unknown entrance, $\widetilde{O}(M + \varepsilon^{-2})$ episodes match an adaptive lower bound up to logarithms, whereas uniform sampling of the same library requires $\Omega(M\varepsilon^{-2})$. Fixed-budget experiments distinguish distribution quality from the costs and best-response errors that determine downstream performance.

1 INTRODUCTION

A policy cover can contain a good way to reach every important state and still induce a poor training distribution. Its guarantee is a maximum over policies: for each state, some member visits it sufficiently often. A learner instead needs one distribution from which to collect training data. Uniformly mixing a large cover can assign little probability to its useful members. In a game, an underrepresented state may support a profitable unilateral deviation from the returned strategy pair.

We study how to convert a supplied policy cover into a useful sampling distribution without identifying the latent-state decoder. Policy-cover constructions address reachability under unknown representations (Misra et al., 2020; Rohatgi & Foster, 2025), and coverage optimization is well developed in single-agent learning (Amortila et al., 2024). The conversion considered here faces two obstacles: the occupancies to be optimized are unobserved, and collecting data to learn their mixture weights can consume the savings in downstream fitting.

The first obstacle has a structural resolution. Constraining every candidate decoder appears conservative: incorrect decoders can partition observations differently. Under common Block emissions, however, each candidate event is a nonnegative linear combination of the same latent occupancies, with coefficients independent of the roll-in policy. Since the true decoder is included, the observable and latent constraints define *the same population feasible set*. Additional candidates affect estimation and computation, but do not worsen the optimal population coverage factor. Calibration can therefore proceed without decoder identification.

The second obstacle is statistical. Uniform absolute-error estimates would require a quadratic dependence on inverse calibration precision. We instead exploit the smaller variance of rare events through a relative-plus-additive concentration bound. The resulting calibration term scales as ε^{-1}, while the sufficient fitting term scales as ε^{-2}. A Bellman-error certificate along true dynamic best responses connects the sampling law to Nash learning, including cover construction in the total cost.

We establish three results:

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.

- **Observable calibration without decoder identification.** For any finite library in an exact Block game, candidate-event constraints preserve latent coverage and admit a simultaneous coverage factor $R \le \min\{M, S\}$. Finite-sample calibration and checked numerical tolerances make this structure usable from observations. A linear program implements the constraints; any solver meeting the tolerances inherits their guarantee.
- **Nash guarantees with complete sample accounting.** Calibration replaces the uniform-library dependence on M by R in the sufficient fitting budget. The general Block result charges cover construction, calibration, and fitting. A hidden-coordinate class makes all operations polynomial despite an exponentially large observation space. Real-successor fitting gives the direct construction; model-based evaluation shares the coverage benefit.
- **Sharp adaptation to an unknown entrance.** A three-stage family admits $\widetilde{O}(M+\varepsilon^{-2})$ total episodes, matching an adaptive sample lower bound up to logarithms. Independent uniform sampling of the same library requires $\Omega(M\varepsilon^{-2})$, under arbitrary downstream learning.

The experiments test both the benefit of adapting the sampling law and the cost of obtaining it. Matched fitting counts, complete budgets, and true best-response residuals distinguish distribution quality from end-to-end performance. Shared-reward tasks allow every comparator to reuse its data and expose the limits of the observed benefit.

2 Setting: Block games and policy covers

Game and learning objective. We study an episodic zero-sum game $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{B}, H, P^\star, r, d_1)$ with a measurable observation space, finite action sets, known rewards $r_h(s, a, b) \in [0, 1]$, and unknown transitions $P_h^\star$. The maximizer receives r_h and the minimizer $-r_h$. Write $A = |\mathcal{A}|$, $B = |\mathcal{B}|$, $J = AB$, and $H_h = H - h + 1$. A Markov product pair (π, ν) has value

$$V_{P,h}^{\pi,\nu}(s) = \mathbb{E}_{P,\pi,\nu}\left[\sum_{t=h}^{H} r_t(S_t, A_t, B_t) \,\middle|\, S_h = s\right], \qquad V_{P,H+1}^{\pi,\nu} = 0.$$

Let $d_{P,h}^{\pi,\nu}$ be its observation occupancy and $V_P^{\pi,\nu} = \mathbb{E}_{d_1} V_{P,1}^{\pi,\nu}$. The target is an ε-Nash product pair, defined by

$$\mathrm{Gap}_P(\pi, \nu) = \sup_{\pi'} V_P^{\pi',\nu} - \inf_{\nu'} V_P^{\pi,\nu'} \le \varepsilon.$$

During training a centralized policy $\omega_h(a, b \mid s)$ may select joint actions; its state–joint-action law is $d_{P,h}^{\omega}\omega_h$. The returned policies must factor as $\pi_h(a \mid s)\nu_h(b \mid s)$.

Common emissions and unknown decoding. The game has S latent states and a decoder $\phi^\star : \mathcal{S} \to [S]$ in a known finite, evaluable class Φ, with $F = |\Phi|$. Its kernel is

$$P_h^\star(\mathrm{d}s' \mid s, a, b) = \sum_{z'} T_h(z' \mid \phi^\star(s), a, b) O_{h+1}(\mathrm{d}s' \mid z').$$

Emission supports are disjoint and decoded by $\phi^\star$. Crucially, $O_h(\cdot \mid z)$ depends only on the latent state and stage, not on the roll-in policy. The initial law is also a mixture of these emissions. The learner observes S_h, can evaluate every candidate decoder, and receives neither latent labels nor the identity of $\phi^\star$. The observation space may be exponentially larger than the latent space.

Supplied covers and charged access. At each stage, a library $\Psi_h = \{\psi_{h,i}\}_{i=1}^{M_h}$, $M_h \le M$, is a $(1, \eta)$ cover if

$$\max_i \Pr_{P^\star, \psi_{h,i}}(Z_h = z) \ge v_h(z) - \eta, \qquad v_h(z) = \sup_{\omega} \Pr_{P^\star, \omega}(Z_h = z).$$

The learner can execute library prefixes and then take a uniform joint action. Cover construction costs N_{cover} episodes and precedes independent calibration data. Centralizing the joint action gives a Block MDP, so the two-context PCE reduction of Rohatgi & Foster (2025) supplies such covers with $M \le H^2 S^2$ under its realizable regression interface. We inherit and charge that provider; other valid providers can be substituted. Appendix A specifies its oracle and probability conditions.

3 OBSERVABLE CALIBRATION OF POLICY COVERS

3.1 CANDIDATE EVENTS PRESERVE THE POPULATION FEASIBLE SET

The ideal constraint would make the mixture occupancy dominate every library member's latent occupancy. Those quantities are unobserved. For each candidate decoder, estimate the observable events

$$D_h(\phi, y, i) = \Pr_{P^\star, \psi_{h,i}}\{\phi(S_h) = y\}, \qquad c_h(\phi, y) = \max_i D_h(\phi, y, i).$$

Write $p_{h,i}(z) = \Pr_{P^\star,\psi_{h,i}}(Z_h = z)$ and $q_h^w(z) = \sum_i w_i p_{h,i}(z)$. The following equivalence explains why constraining incorrect candidates does not require extra population coverage.

Lemma 1 (Observable constraints preserve latent coverage)**.** *Under the common-emission and decoder-inclusion conditions of Section 2, for every fixed stage, $w \in \Delta_{M_h}$, and $t \geq 0$, the following conditions are equivalent:*

$$\sum_i w_i D_h(\phi, y, i) \geq t c_h(\phi, y) \quad (\phi \in \Phi, y \in [S]), \qquad q_h^w(z) \geq t \max_i p_{h,i}(z) \quad (z \in [S]).$$

Every finite library admits such weights with $t \geq 1/\min\{M_h, S\}$; this worst-case bound is tight.

Indeed, common emissions give $D_h(\phi, y, i) = \sum_z a_{\phi,y}(z) p_{h,i}(z)$, where $a_{\phi,y}(z) = \Pr\{\phi(S_h) = y \mid Z_h = z\} \geq 0$ is independent of i. Latent domination therefore implies every candidate-event constraint; the included true decoder gives the converse. Mixing one maximizing member per latent state gives the factor S, while uniform weights give M_h. Lemma 8 proves tightness. No decoder identification or latent occupancy query is needed.

We call a known $R \geq 1$ a *simultaneous coverage factor* if, at every stage, some $w_h^\circ \in \Delta_{M_h}$ satisfies

$$\sum_i w_{h,i}^\circ D_h(\phi, y, i) \geq c_h(\phi, y)/R \quad \text{for all } \phi, y. \tag{1}$$

The lemma supplies $R = \min\{M, S\}$ for any Block-game library, even before requiring it to cover policies outside itself. Additional decoders leave this population feasible set unchanged, but their events still increase estimation and computation costs. In particular, R measures coverage quality, not the number of nonzero mixture weights.

3.2 FINITE-SAMPLE CALIBRATION

Learning weights must cost less than the fitting samples they save. Estimating every occupancy to uniform additive accuracy $O(\eta)$ by a variance-blind bound would cost $O(\eta^{-2})$ samples per member. Bernstein concentration instead uses the small variance of rare events. Collect n_{cal} independent roll-in episodes per member and use the same observations for every decoder. Let $\widehat{D}_h$ be the event frequencies and $\widehat{c}_h(\phi, y) = \max_i \widehat{D}_h(\phi, y, i)$. With $\beta = \eta/(64R)$, maximize $t \in [0, 1]$ over $w \in \Delta_{M_h}$ subject to

$$\sum_i w_i \widehat{D}_h(\phi, y, i) \geq t\widehat{c}_h(\phi, y) - 2\beta \qquad (\phi \in \Phi, y \in [S]). \tag{2}$$

The LP uses only observed events. Executing its mixture and then a uniform joint action gives the fitting law ρ_h^{w}.

Theorem 2 (Learned-mixture coverage)**.** *Condition on $(1, \eta)$ covers satisfying* (1)*. Let*

$$L_{\mathrm{cal}} = \log \frac{2HMFS}{\delta_{\mathrm{w}}}, \qquad n_{\mathrm{cal}} = \left\lceil \frac{320RL_{\mathrm{cal}}}{\eta} \right\rceil.$$

With probability at least $1 - \delta_{\mathrm{w}}$, the calibrated mixtures satisfy, simultaneously for every centralized policy ω and measurable $g \in [0, 1]$,

$$\mathbb{E}_{d^\omega_{P^\star,h}\omega_h} g \leq 2S\eta + \sqrt{5RJ\,\mathbb{E}_{\rho_h^{\mathrm{w}}} g^2}. \tag{3}$$

Calibration costs at most HMn_{cal} episodes. Its additional computation is polynomial in $H, M, F, S, n_{\mathrm{cal}}$ and input precision, plus the cost of policy execution and decoder queries.

From empirical feasibility to true coverage. The simultaneous event $|\widehat{D} - D| \le D/8 + \beta$ makes the LP feasible at $t = 7/(9R)$. It suffices numerically to return weights in the simplex and $7/(10R) \le \widehat{t}_h \le 1$ satisfying (2) with 3β instead of 2β. On the concentration event, every such solution covers states with $v_h(z) \ge 2\eta$ at least $1/(5R)$ as often as any policy. Common emissions and a uniform joint action then yield (3). This guarantee applies to any optimizer meeting the numerical conditions; the prescribed sample count supplies the statistical event. Theorem 9 gives the proof.

For comparison, uniform library mixing replaces $5RJ$ by $2MJ$ (Lemma 36). The remainder $2S\eta$ bounds the probability of states with $v_h(z) < 2\eta$; it can be replaced by their largest total mass under any policy. If every reachable state at a fitted stage has $v_h(z) \ge 2\eta$, this remainder is zero (Corollary 10). This sharper form permits constant-precision calibration in the entrance family.

4 From calibrated data to Nash guarantees

The calibration theorem controls errors under every policy, including the output-dependent best responses defining exploitability. We first use actual successors for fitting, then extend the same argument to model-generated labels. This separates the benefit of sampling weights from the choice of prediction method.

4.1 A best-response Bellman certificate

In kernel P, let $\widehat{Q}_h \in [0, H_h]$ be predictors, let $(\widehat{\pi}_h, \widehat{\nu}_h)$ be their approximate matrix equilibria, and set $\widehat{V}_h = \langle \widehat{\pi}_h, \widehat{Q}_h \widehat{\nu}_h \rangle$. Write $e_h = |\widehat{Q}_h - r_h - P_h \widehat{V}_{h+1}|$. The local residuals obey

$$\max_p \langle p, \widehat{Q}_h \widehat{\nu}_h \rangle - \widehat{V}_h \le \delta_h^+, \qquad \widehat{V}_h - \min_v \langle \widehat{\pi}_h, \widehat{Q}_h v \rangle \le \delta_h^-.$$

Let π^+ and ν^- be true dynamic best responses to the returned opponent policies in this kernel.

Theorem 3 (Nash gap from best-response Bellman errors)**.** *Fix any valid kernel P, bounded returned predictors $\widehat{Q}_h$, and product policies with $\widehat{V}_h = \langle \widehat{\pi}_h, \widehat{Q}_h \widehat{\nu}_h \rangle$ and $\widehat{V}_{H+1} = 0$. With the residuals defined above, these functions and policies satisfy*

$$\begin{aligned} \mathrm{Gap}_P(\widehat{\pi}, \widehat{\nu}) \le & \sum_{h=1}^{H} \mathbb{E}_{d_{P,h}^{\pi^+, \widehat{\nu}} \pi_h^+ \widehat{\nu}_h} [e_h + \delta_h^+(S_h)] \\ & + \sum_{h=1}^{H} \mathbb{E}_{d_{P,h}^{\widehat{\pi}, \nu^-} \widehat{\pi}_h \nu_h^-} [e_h + \delta_h^-(S_h)]. \end{aligned} \tag{4}$$

The inequality follows the realized best-response occupancies directly, with no fixed reference distribution or occupancy-stability term.

The proof iterates the Bellman recursions for the two differences between best-response values and the returned continuation. Those differences need not be nonnegative; their one-sided recursions add to the Nash gap. Appendix I.1 gives the complete argument.

Real-successor fitted minimax. At stage h, regress labels $r_h(X) + \widehat{V}_{h+1}(S_{h+1})$, where $X = (S_h, A_h, B_h) \sim \rho_h^{\mathrm{w}}$ and the successor is observed from $P_h^\star$. Clip predictors to $[0, H_h]$ and solve their local matrix games. Independent stagewise buffers make the continuation, which uses later-stage data, independent of its held-out fitting buffer. With $P = P^\star$, $u_h^2 = \mathbb{E}_{\rho_h^{\mathrm{w}}} e_h^2$, $\kappa_{\mathrm{w}} = 5RJ$, and both local residuals bounded by δ_h, the calibration theorem gives

$$\mathrm{Gap}_{P^\star}(\widehat{\pi}, \widehat{\nu}) \le 4SH^2\eta + 2\sqrt{\kappa_{\mathrm{w}}} \sum_h u_h + 2 \sum_h \delta_h. \tag{5}$$

Thus low fitting error matters through coverage of both actual response laws, not merely through accuracy along the returned strategy pair.

Algorithm 1: Calibrated real-successor Nash learning

Input: cover provider; decoder class Φ; accuracy ε; confidence δ; coverage factor R.

1: Set $\eta = \varepsilon/(32SH^2)$ and $\beta = \eta/(64R)$
2: Construct $(1, \eta)$ covers at failure level $\delta/4$
3: Collect calibration episodes using Theorem 2 with $\delta_{\mathrm{w}} = \delta/4$; solve (2)
4: Collect fresh, independent stagewise (S_h, A_h, B_h, S_{h+1}) buffers using calibrated roll-ins and uniform joint actions, with the real-successor sample counts of Corollary 11
5: Set $\widehat{V}_{H+1} = 0$
6: Working backward, fit $r_h + \widehat{V}_{h+1}(S_{h+1})$, clip to $[0, H_h]$, solve the local matrices, and update $\widehat{V}_h$
7: **return** *the product pair* $(\widehat{\pi}, \widehat{\nu})$.

4.2 Complete sample accounting

Assume realizable finite Bellman classes of size at most $Q_\star$, or the controlled approximation in Proposition 17. Conditional squared-loss ERM on fresh bounded labels supplies $u_h^2 = O(H^2 L_Q/n_{\mathrm{pe}})$, with $L_Q = \log(8HQ_\star/\delta)$; the approximation variant adds its quantization term. These are the fitting conditions of Corollary 18, not consequences of coverage alone. Set $\eta = \varepsilon/(32SH^2)$, $u_h \le \varepsilon/(8H\sqrt{\kappa_{\mathrm{w}}})$, and $\delta_h \le \varepsilon/(8H)$. With the confidence allocation and explicit sample counts in Corollary 11, (5) gives an ε-Nash pair with probability at least $1-\delta$ using

$$N_{\mathrm{env}} = \underbrace{N_{\mathrm{cover}}(\eta, \delta/4)}_{\text{cover construction}} + O\left(\underbrace{HM + \frac{H^3 MRSL_{\mathrm{cal}}}{\varepsilon}}_{\text{calibration}} + \underbrace{\frac{H^5 RJL_Q}{\varepsilon^2}}_{\text{Bellman fitting}}\right). \tag{6}$$

The construction cost is unchanged. Calibration adds an overhead linear in the library size for fixed structural parameters, while fitting depends on R rather than M. Each episode uses at most H transitions; decoder evaluation and local matrix computation are also charged in the appendix. Algorithm 1 states the real-successor construction.

Corollary 4 (General Block-game conversion)**.** *Under the preceding fitting conditions, any supplied $(1, \eta)$ Block-game cover with $M \le H^2S^2$ yields an ε-Nash pair with probability at least $1-\delta$ using*

$$N_{\mathrm{cover}} + O\left(\frac{H^5 S^4 L_{\mathrm{cal}}}{\varepsilon} + \frac{H^5 SJL_Q}{\varepsilon^2}\right), \qquad \eta = \frac{\varepsilon}{32SH^2}.$$

No additional coverage-factor assumption is required.

This uses $R = S$ in (6). Uniform mixing has a sufficient fitting term $O(H^5 MJL_Q/\varepsilon^2)$, hence $O(H^7 S^2 JL_Q/\varepsilon^2)$ for the same worst-case PCE library. The provider cost remains in both totals.

Model-based evaluation uses the same coverage. Alternatively, fit proper kernels $\widehat{P}_h$ by MLE on independent buffers and use pseudo-successors $\widetilde{S}_{h+1} \sim \widehat{P}_h$. For this route write $e_h = |\widehat{Q}_h - r_h - \widehat{P}_h \widehat{V}_{h+1}|$ and $f_h = D_{\mathrm{TV}}(P_h^\star, \widehat{P}_h)$, with $D_{\mathrm{TV}}(P, Q) = \frac{1}{2}\int |\mathrm{d}P - \mathrm{d}Q|$. The true Bellman residual obeys $e_h^\star \le e_h + H_{h+1} f_h$. Applying Theorem 3 in the true kernel gives

$$\mathrm{Gap}_{P^\star}(\widehat{\pi}, \widehat{\nu}) \le \sum_{h=1}^{H} \sum_{\sigma \in \{+,-\}} \mathbb{E}_{\mu_{h,\star}^{\sigma}}[e_h + H_{h+1} f_h + \delta_h^{\sigma}(S_h)], \tag{7}$$

where $\mu_{h,\star}^{\pm}$ are the true dynamic-best-response laws and $H_{H+1} = 0$. Under finite proper-model realizability and the model-route Bellman fitting conditions, both sample bounds above replace L_Q by $L_Q + L_\Theta$, where L_Θ includes the model-class logarithm and MLE optimization tolerance. The full schedule is in Corollary 11; learned-model occupancy transfer is unnecessary. Alternative transfer certificates appear in Appendix J.

When the fitting saving covers calibration. Let $N_{\mathrm{fit}}(\kappa)$ be the sufficient fitting count with common class resolution and confidence allocation: $Hn_{\mathrm{pe}}(\kappa)$ for direct fitting, or $H(n_{\mathrm{mle}}(\kappa) + n_{\mathrm{pe}}(\kappa))$ for model-based evaluation. The calibrated schedule is cheaper whenever

$$n_{\mathrm{cal}} \sum_h M_h < N_{\mathrm{fit}}(2MJ) - N_{\mathrm{fit}}(5RJ). \tag{8}$$

For fixed $M > 5R/2$, this holds at sufficiently small ε: calibration scales as ε^{-1} and the fitting saving as ε^{-2}. The learner can compare these sufficient budgets before collecting calibration data. This comparison retains the same provider cost and its leading exponent when PCE dominates.

5 Sharp adaptation and a computable instance

5.1 An unknown entrance with matching sample bounds

The preceding result compares sufficient schedules. The following family also separates the necessary episode counts of adaptive and fixed uniform sampling. It has an initially uninformative observation and an unknown root action that enters the useful stage-two state. Terminal outcomes reveal which entrance works, while learning its game still requires data.

Proposition 5 (Unknown entrance)**.** *For $M \geq 4$, $0 < \varepsilon \leq 1/80$, and $\delta \in (0, 1/4)$, a three-stage Block-game family with $S = 2$ has M public root actions and an unknown useful entrance. Its action library is an exact cover with $R \leq 2$. Calibration and real-successor fitted minimax return an ε-Nash pair with probability at least $1 - \delta$ using*

$$\widetilde{O}(M + \varepsilon^{-2}) \quad \textit{versus} \quad \Omega(M\varepsilon^{-2})$$

episodes for independent uniform library sampling at success probability $3/4$. Adaptive exploration needs $\Omega(M + \varepsilon^{-2})$ episodes in the worst case on this family at success probability $3/4$. Any data-independent random mixture has worst-case expected useful-state visitation at most $1/M$.

Both intermediate states are deterministically reachable, so Corollary 10 permits constant-precision calibration, costing $\widetilde{O}(M)$ episodes. The same trajectories' public terminal outcomes identify the entrance; fresh observations fit its unknown four-by-two game in $\widetilde{O}(\varepsilon^{-2})$ episodes. The M root actions are counted: the construction exploits deterministic entry rather than treating the general action factor J as constant.

The adaptive lower bound combines $\Omega(M)$ search with $\Omega(\varepsilon^{-2})$ estimation. Uniform sampling reaches the informative state only once per M episodes on average, giving the product lower bound even with arbitrary estimation, planning, and reuse. Proposition 16 proves these claims and counts polynomial LP computation. Return-directed search can attain the same sample order; the result establishes the value of adaptation, not a unique advantage of the LP. When the entrance is public, task-directed sampling needs only $\widetilde{O}(\varepsilon^{-2})$ episodes; the uniform-library separation appears in Proposition 14.

5.2 Hidden coordinates make the operations explicit

Consider $x \in \{0, 1\}^p$ with unknown decoder $\phi_i(x) = x_i$. Conditional on the binary latent state, its coordinate is fixed and the other bits are independent and uniform. A library always has $R \leq 2$, while evaluating its $2p$ candidate events requires no identification of the true coordinate. Two-context regression enumerates p coordinates and fits a four-entry head, so the PCE oracle is implementable.

For any bounded continuation, the true Bellman target has the form $r_h(x, a, b) + g(x_i, a, b)$. Keep the known reward separate and quantize g at resolution τ. Squared-loss ERM reduces to quantized cell means, with

$$\log Q_\star \leq \log p + 2J \log(2 + H/\tau), \qquad u_h^2 \leq 3\tau^2 + O\left(\frac{H^2 \log(Q_\star/\delta)}{n}\right).$$

Taking $\tau \leq \varepsilon/(8\sqrt{6}H\sqrt{\kappa_{\mathrm{w}}})$ leaves half the squared-error budget for estimation. Coordinate fitting costs $O(p(n + J))$ per buffer; a queried policy uses a stored head and a finite matrix solver. Proposition 17 accounts for the recursive query costs without enumerating the 2^p observations.

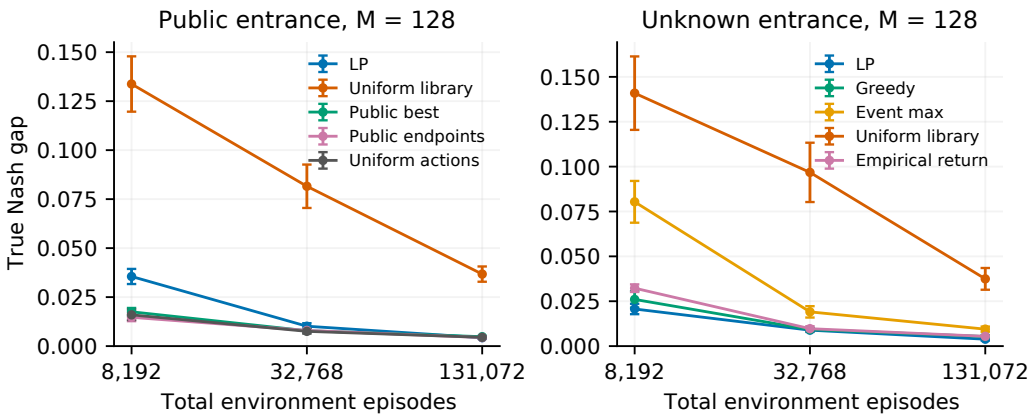

Figure 1: Sampling adaptation avoids entrance dilution; public information also permits effective uncalibrated choices. Both panels use $M = 128$, 16 seeds, complete episode costs, and one-standard-error bars. Unknown-entrance methods learn from trajectories.

With PCE's $M \le 4H^2$, the direct route uses

$$N_{\mathrm{PCE}} + O\left(\frac{H^5 L_{\mathrm{cal}}}{\varepsilon} + \frac{H^5 J L_Q}{\varepsilon^2}\right).$$

Thus its sufficient fitting term improves from H^7 to H^5, with the inherited exploration cost unchanged. For model-based evaluation, assume latent transitions lie on a known denominator-K grid. Proper MLE optimizes each coordinate's binomial rows, with $\Theta_\star \le p(K+1)^{2J}$ and expected $O(p + \log K)$ random-bit cost per simulated successor. Replace L_Q by $L_Q + L_\Theta$ in the display. These class logarithms contain additional action dependence.

Calibration evaluates $O(pn_{\mathrm{cal}} \sum_h M_h)$ decoder events and solves an LP with $M_h + 1$ variables and $2p$ constraints per stage. All trajectory execution is charged. In the entrance family, direct fitting takes $\widetilde{O}(pN_{\mathrm{env}})$ operations plus polynomial LP work; for fixed p and sufficiently small ε, fitting savings dominate both calibration samples and computation. In the experiments, the known reward exposes only the terminal outcome bit; intermediate coordinates remain hidden.

6 Experiments: Distribution Quality and Total Cost

We test when adapting a sampling distribution helps, whether obtaining it is worth its cost, and how coverage relates to response errors. All comparisons use real-successor fitted minimax and exact post-training dynamic best responses. Latent occupancies are evaluator-only inputs. Methods pay for their own provider, calibration, collection, and fitting; saved overhead can fund more fitting data. Error bars are standard errors over seeds. Practical calibration counts are below the theorem's sufficient schedule, so empirical feasibility is not a confidence certificate.

Learning where to sample. The three-stage entrance families use $M \in \{8, 32, 128\}$, 16 seeds, and signal $4\varepsilon = 0.04$. At $M = 128$ and 131,072 episodes, LP and the uncalibrated public-endpoint mixture have mean gaps 0.00430 and 0.00437, versus 0.03675 for uniform library sampling. At 8,192 episodes, calibration consumes half the cap and has a larger gap than the public endpoints (0.03554 versus 0.01469). When the entrance is unknown, LP, greedy, and empirical-return sampling use the same 32 trajectories per root. Their largest-cap mean gaps are 0.00385, 0.00534, and 0.00544, versus 0.03746 for the uniform library. Terminal outcomes select the root, and mistakes enter the gap. Figure 1 supports sampling adaptation while showing that public information can make calibration unnecessary. Appendix F reports the complete settings.

Distribution quality versus total cost. We fix a learned PCE library in a hidden-progress game with $H = 10$, $p = 4$, and eight seeds. Its progress-state transition matrix is $G_H = \left(\begin{smallmatrix} 1 & 1-1/H \\ 0 & 1 \end{smallmatrix}\right)$,

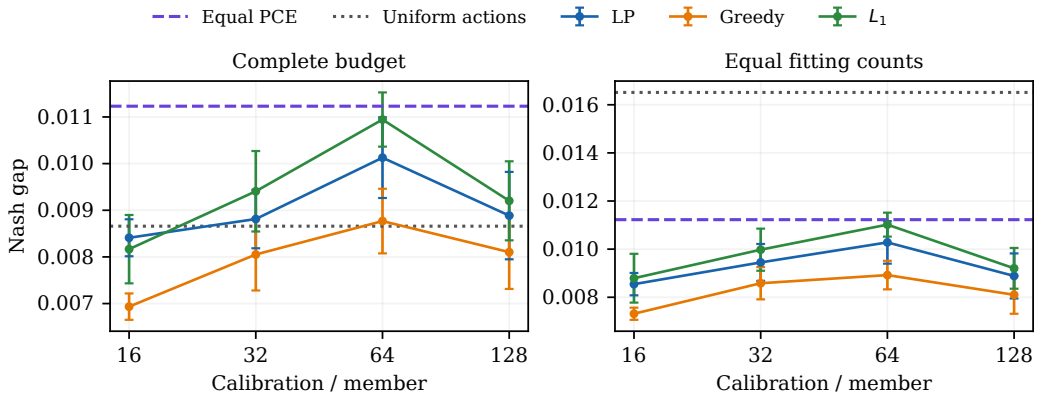

Figure 2: A distributional benefit at matched fitting counts need not yield a better complete-budget result. Left: every method spends its fixed cap. Right: fitting counts are identical and unused budgets remain unspent. Both use the larger cap, eight seeds, and one-standard-error bars; horizontal controls do not calibrate.

with independently permuted rows and columns. Uniform exploration has exponentially decreasing progress visitation, while unilateral responses remain relevant (Appendix B). Two total caps are fixed per seed; calibration varies over 16, 32, 64, and 128 episodes per member. LP, greedy, event max/subset, and the finite-library L_1 objective receive identical observation prefixes and decoder access. Equal PCE omits calibration; uniform actions omit both overheads.

