IDENTIFICATION AND OFF-POLICY EVALUATION IN MARKOV GAMES WITH ADAPTIVE OPPONENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deploying a policy in a multi-agent system can change how the other agents behave. Off-policy evaluation should therefore measure the policy against the opponent response induced by deployment, not against an average opponent in the log. We show that this induced-response value is unidentified without observable response structure and cannot be estimated when the target lacks support on relevant state-joint-action triples or state-response pairs. We introduce a finite response-state model in which the response at stage h depends only on earlier response states. The induced path is therefore obtained by forward recursion, not by solving a simultaneous fixed point. Given analyst-specified response maps $\phi_{1:H}$, our polynomial-time method reconstructs response labels for the logging policies, estimates an opponent-response kernel, recovers the response to each undeployed target in one forward pass, and applies pessimistic off-policy evaluation on the target-reachable support. When overlap is incomplete, we give outer confidence intervals under partial identification for completions that share a response path. We also bound the value error caused by response-model misspecification. Together, these results distinguish structural identification from response recovery and statistical uncertainty.

1 INTRODUCTION

Off-policy evaluation (OPE) estimates the value of a target policy from data generated by one or more logging policies, usually under a fixed environment (Precup et al., 2000; Jiang & Li, 2016). That premise is inappropriate when users, markets, or competing agents adapt to the deployed policy. If a learner deploys policy π and thereby induces opponent policy μ^π, the relevant target is

$$V_{\mathrm{cf}}(\pi) := V^{\pi,\mu^\pi}. \tag{1}$$

We call $V_{\mathrm{cf}}(\pi)$ the *induced-response value*: the value of π against the opponent policy that deployment of π induces. The logs may contain many learner-opponent pairs, yet none need reveal how the opponent would react to an undeployed target. This is a structural problem before it is an optimization problem.

The problem persists even with full ordinary support. Consider a one-step game with one state and two actions per player. Let the response state be $z = \mathbf{1}\{\pi(a_2) \geq 1/2\}$, with $z = 0$ producing opponent action b_1 and $z = 1$ producing b_2. Logging mixes equally between policies with $\pi