Figure 2 compares spending the complete cap with holding fitting counts identical and leaving any remainder unspent. At 64 calibration episodes per member and the larger cap, equal-fitting mean gaps are 0.01028 (LP) and 0.01651 (uniform actions). Under the complete budget they are 0.01013 and 0.00866. A useful distribution at matched fitting counts need not justify the cost of obtaining it. This comparison with uniform actions assesses the complete provider and calibration pipeline; it does not isolate the value of reweighting. The LP episode decomposition is 608,224 for PCE, 17,720 for calibration, and 294,912 for fitting: calibration accounts for only 1.92% of the total.

To isolate calibration within a given library, compare LP with equal PCE and with the same-data optimizers. At this setting, equal PCE has mean gap 0.01123; the paired LP-minus-equal 95% interval $[-0.00427, 0.00206]$ includes zero. Greedy and L_1 also meet the numerical checks in all 256 calibrated stage problems. Their competitive results are consistent with a guarantee based on the checked constraints, rather than on the identity of the optimizer. Event max and subset fail the same empirical check in 80 and 67 stage problems, respectively; all remain in the aggregation.

Coverage versus actual best-response error. Figure 3 connects the coverage theorem to the Nash certificate. We measure both response residuals $B^{\pm} = \sum_h \mathbb{E}_{\mu^{\pm}_{h,\star}} |\widehat{Q}_h - r_h - P^{\star}_h \widehat{V}_{h+1}|$ and the squared error under the training law. LP's minimum visitation ratio rises from 0.539 to 0.592 across the calibration grid, but its gap and response residuals do not decrease monotonically, even with equal fitting counts. At 64 samples, LP raises coverage from 0.423 (equal PCE) to 0.558; its complete-budget $B^+ + B^- = 0.01020$ is close to its gap 0.01013. Coverage controls a worst-case change of measure; the realized gap also depends on where the fitted errors occur. We evaluate B^+ and B^- separately and plot their sum, which, with local matrix residuals, bounds the gap. These diagnostics are used only after fitting.

Shared-data evaluation across rewards. With calibration fixed at 64, we reuse each method's fresh stage buffers for three preselected public rewards: success, failure, and an immediate-action/terminal tradeoff. Collection is counted once and all task fits are timed. Larger-cap worst-task mean gaps are 0.00942 (LP), 0.00790 (greedy), 0.01081 (L_1), 0.01049 (equal PCE), and 0.00764 (uniform actions). The paired LP-minus-equal interval is $[-0.00415, 0.00200]$; success/failure are symmetric controls. The intervals are descriptive and unadjusted for multiple comparisons. Reuse does not establish an LP advantage. Complete time is 232.83 seconds for LP and 0.62 for uniform actions in the same

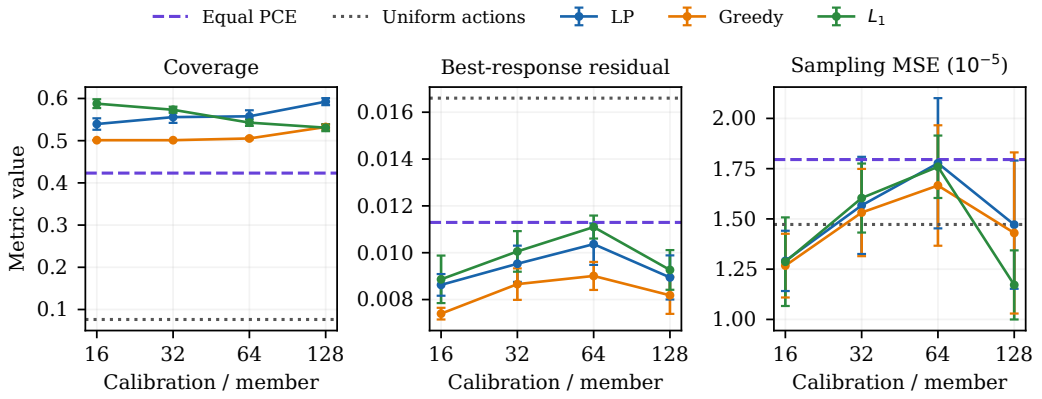

Figure 3: Coverage and best-response Bellman errors need not improve together. Coverage uses the fixed library and learned weights; error panels use identical fitting counts. All panels use the larger cap, eight seeds, and one-standard-error bars.

batched implementation, with provider work dominating LP's total. Corollary 13 permits this reuse for both sampling laws: collection is shared and the fitting logarithm increases by $\log K_{\mathrm{task}}$ for fixed rewards. Appendix G gives all 1,040 scan/task records, failure flags, residuals, and cost conventions.

7 Related Work

Covers and exploration distributions. Policy covers enable exploration with rich observations (Du et al., 2019; Misra et al., 2020). We inherit the two-context PCE construction and its cost from Rohatgi & Foster (2025). Amortila et al. (2024) optimize coverage objectives and give model-based and model-free exploration algorithms under coverability. Mutti et al. (2022) compress policy spaces through occupancy coverage. Our result concerns a supplied finite policy library and an unknown decoder: simultaneous observable-event constraints learn weights whose true coverage guarantee includes their statistical calibration cost. Common emissions make these population constraints equivalent to latent coverage, giving $R \leq S$ for arbitrary supplied Block-game libraries. Appendix E.1 compares an observable convex formulation of finite-library L_1 coverage on the same supplied library, candidate events, and calibration data. In particular, Proposition 3.1 of Amortila et al. (2024) already bounds changes of measure through second moments. Our statistical guarantee accounts for calibration samples and checked tolerances; any optimizer satisfying them inherits the same coverage conclusion.

Equilibrium learning and unilateral coverage. Unilateral concentration supports offline zero-sum learning, whereas coverage of a single played strategy pair is insufficient (Cui & Du, 2022a; Zhong et al., 2022; Cui & Du, 2022b). Ni et al. (2023) combine exploration, representation learning, and planning in broader low-rank games. Our backward certificate converts prediction errors along true best responses into a Nash-gap bound. Learned mixtures improve the cover-to-data step for both transition-MLE evaluation and direct fitted minimax. The auxiliary model-transfer certificate and RepLearn obstruction have different scopes from end-to-end game learners; Appendix M.1 provides a detailed comparison. Appendix P provides an extended related-work discussion, and Appendix Q summarizes structural, computational, and empirical limitations.

8 Conclusion

Common emissions make policy-cover calibration possible without decoder identification. Observable constraints preserve latent coverage, and finite-sample calibration can reduce the sufficient Nash-fitting budget after its own cost is included. The unknown-entrance family admits sample-optimal adaptation up to logarithms. Fixed-budget experiments show why better coverage must be assessed together with provider costs and errors along actual best responses.

AI USE STATEMENT

Generative-AI tools assisted with literature organization, manuscript drafting and consistency checks, and writing, debugging, and revising experimental code. The authors retain full responsibility for the citations, code, algorithms, theoretical claims, and reported results.

REPRODUCIBILITY STATEMENT

Appendix A gives the assumption ledger, Appendix N gives line-by-line pseudocode, Appendices I and B give full proofs. Appendix G specifies the fixed-budget scan and shared-reward protocol; the artifact independently replays all 960 distinct fits. Appendix F specifies the entrance baselines, budgets, seeds, and uncertainty conventions; Appendix H gives the fitting-route controls; Appendix O documents diagnostic corrections. The artifact contains per-run data, source code, and regeneration commands.

REFERENCES

Alekh Agarwal, Sham 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, 2020.

Ahmet Alacaoglu, Luca Viano, Niao He, and Volkan Cevher. A natural actor–critic framework for zero-sum Markov games. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 307–366, 2022.

Noga Alon and Assaf Naor. Approximating the cut-norm via Grothendieck's inequality. *SIAM Journal on Computing*, 35(4):787–803, 2006. doi: 10.1137/S0097539704441629.

Philip Amortila, Dylan J. Foster, and Akshay Krishnamurthy. Scalable online exploration via coverability. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1392–1455, 2024.

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, 2020.

Claire Chen, Yuheng Zhang, Xinyu Liu, Zixuan Xie, Shuze Daniel Liu, and Nan Jiang. Offline two-player zero-sum Markov games with KL regularization. *arXiv preprint arXiv:2605.13025*, 2026. URL https://arxiv.org/abs/2605.13025.

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.

Yuan Cheng, Ruiquan Huang, Yingbin Liang, and Jing Yang. Improved sample complexity for reward-free reinforcement learning under low-rank MDPs. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=jpsw-KuOi7r.

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, 2022a.

Qiwen Cui and Simon S. Du. Provably efficient offline multi-agent reinforcement learning via strategy-wise bonus. In *Advances in Neural Information Processing Systems*, volume 35, 2022b.

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.

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.

Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Exploration is harder than prediction: Cryptographically separating reinforcement learning from supervised learning. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 1953–1967, 2024. doi: 10.1109/FOCS61266.2024.00117.

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.

Ruiquan Huang, Donghao Li, Yingbin Liang, and Jing Yang. Breaking the computational barrier: Provably efficient actor–critic for low-rank MDPs. *arXiv preprint arXiv:2605.01242*, 2026. URL `https://arxiv.org/abs/2605.01242`. Accepted at ICML 2026.

Yassir Jedra, Junghyun Lee, Alexandre Proutiere, and Se-Young Yun. Nearly optimal latent state decoding in block MDPs. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 2805–2904, 2023.

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, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4870–4879, 2020.

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, 2022a.

Gen Li, Yuejie Chi, Yuting Wei, and Yuxin Chen. Minimax-optimal multi-agent RL in Markov games with a generative model. In *Advances in Neural Information Processing Systems*, volume 35, 2022b.

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.

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, 2023a. doi: 10.52202/075280-2916.

Zakaria Mhammedi, Dylan J. Foster, and Alexander Rakhlin. Representation learning with multi-step inverse kinematics: An efficient and optimal approach to rich-observation RL. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24659–24700, 2023b.

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. URL https://jmlr.org/papers/v25/22-0687.html.

Mirco Mutti, Stefano Del Col, and Marcello Restelli. Reward-free policy space compression for reinforcement learning. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3187–3203, 2022. URL https://proceedings.mlr.press/v151/mutti22a.html.

Thanh Nguyen-Tang and Raman Arora. Policy-regret minimization in Markov games with function approximation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 46242–46264, 2025.

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

Andi Nika, Debmalya Mandal, Adish Singla, and Goran Radanovic. Corruption-robust offline two-player zero-sum Markov games. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1243–1251, 2024.

Miruna Oprescu, Andrew Bennett, and Nathan Kallus. Low-rank MDPs with continuous action spaces. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4069–4077, 2024.

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, 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, 2023.

Dhruv Rohatgi and Dylan J. Foster. Necessary and sufficient oracles: Toward a computational taxonomy for reinforcement learning. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 4838–4936, 2025.

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

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 32nd 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.

Andrew J. Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. Reward-free RL is no harder than reward-aware RL in linear Markov decision processes. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22430–22456, 2022.

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.

Sihan Zeng, Thinh Doan, and Justin Romberg. Regularized gradient descent ascent for two-player zero-sum Markov games. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Runyu Zhang, Qinghua Liu, Huan Wang, Caiming Xiong, Na Li, and Yu Bai. Policy optimization for Markov games: Unified framework and faster convergence. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Weitong Zhang, Jiafan He, Dongruo Zhou, Amy Zhang, and Quanquan Gu. Provably efficient representation selection in low-rank Markov decision processes: From online to offline RL. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2488–2497, 2023a.

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, 2023b.

Yulai Zhao, Yuandong Tian, Jason D. Lee, and Simon S. 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.

Han Zhong, Wei Xiong, Jiyuan Tan, Liwei Wang, Tong Zhang, Zhaoran Wang, and Zhuoran Yang. Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27117–27142, 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.

A Complete assumptions and notation

This appendix records the common-emission and finite-fitting conditions used by the learned-mixture results, along with the assumptions for the auxiliary certificates and inherited PCE provider.

A.1 Measurability and probability conventions

The state space $(\mathcal{S}, \mathcal{F}_{\mathcal{S}})$ is standard Borel. All policies and feature maps are measurable, all candidate kernels are dominated by a common σ-finite measure, and all displayed conditional expectations admit regular versions. These conditions ensure that the block likelihood and the policy-induced occupancies are well defined. For distributions P, Q, we use

$$D_{\mathrm{TV}}(P,Q) = \frac{1}{2}\int |\mathrm{d}P - \mathrm{d}Q|, \qquad D^2_{\mathrm{H}}(P,Q) = \int (\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q})^2.$$

Thus $D^2_{\mathrm{TV}}(P,Q) \le D^2_{\mathrm{H}}(P,Q)$. All logarithms are natural.

A.2 Theorem assumptions and sample splitting

The population constraint equivalence requires only common emissions, a finite supplied library, and inclusion of the true decoder. Its $R \le \min(M, S)$ bound is automatic under these assumptions. Converting library coverage to all-policy coverage additionally requires the $(1, \eta)$ cover property. The learned-mixture theorem uses fresh calibration observations and evaluates all candidate decoders; its subsequent fitting buffers are independent conditional on the cover and calibration. A practical numerical feasibility check alone is not a high-probability cover event without the sufficient sample schedule.

Theorems 23 and 3 are deterministic: they require valid measurable kernels, bounded returned functions, finite actions, and measurable best-response selectors. The backward theorem applies to arbitrary returned functions, including fits trained under a different kernel. Theorem 26 adds its explicit true-model transition and learned-model evaluation laws. The direct certificate (7) instead uses true-model response laws for both errors. Neither certificate alone constructs an exploration law.

Corollary 29 requires exact Block structure, policy-independent disjoint emissions, centralized reward-free exploration, a realizable two-context oracle, and the displayed population MLE/PE events. Its L_1 statement does not imply the sharper L_2 statement. The finite-class specialization and the concrete approximate-closure variant supply this second-moment event by their own concentration arguments.

Conditional on the exploration and calibration output, the MLE datasets, PE datasets, and datasets at different stages are independent. In the model route, condition also on all fitted kernels. Backward evaluation uses only later-stage buffers to form its stage-h continuation, leaving the current PE buffer independent. In the real-successor route there is no fitted kernel to condition on. Cover inequalities hold uniformly over centralized Markov policies, including the output-dependent responses selected after fitting.

A.3 Inherited PCE interface

The inherited two-context interface is the following realizable regression problem. A law D on (X_1, X_2) is ϕ-realizable when $X_2 \perp X_1 \mid \phi(X_1)$ and $X_2 \perp X_1 \mid \phi(X_2)$. Given i.i.d. samples from such a law and binary labels satisfying

$$\mathbb{E}[Y \mid X_1, X_2] = f(\phi(X_1), \phi(X_2)), \qquad f : \mathcal{Z} \times \mathcal{Z} \to [0,1],$$

the oracle returns an (optionally improper) predictor with squared error at most α using $N_{2\mathrm{ctx}}(\alpha, \beta)$ labeled examples with probability $1 - \beta$. This is the exact statistical interface used in the sufficiency direction of PCE.

Corollary 29 assumes an exact Φ-decodable Block MG with $|\mathcal{Z}| = S$: conditional observation supports of distinct latent states are disjoint, the latent transition depends only on (z, a, b), and the true decoder lies in Φ. The observation kernel is the policy-independent emission law in Definition 27. Centralized joint-action policies are permitted during reward-free training. The two-context regression oracle satisfies the preceding realizable statistical interface; proper transition MLE supplies $\mathbb{E}_{\rho_h} D^2_{\mathrm{TV}}(P^\star_h, \widehat{P}_h) \le \zeta_h^2$; and average Bellman PE supplies $\mathbb{E}_{\rho_h} e_h \le \varepsilon_{h,\mathrm{B}}$.

Assumption 6 below collects the finite-class realizability, optimizer accuracy, Bellman closure, fresh sample, and two-context sample-efficiency conditions used by Corollary 37. In particular, its Bellman target is $q_h^\star(x) = r_h(x) + \widehat{P}_h \widehat{V}_{h+1}(x)$, and each PE label satisfies $\mathbb{E}[Y_i \mid X_i] = q_h^\star(X_i)$ with $Y_i \in [0, H_h]$. Backward clipping keeps $\widehat{V}_{h+1} \in [0, H_{h+1}]$, so the pseudo-successor construction indeed produces labels in this range. The finite-class condition is used only to turn the abstract MLE and PE accuracy events into the closed-form sample count in (50). The deterministic gap bound (47) remains valid for infinite classes whenever the same population error events are supplied by their statistical oracles.

A.4 Finite-class conditions

Assumption 6 (Finite-class instantiation)**.** *The structural Block-MG, PCE cover-generation, sample-independence, and predictable-continuation conditions of Assumption 28 hold. At each stage, the finite proper transition class contains $P^\star$, has size at most $\Theta_\star$, and is optimized to log-likelihood error ξ_{mle}. For every predictable continuation, the clipped Bellman mean belongs to a finite class $\mathcal{Q}_h \subseteq [0, H_h]^{\mathcal{S}\times\mathcal{A}\times\mathcal{B}}$ of size at most $Q_\star$. PE uses fresh covered contexts, labels in $[0, H_h]$ with that mean, and squared-loss ERM; its output is clipped to $[0, H_h]$. The two-context oracle has labeled-sample efficiency $N_{2\mathrm{ctx}}(\alpha, \beta)$ as defined in* (PCE)*; the polynomial-sample specialization assumes this efficiency is polynomial in inverse accuracy and confidence and in the class description length.*

B Learning a sampling mixture without identifying the decoder

The object being learned. Condition on stagewise $(1, \eta)$ policy covers $\Psi_h = \{\psi_{h,1}, \ldots, \psi_{h,M_h}\}$, $M_h \le M$, constructed before the data used below. Let $F = |\Phi| < \infty$ and suppose each candidate decoder can be evaluated. For $\phi \in \Phi$ and $z \in [S]$, define observable-event probabilities

$$D_h(\phi, z, i) = \Pr_{P^\star, \psi_{h,i}}\{\phi(S_h) = z\}, \qquad c_h(\phi, z) = \max_i D_h(\phi, z, i).$$

The true decoder belongs to Φ, but is not supplied to the algorithm. All probabilities refer to the true environment and roll-in prefixes.

Definition 7 (Simultaneous coverage factor)**.** *The covers have simultaneous coverage factor at most $R \ge 1$ if, for each stage h, there is a vector $w_h^\circ \in \Delta_{M_h}$ such that*

$$\sum_i w_{h,i}^\circ D_h(\phi, z, i) \ge c_h(\phi, z)/R \quad \textit{for every } \phi \in \Phi,\ z \in [S].$$

Zero-probability events impose no restriction. A known upper bound R, rather than the optimizing weights, is an input to the sample schedule.

Lemma 8 (Population constraint equivalence and coverage factor)**.** *Fix a stage in a Block game with policy-independent emissions and $\phi^\star \in \Phi$. For any library, $w \in \Delta_{M_h}$, and $t \ge 0$, the event constraints $\sum_i w_i D_h(\phi, y, i) \ge t c_h(\phi, y)$ for all ϕ, y hold if and only if*

$$q_h^w(z) \ge t \max_i p_i(z) \quad \textit{for all } z, \qquad p_i(z) = \Pr_{\psi_{h,i}}(Z_h = z), \quad q_h^w(z) = \sum_i w_i p_i(z).$$

In particular the optimal simultaneous coverage factor equals the library's optimal latent occupancy factor and is at most $\min\{M_h, S\}$. This upper bound is tight in the worst case. The library need not be a cover of policies outside itself.

Proof. Write $a_{\phi,y}(z) = O_h(\{x : \phi(x) = y\} \mid z)$. Common emissions imply $D_h(\phi, y, i) = \sum_z p_i(z) a_{\phi,y}(z)$ with nonnegative coefficients independent of i. The true decoder's events give necessity. For sufficiency, latent domination gives, for every library member i,

$$\sum_j w_j D_h(\phi, y, j) = \sum_z q_h^w(z) a_{\phi,y}(z) \ge t \sum_z p_i(z) a_{\phi,y}(z) = t D_h(\phi, y, i).$$

Taking the maximum over i proves equivalence. Uniform weights give factor M_h. Alternatively, choose $i_z \in \arg\max_i p_i(z)$ for each latent state and mix these S choices equally, merging repetitions.

Then $q(z) \geq p_{i_z}(z)/S$, giving factor S. For tightness let $k = \min\{M_h, S\}$ policies reach k distinct states deterministically and let any additional members duplicate an occupancy. Every feasible mixture must give all k states probability at least $1/R$, so $R \geq k$. These are population existence statements, not requests for latent-state access by the algorithm. □

Adding candidate decoders containing the true one leaves this population feasible region unchanged. Finite-sample uncertainty and computation still depend on the candidate class: all FS empirical constraints must be evaluated and controlled. A valid general schedule uses $R = \min\{M, S\}$.

Observable calibration and optimization. Set $\beta = \eta/(64R)$ and

$$L_{\mathrm{cal}} = \log \frac{2HMFS}{\delta_{\mathrm{w}}}, \qquad n_{\mathrm{cal}} = \left\lceil \frac{5L_{\mathrm{cal}}}{\beta} \right\rceil.$$

For each (h, i), collect n_{cal} independent episodes with roll-in policy $\psi_{h,i}$. Evaluate every candidate decoder on S_h to form empirical frequencies $\widehat{D}_h(\phi, z, i)$ and $\widehat{c}_h(\phi, z) = \max_i \widehat{D}_h(\phi, z, i)$. No reward, latent-state label, or learned transition model is needed. At each stage solve the linear program

$$\begin{aligned} &\max_{w \in \Delta_{M_h},\, 0 \leq t \leq 1} \quad t \\ &\text{subject to} \quad \sum_i w_i \widehat{D}_h(\phi, z, i) \geq t\widehat{c}_h(\phi, z) - 2\beta \quad (\phi \in \Phi, z \in [S]). \end{aligned} \tag{9}$$

An exact optimizer suffices. For numerical computation it is enough to return $\widehat{w}_h, \widehat{t}_h$ satisfying the simplex constraints, $\widehat{t}_h \geq 7/(10R)$, and all displayed inequalities with 3β in place of 2β; these conditions can be checked directly. Choose a roll-in policy with probabilities $\widehat{w}_h$, then sample a uniform joint action at stage h. Denote the resulting law by ρ_h^{w}. Subsequent MLE and PE buffers are fresh conditional on the cover and calibration.

Theorem 9 (Learned-mixture coverage)**.** *Suppose the conditional covers are $(1, \eta)$ covers in an exact decodable Block game and satisfy Definition 7. With conditional probability at least $1 - \delta_{\mathrm{w}}$, the calibration above admits the stated numerical solution and, simultaneously for every centralized policy ω and measurable $g_h \in [0, 1]$,*

$$\begin{aligned} \mathbb{E}_{d^{\omega}_{P^\star, h}\omega_h} g_h &\leq 2S\eta + 5RJ\, \mathbb{E}_{\rho_h^{\mathrm{w}}} g_h, \\ \mathbb{E}_{d^{\omega}_{P^\star, h}\omega_h} g_h &\leq 2S\eta + \sqrt{5RJ\, \mathbb{E}_{\rho_h^{\mathrm{w}}} g_h^2}. \end{aligned} \tag{10}$$

Calibration uses at most $N_{\mathrm{cal}} = n_{\mathrm{cal}} \sum_h M_h \leq HMn_{\mathrm{cal}}$ episodes. Its computation is polynomial in $H, M, F, S, n_{\mathrm{cal}}$, decoder-query time, and the bit length of the inputs; it does not enumerate observations.

Proof. Condition throughout on the returned covers. Bernstein's inequality for a Bernoulli event of probability D gives, with failure probability at most $2e^{-L_{\mathrm{cal}}}$,

$$|\widehat{D} - D| \leq \sqrt{2DL_{\mathrm{cal}}/n_{\mathrm{cal}}} + 2L_{\mathrm{cal}}/(3n_{\mathrm{cal}}) \leq D/8 + 5L_{\mathrm{cal}}/n_{\mathrm{cal}} \leq D/8 + \beta.$$

A union bound over at most $HMFS$ events makes this simultaneous. The same observations may be used for all decoders; independence between decoder events is not required. Write $q_w = \sum_i w_i D_i$ and $\widehat{q}_w = \sum_i w_i \widehat{D}_i$ for one event. On this event,

$$\tfrac{7}{8} q_w - \beta \leq \widehat{q}_w \leq \tfrac{9}{8} q_w + \beta, \qquad \tfrac{7}{8} c - \beta \leq \widehat{c} \leq \tfrac{9}{8} c + \beta.$$

For the simultaneous witness w°, these inequalities imply

$$\widehat{q}_{w^\circ} \geq \frac{7}{8R} c - \beta \geq \frac{7}{9R} \widehat{c} - \left(1 + \frac{7}{9R}\right) \beta \geq \frac{7}{9R} \widehat{c} - 2\beta.$$

Thus the LP optimum is at least $7/(9R)$, leaving a positive margin for the stated numerical stopping criterion. For any returned feasible numerical solution,

$$\begin{aligned} q_{\widehat{w}} &\geq \tfrac{8}{9}(\widehat{q}_{\widehat{w}} - \beta) \geq \tfrac{8}{9}(\widehat{t}\widehat{c} - 4\beta) \\ &\geq \tfrac{7}{9}\widehat{t}c - \tfrac{8}{9}(\widehat{t} + 4)\beta \geq \frac{49}{90R} c - \frac{40}{9}\beta. \end{aligned}$$

For the true decoder let $v_h(z) = \sup_\omega \Pr_{P^\star,\omega}(Z_h = z)$. The cover guarantee gives $c_h(\phi^\star, z) \geq v_h(z) - \eta$. If $v_h(z) \geq 2\eta$, substitution yields

$$q_{\widehat{w}_h}(\phi^\star, z) \geq \frac{49}{90R}(v_h(z) - \eta) - \frac{5\eta}{72R} \geq \frac{19}{80R} v_h(z) \geq \frac{1}{5R} v_h(z).$$

Common disjoint-support emissions and a uniform joint action therefore bound the density ratio by $5RJ$ on these latent states. The remaining states have total probability at most $2S\eta$ under every policy. Integration proves the first bound. For the second, Cauchy–Schwarz on the nonrare part gives $\mathbb{E}_\mu[g\mathbf{1}_{\mathrm{nonrare}}] \leq \sqrt{5RJ\,\mathbb{E}_{\rho_h^{\mathrm{w}}} g^2}$; the rare part contributes at most $2S\eta$. All statements are uniform over policies and measurable functions, so later data-dependent best responses are included.

Frequency computation uses $Fn_{\mathrm{cal}} \sum_h M_h$ decoder queries. Each LP has $M_h + 1$ variables and FS event constraints. For rational accuracy parameters its coefficients are rational empirical frequencies, so polynomial-time linear programming achieves the required accuracy. Trajectory simulation and policy evaluation during every calibration episode are also charged; no latent observation oracle is invoked. □

Corollary 10 (Reachability-sensitive calibration)**.** *Under the conditions and on the event of Theorem 9, define*

$$\mathcal{R}_h(\eta) = \{z : v_h(z) < 2\eta\}, \qquad b_h(\eta) = \sup_\omega \Pr_{P^\star,\omega}\{Z_h \in \mathcal{R}_h(\eta)\}.$$

Both inequalities in (10) *hold with $b_h(\eta)$ in place of $2S\eta$. In particular, if every reachable latent state at a fitted stage has maximum reachability at least $\lambda > 0$, taking $\eta \leq \lambda/2$ removes the additive term there. Calibration then needs $O(RL_{\mathrm{cal}}/\lambda)$ episodes per member when $\eta = \lambda/2$ and a $(1, \eta)$ cover is available.*

Proof. The proof of Theorem 9 bounds the density ratio by $5RJ$ off $\mathcal{R}_h(\eta)$. For $0 \leq g_h \leq 1$, integration over the remaining states contributes at most $b_h(\eta)$ under any target policy. Also $b_h(\eta) \leq \sum_{z \in \mathcal{R}_h(\eta)} v_h(z) \leq 2S\eta$, recovering the original bounds. If $2\eta \leq \lambda$, the rare set contains only unreachable states and its mass is zero. □

The profile is an analysis quantity, not an additional input oracle. The general schedule below still takes η proportional to ε; a certified reachability lower bound permits the sharper instance-specific schedule.

Corollary 11 (Full cost of weighted equilibrium learning)**.** *Under the finite fitting assumptions of Corollary 37 and Definition 7, replace the uniform roll-in mixtures by* (9)*. Put $\kappa_{\mathrm{w}} = 5RJ$ and use* (48)–(49) *with $\overline{\kappa}$ replaced by κ_{w}. Allocate failure probability $\delta/4$ to each of cover construction, calibration, MLE, and PE; thus the fitting logarithms can use $8H/\delta$ in place of $6H/\delta$. Then the output is an ε-Nash product pair with probability at least $1 - \delta$, using*

$$\begin{aligned} N_{\mathrm{env}} &\leq N_{\mathrm{cover}}(\eta, \delta/4) + n_{\mathrm{cal}} \sum_h M_h + H(n_{\mathrm{mle}} + n_{\mathrm{pe}}) \\ &= N_{\mathrm{cover}}(\eta, \delta/4) + O\left(HM + \frac{H^3 MRSL_{\mathrm{cal}}}{\varepsilon} + \frac{H^5 RJ(L_\Theta + L_Q)}{\varepsilon^2}\right), \end{aligned} \tag{11}$$

where $\eta = \varepsilon/(32SH^2)$. Real-successor fitted minimax satisfies the same guarantee after deleting the MLE samples and L_Θ term.

Proof. On the cover and calibration event, apply (10) directly to the two true-model response laws in (7). This gives

$$\mathrm{Gap}_{P^\star}(\widehat{\pi}, \widehat{\nu}) \leq 4SH^2\eta + 2\sqrt{\kappa_{\mathrm{w}}} \sum_{h<H} H_{h+1}\zeta_h + 2\sqrt{\kappa_{\mathrm{w}}} \sum_h u_h + 2\sum_h \delta_h.$$

Conditional on the covers and weights, independent downstream buffers give the same MLE and PE concentration events as before. The stated schedule bounds the four terms by $\varepsilon/8, \varepsilon(H-1)/(8H), \varepsilon/4, \varepsilon/4$. The calibration count obeys

$$N_{\mathrm{cal}} \leq HM + \frac{10240H^3 MRSL_{\mathrm{cal}}}{\varepsilon}.$$

Each fitting buffer has size $O(H^4RJL/\varepsilon^2)$. Summing over stages and the four failure events proves the claim. For true-successor labels, remove the transition-error term and invoke Corollary 18 on the new collection laws. □

Corollary 12 (General finite-class Block bound)**.** *Under the cover and finite-fitting assumptions of Corollary 11, take $R = S$ without any further coverage-factor assumption. If $M \le H^2S^2$, the complete episode bound is*

$$N_{\mathrm{cover}}(\varepsilon/(32SH^2), \delta/4) + O\left(\frac{H^5S^4L_{\mathrm{cal}}}{\varepsilon} + \frac{H^5SJ(L_\Theta + L_Q)}{\varepsilon^2}\right).$$

The real-successor route deletes the L_Θ term. Computation remains conditional on the supplied provider and fitting implementations.

Proof. Lemma 8 supplies $R \le S$. Substitute this and $M \le H^2S^2$ into (11). The rounding term $HM \le H^3S^2$ is absorbed by the calibration term for $H, S \ge 1$, $\varepsilon \in (0, 1)$, and the displayed confidence log. The probability allocation and all fitting conditions are unchanged. □

Reusing a calibrated distribution.

Corollary 13 (Sharing data across fixed reward tasks)**.** *Fix K_{task} known bounded reward functions on the same dynamics, independently of the collected data. Suppose each task satisfies the fitting assumptions of Corollary 11, with a common upper bound $Q_\star$ on its regression-class size and common accuracy ε. One cover, one calibration table, and one set of independent stagewise real-successor buffers suffice to return ε-Nash pairs for all tasks with probability $1 - \delta$. The episode bound is the real-successor version of* (11)*, with L_Q replaced by $L_Q + \log K_{\mathrm{task}}$. Fitting and local matrix computation are performed for each task. The same data-sharing statement holds for uniform library sampling with its original coverage coefficient.*

Proof. The cover and calibration events concern dynamics and hold for every target policy and bounded function simultaneously. For each fixed task, its stage-h continuation depends only on later-stage buffers, so it is independent of the reused stage-h buffer. Apply the original finite-class regression bound at failure probability $\delta/(4K_{\mathrm{task}})$ per task and union bound over tasks. Independence between tasks is unnecessary. The known reward supplies the labels without further environment calls. The same argument applies to the uniform collection laws. □

This amortizes collection for both samplers; it does not require a baseline to repeat exploration for each reward.

When paying for calibration still saves samples. The unweighted construction has fitting order $H^5MJ(L_\Theta + L_Q)/\varepsilon^2$ before substituting the worst-case cover size. For the same cover, class sizes, and failure allocation, let $N_{\mathrm{fit}}(\kappa) = H(n_{\mathrm{mle}}(\kappa) + n_{\mathrm{pe}}(\kappa))$ denote the explicit sufficient fitting budget. The weighted schedule uses fewer total episodes whenever

$$n_{\mathrm{cal}} \sum_h M_h < N_{\mathrm{fit}}(2MJ) - N_{\mathrm{fit}}(5RJ).$$

For fixed $M > 5R/2$, this holds for all sufficiently small ε: the calibration overhead has order ε^{-1} and the fitting saving has order ε^{-2}, with logarithmic class factors. This compares certified schedules, not a lower bound on every learner. Both sides retain the same cover-construction cost; if inherited PCE dominates, this does not change its leading complexity exponent. The learner can compare these two sufficient budgets immediately after receiving the cover, using a common fitting-class resolution. Choosing the smaller schedule guarantees the smaller certified total count without first paying for calibration in the uniform branch. The next construction shows an actual sampling separation as well.

Proposition 14 (A full-cost separation from uniform library sampling)**.** *For each $M \ge 4$ and sufficiently small $\varepsilon > 0$, there is a three-stage zero-sum hidden-coordinate family, with a public terminal outcome and an explicitly constructible stage-two library of M distinct roll-in policies, for which the following holds. The latent entry mechanism at stage one is known and the sampling interface executes a selected library prefix, followed by an arbitrary stage-two action. The library is an exact cover, has simultaneous coverage factor at most two, and requires no environment episodes*

to construct. The learned-mixture method uses $\widetilde{O}(M + 1/\varepsilon^2)$ episodes, including calibration. Any learner that draws every informative stage-two roll-in independently and uniformly from the supplied library needs $\Omega(M/\varepsilon^2)$ episodes to return an ε-Nash pair with probability at least $3/4$ on both instances. The lower bound restricts the collection law, not the downstream estimator or planner. For fixed p, direct fitting also has computational cost $\widetilde{O}(M + \varepsilon^{-2}) + \mathrm{poly}(M, \log(1/\varepsilon))$; the polynomial term accounts for the calibration LP.

Proof. Use two latent states, initially state zero, $A = 4$, and $B = 2$. Write a maximizer action as $(d, u) \in \{0, 1\}^2$. At stage one, $d = 1$ enters latent state one and $d = 0$ enters state zero; the other action coordinates have no effect on this transition. At stage two, state zero leads to terminal outcome zero. From state one the probability of terminal outcome one is

$$p_\theta(d, u, b) = \tfrac{1}{2} + 4\theta\varepsilon(-1)^d + \tfrac{1}{8}(-1)^{u+b}, \qquad \theta \in \{-1, 1\}.$$

Take $\varepsilon \le 1/80$. These probabilities are bounded away from zero and one. The last term is a coupled matching-pennies game; mixing u or b uniformly gives value zero for that term. All intermediate rewards are zero and the known terminal reward is the public outcome. The game value is $1/2 + 4\varepsilon$.

Define the library's entry probabilities by $r_1 = 1$ and $r_i = (i - 2)/M^3$ for $2 \le i \le M$. These are genuinely different stochastic behaviors at a reachable state, not copies of one policy. Since one policy always enters and another never enters, the stage-two library is an exact max-visitation cover. There is no need to estimate the known latent entry mechanism or terminal reward: entering is optimal under both instances, and only the stage-two matrix must be learned. The policies are generated without knowing θ or the decoder. Hidden-coordinate emissions are as in Proposition 17, except that the terminal bit is public as specified below; hence $R \le 2$. Their computational construction cost is polynomial in M and their description length.

Under uniform library sampling, the probability of the only informative state is $\overline{r} = M^{-1} \sum_i r_i \le 2/M$. Conditional on an informative observation and any chosen (d, u, b), the two Bernoulli means differ by 8ε and are bounded in $[1/4, 3/4]$. Their KL divergence is at most $C\varepsilon^2$ for a universal constant. The chain rule therefore bounds the divergence between N-episode adaptive learning transcripts by $2CN\varepsilon^2/M$. This permits arbitrary stage-two actions and arbitrary reuse of all collected transitions.

For a returned maximizer policy let m be its conditional mean of $(-1)^d$ at stage two in latent state one. Against a uniform minimizer action, its value is at most $1/2 + 4\theta\varepsilon m$, since entering with probability below one only decreases this positive quantity. Its worst-response value is no greater. Thus its Nash gap is at least $4\varepsilon(1 - \theta m)$, and gap at most ε implies $\theta m \ge 3/4$. The sign of m then distinguishes the two instances. We may even give the decoder to this distinguishing test and to the learner; the transition information bound is unchanged. A test correct with probability at least $3/4$ in both instances requires constant total-variation distance and therefore constant KL by Pinsker's inequality. This proves the lower bound.

For the upper bound, both stage-two states have maximum reachability one. Use Corollary 10 with $\eta_0 = 1/4$ on the exact stage-two library and fit its Bellman matrix, with H, S, J, R constant. No rare-state remainder enters this fit. Output the known entering action at stage one. The calibration and fitting guarantees apply to the informative stage-two state; entering is optimal against every opponent, so the full-game Nash gap equals the conditional stage-two gap for these returned policies. The calibration cost is $\widetilde{O}(M)$ and fitting costs $\widetilde{O}(\varepsilon^{-2})$. A finite transition grid containing both instances gives the computable model variant. For example, $\varepsilon = 1/(80k)$ with integer $k \ge 1$ permits $K = 80k$. The true-successor variant needs no model grid. For example, taking $\varepsilon = M^{-2}$ gives an upper bound $\widetilde{O}(M^4)$ against the uniform-library lower bound $\Omega(M^5)$, including calibration. For direct fitting, compute cell means for each candidate coordinate and solve the two resulting constant-size matrix games for the selected coordinate. This takes $\widetilde{O}(pN_{\mathrm{env}})$ operations, including policy execution and observation processing. The remaining calibration LP has polynomial bit complexity in $M, p, \log(1/\varepsilon)$. For fixed p, choosing $\varepsilon = M^{-q}$ with $q > 1$ sufficiently large makes this LP cost lower order than ε^{-2} fitting. The uniform comparator still requires $\Omega(M\varepsilon^{-2})$ environment calls and hence at least that many operations. Thus a resource-saving regime exists after charging computation as well; no practical wall-clock advantage is inferred from this asymptotic example. □

Scope of the separation. Here the entrance is public: always using the entering member already permits $\widetilde{O}(\varepsilon^{-2})$ task-directed fitting, while equally mixing the entering and nonentering endpoints covers both states. Both are legitimate uncalibrated baselines. The proposition isolates the cost of uniform sampling; the unknown-entrance construction in Appendix C additionally requires discovering which member provides coverage.

The comparator samples uniformly from the given library. It is allowed optimal estimation, planning, and data reuse; it is not allowed to change the roll-in distribution. An unrestricted exploration algorithm may also choose an informative mixture. The result establishes the resource cost of the uniform-mixture interface, not a universal lower bound for prior RL algorithms or an advantage unique to transition MLE.

Proposition 15 (Binary hidden-coordinate specialization)**.** *For any roll-in library in the hidden-coordinate family of Proposition 17, $R \le 2$. The same statement holds if terminal emissions use a known public coordinate and every candidate decoder uses that public coordinate at the terminal stage.*

Proof. Both variants have two latent states and common emissions, so Lemma 8 gives $R \le S = 2$ at every stage. Uniform nuisance coordinates make fitting explicitly tractable; they are not required for the coverage-factor bound. □

Consequently, for $S = 2$ and $M \le H^2S^2$, the concrete model route has the full episode bound

$$\mathcal{N}_{\mathrm{PCE}}\left(\frac{\varepsilon}{64H^2}, \frac{\delta}{4}\right) + O\left(\frac{H^5 L_{\mathrm{cal}}}{\varepsilon} + \frac{H^5 J(L_\Theta + L_Q)}{\varepsilon^2}\right).$$

Here $L_{\mathrm{cal}} = \log(8HMp/\delta)$ up to an absolute constant inside the logarithm; $\log \Theta_\star$ and $\log Q_\star$ retain their J-dependent terms. The leading fitting improvement from H^7 to H^5 pays for calibration, but does not change the inherited PCE term.

Computation beyond environment episodes. Let c_{roll} bound one step of executing a supplied roll-in policy and processing its observation, c_ϕ a decoder evaluation, c_r a known-reward query, and $\mathrm{LP}(M, FS, b)$ the cost of one calibration program with b-bit coefficients to the stated accuracy. Set $N_{\mathrm{new}} = N_{\mathrm{cal}} + H(n_{\mathrm{mle}} + n_{\mathrm{pe}})$. Beyond the separately charged cover construction, the coordinate implementation has arithmetic and query cost bounded by

$$\begin{aligned} O\big(&HN_{\mathrm{new}}c_{\mathrm{roll}} + N_{\mathrm{cal}}Fc_\phi + H\,\mathrm{LP}(M, FS, b) \\ &+ Hp(n_{\mathrm{mle}} + n_{\mathrm{pe}} + J) + Hn_{\mathrm{pe}}(p + \log K + TJ + Jc_r)\big). \end{aligned}$$

This includes proper fitting, simulated labels, and all continuation matrix queries. The real-successor variant removes model fitting and simulated-draw costs. Exact bit computation replaces the arithmetic fits by the polynomial bounds in Proposition 17. These costs establish polynomial implementability, not a universal wall-clock improvement; collection, calibration, and fitting time are measured experimentally.

Public terminal outcomes preserve computation. For experiments let $\phi_i(h, x) = x_i$ at $h < H$ and $\phi_i(H, x) = x_1$ for every i (coordinate one is public). The emission fixes the appropriate bit and draws the others fairly. Nonterminal rewards are zero and $r_H(x, a, b) = x_1$. The reward function is thus identical for all possible hidden coordinates, even under arbitrary reward queries. MLE groups current and next bits using this stage-aware decoder; the emission likelihood contribution is still $-(p-1)\log 2$ per observation. Two-context ERM uses its two decoded labels, and Bellman closure still has the form $r + g \circ \phi_i$. Every computational and quantization argument of Proposition 17 therefore remains valid.

A coupled game with difficult uniform exploration. The progress state is one and state zero is absorbing. Start uniformly over these states. For $h < H$, the probability of staying in progress is the matrix

$$G_H = \begin{pmatrix} 1 & 1 - 1/H \\ 0 & 1 \end{pmatrix},$$

with independent row and column permutations at each stage. Only the public terminal outcome is rewarded. Every progress-state local game has an interior saddle, with value $v_H = 1/(1 + 1/H)$. Under uniform actions, progress reaches stage h with probability $\frac{1}{2}[(3 - 1/H)/4]^{h-1}$, whereas its maximum centralized reachability is $1/2$. The uniform coverage ratio therefore decays exponentially with horizon. Against a uniform minimizer, the maximizer reaches the terminal reward with probability $\frac{1}{2}(1 - 1/(2H))^{H-1}$; against a uniform maximizer, the minimizing response gives $\frac{1}{2}(1/2)^{H-1}$. Thus the uniform product pair has exact gap

$$\frac{1}{2}\left[(1 - 1/(2H))^{H-1} - (1/2)^{H-1}\right],$$

which stays bounded away from zero. The rare path is strategically relevant to unilateral deviations, rather than only to joint cooperation.

C When the useful library member is unknown

The public-entry construction isolates the cost of a prescribed uniform sampling law. We now put the useful member's identity in the unknown transition kernel. This distinction matters: shuffling public policy names would leave the original entry probabilities known.

Proposition 16 (Unknown entrance and sample-optimal adaptation)**.** *For $M \geq 4$, $0 < \varepsilon \leq 1/80$, and $\delta \in (0, 1/4)$, there is a three-stage, two-latent-state zero-sum family with an unknown entrance $i_\star \in [M]$ and an unknown coordinate decoder. The first-stage maximizer has M actions; at stage two the action sizes are four and two. The supplied library contains the M distinct deterministic first-stage actions and is an exact stage-two cover with $R \leq 2$. Its construction needs no environment episodes.*

Observable-event calibration, root-action selection from its public terminal outcomes, and independent real-successor fitting return an ε-Nash pair with probability at least $1 - \delta$ using

$$\widetilde{O}(M + \varepsilon^{-2})$$

episodes. Any learner constrained to independent uniform library sampling needs $\Omega(M\varepsilon^{-2})$ episodes for success probability $3/4$. For the same success probability, any unrestricted adaptive learner needs $\Omega(M + \varepsilon^{-2})$ episodes in the worst case on this family. Thus the upper bound is optimal up to logarithmic factors in environment episodes. A mixture chosen before observing the environment has worst-case expected informative-state visitation at most $1/M$.

Proof. Use an initially uniform latent state and common hidden-coordinate emissions. At stage one action i deterministically enters latent state one if $i = i_\star$, and state zero otherwise, independently of the initial state. The minimizer has no effective action there. The public library consists simply of the action labels $1, \ldots, M$; it contains no transition probabilities. State zero at stage two leads to terminal outcome zero. In state one the terminal-one probability is

$$Q_\theta(d, u, b) = \tfrac{1}{2} + 4\theta\varepsilon(-1)^d + \tfrac{1}{8}(-1)^{u+b}, \qquad \theta \in \{-1, 1\}.$$

Only the public terminal outcome is rewarded. Thus both the entrance identity and the favorable stage-two row direction are unknown, while the game value is $v_\star = 1/2 + 4\varepsilon$. The library contains one state-one policy and at least one state-zero policy, giving an exact cover. Lemma 8 gives $R \leq 2$.

For the upper bound, calibrate each library member with independent episodes and uniform stage-two joint actions. The mixture LP uses only the intermediate observations. In the same trajectories, the terminal outcome has mean $1/2$ for $i_\star$ and is identically zero for every other member. Select the root action with largest empirical terminal mean. A positive outcome identifies $i_\star$ without any decoder query; its absence after n trials has probability 2^{-n}. Choosing the calibration count also at least $\lceil \log_2(2/\delta) \rceil$ therefore identifies the root action with failure probability at most $\delta/2$. This root-selection step is shared with the baselines and its trajectory cost is included in calibration; it is not a public entrance oracle.

Here $v_2(0) = v_2(1) = c_2(0) = c_2(1) = 1$. Set $\eta_0 = 1/4$, $R = 2$, and $\beta = 1/512$. Use

$$n_{\mathrm{cal}} = \max\{\lceil 2560 L_{\mathrm{cal}} \rceil, \lceil \log_2(2/\delta) \rceil\}$$

episodes per member, with $\delta_{\mathrm{w}} = \delta/4$ in L_{cal}. On the calibration event the numerical LP conditions imply, for both states,

$$q_{\widehat{w}}(z) \geq \frac{49}{180} - \frac{40}{9 \cdot 512} = \frac{253}{960} > \frac{1}{10}.$$

Corollary 10 removes the rare-state remainder, so constant precision suffices and calibration uses $O(M \log(MF/\delta))$ episodes. On that event, collect $\widetilde{O}(\varepsilon^{-2})$ fresh stage-two samples, using the learned mixture and uniform sampling of the eight joint actions. Finite coordinate-head ERM with mesh $O(\varepsilon)$ and exact local matrix solves yields conditional Nash gap at most ε with failure probability at most $\delta/4$. Allocate $\delta/4$ to calibration, so the three failure events sum to at most δ. Since the selected root always enters the informative state, that conditional gap is the full-game gap. The terminal and initial matrices require no generic regression over the M root actions: the deterministic entrance structure and the charged root-selection step supply the root decision.

For the uniform-library lower bound, fix $i_\star$ and even reveal it and the decoder to the learner. Each episode reaches the only informative state with probability $1/M$. For either instance and any stage-two action, the outcome probabilities lie in $[1/4, 3/4]$ and the two instances differ by 8ε. The transcript KL over N episodes is therefore $O(N\varepsilon^2/M)$, including all reused calibration and fitting observations. Against a uniform stage-two minimizer, the coupling term cancels. If m is the returned maximizer's probability of $d = 0$, averaged over the informative state's emissions, the ε-Nash condition requires $m \geq 7/8$ for $\theta = 1$ and $m \leq 1/8$ for $\theta = -1$. A root choice that sometimes fails to enter only reduces its payoff and cannot weaken these requirements. Thus the returned policy defines a test of θ; binary testing gives the claimed $\Omega(M\varepsilon^{-2})$ lower bound.

For the exploration lower bound, reveal the decoder and stage-two matrix, so a queried root reveals whether it is the entrance. Under a uniform prior on $i_\star$, all unsuccessful queries have the same transcript law. After at most N distinct queries, the chance to have queried or correctly guess the entrance is at most $(N+1)/M$. For any returned maximizer, let r be its probability of choosing the entrance. Its Nash gap is at least $(1-r)v_\star$. Constant accuracy below $1/8$ thus requires $r > 3/4$, identifying the entrance from the returned policy. Success probability $3/4$ on every entrance requires $N = \Omega(M)$.

For the unrestricted estimation lower bound, instead reveal the entrance and decoder but retain the unknown sign θ. Under any adaptive root and stage-two action choices, each episode contributes at most $C\varepsilon^2$ transcript KL: only the informative state's terminal Bernoulli law depends on θ. Thus the same sign test used above requires $N = \Omega(\varepsilon^{-2})$. Both lower bounds apply to the worst-case episode budget of a learner succeeding on this family. Their maximum is at least half their sum, giving $\Omega(M + \varepsilon^{-2})$. Finally, for any data-independent random weights, $\sum_i \mathbb{E}w_i = 1$, so some entrance has $\mathbb{E}w_{i_\star} \leq 1/M$. □

Actions and computation are counted. The M first-stage actions are essential to this construction and are not treated as a constant. Time-dependent action sets can be embedded in the paper's fixed action spaces by duplicating the four stage-two actions with publicly known aliases; the algorithm samples their eight distinct joint representatives. The displayed upper bound uses the specific deterministic root structure proved above, rather than substituting a constant J into the general Block corollary. Root-policy descriptions and root selection take $O(M)$ space and work. For fixed coordinate count p, all observation processing and fitting take $\widetilde{O}(M + \varepsilon^{-2})$ operations; the calibration LP adds $\mathrm{poly}(M, p, \log(1/\delta))$ bit complexity. Taking $\varepsilon = M^{-q}$ for a sufficiently large fixed q also makes the LP lower order than the fitting cost, yielding a full-resource separation from uniform-library sampling.

What this construction establishes. Data are needed to discover which public root action provides the useful coverage. The $\Omega(M)$ search bound does not make the LP the only way to exploit observations. A task-directed method can test each root with $\lceil \log_2(2/\delta) \rceil$ uniform-action episodes, then fit the informative matrix exclusively at the identified root. This gives $O(M \log(1/\delta)) + \widetilde{O}(\varepsilon^{-2})$ episodes, matching the calibrated method up to logarithmic factors. Our empirical-return baseline uses the shared calibration data rather than a separately tuned search budget. Event calibration additionally targets both latent states, supporting reward-free coverage; the return-directed sampler may ignore the zero-value state. After identifying the entrance, mixing it with any other root also covers both states at this order. Hence the instance establishes sample-optimal adaptation, not an

advantage unique to event optimization, a general calibration lower bound, or optimal computational complexity.

D A COMPUTABLE HIDDEN-COORDINATE INSTANCE AND A DIRECT BASELINE

Representation and proper models. Let observations be $x \in \{0,1\}^p$, let $S = 2$, and consider $\Phi = \{\phi_i(x) = x_i : i \in [p]\}$. At stage h, emission $O_{h,i}(\cdot \mid z)$ fixes bit i to z and draws the remaining bits independently and uniformly. These distributions have disjoint supports for each candidate decoder. The unknown true coordinate is in $[p]$; known rewards $r_h(x, a, b) \in [0, 1]$ are efficiently evaluable and need not factor through every candidate coordinate. Each latent transition row is Bernoulli with probability in $\{0, 1/K, \ldots, 1\}$, and the true rows belong to this grid. A stagewise proper model class has size

$$\Theta_\star \le p(K+1)^{2J}.$$

The algorithm never enumerates the 2^p observation space. Sampling a candidate model draws its next bit and $p-1$ fair bits; exact sampling and probability evaluation take polynomial bit time in $p + \log K$, with expected $O(p + \log K)$ random-bit cost.

Proposition 17 (Computable oracle instance)**.** *For the hidden-coordinate family, realizable two-context squared-loss regression, proper transition MLE, and quantized Bellman ERM can all be implemented in time polynomial in the sample size, p, J, H, the input precision, and inverse approximation accuracy. For any bounded predictable continuation, the Bellman class described below has uniform approximation error at most τ and*

$$\log Q_\star \le \log p + 2J \log(2 + H/\tau).$$

With the accuracy schedule of Corollary 37, a constant-factor increase in n_{pe}, and $\tau \le u/\sqrt{6}$, the same ε-Nash guarantee and episode order hold with these explicit class sizes. Computation includes the source PCE cost and the matrix work per policy query, rather than a unit-cost call to an unrestricted oracle.

Proof. For two-context regression enumerate $i \in [p]$, group samples by $(x_i, x_i') \in \{0,1\}^2$, and fit each of the four head entries by its clipped sample mean. Select the coordinate with minimum squared loss. This is exact ERM over $\{f(x_i, x_i') : i \in [p], f \in [0,1]^4\}$, costing $O(pn)$ arithmetic operations. A finite grid of these four entries and bounded-loss uniform convergence give polynomial labeled-sample efficiency in $1/\alpha, \log(p/\beta)$, as required by the realizable PCE interface. The continuous ERM is at least as good empirically as this grid, and its four-parameter heads admit the same uniform-convergence bound by rounding.

For MLE, enumerate the coordinate and group successor-bit counts by (x_i, a, b). Within each row maximize $n_1 \log t + n_0 \log(1-t)$ over grid probabilities t. Concavity shows that only the grid points immediately below and above the empirical frequency need comparison. Empty rows may use a fixed grid point; $0 \log 0 = 0$, and a positive count with zero probability has log likelihood $-\infty$. The emission contribution is the common constant $-n(p-1)\log 2$. Comparing the best rowwise scores for all coordinates therefore returns the exact proper MLE in $O(p(n+J))$ arithmetic operations, up to numerical log precision. Exact candidate comparisons can alternatively use integer likelihood numerators of $O(n \log K)$ bits, since all denominators are common; this gives polynomial bit complexity without a real-arithmetic oracle. An allowed total objective error enters ξ_{mle} in Lemma 20.

For any candidate kernel with coordinate i and any $v \in [0, H_{h+1}]$, its conditional expectation is $P_h v(x, a, b) = g(x_i, a, b)$ for some $g \in [0, H_{h+1}]^{2J}$. Keep the known reward separate and set

$$\mathcal{Q}_h^\tau = \{r_h(x, a, b) + g(x_i, a, b) : i \in [p],\ g \in \mathcal{G}_h^{2J}\},$$

where $\mathcal{G}_h$ covers $[0, H_{h+1}]$ at spacing at most τ and contains both endpoints. These predictors already lie in $[0, H_h]$. Rounding g gives uniform error at most τ and the stated cardinality bound. For each coordinate and each cell, quantize the sample mean of $Y - r_h(X)$ to its nearest grid value, then compare empirical risks. This is exact ERM in $O(p(n+J))$ operations. Keeping r_h outside the cell table is essential when a wrong decoder does not represent the reward. At the terminal stage the target is known exactly.

For completeness, the second-moment lemma extends to this approximate class. In its notation, Bernstein and $2ab \le a^2/2 + 2b^2$ give $|\widehat{\mathbb{E}}Z_q - v_q| \le v_q/2 + cH^2 \log(2Q_\star/\delta')/n$ simultaneously over all candidates, although $q^\star$ need not belong to the class. Compare the ERM with a τ-approximation q_0:

$$\tfrac{1}{2}v_{\widehat{q}} - cH^2L/n \le \widehat{\mathbb{E}}Z_{\widehat{q}} \le \widehat{\mathbb{E}}Z_{q_0} \le \tfrac{3}{2}\tau^2 + cH^2L/n.$$

Thus $v_{\widehat{q}} \le 3\tau^2 + 4cH^2L/n$. Choose the statistical term at most $u^2/2$ and $3\tau^2 \le u^2/2$. This proves the required PE event and completes the finite-class rate with explicitly computable fits. □

Substituted sample complexity. For the uncalibrated uniform-cover baseline with $S = 2$, choose τ proportional to $\varepsilon/(H^2\sqrt{J})$, and let the MLE objective error be zero (or include it explicitly). Then the post-cover episode order is

$$O\left(\frac{H^7 J}{\varepsilon^2}\left[\log\frac{Hp}{\delta} + J\log(K+1) + J\log\left(2 + \frac{H^3\sqrt{J}}{\varepsilon}\right)\right]\right).$$

For real-successor fitted minimax, delete the $J\log(K+1)$ model term. For the main calibrated method, Lemma 8 gives $R \le 2$. Taking τ proportional to $\varepsilon/(H\sqrt{J})$ in Corollary 11 instead gives the full bound

$$N_{\mathrm{PCE}} + O\left(\frac{H^5 L_{\mathrm{cal}}}{\varepsilon} + \frac{H^5 J}{\varepsilon^2}\left[\log\frac{Hp}{\delta} + J\log(K+1) + J\log\left(2 + \frac{H^2\sqrt{J}}{\varepsilon}\right)\right]\right).$$

Its cover confidence is $\delta/4$; the uncalibrated baseline below uses $\delta/3$. These allocations change only constant factors in the inherited provider bound. The H^7 display is retained as the comparator, not the complexity of the learned-mixture construction. The finite-grid uniform-convergence argument above gives the conservative two-context sample efficiency $N_{2\mathrm{ctx}}(\alpha,\beta) = O(\alpha^{-2}[\log p + \log(1/\alpha) + \log(1/\beta)])$. Substituting it into (PCE) yields $\mathcal{N}_{\mathrm{PCE}} = O(\Gamma^{3C_{\mathrm{PCE}}}[\log p + C_{\mathrm{PCE}}\log\Gamma])$, with $\Gamma = HJS/(\eta\delta_{\mathrm{c}})$, $\eta = \varepsilon/(32SH^2)$, and $\delta_{\mathrm{c}} = \delta/3$. This source-theorem bound is polynomial in inverse confidence, not a claimed logarithmic dependence on it. These conservative exponents are separate from the practical PCE settings.

Query-time cost. Returned predictors store a coordinate and $2J$ head entries. A matrix query evaluates J reward and head values and runs the explicit saddle updates, in $O(TJ)$ arithmetic operations plus reward evaluation. Sampling labels at one stage makes n_{pe} continuation-policy queries. These operations and the cost of encoding p bits must be counted; none grows as 2^p. Exact local linear programs are an alternative for the finite diagnostics, with their solve time charged experimentally.

Corollary 18 (Real-successor fitted minimax with uniform cover weights)**.** *Under the same cover and independent stage-buffer conditions, use uniform cover weights, collect true successors and fit labels $Y = r_h(X) + \widehat{V}_{h+1}(S_{h+1})$, $S_{h+1} \sim P_h^\star(\cdot \mid X)$, followed by the same local saddle updates. If $\|\widehat{Q}_h - r_h - P_h^\star\widehat{V}_{h+1}\|_{L_2(\rho_h)} \le u_h$, its gap satisfies*

$$\mathrm{Gap}_{P^\star}(\widehat{\pi},\widehat{\nu}) \le 2\sum_{h=1}^{H}(2SH_h\eta + \sqrt{2MJ}\,u_h + \delta_h).$$

For the hidden-coordinate class this condition is realizable up to the same τ, and the schedule above uses $\mathcal{N}_{\mathrm{PCE}} + O(H^7S^2JL_Q/\varepsilon^2)$ episodes, without MLE.

Proof. Condition on the cover and later-stage fits. Each held-out real successor then has the stated conditional mean. Since $P^\star$ belongs to the coordinate model family, the same $r + g \circ \phi_i$ closure and quantized ERM argument apply. Theorem 3 at $P^\star$ and the Block cover inequality give the bound directly. Substituting the same $\eta, u, \delta_{\mathrm{mat}}$ and counting Hn_{pe} episodes proves the rate. □

This baseline is sufficient for equilibrium learning in the concrete family. The model-based route additionally returns a reusable generative kernel and allows repeated simulated labels or exact integration. Neither our theorem nor the oracle implementation establishes a statistical advantage for that route over real-successor fitting.

Method	Input/structure	Computation	Target and charged cost
Amortila et al. (2024), model route	Realizable model class; bounded coverability	Model estimation and coverage-objective planning	Explore the policy space; charge estimation and exploratory interaction
Amortila et al. (2024), model-free route	Weight-function and value classes	Density-ratio estimation and reward-directed policy optimization	Generate exploratory policies; charge the exploration procedure
Mutti et al. (2022)	Policy space and occupancy-divergence objective	Representative-set optimization; game-theoretic planning reformulation	Compress policies under Rényi-divergence coverage
This work	Paid finite library; common Block emissions; finite evaluable decoder class	Observable event counts and one finite LP per stage; no new exploration-policy search	Convert max-visitation coverage into a sampling law; charge provider, calibration, fitting, and matrix computation

Table 1: Coverage interfaces. These rows solve different exploration problems and do not by themselves establish total-rate dominance.

E Coverage optimization and finite-library conversion

The main comparison concerns the operations needed to construct a sampling distribution. Table 1 distinguishes that problem from generating the original policy library. All costs of the supplied provider remain in our total budget.

Amortila et al. (2024) already connect coverage to second-moment prediction errors in their Proposition 3.1. Our contribution is the finite-library conversion: common emissions make observable-event and latent occupancy constraints equivalent, and calibrated finite counts make those constraints computable without identifying the decoder. The relevant tradeoff is less planning after a cover is supplied, in exchange for exact Block structure and enumerating its candidate events. Statistical logarithms depend on $\log F$, but frequency evaluation and LP size depend on F. Neither a succinct description of an enormous decoder class nor an abstract regression oracle guarantees efficient enumeration. Section 5 provides one explicit implementation of the required fitting operations.

The LP is a polynomial optimization mechanism for the event constraints; it need not be the only successful solver. A heuristic that satisfies the same checked numerical conditions receives the same conditional coverage conclusion on the concentration event. Our greedy baseline uses the same empirical objective and calibration table. Its fixed iteration limit does not itself provide the LP's general optimization guarantee, while its empirical performance is reported without requiring the LP to outperform it on every instance.

E.1 A comparison with identical inputs and access

Fix one supplied library, the same evaluable candidate decoders, and access only to its roll-in observations. Neither method below receives latent labels, occupancies, a transition model, or additional policies. Write $p_i(z)$ for the latent occupancies, $q_w = \sum_i w_i p_i$, and $D_{\phi,y,i}$ for their observable event probabilities. Our population objective maximizes the common domination factor

$$t(w) = \min_{z:\max_i p_i(z)>0} \frac{q_w(z)}{\max_i p_i(z)} = \min_{\phi,y:\max_i D_{\phi,y,i}>0} \frac{\sum_j w_j D_{\phi,y,j}}{\max_i D_{\phi,y,i}}.$$

Given the event table, this is a finite matrix game with entries $D_{\phi,y,i}/\max_j D_{\phi,y,j}$, so standard optimization suffices. The statistical task is to obtain a reliable table and transfer its constraints to true coverage.

The finite-library, state-occupancy specialization of the regularized L_1 objective of Amortila et al. (2024) is

$$\mathcal{C}_\alpha(w) = \max_i \sum_z \frac{p_i(z)^2}{q_w(z) + \alpha p_i(z)}, \qquad \alpha > 0.$$

Appending the same uniform joint action to every prefix leaves this expression unchanged. This specialization also admits an observable implementation under the present assumptions.

Proposition 19 (Observable finite-library L_1 objective)**.** *Under the common-emission and decoder-inclusion conditions of Lemma 8, for every $w \in \Delta_M$,*

$$\mathcal{C}_\alpha(w) = \max_{\phi\in\Phi, i\in[M]} \sum_y \frac{D^2_{\phi,y,i}}{\sum_j w_j D_{\phi,y,j} + \alpha D_{\phi,y,i}},$$

where $0/0 = 0$. The right-hand side is convex in w and can be estimated using the same calibration frequency table as our LP.

Proof. Let $a_y(z) = \Pr\{\phi(S_h) = y \mid Z_h = z\}$, so $D_{\phi,y,i} = \sum_z a_y(z)p_i(z)$ and $\sum_y a_y(z) = 1$. Weighted Cauchy–Schwarz gives

$$\frac{(\sum_z a_y(z)p_i(z))^2}{\sum_z a_y(z)(q_w(z) + \alpha p_i(z))} \le \sum_z \frac{a_y(z)p_i(z)^2}{q_w(z) + \alpha p_i(z)}.$$

Sum over y. Equality holds for the included true decoder, proving the identity after maximizing over ϕ, i. Each summand is a nonnegative constant divided by a positive affine function, hence convex; zero numerators can be omitted. Taking a maximum preserves convexity. □

Thus lack of latent labels does not prevent a finite-library L_1 comparator. Our experiments minimize its empirical version at fixed $\alpha = 0.01$ with exactly the same observations, charging its own optimization time. This is a specialization of the coverage objective, not a reimplementation of the original full exploration algorithm.

The two objectives need not select the same weights. If $t(w) > 0$, $\mathcal{C}_\alpha(w) \le 1/t(w)$, but bounded average coverage need not bound worst-case domination. For example, let $p_1 = (r, 1-r)$, $p_2 = (0, 1)$, and $w = (r, 1-r)$, with $0 < r \le 1/2$. Then $1/t(w) = 1/r$, whereas $\mathcal{C}_0(w) = \max\{2/(1+r), 1/(1-r^2)\} \le 2$; positive regularization only decreases this objective. This distinguishes the coverage requirements, not their universal merit.

Our guarantee supplies a relative-plus-additive calibration schedule, checkable tolerances, and a resulting Nash budget for the domination constraints. It does not prove that the L_1 comparator needs more samples or computation. If any comparator meets the same numerical conditions, it inherits the same conditional guarantee. The original PCE cost is common to both methods and remains charged; its two-context regression oracle does not by itself provide enumeration of Φ. Both finite-table implementations here explicitly require that access.

F Public information, shared calibration, and unknown coverage

The comparison contains 2,016 method–budget–seed records: 720 public-entrance records, 1,008 unknown-entrance records, and 288 coupled PCE records. The first two families use seeds $0, \dots, 15$ and the PCE family uses seeds $0, \dots, 7$. A separate two-seed implementation pilot is excluded. All methods fit real-successor labels and solve their local matrices exactly. Thus the comparisons isolate collection laws rather than differences between model-based and direct fitting.

F.1 Reweighting the same observations

For each stage let D be the matrix of empirical candidate-event frequencies, with one column per library member, and let $c_e = \max_i D_{ei}$. The LP maximizes

$$f(w) = \min\left\{1, \min_{e:c_e>0} \frac{(Dw)_e + 2\beta}{c_e}\right\}, \qquad w \in \Delta_M,$$

with $\beta = 0.005$. We compare the following observable rules.

- **Event max:** select the first maximizing column for each event and mix these selected columns, retaining multiplicities.
- **Event subset:** deduplicate the selected columns and mix the resulting subset uniformly.

- **Greedy:** start at the uniform mixture. At iteration $k = 1, \ldots, 256$, choose the column maximizing $f((1-\gamma_k)w + \gamma_k e_i)$, where $\gamma_k = 1/(k+1)$. Return the best iterate, including the initial uniform mixture.

These rules and the LP receive exactly the same calibration frequency table. Each hypothetical run is charged its full calibration collection cost and its own optimization time. Greedy optimizes the same objective; we do not claim that 256 iterations meet the theorem's numerical tolerances. Any output that passes those tolerances inherits the conditional coverage guarantee, independently of which solver produced it. In the 96 calibrated stage problems across the PCE runs, LP and greedy both attain empirical objective at least $7/(10R) = 0.35$ with $R = 2$; event max and event subset do so in 69 cases, and equal weights in 93. These are numerical checks at the experimental slack, not evidence that the practical calibration counts satisfy the theorem's concentration schedule. They further support distinguishing the event constraints from the particular optimizer used to enforce them.

F.2 Public entrance: use the information already available

We use the distinct-policy family in Proposition 14, with $p = 6$, $M \in \{8, 32, 128\}$, and stage-two signal $4\varepsilon = 0.04$. Total episode caps are $8,192$, $32,768$, and $131,072$. The LP pays $32M$ calibration episodes. Uniform-library sampling, uniform actions, and the following two baselines spend the whole cap on real-successor fitting:

- **Public best:** use the library member with largest public entry probability.
- **Public endpoints:** use a mixture of the largest and smallest public entry probabilities that gives latent-state visitation $1/2$ to each state.

The policies' public descriptions suffice for these choices; neither baseline reads a hidden state or estimates an unknown transition. Public best has zero minimum two-state coverage, because it ignores the zero-value state. Its small Nash gap illustrates why a reward-free coverage objective can be stronger than this task requires.

At $M = 128$ and $131,072$ episodes, LP, public best, and public endpoints have respective gaps 0.00430 ± 0.00059, 0.00480 ± 0.00041, and 0.00437 ± 0.00067. At $8,192$ episodes, they have 0.03554 ± 0.00386, 0.01751 ± 0.00196, and 0.01469 ± 0.00197. Calibration consumes half of the small cap and provides no information needed by the public alternatives. This family supports the lower bound against uniform library sampling, but not a claim that calibration is necessary when the entrance is public.

F.3 Unknown entrance: the transition kernel carries the uncertainty

We implement Proposition 16 with $p = 6$ and the same M, caps, and signal. The entrance is drawn using a separate seed stream; the hidden coordinate cycles over the six candidates and the sign alternates across seeds. The learner sees the M root-action labels, candidate decoders, and sampled observation–action–outcome tuples. It does not receive entry probabilities or entrance metadata. The uninformative initial latent draw is marginalized by the simulator. The M root actions are part of the instance size.

LP, event max, event subset, and greedy receive the same 32 episodes per root action. Their returned root policy selects the largest empirical terminal return from these trajectories. The **empirical-return** baseline uses the same root choice and the same charged calibration data, then samples that root exclusively for fitting. **Uniform library** pays no calibration and estimates its root decision from its own fitting trajectories. **Fixed first** commits to public root label zero before seeing any data and uses it for collection. No method is given the true entrance as a training input. Fresh fitting buffers are generated separately for each budget.

The evaluator integrates over the true emission distribution and computes both unilateral best responses. An incorrect returned root therefore contributes to the full-game Nash gap; it is not removed by conditioning the evaluation on a successful entrance. At $M = 128$ and $131,072$ episodes, LP attains 0.00385 ± 0.00073, uniform library 0.03746 ± 0.00605, greedy 0.00534 ± 0.00079, and empirical return 0.00544 ± 0.00038. The latter samples only the useful state and has zero minimum

two-state coverage. These results establish a benefit from learning where to collect data, while showing that a task-directed alternative can also exploit the calibration trajectories.

F.4 Coupled PCE covers under complete budget caps

The hidden-progress game and PCE settings match Appendix H: $H \in \{6, 10\}$, $p = 4$, regression batch 64, test batch 24, 24 candidate centers, $\gamma = 0.15$, $\gamma' = 0.4$, and $2H$ exploration rounds. Each relevant cover member receives 64 calibration episodes. For each seed and reference allocation $n \in \{1,024, 4,096, 16,384\}$, the common episode cap is

$$N_{\mathrm{cap}} = N_{\mathrm{PCE}} + N_{\mathrm{cal}} + 2(H-1)n.$$

All calibrated methods pay $N_{\mathrm{PCE}} + N_{\mathrm{cal}}$; the equal mixture pays N_{PCE}; uniform actions pay neither. Each spends its remainder on real-successor fitting, with fewer than $H-1$ unused episodes due to integer stagewise buffers. The factor two defines the inherited budget grid and does not charge two episodes per direct label.

The same cover and calibration seeds are used across reweighting rules. Every mixture, including the equal mixture, uses categorical sampling. The model/direct control implementation in Appendix H uses integer indexing for the equal mixture: it induces the same distribution but consumes a different random stream. Those control tables therefore remain separate, and their equal-mixture gap means are not substituted into these comparisons.

At $H = 10$, the mean minimum visitation ratios are 0.573 (LP), 0.505 (greedy), 0.281 (event max), 0.277 (event subset), 0.423 (equal), and 0.076 (uniform actions). At the largest cap, their gaps are respectively 0.00829, 0.00700, 0.01166, 0.01045, 0.00856, and 0.00699. PCE costs 608,224 episodes and calibration 17,720 on average. LP takes 344.41 seconds including construction, calibration, optimization, and fitting; uniform actions take 162.93 seconds. Thus improved worst-case coverage does not imply the best observed gap or runtime after all costs are counted.

Uncertainty, costs, and reproducibility. Tables report means and standard errors across independent seeds. The artifact also stores within-seed LP-minus-comparator differences and two-sided 95% Student intervals, without a multiple-comparison correction. The largest-budget LP-minus-equal intervals are $[-0.00559, 0.00099]$ for $H = 6$ and $[-0.00289, 0.00235]$ for $H = 10$. Both contain zero. We make no familywise significance claim, and no seeds were added in response to the observed comparisons. Runtime includes the provider, calibration collection, each method's weight optimization, fitting, and local matrix solves; exact diagnostic evaluation is excluded for all methods. Shared data are reused to make the comparison paired, but their cost is charged in full to each method. The exporter checks row uniqueness, seed counts, episode decomposition, and integer remainders across the recorded configurations.

F.5 Complete public-entrance results

Table 2: Public entrance: 16 seeds; complete episode budgets; gap means $\pm$ standard errors. Coverage is the minimum visitation ratio; seconds include output matrix computation.

M	Episodes	Method	Nash gap	Coverage	Seconds
8	8,192	Uniform library	0.0303 ± 0.0079	0.1301	0.012
8	8,192	LP	0.0185 ± 0.0020	0.4901	0.037
8	8,192	Public best	0.0175 ± 0.0020	0.0000	0.015
8	8,192	Public endpoints	0.0147 ± 0.0020	0.5000	0.012
8	8,192	Uniform actions	0.0159 ± 0.0026	0.5000	0.012
8	32,768	Uniform library	0.0204 ± 0.0023	0.1301	0.029
8	32,768	LP	0.0102 ± 0.0012	0.4901	0.051
8	32,768	Public best	0.0078 ± 0.0008	0.0000	0.030
8	32,768	Public endpoints	0.0080 ± 0.0009	0.5000	0.028
8	32,768	Uniform actions	0.0075 ± 0.0012	0.5000	0.027
8	131,072	Uniform library	0.0082 ± 0.0012	0.1301	0.096
8	131,072	LP	0.0044 ± 0.0005	0.4901	0.117

Continued on next page

Table 2 continued

M	Episodes	Method	Nash gap	Coverage	Seconds
8	131,072	Public best	0.0048 ± 0.0004	0.0000	0.095
8	131,072	Public endpoints	0.0044 ± 0.0007	0.5000	0.093
8	131,072	Uniform actions	0.0045 ± 0.0003	0.5000	0.092
32	8,192	Uniform library	0.1215 ± 0.0188	0.0317	0.012
32	8,192	LP	0.0183 ± 0.0023	0.4939	0.061
32	8,192	Public best	0.0175 ± 0.0020	0.0000	0.015
32	8,192	Public endpoints	0.0147 ± 0.0020	0.5000	0.012
32	8,192	Uniform actions	0.0159 ± 0.0026	0.5000	0.012
32	32,768	Uniform library	0.0327 ± 0.0041	0.0317	0.028
32	32,768	LP	0.0108 ± 0.0014	0.4939	0.077
32	32,768	Public best	0.0078 ± 0.0008	0.0000	0.031
32	32,768	Public endpoints	0.0080 ± 0.0009	0.5000	0.027
32	32,768	Uniform actions	0.0075 ± 0.0012	0.5000	0.027
32	131,072	Uniform library	0.0184 ± 0.0023	0.0317	0.096
32	131,072	LP	0.0042 ± 0.0005	0.4939	0.142
32	131,072	Public best	0.0048 ± 0.0004	0.0000	0.096
32	131,072	Public endpoints	0.0044 ± 0.0007	0.5000	0.092
32	131,072	Uniform actions	0.0045 ± 0.0003	0.5000	0.092
128	8,192	Uniform library	0.1338 ± 0.0141	0.0078	0.013
128	8,192	LP	0.0355 ± 0.0039	0.4990	0.205
128	8,192	Public best	0.0175 ± 0.0020	0.0000	0.015
128	8,192	Public endpoints	0.0147 ± 0.0020	0.5000	0.012
128	8,192	Uniform actions	0.0159 ± 0.0026	0.5000	0.012
128	32,768	Uniform library	0.0816 ± 0.0111	0.0078	0.029
128	32,768	LP	0.0101 ± 0.0016	0.4990	0.221
128	32,768	Public best	0.0078 ± 0.0008	0.0000	0.030
128	32,768	Public endpoints	0.0080 ± 0.0009	0.5000	0.028
128	32,768	Uniform actions	0.0075 ± 0.0012	0.5000	0.028
128	131,072	Uniform library	0.0367 ± 0.0039	0.0078	0.098
128	131,072	LP	0.0043 ± 0.0006	0.4990	0.286
128	131,072	Public best	0.0048 ± 0.0004	0.0000	0.095
128	131,072	Public endpoints	0.0044 ± 0.0007	0.5000	0.092
128	131,072	Uniform actions	0.0045 ± 0.0003	0.5000	0.092

F.6 Complete unknown-entrance results

Table 3: Unknown entrance: 16 seeds; complete episode budgets; gap means $\pm$ standard errors. Coverage is the minimum visitation ratio; seconds include output matrix computation.

M	Episodes	Method	Nash gap	Coverage	Seconds
8	8,192	Uniform library	0.0475 ± 0.0070	0.1250	0.010
8	8,192	Event max	0.0307 ± 0.0042	0.2083	0.012
8	8,192	Event subset	0.0341 ± 0.0060	0.1677	0.012
8	8,192	Fixed first	0.6650 ± 0.0000	0.0000	0.007
8	8,192	Greedy	0.0205 ± 0.0020	0.4995	0.019
8	8,192	LP	0.0201 ± 0.0018	0.5000	0.019
8	8,192	Empirical return	0.0189 ± 0.0014	0.0000	0.014
8	32,768	Uniform library	0.0172 ± 0.0024	0.1250	0.028
8	32,768	Event max	0.0118 ± 0.0016	0.2083	0.029
8	32,768	Event subset	0.0134 ± 0.0018	0.1677	0.030
8	32,768	Fixed first	0.6650 ± 0.0000	0.0000	0.024
8	32,768	Greedy	0.0081 ± 0.0011	0.4995	0.036
8	32,768	LP	0.0071 ± 0.0010	0.5000	0.035
8	32,768	Empirical return	0.0083 ± 0.0006	0.0000	0.032
8	131,072	Uniform library	0.0078 ± 0.0009	0.1250	0.101
8	131,072	Event max	0.0085 ± 0.0010	0.2083	0.101
8	131,072	Event subset	0.0068 ± 0.0011	0.1677	0.100
8	131,072	Fixed first	0.6650 ± 0.0000	0.0000	0.095

Continued on next page

Table 3 continued

M	Episodes	Method	Nash gap	Coverage	Seconds
8	131,072	Greedy	0.0039 ± 0.0004	0.4995	0.107
8	131,072	LP	0.0040 ± 0.0004	0.5000	0.109
8	131,072	Empirical return	0.0044 ± 0.0004	0.0000	0.102
32	8,192	Uniform library	0.0744 ± 0.0121	0.0312	0.010
32	8,192	Event max	0.0611 ± 0.0091	0.0990	0.012
32	8,192	Event subset	0.0447 ± 0.0060	0.1026	0.012
32	8,192	Fixed first	0.6650 ± 0.0000	0.0000	0.007
32	8,192	Greedy	0.0152 ± 0.0027	0.4999	0.016
32	8,192	LP	0.0176 ± 0.0030	0.5000	0.014
32	8,192	Empirical return	0.0214 ± 0.0021	0.0000	0.015
32	32,768	Uniform library	0.0396 ± 0.0068	0.0312	0.028
32	32,768	Event max	0.0214 ± 0.0028	0.0990	0.030
32	32,768	Event subset	0.0186 ± 0.0025	0.1026	0.030
32	32,768	Fixed first	0.6650 ± 0.0000	0.0000	0.024
32	32,768	Greedy	0.0092 ± 0.0014	0.4999	0.034
32	32,768	LP	0.0084 ± 0.0010	0.5000	0.033
32	32,768	Empirical return	0.0096 ± 0.0008	0.0000	0.033
32	131,072	Uniform library	0.0130 ± 0.0020	0.0312	0.103
32	131,072	Event max	0.0098 ± 0.0017	0.0990	0.103
32	131,072	Event subset	0.0082 ± 0.0010	0.1026	0.102
32	131,072	Fixed first	0.6650 ± 0.0000	0.0000	0.096
32	131,072	Greedy	0.0042 ± 0.0007	0.4999	0.105
32	131,072	LP	0.0036 ± 0.0006	0.5000	0.106
32	131,072	Empirical return	0.0047 ± 0.0005	0.0000	0.103
128	8,192	Uniform library	0.1409 ± 0.0205	0.0078	0.010
128	8,192	Event max	0.0804 ± 0.0116	0.1042	0.018
128	8,192	Event subset	0.0811 ± 0.0135	0.0888	0.018
128	8,192	Fixed first	0.6650 ± 0.0000	0.0000	0.007
128	8,192	Greedy	0.0260 ± 0.0046	0.4998	0.024
128	8,192	LP	0.0207 ± 0.0029	0.5000	0.021
128	8,192	Empirical return	0.0323 ± 0.0021	0.0000	0.021
128	32,768	Uniform library	0.0968 ± 0.0165	0.0078	0.029
128	32,768	Event max	0.0191 ± 0.0032	0.1042	0.037
128	32,768	Event subset	0.0180 ± 0.0022	0.0888	0.037
128	32,768	Fixed first	0.6650 ± 0.0000	0.0000	0.024
128	32,768	Greedy	0.0092 ± 0.0009	0.4998	0.042
128	32,768	LP	0.0089 ± 0.0010	0.5000	0.040
128	32,768	Empirical return	0.0097 ± 0.0014	0.0000	0.039
128	131,072	Uniform library	0.0375 ± 0.0061	0.0078	0.105
128	131,072	Event max	0.0094 ± 0.0014	0.1042	0.110
128	131,072	Event subset	0.0115 ± 0.0015	0.0888	0.109
128	131,072	Fixed first	0.6650 ± 0.0000	0.0000	0.095
128	131,072	Greedy	0.0053 ± 0.0008	0.4998	0.113
128	131,072	LP	0.0039 ± 0.0007	0.5000	0.114
128	131,072	Empirical return	0.0054 ± 0.0004	0.0000	0.110

F.7 Complete shared-calibration PCE results

Table 4: Coupled PCE games: eight seeds; shared calibration; all collection, fitting, and output matrix costs. Gaps are means ± standard errors.

H	n	Method	Nash gap	Coverage	Episodes	Seconds
6	1,024	LP	0.0279 ± 0.0027	0.5635	146,800.0	32.05
6	1,024	Greedy	0.0255 ± 0.0032	0.5202	146,800.0	32.06
6	1,024	Event max	0.0347 ± 0.0064	0.3008	146,800.0	32.03
6	1,024	Event subset	0.0292 ± 0.0030	0.3270	146,800.0	32.04
6	1,024	Uniform library	0.0262 ± 0.0020	0.4758	146,798.5	32.03
6	1,024	Uniform actions	0.0084 ± 0.0011	0.2517	146,798.1	17.08
6	4,096	LP	0.0129 ± 0.0013	0.5635	177,520.0	38.21

Continued on next page

Table 4 continued

H	n	Method	Nash gap	Coverage	Episodes	Seconds
6	4,096	Greedy	0.0112 ± 0.0012	0.5202	177,520.0	38.23
6	4,096	Event max	0.0170 ± 0.0024	0.3008	177,520.0	38.20
6	4,096	Event subset	0.0171 ± 0.0024	0.3270	177,520.0	38.20
6	4,096	Uniform library	0.0144 ± 0.0014	0.4758	177,518.5	38.19
6	4,096	Uniform actions	0.0076 ± 0.0006	0.2517	177,518.1	20.66
6	16,384	LP	0.0056 ± 0.0007	0.5635	300,400.0	62.85
6	16,384	Greedy	0.0047 ± 0.0002	0.5202	300,400.0	62.88
6	16,384	Event max	0.0077 ± 0.0007	0.3008	300,400.0	62.93
6	16,384	Event subset	0.0077 ± 0.0009	0.3270	300,400.0	62.80
6	16,384	Uniform library	0.0079 ± 0.0012	0.4758	300,398.5	62.85
6	16,384	Uniform actions	0.0058 ± 0.0004	0.2517	300,398.1	34.97
10	1,024	LP	0.0335 ± 0.0021	0.5726	644,376.0	250.01
10	1,024	Greedy	0.0313 ± 0.0020	0.5045	644,376.0	250.05
10	1,024	Event max	0.0386 ± 0.0019	0.2807	644,376.0	250.02
10	1,024	Event subset	0.0401 ± 0.0017	0.2769	644,376.0	250.01
10	1,024	Uniform library	0.0294 ± 0.0031	0.4232	644,372.5	250.04
10	1,024	Uniform actions	0.0100 ± 0.0012	0.0763	644,371.9	114.22
10	4,096	LP	0.0164 ± 0.0009	0.5726	699,672.0	268.91
10	4,096	Greedy	0.0151 ± 0.0009	0.5045	699,672.0	268.95
10	4,096	Event max	0.0215 ± 0.0019	0.2807	699,672.0	268.94
10	4,096	Event subset	0.0198 ± 0.0016	0.2769	699,672.0	268.93
10	4,096	Uniform library	0.0167 ± 0.0016	0.4232	699,668.5	268.91
10	4,096	Uniform actions	0.0090 ± 0.0012	0.0763	699,667.9	123.97
10	16,384	LP	0.0083 ± 0.0006	0.5726	920,856.0	344.41
10	16,384	Greedy	0.0070 ± 0.0007	0.5045	920,856.0	344.59
10	16,384	Event max	0.0117 ± 0.0011	0.2807	920,856.0	344.77
10	16,384	Event subset	0.0105 ± 0.0015	0.2769	920,856.0	344.53
10	16,384	Uniform library	0.0086 ± 0.0007	0.4232	920,852.5	344.66
10	16,384	Uniform actions	0.0070 ± 0.0009	0.0763	920,851.9	162.93

G FIXED CALIBRATION BUDGETS, RESPONSE ERRORS, AND SHARED TASKS

The study contains 704 scan records and 336 reward-task records. All settings, seeds, and tasks were fixed before formal execution. We use $H = 10$, $p = 4$, and seeds $0, \ldots, 7$ in the same coupled hidden-progress game and with the same PCE settings as Appendix F. The provider's mean episode count is 608,224. Its returned predictors are serialized without executable objects, and their actions are checked against the original policies. The library is then fixed throughout all comparisons for that seed.

G.1 COMPLETE AND EQUAL-FITTING BUDGETS

Let $m_{\mathrm{cal}} = \sum_{h=2}^{H-1} M_h$. The cap is fixed as

$$B_n = N_{\mathrm{PCE}} + 64 m_{\mathrm{cal}} + 2(H-1)n, \qquad n \in \{4,096, 16,384\}.$$

The number 64 in this definition is unchanged when calibration varies over $c \in \{16, 32, 64, 128\}$. A calibrated method spends $N_{\mathrm{PCE}} + c m_{\mathrm{cal}}$ on its overhead and uses $\lfloor (B_n - N_{\mathrm{PCE}} - c m_{\mathrm{cal}})/(H-1) \rfloor$ fitting episodes per nonterminal stage. Equal PCE pays only its provider; uniform actions pay neither overhead. Rounding leaves fewer than $H - 1$ unused episodes. At the larger cap, the mean LP decomposition at $c = 64$ is 608,224 provider, 17,720 calibration, and 294,912 fitting episodes. The calibration share is 1.92% of the total.

The diagnostic slice instead gives every method

$$n_{\mathrm{equal}} = \left\lfloor \frac{B_n - N_{\mathrm{PCE}} - 128 m_{\mathrm{cal}}}{H-1} \right\rfloor$$

fitting episodes per stage. It reports the unspent remainder; it is an equal-fitting comparison, not an equal-spend efficiency claim.

Each member's calibration counts are prefixes of the same saved full observation–action trajectories. All five reweighting rules see those same observations. LP, greedy, event max, and event subset use the definitions in Appendix F, with $\beta = 0.005$. The L_1 comparator minimizes the empirical objective in Proposition 19, with $\alpha = 0.01$, through a convex epigraph program (SLSQP, tolerance 10^{-10}, at most 2,000 iterations). It receives no true decoder or additional trajectories.

Across the 256 distinct calibrated stage problems per method, LP, greedy, and L_1 each meet the empirical $t \geq 0.35$ check in every case. Event max fails this check in 80 problems and event subset in 67. These failures are retained in the results. The counts remain practical counts rather than the theorem's statistical schedule. The protocol also retains empty provider stages and optimizer exceptions, with declared uniform/equal fallbacks and all attempted work charged. Neither fallback occurred in the formal batch. A reduced-budget pilot at $H = 3$ and seed 100 exposed empty provider stages; its failure and corrected implementation checks are retained separately and excluded from tables.

G.2 Residuals under the actual dynamic responses

For each fitted pair the evaluator constructs the true finite observation game. With $\widehat{V}_h = \widehat{\pi}_h^\top \widehat{Q}_h \widehat{\nu}_h$, define $e_h^\star = |\widehat{Q}_h - r_h - P_h^\star \widehat{V}_{h+1}|$ and report

$$B^+ = \sum_h \mathbb{E}_{\mu^+_{h,\star}} e_h^\star, \qquad B^- = \sum_h \mathbb{E}_{\mu^-_{h,\star}} e_h^\star, \qquad \sum_h \mathbb{E}_{\rho_h} (e_h^\star)^2.$$

The two laws use true dynamic best responses to the returned opponent policies, rather than only the played pair. The local numerical matrix residuals are also measured. Every returned pair satisfies the deterministic backward certificate within 10^{-8} numerical tolerance. No evaluator quantity selects weights, calibration counts, or predictors.

Figure 3 in the main text summarizes these diagnostics.

LP's larger-cap visitation ratios at $c = 16, 32, 64, 128$ are $0.539, 0.556, 0.558, 0.592$, whereas its complete-budget gaps are $0.00841, 0.00881, 0.01013, 0.00889$. In the equal-fitting slice the gaps are $0.00854, 0.00945, 0.01028, 0.00889$. Thus simply removing the fitting allocation difference does not make gap monotone in calibration. At $c = 64$, LP's $B^+ + B^-$ is 0.01020 and the exact gap is 0.01013; the response certificate is close to the observed gap here. The worst-case visitation ratio alone does not describe the error locations that enter that certificate.

G.3 One collected dataset, three public rewards

Calibration is fixed at 64, independently of scan outcomes. We collect fresh fitting data for this phase. The preselected rewards are

	$h < H$	$h = H$
success	0	x_1
failure	0	$1 - x_1$
action tradeoff	$\frac{1}{2(H-1)} \mathbf{1}\{a = b\}$	$x_1/2$

where x_1 is the public terminal outcome bit. These queries are identical for every hidden coordinate. Each method reuses its own independent stagewise buffers across all three tasks. Known reward queries relabel them, and each task is fitted backward independently. Every comparator enjoys the same reuse rights. The episode total is counted once per method and budget, not summed over task rows; all three fitting and matrix-solution times are counted. Corollary 13 states the corresponding finite-task data-sharing guarantee.

We take the maximum of the three gaps within each seed before reporting its mean and standard error. At the larger cap the means are 0.00942 (LP), 0.00790 (greedy), 0.01081 (L_1), 0.01049 (equal), and 0.00764 (uniform actions). The paired LP-minus-equal 95% interval is $[-0.00415, 0.00200]$; LP-minus-greedy is $[0.00056, 0.00248]$. These are descriptive intervals without multiplicity correction. Success and failure have matching gaps to numerical precision in this family, while the action-tradeoff gaps are smaller. Thus these tasks verify legal data reuse, but do not supply three independent hard instances or establish a calibration advantage under task reuse.

G.4 COMPUTATION, PRESERVATION, AND REPLAY

This study uses the same vectorized simulator and integer transition histograms for every method. Selectors are queried on observed inputs; the simulator returns no latent labels. At the fixed experimental width $2^p = 16$, histograms preserve all squared-loss sufficient statistics. Tests compare their fits against the original tuple implementation; the polynomial coordinate implementation for general p remains available. This benchmark acceleration does not assert logarithmic computation in observation count for dense histograms.

Complete time is the sum of measured provider, applicable calibration prefix, frequency computation, weight optimization, fitting-prefix collection, and all task fitting and matrix-solution times. PCE is run with its unchanged implementation. Shared execution saves research compute but is charged in full to each hypothetical method. Exact evaluation and export I/O are excluded throughout. Times are compared within this study, not against the scalar-collection study in Appendix F. The artifact retains raw calibration trajectories, integer fitting counts, returned heads, failure flags, and fixed protocols. All 1,040 records passed budget and response-certificate checks; all 960 distinct fits were replayed from stored sufficient statistics.

G.5 COMPLETE SCAN AND TASK RESULTS

Table 5: Complete budget, reference allocation 4,096, eight seeds. Errors are standard errors; BR is the sum of the two best-response Bellman residuals. Episodes are in thousands; times include all charged training operations.

Method	n_{cal}	Coverage	Gap	BR	Episodes	Seconds
LP	16	0.539	0.01408 ± 0.00080	0.01415 ± 0.00080	699.7	231.76
LP	32	0.556	0.01562 ± 0.00093	0.01569 ± 0.00093	699.7	232.09
LP	64	0.558	0.01815 ± 0.00131	0.01824 ± 0.00130	699.7	232.50
LP	128	0.592	0.01878 ± 0.00085	0.01886 ± 0.00085	699.7	232.93
Greedy	16	0.501	0.01263 ± 0.00095	0.01273 ± 0.00095	699.7	231.91
Greedy	32	0.501	0.01245 ± 0.00114	0.01252 ± 0.00112	699.7	232.23
Greedy	64	0.505	0.01576 ± 0.00106	0.01581 ± 0.00107	699.7	232.65
Greedy	128	0.532	0.01674 ± 0.00091	0.01684 ± 0.00092	699.7	233.04
L_1	16	0.588	0.01572 ± 0.00075	0.01582 ± 0.00076	699.7	232.84
L_1	32	0.573	0.01783 ± 0.00097	0.01791 ± 0.00096	699.7	233.12
L_1	64	0.543	0.01975 ± 0.00172	0.01983 ± 0.00171	699.7	233.55
L_1	128	0.531	0.02200 ± 0.00176	0.02207 ± 0.00177	699.7	233.64
Event max	16	0.253	0.01798 ± 0.00144	0.01805 ± 0.00144	699.7	231.86
Event max	32	0.300	0.01902 ± 0.00128	0.01911 ± 0.00128	699.7	232.24
Event max	64	0.271	0.01969 ± 0.00096	0.01976 ± 0.00096	699.7	232.66
Event max	128	0.286	0.02311 ± 0.00281	0.02317 ± 0.00280	699.7	233.04
Event subset	16	0.269	0.01838 ± 0.00160	0.01847 ± 0.00160	699.7	231.86
Event subset	32	0.323	0.01819 ± 0.00121	0.01826 ± 0.00121	699.7	232.23
Event subset	64	0.268	0.01844 ± 0.00118	0.01851 ± 0.00117	699.7	232.66
Event subset	128	0.312	0.02230 ± 0.00333	0.02238 ± 0.00333	699.7	233.04
Equal PCE	0	0.423	0.02096 ± 0.00172	0.02103 ± 0.00172	699.7	232.38
Uniform actions	0	0.076	0.00879 ± 0.00072	0.00888 ± 0.00072	699.7	0.45

Table 6: Equal fitting counts, reference allocation 4,096, eight seeds. Errors are standard errors; BR is the sum of the two best-response Bellman residuals. Episodes are in thousands; times include all charged training operations.

Method	n_{cal}	Coverage	Gap	BR	Episodes	Seconds
LP	16	0.539	0.01761 ± 0.00125	0.01769 ± 0.00125	668.7	231.69
LP	32	0.556	0.01955 ± 0.00123	0.01962 ± 0.00123	673.1	232.03
LP	64	0.558	0.01911 ± 0.00106	0.01919 ± 0.00107	681.9	232.46
LP	128	0.592	0.01878 ± 0.00085	0.01886 ± 0.00085	699.7	232.93
Greedy	16	0.501	0.01595 ± 0.00125	0.01602 ± 0.00124	668.7	231.80

Method	n_{cal}	Coverage	Gap	BR	Episodes	Seconds
Greedy	32	0.501	0.01593 ± 0.00108	0.01599 ± 0.00108	673.1	232.13
Greedy	64	0.505	0.01602 ± 0.00113	0.01608 ± 0.00113	681.9	232.57
Greedy	128	0.532	0.01674 ± 0.00091	0.01684 ± 0.00092	699.7	233.04
L_1	16	0.588	0.01921 ± 0.00091	0.01929 ± 0.00091	668.7	232.44
L_1	32	0.573	0.02313 ± 0.00147	0.02319 ± 0.00146	673.1	232.73
L_1	64	0.543	0.02218 ± 0.00143	0.02224 ± 0.00143	681.9	233.16
L_1	128	0.531	0.02200 ± 0.00176	0.02207 ± 0.00177	699.7	233.64
Event max	16	0.253	0.02260 ± 0.00097	0.02266 ± 0.00096	668.7	231.76
Event max	32	0.300	0.01997 ± 0.00132	0.02002 ± 0.00132	673.1	232.13
Event max	64	0.271	0.02167 ± 0.00169	0.02174 ± 0.00170	681.9	232.56
Event max	128	0.286	0.02311 ± 0.00281	0.02317 ± 0.00280	699.7	233.04
Event subset	16	0.269	0.02176 ± 0.00109	0.02184 ± 0.00109	668.7	231.76
Event subset	32	0.323	0.02012 ± 0.00120	0.02019 ± 0.00119	673.1	232.13
Event subset	64	0.268	0.02032 ± 0.00120	0.02039 ± 0.00120	681.9	232.56
Event subset	128	0.312	0.02230 ± 0.00333	0.02238 ± 0.00333	699.7	233.04
Equal PCE	0	0.423	0.02424 ± 0.00204	0.02432 ± 0.00204	664.2	231.97
Uniform actions	0	0.076	0.03518 ± 0.00288	0.03525 ± 0.00288	56.0	0.04

Table 7: Complete budget, reference allocation 16,384, eight seeds. Errors are standard errors; BR is the sum of the two best-response Bellman residuals. Episodes are in thousands; times include all charged training operations.

Method	n_{cal}	Coverage	Gap	BR	Episodes	Seconds
LP	16	0.539	0.00841 ± 0.00040	0.00848 ± 0.00040	920.9	232.21
LP	32	0.556	0.00881 ± 0.00063	0.00888 ± 0.00063	920.9	232.48
LP	64	0.558	0.01013 ± 0.00086	0.01020 ± 0.00086	920.9	232.85
LP	128	0.592	0.00889 ± 0.00094	0.00894 ± 0.00094	920.9	233.25
Greedy	16	0.501	0.00693 ± 0.00028	0.00701 ± 0.00028	920.9	232.73
Greedy	32	0.501	0.00805 ± 0.00077	0.00812 ± 0.00077	920.9	232.92
Greedy	64	0.505	0.00877 ± 0.00069	0.00886 ± 0.00069	920.9	233.35
Greedy	128	0.532	0.00810 ± 0.00079	0.00818 ± 0.00079	920.9	233.66
L_1	16	0.588	0.00817 ± 0.00073	0.00824 ± 0.00073	920.9	235.54
L_1	32	0.573	0.00941 ± 0.00086	0.00946 ± 0.00086	920.9	235.93
L_1	64	0.543	0.01094 ± 0.00058	0.01103 ± 0.00059	920.9	236.38
L_1	128	0.531	0.00920 ± 0.00085	0.00926 ± 0.00085	920.9	236.12
Event max	16	0.253	0.01073 ± 0.00066	0.01080 ± 0.00065	920.9	232.61
Event max	32	0.300	0.01000 ± 0.00094	0.01011 ± 0.00094	920.9	233.01
Event max	64	0.271	0.01217 ± 0.00117	0.01225 ± 0.00116	920.9	233.43
Event max	128	0.286	0.00987 ± 0.00062	0.00995 ± 0.00062	920.9	233.73
Event subset	16	0.269	0.00984 ± 0.00054	0.00991 ± 0.00054	920.9	232.61
Event subset	32	0.323	0.01084 ± 0.00108	0.01090 ± 0.00108	920.9	233.00
Event subset	64	0.268	0.01096 ± 0.00099	0.01104 ± 0.00100	920.9	233.43
Event subset	128	0.312	0.01014 ± 0.00099	0.01023 ± 0.00099	920.9	233.72
Equal PCE	0	0.423	0.01123 ± 0.00086	0.01131 ± 0.00086	920.9	234.95
Uniform actions	0	0.076	0.00866 ± 0.00095	0.00873 ± 0.00094	920.9	0.59

Table 8: Equal fitting counts, reference allocation 16,384, eight seeds. Errors are standard errors; BR is the sum of the two best-response Bellman residuals. Episodes are in thousands; times include all charged training operations.

Method	n_{cal}	Coverage	Gap	BR	Episodes	Seconds
LP	16	0.539	0.00854 ± 0.00046	0.00863 ± 0.00046	889.8	232.15
LP	32	0.556	0.00945 ± 0.00077	0.00953 ± 0.00077	894.3	232.42
LP	64	0.558	0.01028 ± 0.00088	0.01037 ± 0.00089	903.1	232.81

Method	n_{cal}	Coverage	Gap	BR	Episodes	Seconds
LP	128	0.592	0.00889 ± 0.00094	0.00894 ± 0.00094	920.9	233.25
Greedy	16	0.501	0.00732 ± 0.00025	0.00740 ± 0.00025	889.8	232.61
Greedy	32	0.501	0.00858 ± 0.00067	0.00866 ± 0.00067	894.3	232.83
Greedy	64	0.505	0.00892 ± 0.00059	0.00901 ± 0.00060	903.1	233.28
Greedy	128	0.532	0.00810 ± 0.00079	0.00818 ± 0.00079	920.9	233.66
L_1	16	0.588	0.00879 ± 0.00101	0.00886 ± 0.00101	889.8	235.14
L_1	32	0.573	0.00998 ± 0.00088	0.01006 ± 0.00087	894.3	235.53
L_1	64	0.543	0.01102 ± 0.00050	0.01110 ± 0.00049	903.1	236.00
L_1	128	0.531	0.00920 ± 0.00085	0.00926 ± 0.00085	920.9	236.12
Event max	16	0.253	0.01129 ± 0.00057	0.01137 ± 0.00058	889.8	232.51
Event max	32	0.300	0.01112 ± 0.00111	0.01118 ± 0.00112	894.3	232.91
Event max	64	0.271	0.01239 ± 0.00106	0.01246 ± 0.00106	903.1	233.33
Event max	128	0.286	0.00987 ± 0.00062	0.00995 ± 0.00062	920.9	233.73
Event subset	16	0.269	0.01005 ± 0.00072	0.01010 ± 0.00072	889.8	232.50
Event subset	32	0.323	0.01158 ± 0.00125	0.01163 ± 0.00125	894.3	232.90
Event subset	64	0.268	0.01106 ± 0.00085	0.01113 ± 0.00084	903.1	233.33
Event subset	128	0.312	0.01014 ± 0.00099	0.01023 ± 0.00099	920.9	233.72
Equal PCE	0	0.423	0.01122 ± 0.00070	0.01129 ± 0.00071	885.4	234.54
Uniform actions	0	0.076	0.01651 ± 0.00136	0.01660 ± 0.00136	277.2	0.18

Table 9: Shared-data reward tasks at reference allocation 4,096, eight seeds. Every method pays once for its own collection and for all three fits. Worst gap is maximized over tasks before averaging over seeds. All gaps are in units of 10^{-3}.

Method	Success	Failure	Tradeoff	Worst	Seconds
LP	14.64 ± 1.35	14.64 ± 1.35	4.01 ± 0.78	14.64 ± 1.35	232.50
Greedy	13.73 ± 1.45	13.73 ± 1.45	3.85 ± 0.69	13.73 ± 1.45	232.62
L_1	19.31 ± 1.23	19.31 ± 1.23	5.13 ± 0.40	19.31 ± 1.23	233.25
Event max	19.06 ± 0.78	19.06 ± 0.78	4.45 ± 0.47	19.06 ± 0.78	232.59
Event subset	18.26 ± 0.95	18.26 ± 0.95	4.59 ± 0.72	18.26 ± 0.95	232.59
Equal PCE	15.44 ± 1.22	15.44 ± 1.22	3.71 ± 0.43	15.44 ± 1.22	232.37
Uniform actions	9.27 ± 1.29	9.27 ± 1.29	2.29 ± 0.49	9.27 ± 1.29	0.47

Table 10: Shared-data reward tasks at reference allocation 16,384, eight seeds. Every method pays once for its own collection and for all three fits. Worst gap is maximized over tasks before averaging over seeds. All gaps are in units of 10^{-3}.

Method	Success	Failure	Tradeoff	Worst	Seconds
LP	9.42 ± 0.91	9.42 ± 0.91	1.95 ± 0.40	9.42 ± 0.91	232.83
Greedy	7.90 ± 0.95	7.90 ± 0.95	1.69 ± 0.37	7.90 ± 0.95	233.27
L_1	10.81 ± 0.92	10.81 ± 0.92	1.97 ± 0.22	10.81 ± 0.92	235.67
Event max	9.90 ± 1.59	9.90 ± 1.59	2.04 ± 0.47	9.90 ± 1.59	233.26
Event subset	9.57 ± 1.24	9.57 ± 1.24	1.80 ± 0.38	9.57 ± 1.24	233.26
Equal PCE	10.49 ± 0.71	10.49 ± 0.71	2.07 ± 0.27	10.49 ± 0.71	234.77
Uniform actions	7.64 ± 0.86	7.64 ± 0.86	1.59 ± 0.32	7.64 ± 0.86	0.62

H MODEL-BASED AND REAL-SUCCESSOR FITTING CONTROLS

These controls compare model-based and real-successor fitting. The same-calibration comparisons appear in Appendices F and G; the datasets are not pooled. Control-study timers stop at the callable policy; the other two studies also materialize output matrices before stopping the timer.

Implementation and data provenance. This control study contains 756 formal result rows: 288 PCE-game rows, 432 library-game rows, 24 matched-context controls, and 12 reward-decoded controls. They were run on the same CPU server in the `acmg` conda environment, with Python 3.11.15 and NumPy 2.4.6. Linear programs use SciPy HiGHS; the two-action interior formula is checked against the general local LP. BLAS and OpenMP use one thread per worker. Collection, calibration, and fitting wall times are recorded; exact latent-space evaluation and plotting are excluded from training time.

`run_mixture_revision.py` generates the PCE comparisons, `run_library_mixture.py` generates the library comparisons, and `run_mixture_controls.py` generates the additional controls. `export_mixture_revision.py` regenerates the paper figures, tables and grouped summaries from per-run files. The formal directories are `main`, `library_final`, and `controls` under `results/fitting_routes` in the artifact. Pilot outputs and an earlier library timing run are retained but excluded from the paper aggregation.

Observable calibration. For each noninitial, nonterminal stage, execute every returned cover policy for the stated number of independent episodes. Only the reached observations are used to estimate the two events for every candidate coordinate. Solve (2) with $\beta = 0.005$ and check the simplex constraints, LP feasibility, and $t \geq 7/(10R)$ for $R = 2$. This empirical check alone is not a confidence guarantee; the formal theorem additionally requires its sufficient calibration count. At the initial stage every roll-in policy has the same state law, so no calibration is needed. At the terminal

M	Episodes	Learned mixture	Uniform library	Uniform actions
8	8,192	0.0185 ± 0.0020	0.0303 ± 0.0079	0.0159 ± 0.0026
8	32,768	0.0102 ± 0.0012	0.0204 ± 0.0023	0.0075 ± 0.0012
8	131,072	0.0044 ± 0.0005	0.0082 ± 0.0012	0.0045 ± 0.0003
32	8,192	0.0183 ± 0.0023	0.1215 ± 0.0188	0.0159 ± 0.0026
32	32,768	0.0108 ± 0.0014	0.0327 ± 0.0041	0.0075 ± 0.0012
32	131,072	0.0042 ± 0.0005	0.0184 ± 0.0023	0.0045 ± 0.0003
128	8,192	0.0355 ± 0.0039	0.1338 ± 0.0141	0.0159 ± 0.0026
128	32,768	0.0101 ± 0.0016	0.0816 ± 0.0111	0.0075 ± 0.0012
128	131,072	0.0043 ± 0.0006	0.0367 ± 0.0039	0.0045 ± 0.0003

Table 11: Complete library-game results: mean Nash gap $\pm$ standard error over 16 seeds. Total episodes include the learned method's $32M$ calibration samples. All policies in each library are distinct.

stage the reward matrix is known exactly and both fitting and transition residuals are zero, so terminal calibration and fitting are omitted. These optimizations reduce the actual count below the theorem's conservative HMn_{cal} bound.

Each downstream buffer is independent of the calibration data. The algorithm never uses evaluator occupancies, the true coordinate index, or the known latent transition table to choose its weights or fitted predictions. Public terminal rewards use coordinate zero in code; candidate decoder i uses coordinate i at intermediate stages and coordinate zero at the last stage. Testing every reward query under different hidden coordinate choices confirms that the reward interface does not identify the intermediate decoder.

H.1 Coupled library game

We implement the three-stage construction of Proposition 14. There are four maximizer actions (d, u) and two minimizer actions b. The known latent entry mechanism maps the stage-one decision d to the next latent state. From latent state one at stage two, the terminal success probability is

$$\tfrac{1}{2} + 0.04\theta(-1)^d + \tfrac{1}{8}(-1)^{u+b}, \qquad \theta \in \{-1, 1\}.$$

State zero leads to failure; all earlier rewards are zero. The exact game value is 0.54. Six-bit observations contain an unknown state coordinate, and the terminal outcome uses the public coordinate.

For $M \in \{8, 32, 128\}$, the library's stage-one entry probabilities are 1 and $(i-2)/M^3$ for $2 \le i \le M$. Each policy thus differs in behavior at an actually visited state. Library order and candidate decoder order are independently shuffled. Sixteen seeds alternate θ and cycle the hidden coordinate; every seed is included. The library is generated by these probabilities without environment episodes or latent-state access. The learned method spends $32M$ episodes on calibration, then uses its selected mixture. The uniform library spends all episodes on true-successor fitting. Both use the same finite coordinate-head regression and exact local matrix solves. The root action is known to enter under either instance, so only the stage-two Bellman matrix must be estimated.

Total budgets are 8,192, 32,768, and 131,072 episodes. The simulator vectorizes independent rollouts but returns only observation–action– public-outcome tuples; a test checks this sampler against the explicit entry and transition probabilities. Episode counts are unchanged by vectorization, and each budget has its own measured collection time. A uniform-action reference reaches the informative state with probability $1/2$. It is intentionally included to show the scope of the separation: it changes the library sampling law and is not subject to the uniform-library lower bound.

H.2 PCE in coupled hidden-progress games

The transition and gap formulas are derived in Appendix B. We use $H = 6, 10$, four-bit observations, and eight seeds per horizon. At each stage independent row and column permutations randomize the progress matrix; the hidden coordinate cycles with the seed. Initial progress probability is $1/2$, so the marginal initial observation is uniform and does not expose the hidden coordinate. Candidate decoder order is shuffled independently. The model grid denominator is $K = H$, which contains every true transition probability, and Bellman head spacing is at most 10^{-4}.

H	n	Fit	Budget	Learned PCE	Equal PCE	Uniform
6	1,024	model	146,800	0.0369 ± 0.0012	0.0377 ± 0.0057	0.0096 ± 0.0013
6	1,024	real	146,800	0.0279 ± 0.0027	0.0268 ± 0.0043	0.0083 ± 0.0004
6	4,096	model	177,520	0.0211 ± 0.0021	0.0181 ± 0.0026	0.0087 ± 0.0007
6	4,096	real	177,520	0.0129 ± 0.0013	0.0151 ± 0.0015	0.0075 ± 0.0005
6	16,384	model	300,400	0.0089 ± 0.0014	0.0125 ± 0.0015	0.0061 ± 0.0009
6	16,384	real	300,400	0.0056 ± 0.0007	0.0074 ± 0.0008	0.0051 ± 0.0007
10	1,024	model	644,376	0.0554 ± 0.0070	0.0409 ± 0.0038	0.0171 ± 0.0013
10	1,024	real	644,376	0.0335 ± 0.0021	0.0255 ± 0.0014	0.0105 ± 0.0013
10	4,096	model	699,672	0.0261 ± 0.0027	0.0253 ± 0.0018	0.0152 ± 0.0015
10	4,096	real	699,672	0.0164 ± 0.0009	0.0169 ± 0.0011	0.0099 ± 0.0010
10	16,384	model	920,856	0.0123 ± 0.0013	0.0135 ± 0.0012	0.0142 ± 0.0010
10	16,384	real	920,856	0.0083 ± 0.0006	0.0090 ± 0.0007	0.0084 ± 0.0006

Table 12: All coupled hidden-progress results: mean gap $\pm$ standard error over eight seeds. "Budget" is the mean complete episode cap; actual counts differ only through the stated integer remainders. Model fitting uses MLE and one pseudo-successor per PE context.

PCE uses all $2H$ rounds, 64 samples per regression call, 24 tests, 24 candidate centers, $\gamma = 0.15$, and $\gamma' = 0.4$. No privileged cover replaces its output. We calibrate each relevant returned policy with 64 episodes. Each PCE cover and its calibration are shared across comparisons for the same seed, but their entire construction cost is charged to each hypothetical method that uses them. Downstream data are fresh. For a reference allocation $n \in \{1024, 4096, 16384\}$, the common budget cap is

$$N_{\text{cap}} = N_{\text{PCE}} + N_{\text{cal}} + 2(H-1)n.$$

Learned-weight model fitting uses n MLE and n PE episodes per nonterminal stage. Its real-successor counterpart uses $2n$. Equal-weight PCE uses the saved calibration episodes for fitting; uniform-action collection uses the saved PCE and calibration episodes. Each method takes the floor of its available budget divided by its number of stagewise buffers. Thus every method respects the same cap, with fewer than $2(H-1)$ unused episodes.

The coverage diagnostic is $\min_{h<H, z:v_h(z)>0} q_h(z)/v_h(z)$, evaluated exactly after training; the terminal stage is excluded because its reward is known and its Bellman error is zero. The mean ratios for equal and learned weights are 0.47579 versus 0.56352 at $H = 6$, and 0.42318 versus 0.57258 at $H = 10$. The cover's maximum visitation is unchanged by reweighting; it is the collection mixture that improves.

At the largest allocation, the paired learned-minus-equal mean gap differences are -0.00360 ± 0.00189 (model) and -0.00185 ± 0.00114 (real) at $H = 6$, and -0.00121 ± 0.00167 and -0.00074 ± 0.00106 at $H = 10$. These are standard errors of within-seed differences, preserving the shared cover. All four two-sided 95% Student intervals include zero. The data support a coverage improvement and a budget-dependent change in mean gap, rather than a statistically established PCE performance advantage at these seed counts.

For $H = 10$ at the largest allocation, mean collection-plus-fitting times are 336.29, 333.17, and 151.51 seconds for learned-weight PCE, equal-weight PCE, and uniform actions on the model route. The corresponding real-successor times are 334.63, 331.51, and 148.12 seconds. This implementation therefore does not establish a wall-clock speedup over uniform actions. Per-run times for every configuration are stored with the raw results; machine scheduling makes them descriptive rather than a hardware-independent complexity comparison.

H.3 Matched contexts and the known-reward baseline

The pseudo-successor control uses $H = 6$ and six seeds, with the same PCE and calibration settings. Hold the cover, weights, MLE buffer, and 512 PE contexts per stage fixed, then use 1, 4, or 16 simulated successors per context. Each target averages those successor values. The real-successor control uses the same number of contexts and retains their true successor. Its fitting environment cost is half the model route's, because it does not need an MLE buffer. The continuation is refitted backward for each variant, as in the actual algorithm.

Labels	Draws/context	Fit episodes	Fit seconds	Nash gap
model	1	5,120	1.073	0.0527 ± 0.0065
model	4	5,120	1.183	0.0392 ± 0.0078
model	16	5,120	1.691	0.0223 ± 0.0069
real	0	2,560	0.517	0.0553 ± 0.0054

Table 13: Matched-context control at $H = 6$, with six seeds and 512 PE contexts per nonterminal stage. Fit seconds include buffer collection and fitting, but omit the common cover and calibration cost. The final column is mean gap $\pm$ standard error.

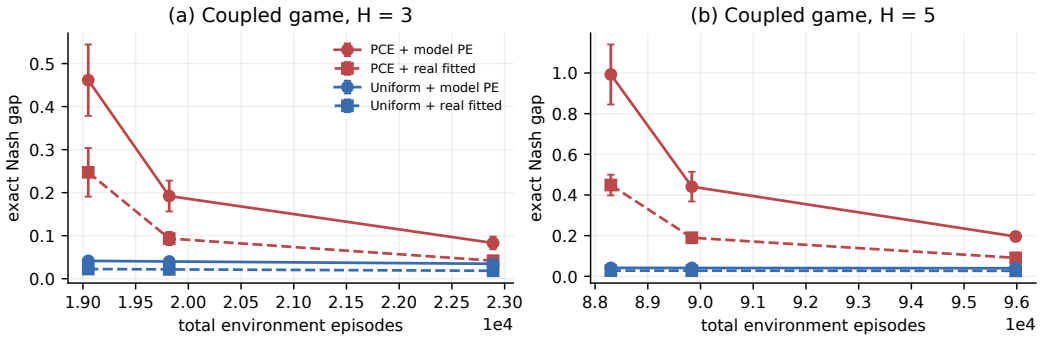

Figure 4: Recurrent-reward control: complete-budget comparison with 12 seeds and 1.96 standard-error bars. The reward-decoded baseline above uses the same largest-budget caps. This easy-coverage game is distinct from the public-terminal hidden-progress experiments.

More pseudo-successors reduce conditional simulation noise but do not repair a wrong model or a poorly covered context law. These controls therefore explain part of the model-route error without identifying a model-bias contribution by subtraction. They also do not imply a matched-cost advantage over real-successor fitting.

For the recurrent-reward game in Appendix O, the known reward $r_h(x, a, b) = \phi^\star(x)$ is itself a valid decoder. The new baseline queries that reward on each observed context and successor, fits denominator-20 two-state transition MLEs, and solves the resulting two-state game by exact backward minimax planning. It uses no true-coordinate index or privileged latent-state interface. For each of the 12 original $H = 5$ seeds it receives the exact previous largest-budget cap and uses its integer number of stagewise buffers; fewer than four episodes remain unused. Its mean gap is 0.0028129 ± 0.0028129: eleven seeds are below 10^{-12} and one has gap 0.0337549. The nonzero seed is retained.

Uniform exploration in that game satisfies

$$q_{h+1}(1) = \frac{19}{40} + \frac{1}{16} q_h(1), \quad q_1(1) = \frac{1}{2}, \qquad q_h(1) = \frac{38}{75} - \frac{1}{150} 16^{-(h-1)}.$$

Thus its uniform joint-action law dominates every centralized policy with coefficient at most $300/37$. At stage five its exact coefficient is 7.06369. Even the ideal state mixture, while retaining uniform target actions, achieves only 6.94735, an improvement of about 1.65%. This is an explanation of the easy-control regime, not a general claim about policy-cover methods or unknown representations.

I COMPLETE PROOFS

I.1 THE POST-COVER THEOREMS AND THEIR TIGHTNESS

Proof of Theorem 23. For any fixed product pair (π, ν), the simulation identity under the true occupancy is

$$V_{P^\star}^{\pi,\nu} - V_{\widehat{P}}^{\pi,\nu} = \sum_{h=1}^{H-1} \mathbb{E}_{d_{P^\star,h}^{\pi,\nu}\pi_h\nu_h}\left[(P_h^\star - \widehat{P}_h)V_{\widehat{P},h+1}^{\pi,\nu}\right].$$

Since the continuation lies in $[0, H_{h+1}]$,

$$\sup_\pi \left|V_{P^\star}^{\pi,\widehat{\nu}} - V_{\widehat{P}}^{\pi,\widehat{\nu}}\right| \le \mathfrak{E}_H^+(\widehat{P};\widehat{\nu}),$$
$$\sup_\nu \left|V_{P^\star}^{\widehat{\pi},\nu} - V_{\widehat{P}}^{\widehat{\pi},\nu}\right| \le \mathfrak{E}_H^-(\widehat{P};\widehat{\pi}).$$

Expand and cancel the played value before comparing kernels:

$$\begin{aligned}\mathrm{Gap}_{P^\star}(\widehat{\pi},\widehat{\nu}) &= \sup_\pi V_{P^\star}^{\pi,\widehat{\nu}} - \inf_\nu V_{P^\star}^{\widehat{\pi},\nu} \\ &\le \sup_\pi V_{\widehat{P}}^{\pi,\widehat{\nu}} - \inf_\nu V_{\widehat{P}}^{\widehat{\pi},\nu} + \mathfrak{E}_H^+(\widehat{P};\widehat{\nu}) + \mathfrak{E}_H^-(\widehat{P};\widehat{\pi}) \\ &= \mathrm{Gap}_{\widehat{P}}(\widehat{\pi},\widehat{\nu}) + \mathfrak{E}_H^+(\widehat{P};\widehat{\nu}) + \mathfrak{E}_H^-(\widehat{P};\widehat{\pi}).\end{aligned}$$

This is (31). □

Proof of Theorem 3. Define the best-response continuation values

$$U_h(s) = \sup_{\pi_{h:H}} V_{P,h}^{\pi,\widehat{\nu}}(s), \qquad L_h(s) = \inf_{\nu_{h:H}} V_{P,h}^{\widehat{\pi},\nu}(s),$$

with $U_{H+1} = L_{H+1} = \widehat{V}_{H+1} = 0$. Because the action sets are finite, choose measurable Markov selectors $\pi_h^+(s)$ and $\nu_h^-(s)$ attaining the Bellman maximum and minimum, respectively. Set $D_h^+ = U_h - \widehat{V}_h$ and $D_h^- = \widehat{V}_h - L_h$. These functions need not be nonnegative statewise; the following one-sided recursions do not take absolute values of them.

At a fixed (h, s), Bellman optimality and the definition of e_h give

$$\begin{aligned}D_h^+(s) &= \left\langle \pi_h^+, \{r_h + P_h U_{h+1}\}\widehat{\nu}_h \right\rangle(s) - \widehat{V}_h(s) \\ &\le \mathbb{E}_{\pi_h^+,\widehat{\nu}_h}\left[e_h(s,A_h,B_h) + P_h D_{h+1}^+(s,A_h,B_h)\right] + \delta_h^+(s).\end{aligned}$$

The last step uses $\langle \pi_h^+, \widehat{Q}_h\widehat{\nu}_h\rangle \le \max_p \langle p, \widehat{Q}_h\widehat{\nu}_h\rangle$ and (35). Taking the initial expectation and repeatedly applying the tower property along $(\pi^+, \widehat{\nu}, P)$ yields

$$\mathbb{E}_{d_1} D_1^+ \le \sum_{h=1}^H \mathbb{E}_{d_{P,h}^{\pi^+,\widehat{\nu}}\pi_h^+\widehat{\nu}_h}\left[e_h + \delta_h^+(S_h)\right]. \tag{12}$$

The minimizer calculation is symmetric but reverses the Bellman comparison:

$$\begin{aligned}D_h^-(s) &= \widehat{V}_h(s) - \left\langle \widehat{\pi}_h, \{r_h + P_h L_{h+1}\}\nu_h^- \right\rangle(s) \\ &\le \mathbb{E}_{\widehat{\pi}_h,\nu_h^-}\left[e_h(s,A_h,B_h) + P_h D_{h+1}^-(s,A_h,B_h)\right] + \delta_h^-(s).\end{aligned}$$

Thus

$$\mathbb{E}_{d_1} D_1^- \le \sum_{h=1}^H \mathbb{E}_{d_{P,h}^{\widehat{\pi},\nu^-}\widehat{\pi}_h\nu_h^-}\left[e_h + \delta_h^-(S_h)\right]. \tag{13}$$

Finally, zero-sum cancellation gives

$$\mathrm{Gap}_P(\widehat{\pi},\widehat{\nu}) = \mathbb{E}_{d_1}[U_1 - L_1] = \mathbb{E}_{d_1}[D_1^+ + D_1^-].$$

Adding (12) and (13) proves (36). Under the stated domination conditions, $\mathbb{E}_d e_h \le C_h^\pm \mathbb{E}_{\rho_h} e_h \le C_h^\pm \varepsilon_{h,\mathrm{B}}$; the uniform residual bounds then prove (37). □

Proof of Theorem 26. Apply Theorem 3 in the learned kernel $P = \widehat{P}$. By the definitions in Appendix J,

$$\mathrm{Gap}_{\widehat{P}}(\widehat{\pi}, \widehat{\nu}) \le \mathfrak{B}_H^+ + \mathfrak{B}_H^-.$$

Substitution into Theorem 23 proves (24).

For the density-ratio statement, Cauchy–Schwarz gives, for every maximizer deviation π,

$$\mathbb{E}_{\mu^+_{h,\mathrm{tr}}(\pi)} D_{\mathrm{TV}}(P^\star_h, \widehat{P}_h) \le \sqrt{C^+_{h,\mathrm{tr}}} \left(\mathbb{E}_{\rho_h} D^2_{\mathrm{TV}}(P^\star_h, \widehat{P}_h) \right)^{1/2} \le \sqrt{C^+_{h,\mathrm{tr}}}\, \zeta_h.$$

Taking the supremum and summing with weights H_{h+1} proves the plus transition term; the minus term is identical. For evaluation, direct change of measure gives $\mathbb{E}_{\mu^+_{h,\mathrm{pe}}} e_h \le C^+_{h,\mathrm{pe}} \mathbb{E}_{\rho_h} e_h$ and its minimizer analogue. The uniform residual bounds integrate to $\bar{\delta}^{\pm}_h$. Summing these four inequalities proves (40). □

Proof of Corollary 33. For the average strategies of entropic no-regret dynamics on the fixed matrix $\widehat{Q}_h(s)$, the two residuals in (35) are uniformly at most

$$\delta_T \le c_0 H \sqrt{\frac{\log A + \log B}{T}}$$

for an absolute constant c_0; clipping makes the payoff range at most H. Pointwise PE gives $e_h \le \varepsilon_{\mathrm{B}}$ under both unilateral laws. Substitution into (36) gives $2H(\varepsilon_{\mathrm{B}} + \delta_T)$, which is (39) after absorbing constants. □

Proof of Proposition 31. Consider a horizon-two game with a unique initial state, actions $\mathcal{A} = \{a_0, a_1\}$, $\mathcal{B} = \{b_0, b_1\}$, and terminal states g, ℓ. The only reward is $r_2(g) = 1$ and $r_2(\ell) = 0$. Consequently, the value of a stage-one action pair is its probability of transitioning to g. Let the learned kernel have good-state probability

$$\widehat{p}(g \mid a, b) = \frac{1}{2} \quad \text{for every } (a, b),$$

and let the true probabilities, with rows indexed by (a_0, a_1) and columns by (b_0, b_1), be

$$[P^\star(g \mid a, b)]_{a,b} = \begin{pmatrix} \frac{1}{2} & 0 \\ 1 & \frac{1}{2} \end{pmatrix}.$$

These kernels have a two-dimensional factorization: use the feature $(p(g \mid a, b), 1 - p(g \mid a, b))$ and the two terminal point masses. Return $\widehat{\pi} = \delta_{a_0}$ and $\widehat{\nu} = \delta_{b_0}$. The learned payoff matrix is constant, hence $\mathrm{Gap}_{\widehat{P}}(\widehat{\pi}, \widehat{\nu}) = 0$. In the true game,

$$\sup_\pi V^{\pi, \widehat{\nu}}_{P^\star} = 1, \qquad \inf_\nu V^{\widehat{\pi}, \nu}_{P^\star} = 0,$$

so the true gap is one. Against b_0, the largest Bernoulli transition TV error is attained by a_1 and equals $1/2$; against a_0, the largest error is attained by b_1 and also equals $1/2$. Because the sole continuation weight is one, $\mathfrak{E}^+_H = \mathfrak{E}^-_H = 1/2$. This proves equality. To isolate the two sides, instead use the true good-state probability matrices

$$\begin{pmatrix} 1/2 & 1/2 \\ 1 & 1/2 \end{pmatrix}, \qquad \begin{pmatrix} 1/2 & 0 \\ 1/2 & 1/2 \end{pmatrix}$$

with the same learned kernel and output pair. Their true gaps are both $1/2$, while their unilateral error pairs are $(1/2, 0)$ and $(0, 1/2)$, respectively. Thus each family is necessary even when the other coefficient is allowed to increase. In the equality example, if either coefficient were strictly below one while the other remained one, the right-hand side would be strictly below the true gap. □

Proof of Corollary 32. Every product pair appearing in either unilateral error is a centralized policy in Π_J. Hence $\mathfrak{E}^+_H \le \mathfrak{E}_H \le \varepsilon_{\mathrm{sys}}$ and $\mathfrak{E}^-_H \le \mathfrak{E}_H \le \varepsilon_{\mathrm{sys}}$. Apply Theorem 23. □

I.2 Block coverage and the finite-sample corollary

Proof of Lemma 36. Fix a stage and write $p^\omega(z) = \mathbb{P}_{P^\star,\omega}(Z_h = z)$ and $m(z) = \sup_{\omega'} p^{\omega'}(z)$. The $(1,\eta)$-cover property gives, for each latent state z, a policy $\psi_z \in \Psi_h$ with $p^{\psi_z}(z) \geq m(z) - \eta$. If $q(z) = M^{-1} \sum_{\psi \in \Psi_h} p^\psi(z)$, then

$$q(z) \geq \frac{m(z) - \eta}{M}.$$

Split the latent states into $\mathcal{R} = \{z : m(z) < 2\eta\}$ and its complement. For every target policy, $p^\omega(\mathcal{R}) \leq 2S\eta$. On $\mathcal{R}^c$, $q(z) \geq m(z)/(2M) \geq p^\omega(z)/(2M)$. After the exploratory joint action is made uniform, the target state–action law is therefore dominated by ρ_h with ratio $2MJ$ on $\mathcal{R}^c$. The emission law conditional on z is the same under every policy, so the same ratio holds for observations. Thus, for $g_h \in [0,1]$,

$$\mathbb{E}_{d^\omega_{P^\star,h}\omega_h} g_h \leq 2S\eta + 2MJ\mathbb{E}_{\rho_h} g_h,$$

which is (44). On $\mathcal{R}^c$, Cauchy–Schwarz under ρ_h instead gives

$$\mathbb{E}_{d^\omega_{P^\star,h}\omega_h}[g_h \mathbf{1}_{\mathcal{R}^c}] \leq \sqrt{2MJ\mathbb{E}_{\rho_h} g_h^2},$$

proving (45). □

Lemma 20 (Finite-class transition and Bellman fits)**.** *Under Assumption 6, fix a stage h. If the proper MLE uses n samples from ρ_h, then, with probability at least $1-\delta'$,*

$$\mathbb{E}_{\rho_h} D^2_{\mathrm{TV}}(P^\star_h, \widehat{P}_h) \leq \frac{2\log(\Theta_\star/\delta') + \xi_{\mathrm{mle}}}{n}. \tag{14}$$

If squared-loss PE uses n fresh samples, then, for a universal constant c_{pe}, with probability at least $1-\delta'$,

$$\mathbb{E}_{\rho_h}(\widehat{Q}_h - q^\star_h)^2 \leq \frac{c_{\mathrm{pe}} H^2 \log(2Q_\star/\delta')}{n}. \tag{15}$$

Both statements remain valid conditional on the past when the stage-h sampling law and continuation are predictable.

Proof. For the MLE claim, regard one observation as a joint draw $(X,Y) \sim \rho_h(x) P^\star_h(y \mid x)$. The factor $\rho_h(x)$ is common to every candidate, so maximizing the joint likelihood is the same as maximizing the conditional transition likelihood. Apply the square-root likelihood argument of Lemma 21 simultaneously to the finite class Θ_h. Since all n samples have the same joint law and the selected model is within ξ_{mle} of the true log likelihood, it gives

$$n\, D^2_{\mathrm{H}}(\rho_h P^\star_h, \rho_h \widehat{P}_h) \leq 2\log(\Theta_\star/\delta') + \xi_{\mathrm{mle}}.$$

The common marginal implies $D^2_{\mathrm{H}}(\rho_h P^\star_h, \rho_h \widehat{P}_h) = \mathbb{E}_{\rho_h} D^2_{\mathrm{H}}(P^\star_h, \widehat{P}_h)$. Now use $D^2_{\mathrm{TV}} \leq D^2_{\mathrm{H}}$, proving (14).

For PE, condition on the continuation and write $q^\star(x) = \mathbb{E}[Y \mid X = x]$. For each $q \in \mathcal{Q}_h$, let

$$Z_q = (q(X) - Y)^2 - (q^\star(X) - Y)^2, \qquad v_q = \mathbb{E}_{\rho_h}(q - q^\star)^2.$$

Then $\mathbb{E} Z_q = v_q$. Since $q, q^\star, Y \in [0,H]$,

$$Z_q^2 = (q - q^\star)^2 (q + q^\star - 2Y)^2 \leq 4H^2 (q - q^\star)^2,$$

and hence $\mathbb{E} Z_q^2 \leq 4H^2 v_q$, while $|Z_q| \leq 2H^2$. Bernstein's inequality and a union bound over $\mathcal{Q}_h$ imply, simultaneously for every q, that its empirical excess loss differs from v_q by at most

$$c_1 H \sqrt{\frac{v_q \log(2Q_\star/\delta')}{n}} + c_2 H^2 \frac{\log(2Q_\star/\delta')}{n}.$$

The ERM has nonpositive empirical excess loss relative to the realizable $q^\star$. Moving the preceding deviation terms to the other side and using $2ab \leq a^2/2 + 2b^2$ yields

$$\mathbb{E}_{\rho_h}(\widehat{Q}_h - q^\star)^2 \leq \frac{c_{\mathrm{pe}} H^2 \log(2Q_\star/\delta')}{n}$$

for a universal c_{pe}. This proves (15); Cauchy–Schwarz also gives the corresponding first-moment bound. The conditional version follows by applying the same argument after fixing the predictable past. □

Proof of Corollary 29. The pointwise inequality $e_h^\star \le e_h + H_{h+1} D_{\mathrm{TV}}(P_h^\star, \widehat{P}_h)$ and Theorem 3 in $P^\star$ give (7). Each of its two response policies is a centralized Markov policy. The cover event is uniform over such policies, so its use for data-dependent best responses requires no extra union bound. Lemma 36 bounds either transition expectation by $2S\eta + \sqrt{2MJ}\zeta_h$, either Bellman expectation by $2SH_h\eta + 2MJ\varepsilon_{h,\mathrm{B}}$, and the matrix term by δ_h. Summation proves (47) and (27). On the additional second-moment event, apply the same lemma to e_h/H_h and replace $2MJ\varepsilon_{h,\mathrm{B}}$ by $\sqrt{2MJ}u_h$. Since $\sum_{h<H} H_{h+1} + \sum_{h\le H} H_h = H^2$, the rare-state terms sum to $4SH^2\eta$, proving (28). The PCE costs are polynomial in $(H, S, J, 1/\eta, 1/\delta)$ and its two-context sample efficiency as specified by (PCE). □

Proof of Corollary 37. Allocate $\delta_{\mathrm{fail}}/3$ to each of PCE, MLE, and PE. Apply Lemma 20 with $\delta' = \delta_{\mathrm{fail}}/(3H)$ at each stage. The sample sizes give $\|f_h\|_{L_2(\rho_h)} \le \zeta$ and $\|e_h\|_{L_2(\rho_h)} \le u$. The matrix regret guarantee gives $\delta_h^{\pm} \le c_0 H\sqrt{(\log A + \log B)/T} \le \varepsilon/(8H)$. Substituting the tolerances in (28) gives

$$\mathrm{Gap}_{P^\star}(\widehat{\pi}, \widehat{\nu}) \le \frac{\varepsilon}{8} + \frac{\varepsilon(H-1)}{8H} + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} < \varepsilon.$$

Here the transition sum is $H(H-1)/2$ and $2MJ \le \overline{\kappa} = 2H^2S^2J$. The source PCE theorem gives exactly (PCE). The independent post-cover buffers cost at most $H(n_{\mathrm{mle}} + n_{\mathrm{pe}})$ episodes, including the conservative terminal allocation. Finally,

$$\zeta^{-2} = \frac{128H^6S^2J}{\varepsilon^2}, \qquad H^2u^{-2} = \frac{128H^6S^2J}{\varepsilon^2}$$

proves (50). A trajectory uses at most H environment transitions; episode and transition counts are not identical. The matrix work is per queried observation, as stated in the corollary. □

I.3 Limits of prediction-only interfaces

Proof of Proposition 30. Given a rank-d MDP $M = (\mathcal{S}, \mathcal{A}, H, P^\star, r, d_1)$, construct a zero-sum Markov game with $\mathcal{B} = \{b_0\}$, the same transition kernel $P^\star(\cdot \mid s, a, b_0) = P^\star(\cdot \mid s, a)$, maximizer reward $r(s, a, b_0) = r(s, a)$, and minimizer reward $-r$. The low-rank factorization is unchanged after appending the dummy action. There is only one minimizer policy, so the second unilateral exploitability term is zero and the first is exactly

$$\sup_\pi V_M^\pi - V_M^{\widehat{\pi}}.$$

The interaction transcript and every supervised-learning or policy-evaluation query are identical to those in the original MDP. Thus an ε-gap policy for the constructed game is an ε-optimal MDP policy with the same resources. □

Proof of Lemma 35. Fix a policy ω and a stage h, and abbreviate $e_h(s, j) = D_{\mathrm{TV}}(P_h^\star(\cdot \mid s, j), \widehat{P}_h(\cdot \mid s, j))$. The density-ratio assumption and Cauchy–Schwarz give

$$\mathbb{E}_{d^\omega_{P^\star,h}\omega_h} e_h \le \sqrt{\mathbb{E}_{d^\omega_{P^\star,h}\omega_h} e_h^2} \le \sqrt{C_h \mathbb{E}_{\rho_h} e_h^2} \le \sqrt{C_h}\,\zeta_h.$$

Multiply by H_{h+1}, sum over h, and take the supremum over ω to obtain (43).

The obstruction uses an initial state with two actions L, R and two terminal states u, v. Both rank-two models send L deterministically to u; under R, the first sends the process to u and the second to v. A dataset containing only L transitions has identical likelihood under both models, so both attain zero excess MLE risk. The policy choosing R sees transition total variation one. Thus no bound on $\mathfrak{E}_H$ can be a function of the on-data MLE excess risk alone. □

I.4 Finite-encoding separation for the RepLearn discriminator

Proof of Proposition 34. We give a reduction with rational inputs and an exact equal-norm identity. Let $G = (V, E)$ be a nonempty unweighted simple graph, with $n = |V|$, $m = |E|$, and maximum

degree $D_{\max}$. Let $M \in \{-1, 0, 1\}^{m \times n}$ be an oriented incidence matrix: the row for edge $\{i, j\}$ is $e_i - e_j$. For every sign vector y, $\|My\|_1$ is twice the number of edges cut by y, so

$$\|M\|_{\infty \to 1} = \max_{x \in \{-1,1\}^m, y \in \{-1,1\}^n} x^\top My = 2\,\mathrm{MaxCut}(G). \tag{16}$$

This is the incidence-matrix reduction underlying the norm hardness in Alon & Naor (2006, Proposition 2.2).

Set $c = 2D_{\max} + 1$. Construct an integer matrix B by adding one row $e_i + e_j$ for every edge, and $c - 2\deg(i)$ copies of e_i for each vertex. These multiplicities are positive integers. If r is the number of rows of B, then $m + r = nc - 2m = O(n^2)$. The two rows associated with an edge contribute

$$(e_i - e_j)(e_i - e_j)^\top + (e_i + e_j)(e_i + e_j)^\top = 2e_ie_i^\top + 2e_je_j^\top.$$

The remaining rows complete the diagonal, giving the exact identity

$$C = \begin{pmatrix} M \\ B \end{pmatrix}, \qquad C^\top C = cI_n. \tag{17}$$

Take the complete equally weighted empirical design on $N = 2(m + r)$ paired predecessor/successor contexts. For sign vectors x, y, set

$$r_x = (x, 0_r, -x, 0_r), \qquad z_y = (Cy, -Cy).$$

Then

$$\mathbf{1}^\top r_x = \mathbf{1}^\top z_y = 0, \quad \|z_y\|_2^2 = 2cn, \quad r_x^\top z_y = 2x^\top My. \tag{18}$$

Choose $\beta = (4cn)^{-1}$ and $\alpha = (8mcn)^{-1}$. Define $f_x = \frac{1}{2}\mathbf{1} + \alpha r_x$, the current class $\mathcal{H}_0 = \{a\mathbf{1}\}$, and alternative classes $\mathcal{H}_y = \{a\mathbf{1} + b\beta z_y : a, b \in \mathbb{R}\}$. Centering and orthogonal projection give

$$\begin{aligned} \min_{g \in \mathcal{H}_0} \|g - f_x\|_2^2 - \min_{g \in \mathcal{H}_y} \|g - f_x\|_2^2 &= \alpha^2 \frac{(r_x^\top z_y)^2}{\|z_y\|_2^2} \\ &= \frac{2\alpha^2}{cn}(x^\top My)^2. \end{aligned} \tag{19}$$

We realize these classes in the one-action, dimension-two RepLearn syntax. On predecessor contexts the current feature equals $(1/2, 0)$. For each y, include a map with predecessor values $(1/2, \beta z_y(i))$ and zero successor values. For each x, include a map with predecessor values $(1/2, 0)$ and successor values $(f_x(i), 0)$. Use unit-ball discriminator weights θ, so $f(s) = \theta^\top \phi(s)$ ranges over tf_x, $|t| \le 1$, and zero. Since $|z_y(i)| \le \sqrt{cn}$, all feature norms are at most one; labels are bounded by one in absolute value. The predecessor maps realize the stated predictor classes with coefficients $(2a, b)$. For labels tf_x, the unconstrained optimum has $2a = t$ and

$$|b^\star| \le \frac{\alpha}{\beta} \frac{\|r_x\|_2}{\|z_y\|_2} = \frac{1}{2\sqrt{mcn}} \le \frac{1}{2}.$$

It therefore lies in the alternative weight ball of radius $2\sqrt{d}$ for $d = 2$, and also in the larger current-regression ball. Scaling the label by t multiplies the improvement by t^2, so its maximum is attained at $|t| = 1$. Consequently,

$$\Delta_D = \frac{2\alpha^2}{cn}\|M\|_{\infty \to 1}^2 = \frac{8\alpha^2}{cn}\,\mathrm{MaxCut}(G)^2. \tag{20}$$

Every entry and normalizing constant is rational with polynomial bit length, and each feature map has a succinct description by its sign vector and graph. Different integer cut values give optimum values separated by at least $8\alpha^2/(cn) = 1/(8m^2c^3n^3)$. An additive approximation smaller than one third of this inverse-polynomial spacing recovers the integer optimal cut value. This proves NP-hardness. For average empirical loss, divide the spacing by N; it remains inverse polynomial.

It remains to verify the two constituent ERMs on this same complete design, for any rational response vector q. For the scaled label class, $\|f_x\|_2^2 = N/4 + 2m\alpha^2$ is constant. Eliminating $t \in [-1, 1]$ reduces the objective to maximizing $|q^\top f_x|$. Writing this correlation as a fixed intercept plus a linear form in x shows that one of the two coordinatewise sign choices is optimal.

For the representation-predictor union, write $q = (q^+, q^-)$ in its two halves. The centered feature z_y is orthogonal to the constant vector and has length independent of y. After optimizing the two regression coefficients, the objective is nonincreasing in

$$|q^\top z_y| = |y^\top C^\top (q^+ - q^-)|.$$

This monotonicity also holds under the Euclidean weight ball, since its feasible coefficient set is the same for every y and is symmetric in the slope. Thus a coordinatewise sign choice again selects an optimal representation. The remaining regression is a convex problem in two variables, solvable to any rational objective tolerance in polynomial time in the input bit length and the logarithm of inverse tolerance. The label ERM has only a scalar coefficient. These are the constituent oracles claimed in the proposition; arbitrary reweighting of the design is outside this claim. Polynomially many calls to these routines cannot compute Δ_D to the stated accuracy unless $\mathsf{P} = \mathsf{NP}$. □

I.5 A LIKELIHOOD CONCENTRATION LEMMA

We first prove a generic conditional-density lemma.

Lemma 21 (Square-root likelihood martingale). *Let $Y_i \mid \mathcal{F}_{i-1} \sim p_i^\star(\cdot)$, where the context determining $p_i^\star$ is $\mathcal{F}_{i-1}$-measurable. For a fixed candidate θ, let p_i^θ be another predictable conditional density and set*

$$A_i(\theta) = \int \sqrt{p_i^\star(y) p_i^\theta(y)}\, \mathrm{d}y.$$

Define

$$R_i(\theta) = \begin{cases} \sqrt{p_i^\theta(Y_i)/p_i^\star(Y_i)}/A_i(\theta), & A_i(\theta) > 0, \\ 1, & A_i(\theta) = 0, \end{cases} \qquad M_n(\theta) = \prod_{i=1}^n R_i(\theta).$$

Then $M_n(\theta)$ is a nonnegative martingale with $M_0 = 1$. Consequently, with probability at least $1 - \delta$, simultaneously for all n and all $\theta \in \Theta$ such that $A_i(\theta) > 0$ for $i \le n$,

$$\frac{1}{2} \sum_{i=1}^n \log \frac{p_i^\theta(Y_i)}{p_i^\star(Y_i)} - \sum_{i=1}^n \log A_i(\theta) \le \log \frac{|\Theta|}{\delta}. \tag{21}$$

Proof. Conditional on $\mathcal{F}_{i-1}$, the conditional expectation of $R_i(\theta)$ is one. For $A_i(\theta) > 0$, this follows from

$$\mathbb{E}\left[\sqrt{p_i^\theta(Y_i)/p_i^\star(Y_i)} \,\middle|\, \mathcal{F}_{i-1} \right] = \int \sqrt{p_i^\theta(y) p_i^\star(y)}\, \mathrm{d}y = A_i(\theta),$$

and it is immediate from the definition when $A_i(\theta) = 0$. Therefore $\mathbb{E}[M_i(\theta) \mid \mathcal{F}_{i-1}] = M_{i-1}(\theta)$. Ville's inequality gives $\mathbb{P}(\sup_n M_n(\theta) > |\Theta|/\delta) \le \delta/|\Theta|$. A union bound over θ, followed by taking logarithms, proves (21). If $A_i(\theta) = 0$, then $p_i^\theta(Y_i) = 0$ $p_i^\star$-almost surely. Such a candidate has log likelihood $-\infty$ and cannot be an approximate likelihood maximizer relative to the true model at any finite tolerance. Hence the positive-affinity part of the event covers every candidate used below. □

J MODEL TRANSFER AND ALTERNATIVE EVALUATION INTERFACES

Low-rank normalization.

Assumption 22 (Low-rank realizability and normalization). *For each h, unknown maps $\phi_h^\star : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}^d$ and $\mu_h^\star : \mathcal{S} \to \mathbb{R}^d$ satisfy*

$$P_h^\star(s' \mid s, a, b) = \langle \phi_h^\star(s, a, b), \mu_h^\star(s') \rangle, \qquad \|\phi_h^\star(s, a, b)\|_2 \le 1,$$

and $\| \int \mu_h^\star(s) g(s)\, \mathrm{d}s \|_2 \le \sqrt{d}$ for every measurable $g : \mathcal{S} \to [0, 1]$. A supplied model class contains the true model, and every candidate is a valid normalized transition kernel.

The representation need not be identifiable. The backward and model-transfer certificates use only valid kernels and bounded rewards; low rank enters through prediction guarantees and the construction of exploration laws. Appendix A states the common measurability and sampling conventions.

$$\widehat{M} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} \qquad M^\star = \begin{pmatrix} \boxed{1/2} & 0 \\ 1 & 1/2 \end{pmatrix}$$

Returned pair: first row, first column. Played-pair error: 0.
Maximizer deviation: $1/2 \to 1$. Minimizer deviation: $1/2 \to 0$.

Figure 5: Prediction on the played pair does not certify equilibrium. Entries are probabilities of the terminal reward 1. The learned-model gap is zero, whereas the true gap is one. Each unilateral transition-error term equals $1/2$.

Correct on the played action, wrong about deviations. Consider two terminal states, good and bad, with rewards 1 and 0; the initial reward is zero. Figure 5 specifies the probability of reaching the good state for each first-stage action pair. Both kernels have rank at most two. Return the first row and first column. The pair is an equilibrium of the learned model, and its played transition is predicted exactly. In the true game, the maximizer gains by switching to the second row, and the minimizer gains by switching to the second column. The resulting Nash gap is $1 - 0 = 1$.

The example identifies a failure of a returned-policy certificate. It does not imply that every learner fails on this game. Removing either unobserved discrepancy leaves a one-sided example with gap $1/2$, which also establishes the separate sharpness of the coefficients in Theorem 23.

K From transition prediction to an equilibrium certificate

Fix a learned kernel $\widehat{P}$ and returned pair $(\widehat{\pi}, \widehat{\nu})$. Write $d\pi\nu$ for the state–joint-action law obtained from a state law d and independent actions. The two families of true-game unilateral laws are

$$\mu^+_{h,\mathrm{tr}}(\pi) = d^{\pi,\widehat{\nu}}_{P^\star,h}\pi_h\widehat{\nu}_h, \qquad \mu^-_{h,\mathrm{tr}}(\nu) = d^{\widehat{\pi},\nu}_{P^\star,h}\widehat{\pi}_h\nu_h.$$

We use $D_{\mathrm{TV}}(P,Q) = \frac{1}{2}\int |\mathrm{d}P - \mathrm{d}Q|$. For a learned kernel $\widehat{P}$ and output pair $(\widehat{\pi}, \widehat{\nu})$, define the two transition errors exposed by a one-player deviation:

$$\mathfrak{E}^+_H(\widehat{P};\widehat{\nu}) = \sup_\pi \sum_{h=1}^{H-1} H_{h+1}\mathbb{E}_{\mu^+_{h,\mathrm{tr}}(\pi)} D_{\mathrm{TV}}(P^\star_h, \widehat{P}_h),$$
$$\mathfrak{E}^-_H(\widehat{P};\widehat{\pi}) = \sup_\nu \sum_{h=1}^{H-1} H_{h+1}\mathbb{E}_{\mu^-_{h,\mathrm{tr}}(\nu)} D_{\mathrm{TV}}(P^\star_h, \widehat{P}_h). \tag{22}$$

Theorem 23 (Sharp unilateral-cover-to-Nash decomposition)**.** *For every valid $\widehat{P}$ and every Markov product pair,*

$$\mathrm{Gap}_{P^\star}(\widehat{\pi},\widehat{\nu}) \le \mathrm{Gap}_{\widehat{P}}(\widehat{\pi},\widehat{\nu}) + \mathfrak{E}^+_H(\widehat{P};\widehat{\nu}) + \mathfrak{E}^-_H(\widehat{P};\widehat{\pi}). \tag{23}$$

Each coefficient is one, and no error term for the played pair is required.

Why only two errors appear. For a fixed pair, the simulation identity gives

$$V^{\pi,\nu}_{P^\star} - V^{\pi,\nu}_{\widehat{P}} = \sum_{h=1}^{H-1} \mathbb{E}_{d^{\pi,\nu}_{P^\star,h}\pi_h\nu_h}[(P^\star_h - \widehat{P}_h)V^{\pi,\nu}_{\widehat{P},h+1}].$$

The continuation lies in $[0, H_{h+1}]$, so total variation controls each summand. The played value has already canceled in the definition of the Nash gap before comparing the kernels. Applying simulation to the two remaining deviation families proves the theorem; Appendix I.1 gives the full calculation.

Figure 5 attains equality with $\mathfrak{E}^+_H = \mathfrak{E}^-_H = 1/2$. Proposition 31 also isolates each side, so either unit coefficient is necessary in this universal certificate. This sharpness concerns the error decomposition: it does not make a particular density-ratio condition necessary for learning every game.

K.1 THE COVERAGE INTERFACE

Apply Theorem 3 in $P = \widehat{P}$, with best responses π^+, ν^-, and define

$$\mu^+_{h,\mathrm{pe}} = d^{\pi^+,\widehat{\nu}}_{\widehat{P},h}\pi^+_h\widehat{\nu}_h, \qquad \mu^-_{h,\mathrm{pe}} = d^{\widehat{\pi},\nu^-}_{\widehat{P},h}\widehat{\pi}_h\nu^-_h.$$

For probability laws μ, ρ, write $\mu \ll_C \rho$ when $\mu \ll \rho$ and $\|\mathrm{d}\mu/\mathrm{d}\rho\|_\infty \le C$.

Definition 24 (Output-dependent unilateral cover)**.** *A family $\rho = (\rho_1, \dots, \rho_H)$ on $\mathcal{S} \times \mathcal{A} \times \mathcal{B}$ is a unilateral cover for $(\widehat{P}, \widehat{\pi}, \widehat{\nu})$ with coefficients $C^{\pm}_{h,\mathrm{tr}}, C^{\pm}_{h,\mathrm{pe}}$ if the following absolute-continuity and Radon–Nikodym bounds hold:*

$$\mu^+_{h,\mathrm{tr}}(\pi) \ll_{C^+_{h,\mathrm{tr}}} \rho_h \quad \text{for every } \pi, \qquad \mu^-_{h,\mathrm{tr}}(\nu) \ll_{C^-_{h,\mathrm{tr}}} \rho_h \quad \text{for every } \nu,$$
$$\mu^+_{h,\mathrm{pe}} \ll_{C^+_{h,\mathrm{pe}}} \rho_h, \qquad \mu^-_{h,\mathrm{pe}} \ll_{C^-_{h,\mathrm{pe}}} \rho_h.$$

The first two bounds form the transition cover*; the last two form the* evaluation cover*. Uniform domination of centralized Markov occupancies under both relevant kernels is a stronger, output-independent sufficient condition.*

The transition laws concern the true model; the evaluation laws concern learned-model best responses. The four coefficients are analysis certificates rather than numerical inputs to the solver. We also allow approximate domination through explicit remainders, as in the Block construction in Section 4.

K.2 INDEPENDENT FITTING AND BELLMAN EVALUATION

Prediction interfaces. Proper transition MLE returns a normalized kernel from a fixed class, supporting conditional sampling or integration. Average Bellman PE estimates $r_h + P_h v$ for a predictable continuation, with stated $L_1(\rho_h)$ or $L_2(\rho_h)$ error. Ordinary one-context SL minimizes a sample-decomposable risk over a fixed one-head class; it does not itself choose a covering exploration policy.

Algorithm 2 uses a separate PE buffer at each stage, withheld from all later-stage fits. Conditional on the cover and fitted model, its contexts are independent of the later-stage continuation. The label $r_h(X) + \widehat{V}_{h+1}(\widetilde{S}_{h+1})$, with $\widetilde{S}_{h+1} \sim \widehat{P}_h(\cdot \mid X)$, has the required conditional mean $r_h + \widehat{P}_h\widehat{V}_{h+1}$.

Remark 25 (Pseudo-successors and model error)**.** *Conditional on $\widehat{P}_h$ and $\widehat{V}_{h+1}$, pseudo-successor sampling is unbiased for the learned-model Bellman target; its Monte Carlo noise enters $\varepsilon_{h,\mathrm{B}}$, not model bias, and exact integration may replace sampling. A real successor targets $P^\star_h\widehat{V}_{h+1}$ and defines the directly analyzed fitted-minimax route (Section 4).*

The matrix updates are evaluated at each queried observation using the returned predictor; they do not enumerate a rich observation space. Entropic no-regret updates give a local residual of order $H_h\sqrt{(\log A + \log B)/T}$. Appendix N specifies these updates and their query cost.

K.3 A COMPOSABLE NASH GUARANTEE

For Algorithm 2 run in $\widehat{P}$, define

$$\mathfrak{B}^{\pm}_H = \sum_{h=1}^{H} \mathbb{E}_{\mu^{\pm}_{h,\mathrm{pe}}}[e_h + \delta^{\pm}_h(S_h)].$$

Theorem 26 (Post-cover MLE/PE-to-Nash reduction)**.** *For every true kernel $P^\star$, valid learned kernel $\widehat{P}$, and output of Algorithm 2,*

$$\mathrm{Gap}_{P^\star}(\widehat{\pi}, \widehat{\nu}) \le \mathfrak{E}^+_H + \mathfrak{E}^-_H + \mathfrak{B}^+_H + \mathfrak{B}^-_H. \tag{24}$$

Theorem 26 follows by applying the backward bound in the learned model and then Theorem 23. Its density-ratio specialization, equation (40) in Appendix M, separates transition, evaluation, and matrix errors. For signed rewards, continuation reward spans replace H_{h+1} (Appendix O.1).

Algorithm 2: POSTCOVER-PE

Input: covered samplers $\{\rho_h\}_{h=1}^H$; proper transition-MLE and average Bellman-PE classes; sample sizes $n_{\mathrm{mle}}, n_{\mathrm{pe}}$; matrix iterations T.
Output: a Markov product pair $(\widehat{\pi}, \widehat{\nu})$.

for $h = 1, \dots, H$ **do**
  Draw n_{mle} independent covered transitions from ρ_h and store them in $\mathcal{D}_h^{\mathrm{mle}}$
  Draw n_{pe} fresh covered contexts from ρ_h and store them in $\mathcal{D}_h^{\mathrm{pe}}$
  Fit $\widehat{P}_h$ by proper conditional MLE on $\mathcal{D}_h^{\mathrm{mle}}$
end
Set $\widehat{V}_{H+1} = 0$
for $h = H, H-1, \dots, 1$ **do**
  foreach $X = (S_h, A_h, B_h) \in \mathcal{D}_h^{\mathrm{pe}}$ **do**
    Draw $\widetilde{S}_{h+1} \sim \widehat{P}_h(\cdot \mid X)$ and form $Y = r_h(X) + \widehat{V}_{h+1}(\widetilde{S}_{h+1})$
  end
  Call average Bellman PE on $\{(X, Y)\}$ to obtain $\widehat{Q}_h^{\mathrm{raw}}$, and set $\widehat{Q}_h = \mathrm{clip}_{[0,H_h]}\, \widehat{Q}_h^{\mathrm{raw}}$
  Run T entropic no-regret iterations on each matrix $\widehat{Q}_h(s)$, and average both players' iterates to obtain $(\widehat{\pi}_h, \widehat{\nu}_h)$
  Set $\widehat{V}_h(s) = \langle \widehat{\pi}_h(\cdot \mid s), \widehat{Q}_h(s)\widehat{\nu}_h(\cdot \mid s) \rangle$
end
return $(\widehat{\pi}, \widehat{\nu})$.

L UNIFORM-MIXTURE BLOCK BASELINE

We record the uniform-mixture composition used as the baseline for the learned-weight construction, including its cover-provider assumptions.

L.1 WHAT COMPUTATION DOES A COVER PROVIDER REQUIRE?

Setting $B = 1$ preserves the MDP interaction transcript and prediction queries, so the one-context oracle separation of Golowich et al. (2024) remains relevant to all-game guarantees. A narrower obstruction concerns RepLearn: Proposition 34 gives succinct dimension-two instances on a fixed complete equally weighted design where the constituent one-head squared-loss ERMs are tractable, but the pure-witness discriminator value is NP-hard to approximate to inverse-polynomial accuracy. This result concerns that discriminator, not arbitrary cover providers or composite proper-MLE interfaces. Appendix M.5 develops the reductions and their precise oracle scope.

L.2 BLOCK STRUCTURE AND A POLICY-COVER PROVIDER

Definition 27 (Decodable Block Markov game)**.** *There is a finite latent set $\mathcal{Z}$, $|\mathcal{Z}| = S$, and an unknown decoder $\phi^\star : \mathcal{S} \to \mathcal{Z}$ in a known class Φ. The kernel factors as*

$$P_h^\star(\mathrm{d}s' \mid s, a, b) = \sum_{z' \in \mathcal{Z}} T_h(z' \mid \phi^\star(s), a, b) O_{h+1}(\mathrm{d}s' \mid z').$$

The emissions $O_h(\cdot \mid z)$ have disjoint supports decoded as z and depend only on the latent state and stage; d_1 is a mixture of $O_1(\cdot \mid z)$. During reward-free exploration a central learner may choose the joint action $j = (a, b)$; the final policies must factor as $\pi_h(a \mid s)\nu_h(b \mid s)$.

Centralizing (a, b) as $j \in [J]$ produces a Block MDP. The two-context PCE reduction of Rohatgi & Foster (2025) returns stagewise policy families Ψ_h. Roll in with a uniformly chosen member and draw a uniform joint action at stage h to implement ρ_h. We call its composition with POSTCOVER-PE BLOCKMG-PE. The final output remains a product policy pair.

Assumption 28 (Structured cover and population oracle accuracy)**.** *The game is an exact Φ-decodable Block MG with $|\mathcal{Z}| = S$, disjoint latent emission supports, and centralized joint-action access during reward-free training. The realizable two-context oracle satisfies Appendix A, so PCE returns stagewise*

$(1,\eta)$-policy covers of size $M \le H^2S^2$. On the induced law ρ_h, proper MLE and average Bellman PE satisfy

$$\mathbb{E}_{\rho_h} D^2_{\mathrm{TV}}(P^\star_h, \widehat{P}_h) \le \zeta_h^2, \qquad \mathbb{E}_{\rho_h}|\widehat{Q}_h - r_h - \widehat{P}_h \widehat{V}_{h+1}| \le \varepsilon_{h,\mathrm{B}}.$$

Conditional on the cover, all MLE and PE buffers use independent draws. Each continuation excludes its held-out PE buffer and is independent of that buffer given the cover and fitted model; both matrix residuals are at most δ_h.

For the explicit sample bound, Assumption 6 in Appendix A requires a realizable finite proper model class of size $\Theta_\star$, finite Bellman closure of size $Q_\star$, and squared-loss ERM on fresh bounded labels. The coordinate class in Appendix D replaces exact Bellman closure by a controlled quantization error.

For every centralized policy ω and measurable $g_h \in [0,1]$, Lemma 36 gives

$$\begin{aligned}\mathbb{E}_{d^{\omega}_{P^\star,h}\omega_h} g_h &\le 2S\eta + 2MJ\mathbb{E}_{\rho_h} g_h,\\ \mathbb{E}_{d^{\omega}_{P^\star,h}\omega_h} g_h &\le 2S\eta + \sqrt{2MJ\,\mathbb{E}_{\rho_h} g_h^2}.\end{aligned} \tag{25}$$

The remainder accounts for latent states with maximum visitation below 2η. This approximate domination need not satisfy the exact absolute-continuity conditions of Definition 24. Instead, it controls the true-model certificate (7) with an explicit rare-state remainder.

L.3 Direct transfer and second-moment evaluation

Put $\kappa = 2MJ$, $u_h = (\mathbb{E}_{\rho_h} e_h^2)^{1/2}$, and

$$\Delta_{\mathrm{sys}} = \sum_{h=1}^{H-1} H_{h+1}\big(2S\eta + \sqrt{\kappa}\,\zeta_h\big).$$

Apply the cover inequality directly to the true-model response laws in (7). For PE, either its population L_1 guarantee or an L_2 guarantee can be used:

$$\mathbb{E}_{\mu^{\pm}_{h,\star}} e_h \le 2SH_h\eta + \min\{\kappa\varepsilon_{h,\mathrm{B}}, \sqrt{\kappa}\,u_h\}. \tag{26}$$

The second term requires a second-moment event, not merely the definition of an average L_1 oracle. Lemma 20 proves that finite-class squared-loss PE supplies $u_h^2 \le c_{\mathrm{pe}} H^2 \log(2Q_\star/\delta')/n_{\mathrm{pe}}$. No learned-occupancy transfer is needed.

Corollary 29 (Block-structured end-to-end resolution)**.** *Under Assumption 28,* BlockMG-PE *returns a Markov product pair satisfying*

$$\mathrm{Gap}_{P^\star}(\widehat{\pi},\widehat{\nu}) \le 2\Delta_{\mathrm{sys}} + 2\sum_{h=1}^{H}\big(2SH_h\eta + 2MJ\varepsilon_{h,\mathrm{B}} + \delta_h\big). \tag{27}$$

If additionally $\mathbb{E}_{\rho_h} e_h^2 \le u_h^2$, the sharper bound is

$$\begin{aligned}\mathrm{Gap}_{P^\star}(\widehat{\pi},\widehat{\nu}) \le{}& 4SH^2\eta + 2\sqrt{\kappa}\sum_{h=1}^{H-1} H_{h+1}\zeta_h\\ &+ 2\sqrt{\kappa}\sum_{h=1}^{H} u_h + 2\sum_{h=1}^{H}\delta_h.\end{aligned} \tag{28}$$

Under Assumption 6, the accuracy schedule in Appendix M returns an ε-Nash pair with probability at least $1-\delta$ using

$$\mathcal{N}_{\mathrm{PCE}}\left(\frac{\varepsilon}{32SH^2}, \frac{\delta}{3}\right) + O\left(\frac{H^7S^2J(L_\Theta + L_Q)}{\varepsilon^2}\right) \tag{29}$$

environment episodes, where L_Θ, L_Q contain logarithmic class-size and confidence factors, and L_Θ also contains the MLE optimization error ξ_{mle}, as specified in Appendix M.

The factor $J = AB$ comes from centralized exploration and target-action sampling. The game-specific step is the conversion from covered transition and Bellman predictions to a product-Nash guarantee. The source PCE sample and oracle costs are charged explicitly in (PCE).

M AUXILIARY REDUCTIONS AND ORACLE ANALYSIS

This section records auxiliary reductions and provider assumptions. A sharp simulation theorem identifies two output-dependent model errors, and a backward recursion propagates average PE error along the corresponding unilateral best-response occupancies. The oracle analysis then compares one-context prediction, the RepLearn discriminator, and the two-context PCE construction used for decodable Block MGs.

M.1 LOW-RANK COMPARISON FOR THE AUXILIARY CERTIFICATES

Table 14 separates statistical scope from computational access. It does not assert rate dominance because the rows use different model classes and cover providers.

Table 14: Auxiliary low-rank comparison. Here J denotes the joint-action size. Rates suppress logarithmic factors and fix the number of players.

Result	Structure	Exploration/computation	Guarantee	Episode complexity or boundary
Ni et al. (2023), model based	Unknown low-rank general-sum MG	Model MLE, optimistic learned-game planning, equilibrium oracle	NE/CE/CCE	$\widetilde{O}(H^6d^4J^2/\varepsilon^2)$, with logarithmic model-class dependence
Ni et al. (2023), model free	Block general-sum MG	Representation and discriminator objectives with equilibrium planning	NE/CE/CCE	$\widetilde{O}(H^6d^4J^3/\varepsilon^2)$
Theorem 26 (ours)	Any zero-sum MG with a supplied output-dependent cover	Proper transition MLE, average PE, explicit matrix updates	Product Nash gap	Exact four-term reduction; cover generation is modular and not assumed free
Corollary 29 (ours)	Exact decodable Block zero-sum MG	Uniform PCE mixture, MLE/PE, true-model backward analysis	ε-Nash product pair	$\mathcal{N}_{\mathrm{PCE}} + \widetilde{O}(H^7S^2J/\varepsilon^2)$
Corollary 18	Hidden-coordinate Block game	PCE, real-successor ERM, local saddle updates	Product Nash gap	$\mathcal{N}_{\mathrm{PCE}} + O(H^7S^2JL_Q/\varepsilon^2)$; no MLE

The guarantees of Ni et al. (2023) cover broader game classes and charge exploration directly in their rates. They couple exploration to representation learning and optimistic equilibrium planning. Theorem 26 instead conditions on realized unilateral coverage and identifies proper MLE, average PE, and explicit matrix updates as sufficient downstream computation. Corollary 29 obtains an end-to-end result only after paying for the structured two-context PCE provider.

M.2 WHY THE GENERAL ORACLE QUESTION ALREADY APPEARS IN MDPS

Proposition 30 (Any all-game theorem contains the unknown-representation MDP problem)**.** *Let $\mathfrak{A}$ learn an ε-Nash policy with polynomial samples and polynomially many SL/PE calls for every unknown-representation rank-d zero-sum Markov game. Then the same resources yield an ε-optimal policy for every unknown-representation rank-d MDP.*

The reduction sets $\mathcal{B}$ to a singleton and gives the minimizer reward $-r$. It preserves the horizon, rank, state space, action size, episodic access, and every SL/PE query. Consequently, the Markov-game target cannot be easier than the corresponding minimization-oracle question for low-rank MDPs. The single-agent actor–critic guarantee of Huang et al. (2026) supplies the matching MLE/PE interface under an external sampling distribution that covers the relevant occupancies.

M.3 SHARP OUTPUT-DEPENDENT UNILATERAL COVER-TO-NASH BRIDGE

For an output pair $(\widehat{\pi}, \widehat{\nu})$, define only the two model errors that can be exposed by a unilateral deviation:

$$\mathfrak{E}_H^+(\widehat{P}; \widehat{\nu}) = \sup_{\pi} \sum_{h=1}^{H-1} H_{h+1} \mathbb{E}_{d^{\pi,\widehat{\nu}}_{P^\star,h} \pi_h \widehat{\nu}_h} D_{\mathrm{TV}}(P_h^\star, \widehat{P}_h),$$
$$\mathfrak{E}_H^-(\widehat{P}; \widehat{\pi}) = \sup_{\nu} \sum_{h=1}^{H-1} H_{h+1} \mathbb{E}_{d^{\widehat{\pi},\nu}_{P^\star,h} \widehat{\pi}_h \nu_h} D_{\mathrm{TV}}(P_h^\star, \widehat{P}_h). \tag{30}$$

Here and below, a state occupancy followed by two policies denotes the induced state–joint-action law.

Theorem 23 applies to every valid learned kernel $\widehat{P}$ and Markov pair $(\widehat{\pi}, \widehat{\nu})$, and states

$$\mathrm{Gap}_{P^\star}(\widehat{\pi}, \widehat{\nu}) \le \mathrm{Gap}_{\widehat{P}}(\widehat{\pi}, \widehat{\nu}) + \mathfrak{E}_H^+(\widehat{P}; \widehat{\nu}) + \mathfrak{E}_H^-(\widehat{P}; \widehat{\pi}). \tag{31}$$

Consequently, learned-model gap $\varepsilon_{\mathrm{plan}}$ and unilateral errors $\varepsilon_+, \varepsilon_-$ imply true gap at most $\varepsilon_{\mathrm{plan}} + \varepsilon_+ + \varepsilon_-$. Each coefficient in (31) is one: no error for the played pair is needed.

Proposition 31 (The two unit coefficients are jointly tight)**.** *There is a horizon-two rank-two zero-sum Markov game, a rank-two learned kernel, and a deterministic output pair for which*

$$\mathrm{Gap}_{P^\star} = 1, \qquad \mathrm{Gap}_{\widehat{P}} = 0, \qquad \mathfrak{E}_H^+ = \mathfrak{E}_H^- = \frac{1}{2}.$$

Hence equality holds in (31)*. In particular, neither coefficient one can be replaced by any smaller universal constant, even if the other coefficient remains one.*

The improvement over a four-value comparison follows from zero-sum cancellation. Indeed, $\mathrm{Gap}_P(\widehat{\pi}, \widehat{\nu}) = \sup_\pi V_P^{\pi,\widehat{\nu}} - \inf_\nu V_P^{\widehat{\pi},\nu}$, so the played value cancels before models are compared. This makes the bridge output-dependent: exploration targets precisely the policies that arise as one-sided deviations from the returned pair.

For comparison, regard a joint action $j = (a, b)$ as one centralized action and let Π_J contain all centralized Markov policies. Define

$$\mathfrak{E}_H(\widehat{P}) = \sup_{\omega \in \Pi_J} \sum_{h=1}^{H-1} H_{h+1} \mathbb{E}_{d^{\omega}_{P^\star,h} \omega_h} D_{\mathrm{TV}}(P_h^\star, \widehat{P}_h). \tag{32}$$

Every product pair and unilateral deviation belongs to Π_J, so Theorem 23 immediately gives the following centralized statement.

Corollary 32 (Centralized policy-uniform cover)**.** *If* $\mathfrak{E}_H(\widehat{P}) \le \varepsilon_{\mathrm{sys}}$ *and* $\mathrm{Gap}_{\widehat{P}}(\widehat{\pi}, \widehat{\nu}) \le \varepsilon_{\mathrm{plan}}$*, then*

$$\mathrm{Gap}_{P^\star}(\widehat{\pi}, \widehat{\nu}) \le \varepsilon_{\mathrm{plan}} + 2\varepsilon_{\mathrm{sys}}. \tag{33}$$

The zero-sum cancellation yields the coefficient two.

M.4 NASH GAP FROM BEST-RESPONSE BELLMAN ERRORS

For a known kernel P, a reference law ρ_h on $\mathcal{S} \times \mathcal{A} \times \mathcal{B}$, and a bounded continuation v, define an average one-step Bellman-evaluation call to return $\widehat{Q}_h$ with

$$\mathbb{E}_{\rho_h} \left| \widehat{Q}_h - \{r_h + P_h v\} \right| \le \varepsilon_{h,\mathrm{B}}. \tag{34}$$

The call evaluates the supplied continuation by using $v(S_{h+1})$ as an artificial terminal reward.

Starting from $\widehat{V}_{H+1} = 0$, process stages backward. At stage h, use $v = \widehat{V}_{h+1}$, compute approximate saddle actions $(\widehat{\pi}_h, \widehat{\nu}_h)$ for the fixed matrix $\widehat{Q}_h(s)$, and set $\widehat{V}_h(s) = \langle \widehat{\pi}_h, \widehat{Q}_h \widehat{\nu}_h \rangle(s)$. Let $\delta_h^+(s), \delta_h^-(s)$ be any certified local residuals satisfying

$$\max_p \langle p, \widehat{Q}_h \widehat{\nu}_h \rangle(s) - \widehat{V}_h(s) \le \delta_h^+(s),$$
$$\widehat{V}_h(s) - \min_v \langle \widehat{\pi}_h, \widehat{Q}_h v \rangle(s) \le \delta_h^-(s). \tag{35}$$

Write $e_h = |\widehat{Q}_h - (r_h + P_h \widehat{V}_{h+1})|$.

Let π^+ be a Markov best response to $\widehat{\nu}$ in P, and let ν^- be a Markov best response to $\widehat{\pi}$. Theorem 3 gives the deterministic inequality

$$\begin{aligned}\mathrm{Gap}_P(\widehat{\pi}, \widehat{\nu}) \le & \sum_{h=1}^{H} \mathbb{E}_{d_{P,h}^{\pi^+, \widehat{\nu}} \pi_h^+ \widehat{\nu}_h} \left[e_h + \delta_h^+(S_h) \right] \\ & + \sum_{h=1}^{H} \mathbb{E}_{d_{P,h}^{\widehat{\pi}, \nu^-} \widehat{\pi}_h \nu_h^-} \left[e_h + \delta_h^-(S_h) \right]. \end{aligned} \tag{36}$$

In particular, if the two unilateral state–joint-action laws are dominated by ρ_h with ratios C_h^+ and C_h^-, respectively, and $\|\delta_h^\pm\|_\infty \le \delta_h^\pm$, then

$$\mathrm{Gap}_P(\widehat{\pi}, \widehat{\nu}) \le \sum_{h=1}^{H} \left[(C_h^+ + C_h^-)\varepsilon_{h,\mathrm{B}} + \delta_h^+ + \delta_h^- \right]. \tag{37}$$

The deterministic inequality follows the two realized dynamic best responses. The dominated-law corollary converts those occupancies to the reference laws used by average PE and thereby identifies its exploration target.

For a known model P, define a pointwise one-step Bellman-evaluation interface $\mathsf{PE}_\infty(\varepsilon_{\mathrm{B}})$: given a bounded measurable continuation v, it returns $\widehat{Q}_h$ satisfying

$$\left\| \widehat{Q}_h - \{ r_h + P_h v \} \right\|_\infty \le \varepsilon_{\mathrm{B}}. \tag{38}$$

This interface predicts the supplied continuation and can be written as an ordinary PE query by using $v(S_{h+1})$ as an artificial state-only terminal reward and rescaling to the oracle's reward range. The distinction from the average-PE interface in Appendix K.2 is the norm: average $L_1(\rho_h)$ PE implies (38) only under an appropriate coverage or uniform-accuracy guarantee.

Run the same backward procedure and use T entropic no-regret iterations on each fixed matrix.

Corollary 33 (Pointwise-PE backward planning)**.** *The procedure uses H pointwise PE calls and explicit matrix no-regret updates and returns a Markov pair satisfying*

$$\mathrm{Gap}_P(\widehat{\pi}, \widehat{\nu}) \le 2H\varepsilon_{\mathrm{B}} + cH^2 \sqrt{\frac{\log A + \log B}{T}} \tag{39}$$

for an absolute constant c.

The proof in Appendix I.1 is a backward induction. It propagates local saddle errors along the two best-response trajectories.

Equations (30) and (36) turn the informal phrase "better exploration" into two precise missing objects. Model-based reward-free exploration establishes policy-uniform system identification through learned-model planning (Cheng et al., 2023). SpanRL constructs a policy cover under its reachability condition by interleaving policy optimization with the max–min RepLearn representation objective (Mhammedi et al., 2023a).

Density-ratio specialization of Theorem 26. In particular, suppose ρ is a unilateral cover in the sense of Definition 24, $\mathbb{E}_{\rho_h} D_{\mathrm{TV}}^2(P_h^\star, \widehat{P}_h) \le \zeta_h^2$, $\mathbb{E}_{\rho_h} e_h \le \varepsilon_{h,\mathrm{B}}$, and $\|\delta_h^\pm\|_\infty \le \bar{\delta}_h^\pm$. Then

$$\begin{aligned}\mathrm{Gap}_{P^\star}(\widehat{\pi}, \widehat{\nu}) \le & \sum_{h=1}^{H-1} H_{h+1} \big(\sqrt{C_{h,\mathrm{tr}}^+} + \sqrt{C_{h,\mathrm{tr}}^-} \big) \zeta_h \\ & + \sum_{h=1}^{H} \big[(C_{h,\mathrm{pe}}^+ + C_{h,\mathrm{pe}}^-)\varepsilon_{h,\mathrm{B}} + \bar{\delta}_h^+ + \bar{\delta}_h^- \big]. \end{aligned} \tag{40}$$

M.5 AUDITING THE TWO CANDIDATE ORACLE BRIDGES

We first examine whether RepLearn itself reduces to ordinary squared-loss SL. For a dataset $D = \{(x_i, a_i, x_i')\}_{i=1}^n$, write

$$L_D(\phi, w, f) = \sum_{i=1}^{n} \big(\phi(x_i, a_i)^\top w - f(x_i')\big)^2.$$

The feature-selection line of RepLearn is an ordinary multi-task SL call: after augmenting each example by a task index $\ell \in [t]$, use the predictor class

$$\mathcal{G}_t = \big\{(x, a, \ell) \mapsto \phi(x, a)^\top w_\ell : \phi \in \Phi,\ (w_1, \dots, w_t) \in \mathcal{W}^t\big\}$$

and squared-loss labels $f^{(\ell)}(x')$. Its empirical risk is exactly $t^{-1} \sum_{\ell=1}^t L_D(\phi, w_\ell, f^{(\ell)})$.

The discriminator line is different. With the current feature ϕ_0, it asks for

$$\Delta_D = \max_{f \in \mathcal{F},\, \widetilde{\phi} \in \Phi} \left\{ \min_{w} L_D(\phi_0, w, f) - \min_{\widetilde{w}} L_D(\widetilde{\phi}, \widetilde{w}, f) \right\}. \tag{41}$$

For fixed w, a shifted difference of the two sample losses gives an SL best response over $(f, \widetilde{\phi}, \widetilde{w})$, provided the interface allows this composite class. However, no-regret dynamics with those best responses compute the wrong saddle value in general:

$$\max_{z \in \{-1,1\}} \min_{w \in [-1,1]} (w - z)^2 = 0, \qquad \min_{w \in [-1,1]} \max_{z \in \{-1,1\}} (w - z)^2 = 1.$$

Here the alternative predictor fits each pure label z exactly. Thus convexifying the discriminator forces one shared current-regression weight across a mixture, whereas (41) refits w separately for each pure discriminator.

The same mismatch yields a computational separation.

Proposition 34 (Natural-ERM separation for the RepLearn discriminator)**.** *There is a family of succinct empirical instances of* (41) *with representation dimension two and bounded labels and features for which (i) on the fixed complete equally weighted empirical design, one-head squared-loss ERM over the discriminator-label and representation-predictor classes, for any rational response vector, is solvable to tolerance $\tau > 0$ in time polynomial in the input bit length and $\log(1/\tau)$, but (ii) computing Δ_D to inverse-polynomial additive accuracy is NP-hard. Consequently, unless $\mathsf{P} = \mathsf{NP}$, the discriminator cannot generically be implemented merely by alternating the natural constituent one-head regression routines.*

The proof in Appendix I.4 starts from the integer MAX-CUT incidence construction (Alon & Naor, 2006), completes its Gram matrix exactly, and embeds the resulting regressions into the one-action specialization of SpanRL's discriminator form $f(x) = \max_a \theta^\top \phi(x, a)$. The constituent-ERM claim concerns this fixed design, rather than arbitrary weighted samples or composite proper-MLE interfaces.

We next ask whether MLE itself supplies the missing unilateral cover. The strongest direct change-of-measure statement is the following conditional bridge.

Lemma 35 (Covered MLE implies policy-uniform system identification)**.** *Let ρ_h be the state–joint-action law used to fit $\widehat{P}_h$, and suppose that, for every centralized policy ω,*

$$d^{\omega}_{P^\star, h}(s)\omega_h(j \mid s) \le C_h \rho_h(s, j) \quad \textit{for all } (s, j), \qquad \mathbb{E}_{\rho_h} D^2_{\mathrm{TV}}(P^\star_h, \widehat{P}_h) \le \zeta_h^2. \tag{42}$$

Then

$$\mathfrak{E}_H(\widehat{P}) \le \sum_{h=1}^{H-1} H_{h+1} \sqrt{C_h}\, \zeta_h. \tag{43}$$

The lemma maps a unilateral density cover and covered MLE risk to policy-uniform identification. A horizon-two rank-two example demonstrates the role of the cover: two models agree on all transitions collected under one initial action and differ in total variation by one under the other, yielding zero on-law excess risk and $\mathfrak{E}_H = 1$. Cover generation therefore requires policy selection along likelihood-width directions, as implemented by learned-model planning or a representation discriminator in existing algorithms.

M.6 Block Markov game instantiation

Suppose observations $x \in \mathcal{S}$ admit an unknown decoder $\phi^\star : x \mapsto z \in \mathcal{Z}$, $|\mathcal{Z}| = S$, from a known class Φ; conditional on the latent state, emission supports are disjoint, and the latent transition depends only on (z, a, b). Centralizing $j = (a, b)$ makes this a Φ-decodable Block MDP with $J = AB$ actions. The PCE reduction of Rohatgi & Foster (2025) uses only two-context regression and episodic interaction to return, for each stage, a $(1, \eta)$-policy cover Ψ_h of size at most $M \le H^2S^2$, with polynomial sample and oracle complexity and without a reachability assumption. The cover policies may use centralized joint actions during reward-free training; the returned game policies remain product policies.

For each h, let ρ_h roll in a uniformly chosen $\psi \in \Psi_h$ and choose J_h uniformly. The following elementary conversion is what the general low-rank argument lacks.

Lemma 36 (Block policy cover gives approximate density coverage)**.** *For every centralized policy ω and every measurable $g_h \in [0, 1]$,*

$$\mathbb{E}_{d^{\omega}_{P^\star,h}\omega_h} g_h \le 2S\eta + 2MJ\,\mathbb{E}_{\rho_h} g_h, \tag{44}$$

$$\mathbb{E}_{d^{\omega}_{P^\star,h}\omega_h} g_h \le 2S\eta + \sqrt{2MJ\,\mathbb{E}_{\rho_h} g_h^2}. \tag{45}$$

The additive term contains only latent states whose maximum possible visitation is below 2η.

Assume a proper transition MLE returns $\widehat{P}$ with $\mathbb{E}_{\rho_h} D^2_{\mathrm{TV}}(P^\star_h, \widehat{P}_h) \le \zeta_h^2$, and set

$$\Delta_{\mathrm{sys}} = \sum_{h=1}^{H-1} H_{h+1} \left(2S\eta + \sqrt{2MJ}\,\zeta_h\right). \tag{46}$$

Lemma 36 implies $\mathfrak{E}_H(\widehat{P}) \le \Delta_{\mathrm{sys}}$.

For Corollary 29, assume that PCE has access to a polynomial-sample two-context regression oracle, Algorithm 3's proper MLE satisfies the displayed $L_2(\rho_h)$ event, and each backward Bellman PE call in $\widehat{P}$ has $\mathbb{E}_{\rho_h} e_h \le \varepsilon_{h,\mathrm{B}}$ with clipped error $e_h \le H$. If the local matrix residuals are at most δ_h on both sides, the returned pair obeys

$$\mathrm{Gap}_{P^\star}(\widehat{\pi}, \widehat{\nu}) \le 2\Delta_{\mathrm{sys}} + 2\sum_{h=1}^{H} \left(2SH_h\eta + 2MJ\varepsilon_{h,\mathrm{B}} + \delta_h\right). \tag{47}$$

Consequently, polynomially small $\eta, \zeta_h, \varepsilon_{h,\mathrm{B}}, \delta_h$ give a polynomial-sample ε-Nash guarantee using two-context SL, proper MLE, average PE, and explicit matrix updates.

For an explicit rate, let $N_{2\mathrm{ctx}}(\alpha, \beta)$ denote the labeled-sample complexity of two-context regression at excess risk α and failure probability β. The PCE theorem of Rohatgi & Foster (2025) gives a universal constant C_{PCE} such that, with

$$\Gamma(\eta, \delta_{\mathrm{c}}) = \frac{HJS}{\eta\delta_{\mathrm{c}}}, \qquad \mathcal{N}_{\mathrm{PCE}}(\eta, \delta_{\mathrm{c}}) = \Gamma(\eta, \delta_{\mathrm{c}})^{C_{\mathrm{PCE}}} N_{2\mathrm{ctx}}\left(\Gamma(\eta, \delta_{\mathrm{c}})^{-C_{\mathrm{PCE}}}, \Gamma(\eta, \delta_{\mathrm{c}})^{-C_{\mathrm{PCE}}}\right), \tag{PCE}$$

both its episodic sample complexity and its two-context oracle time are at most $\mathcal{N}_{\mathrm{PCE}}$, and its cover has size at most $\overline{M} = H^2S^2$. Oracle time charges each regression query linearly in its input length, as in the cited computational model; it does not include the unrestricted cost of solving the regression instance.

Corollary 37 (Explicit finite-class Block-MG complexity)**.** *Suppose Assumption 6 holds and let $\varepsilon, \delta_{\mathrm{fail}} \in (0, 1)$. Put $J = AB$, $\overline{M} = H^2S^2$, $\overline{\kappa} = 2\overline{M}J$, and choose*

$$\eta = \frac{\varepsilon}{32SH^2}, \qquad \zeta = \frac{\varepsilon}{8H^2\sqrt{\overline{\kappa}}}, \qquad u = \frac{\varepsilon}{8H\sqrt{\overline{\kappa}}}, \qquad \delta_{\mathrm{mat}} = \frac{\varepsilon}{8H}. \tag{48}$$

Let

$$L_\Theta = 2\log\frac{6H\Theta_\star}{\delta_{\mathrm{fail}}} + \xi_{\mathrm{mle}}, \qquad L_Q = \log\frac{6HQ_\star}{\delta_{\mathrm{fail}}},$$

and use, at every stage,

$$n_{\mathrm{mle}} = \left\lceil \frac{L_\Theta}{\zeta^2} \right\rceil, \qquad n_{\mathrm{pe}} = \left\lceil \frac{c_{\mathrm{pe}} H^2 L_Q}{u^2} \right\rceil, \qquad T = \max\left\{1, \left\lceil \frac{64 c_0^2 H^4 (\log A + \log B)}{\varepsilon^2} \right\rceil \right\}, \tag{49}$$

where c_{pe} and c_0 are universal constants from finite-class squared-loss regression and entropic matrix regret, respectively. Then Algorithm 3 returns an ε-Nash pair with probability at least $1 - \delta_{\mathrm{fail}}$. A sample-split implementation uses at most

$$\begin{aligned} N_{\mathrm{env}} &\le \mathcal{N}_{\mathrm{PCE}}(\eta, \delta_{\mathrm{fail}}/3) + H(n_{\mathrm{mle}} + n_{\mathrm{pe}}) \\ &= \mathcal{N}_{\mathrm{PCE}}(\eta, \delta_{\mathrm{fail}}/3) + O\left(\frac{H^7 S^2 J \, (L_\Theta + L_Q)}{\varepsilon^2} \right) \end{aligned} \tag{50}$$

episodic trajectories. It makes one PCE call, one joint proper-MLE call (equivalently H stagewise fits), and H average-PE calls. Each queried policy matrix requires T update rounds; the total matrix work therefore depends on the number of continuation and output-policy queries. There is no enumeration of the observation space.

Composition with the game solver. The PCE cover and its cost in (PCE) are inherited after centralizing $j = (a, b)$. Apply Theorem 3 directly in the true kernel, bound its residual by learned-model PE error plus one-step transition error, and use Lemma 36 for both terms. This gives Corollary 29 without passing through learned-model occupancies. Theorem 23 remains a sharp independent model-transfer result, while Theorem 26 supplies an alternative certificate when a learned-model evaluation cover is given. The centralized cover contributes $J = AB$; this result does not establish an end-to-end $A + B$ action dependence. The concrete implementation and real-successor comparison appear in Appendix D.

N Uniform baseline and shared downstream routines

We expand the uniform Block-MG baseline, then specify its matrix queries and the real-successor alternative. For Algorithm 1, replace each uniform cover draw below by its independently calibrated weights and apply the real-successor modification specified below. The local matrix routines are shared by both fitting routes.

PCE is the inherited reward-free cover reduction; after cover construction, all optimization consists of proper MLE, average PE, and explicit finite matrix updates.

Explicit matrix queries. For a fixed matrix $Q \in [0, H_h]^{A \times B}$, initialize $p_1 = \mathrm{Unif}(\mathcal{A})$, $v_1 = \mathrm{Unif}(\mathcal{B})$. At each round use

$$p_{t+1}(a) \propto p_t(a) \exp\{\gamma (Q v_t)_a\}, \qquad v_{t+1}(b) \propto v_t(b) \exp\{-\gamma (p_t^\top Q)_b\}.$$

Return the two separately averaged strategies $\bar{p}, \bar{v}$. With γ of order $\sqrt{(\log A + \log B)/(H_h^2 T)}$, the sum of external regrets divided by T is at most $c_0 H_h \sqrt{(\log A + \log B)/T}$, after enlarging the universal constant. The matrix duality gap $\max_p p^\top Q \bar{v} - \min_v \bar{p}^\top Q v$ is bounded by this sum. Both nonnegative residuals around $\bar{p}^\top Q \bar{v}$ are at most that gap. Each round costs $O(AB)$; this is a per-query operation, not a loop over an infinite observation space.

Real-successor fitted minimax. Omit all MLE buffers and fits. In the PE buffer retain (S_h, J_h, S_{h+1}), and use this recorded true successor in the backward label. Keep stagewise sample splitting and the same matrix procedure. Corollary 18 applies when the class approximates the true Bellman mean. In the coordinate instance both routes fit $r_h + g \circ \phi_i$ by the explicit cellwise ERM. With known rewards, the terminal predictor is r_H and both routes can omit terminal samples; the main sample count conservatively includes them.

O Diagnostic controls and data corrections

The main experiments concern observable mixture calibration. Bridge, blind-spot, deep-chain, and active-occupancy diagnostics serve as implementation checks; their raw outputs and scripts are in

Algorithm 3: BLOCKMG-PE: detailed implementation

Input: cover accuracy η; sample sizes $n_{\mathrm{mle}}, n_{\mathrm{pe}}$; PE tolerances $\{\varepsilon_{h,\mathrm{B}}\}$; matrix iterations T; decoder class Φ; two-context oracle.

Output: a Markov product pair $(\widehat{\pi}, \widehat{\nu})$.

Replace each (a, b) by the centralized exploration action $j \in [J]$
Run PCE at accuracy η, retain covers $\Psi_1, \ldots, \Psi_H$, and pad them to common size $M \le H^2 S^2$
for $h = 1, \ldots, H$ **do**
  for $i = 1, \ldots, n_{\mathrm{mle}}$ **do**
    Draw $\psi \sim \mathrm{Unif}(\Psi_h)$, roll in through stage $h-1$, draw $J_h \sim \mathrm{Unif}([J])$, observe S_{h+1}, and append (S_h, J_h, S_{h+1}) to $\mathcal{D}_h^{\mathrm{mle}}$
  end
  for $i = 1, \ldots, n_{\mathrm{pe}}$ **do**
    Repeat with an independent episode and append only (S_h, J_h) to $\mathcal{D}_h^{\mathrm{pe}}$
  end
  Let ρ_h denote the common population context law
  Fit $\widehat{P}_h$ by proper MLE on $\mathcal{D}_h^{\mathrm{mle}}$
end
Set $\widehat{V}_{H+1} = 0$
for $h = H, H-1, \ldots, 1$ **do**
  foreach $X = (S_h, J_h) \in \mathcal{D}_h^{\mathrm{pe}}$ **do**
    Draw $\widetilde{S}_{h+1} \sim \widehat{P}_h(\cdot \mid X)$ and label X by $r_h(X) + \widehat{V}_{h+1}(\widetilde{S}_{h+1})$
  end
  Call average Bellman PE to obtain $\widehat{Q}_h^{\mathrm{raw}}$, and set $\widehat{Q}_h = \mathrm{clip}_{[0,H_h]}\, \widehat{Q}_h^{\mathrm{raw}}$
  At a policy query s, run T entropic matrix no-regret iterations on $\widehat{Q}_h(s)$ and average both players' iterates to obtain $(\widehat{\pi}_h, \widehat{\nu}_h)$
  Set $\widehat{V}_h = \langle \widehat{\pi}_h, \widehat{Q}_h \widehat{\nu}_h \rangle$
end
return $(\widehat{\pi}, \widehat{\nu})$ *and the bound in* (47).

the artifact. They do not establish a performance advantage of the calibrated sampler. We document the metric corrections and recurrent-reward control needed to interpret the fitting comparisons in Appendix H.

Recurrent-reward control. The recurrent-reward hidden-coordinate game uses transition-to-state-one matrices $\left(\begin{smallmatrix} .85 & .25 \\ .15 & .65 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} .75 & .20 \\ .30 & .90 \end{smallmatrix}\right)$ in latent states zero and one, with reward $r_h(x, a, b) = \phi^\star(x)$. The initial latent state is uniform. It uses $p = 4$, $H \in \{3, 5\}$, 12 seeds, denominator-20 proper MLE, quantized coordinate-head PE, and exact local saddles. PCE has $2H$ rounds, 64 regression samples, 24 tests and centers, $\gamma = .15$, and $\gamma' = .4$. Complete budgets are $N_{\mathrm{PCE}} + 2(H-1)n$ for $n \in \{64, 256, 1024\}$. At $H = 5, n = 1024$, the mean cap is 95,978.67 episodes. The model/real PCE gaps are $0.19578/0.09072$, and the uniform-action counterparts are $0.03968/0.02679$. The legal reward-decoded baseline and its nonzero-gap seed are reported in Appendix H. Its known reward reveals a decoder, so it is an easy control rather than a hidden-representation exploration test.

Legacy PCE metric. The 270-row legacy PCE sanity dataset is preserved unchanged. Its old cover deficit subtracted the sum of member occupancies, whereas the cover definition requires their maximum. That deficit is withdrawn: per-member occupancies were not saved, so it cannot be repaired from the old aggregate. The corrected metric is invariant to duplicating a policy and has a regression test. Cover construction cost is charged only to methods that use it, and the privileged latent cover remains an unpriced diagnostic. Smoothed historical estimators do not verify the new finite-class MLE theorem. The artifact records the original data and the separate accounting audit.

O.1 Reward-span extension

Theorems in the main text use rewards in $[0, 1]$. Historical deep-chain diagnostics instead use terminal rewards $(+1, -1, 0)$ and zero earlier rewards. For rewards $r_h \in [a_h, a_h + L]$, a continuation has span at most LH_{h+1}, so

$$|(P_h^\star - \widehat{P}_h)V| \le LH_{h+1} D_{\mathrm{TV}}(P_h^\star, \widehat{P}_h).$$

Theorem 23 therefore multiplies its two transition terms by L; Theorem 3 keeps the actual Bellman and matrix residuals unchanged. For the signed deep chain $L = 2$. Historical exports used unit-span TV weights; the corrected certificate is $\mathrm{Gap}_{\widehat{P}} + 2\mathfrak{E}_H^+ + 2\mathfrak{E}_H^-$. The correction preserves gaps, policies, seed identifiers, and PE errors. Sample-mean backward fits and local saddles remain in $[-1, 1]$, so clipping to that interval is the identity in this benchmark. For fixed policies, replacing terminal reward by $(r_H + 1)/2$ halves the gap and its standard error; it does not describe retraining under a changed reward scale. The artifact retains both the raw signed data and the separate certificate-correction table.

P Extended related work

Tabular and structured zero-sum Markov games. Early dynamic-programming formulations trace to stochastic and Markov games (Shapley, 1953; Littman, 1994). Modern tabular analyses establish polynomial self-play guarantees and near-optimal dependence on the horizon, state space, and action sizes (Bai & Jin, 2020; Bai et al., 2020; Liu et al., 2021). Generative-model algorithms attain minimax rates under stronger sampling access (Li et al., 2022b), while model-free refinements improve sample efficiency under episodic interaction (Feng et al., 2024). For known structure, fictitious-play and no-regret methods exploit factored or single-controller transitions (Qiu et al., 2021); our setting instead has an unknown predictive representation shared by both players' actions.

Function approximation and policy optimization in games. Known linear features support optimistic equilibrium learning in simultaneous- move and linear-mixture games (Xie et al., 2020; Chen et al., 2022), and kernelized or neural classes have been treated through confidence bounds and correlated-equilibrium subroutines (Li et al., 2022a). More general complexity measures include posterior-sampling and offline frameworks for general function approximation (Qiu et al., 2023; Zhang et al., 2023b). Actor-based analyses include natural actor–critic, optimistic policy optimization, statewise regret, regularized gradient dynamics, and spectral-embedding policy optimization for continuous-state zero-sum games (Alacaoglu et al., 2022; Zhao et al., 2022; Zhang et al., 2022; Zeng et al., 2022; Zhou et al., 2024). Recent results also study independent function approximation under local access and policy regret against adaptive opponents (Fan et al., 2024; Nguyen-Tang & Arora, 2025). Our analysis focuses on the complementary interface between representation-dependent coverage and MLE/average-PE equilibrium computation.

Offline games and unilateral coverage. Offline zero-sum learning makes the deviation-coverage issue explicit. Single-strategy coverage is insufficient in general, whereas unilateral concentrability supports pessimistic equilibrium learning (Cui & Du, 2022a). Data-dependent pessimism, relative uncertainty, and strategy-wise bonuses sharpen this principle under linear or tabular structure (Zhong et al., 2022; Cui & Du, 2022b). Extensions cover general function approximation and corrupted datasets (Zhang et al., 2023b; Nika et al., 2024); KL-regularized offline self-play provides a recent alternative to explicit pessimism under unilateral concentrability (Chen et al., 2026). Our learned collection law covers all centralized policies, hence both unilateral deviation families, and is selected before the independent fitting buffers are collected.

Unknown low-rank representations. FLAMBE initiated model-based representation learning for low-rank MDPs (Agarwal et al., 2020); subsequent work developed model-free, online/offline, density-feature, and reward-free variants (Modi et al., 2024; Uehara et al., 2022; Huang et al., 2023; Cheng et al., 2023; Mhammedi et al., 2023a). The model has also been studied with representation selection and continuous actions (Zhang et al., 2023a; Oprescu et al., 2024). The closest game result is Ni et al. (2023), which learns equilibria in unknown-representation general-sum low-rank games using model-based or model-free representation learning followed by equilibrium planning. We instead characterize a coverage interface under which proper transition estimation and average PE support equilibrium computation.

The model-based and model-free rates of Ni et al. (2023) include exploration, whereas our displayed post-cover rate must be added to the source PCE cost. Their integrated representation learning and planning also use different oracle access; these displays do not establish rate dominance. Appendix M.1 compares the exact rates and assumptions; a general low-rank cover provider remains open.

Rich observations, policy covers, and oracle complexity. Bellman and witness rank formalize strategic exploration with rich observations (Jiang et al., 2017; Sun et al., 2019). For Block MDPs, latent decoding, kinematic abstraction, and reward-free exploration construct policy covers (Du et al., 2019; Misra et al., 2020; Mhammedi et al., 2023b; Jedra et al., 2023; Jin et al., 2020; Wagenmaker et al., 2022); coverability extends this view beyond Block models (Amortila et al., 2024). Prediction-only oracles can nevertheless be weaker than reward-directed exploration (Golowich et al., 2024). Rohatgi & Foster (2025) identify two-context regression as sufficient—and, under regularity, necessary—for Block-MDP exploration, while giving low-rank hardness evidence. We learn a sampling mixture over their finite cover and convert its coverage into a Nash guarantee. The population equivalence and the comparison of observable optimization interfaces appear in Appendices B and E. The MLE/PE actor–critic interface of Huang et al. (2026) is the single-agent counterpart; unilateral coverage is its two-player analogue.

Q Limitations and directions for further work

Structural and computational assumptions. The learned-mixture guarantee requires exact Block structure with common, policy-independent emissions and a finite evaluable decoder class containing the true decoder. These assumptions imply $R \leq \min\{M, S\}$ for any library. They do not make an arbitrarily large decoder class computationally tractable: the statistical cost depends logarithmically on its size, but event evaluation and LP size depend linearly on it. Finite-fitting guarantees require Bellman closure or the controlled approximation proved for the coordinate class; the model-based route additionally requires realizable proper models. Extending the population equivalence and calibration guarantee to misspecified decoders or approximately common emissions would require additional error terms.

Which costs improve. Calibration improves the sampling law of an available cover and retains the provider's full construction cost. It does not change the leading exponent when inherited PCE dominates. The episode comparisons and operation bounds are distinct: polynomial implementability alone does not imply a practical runtime gain. The three-stage examples identify regimes where fitting savings also dominate calibration computation, whereas the measured PCE experiments show no overall runtime advantage.

Scope of the sampling separations. The uniform-library lower bounds constrain roll-ins to independent uniform draws while allowing arbitrary downstream learning. Their product dependence does not apply to adaptive exploration. A public entrance permits uncalibrated task-directed sampling; when the entrance is unknown, the example counts its M root actions and admits a return-directed search with $O(M \log(1/\delta)) + \widetilde{O}(\varepsilon^{-2})$ episodes. Constant-precision event calibration attains the same sample order, matching the separate adaptive lower bound on that family up to logarithms. This instance has no rare-state calibration difficulty and does not establish necessity of the general ε^{-1} calibration term. Other optimizers that meet the event constraints can also use the same coverage guarantee. No statistical advantage of transition MLE over real-successor fitted minimax is established.

Empirical evidence. The synthetic experiments use practical PCE and calibration budgets below the theorems' sufficient counts. A feasible empirical LP is therefore not, by itself, a high-probability coverage certificate. Strong worst-case coverage need not minimize a task-specific Nash gap, and calibration leaves fewer samples for fitting at a fixed budget. The calibration scan shows nonmonotone response errors and Nash gaps. Shared-data reuse is allowed for every method; the three predefined rewards include a symmetric success/failure pair, so they are not three independent hard instances. Largest-budget LP-versus-equal gap intervals include zero, and regularized coverage is competitive on the same data. Adaptive calibration in broader environments remains open.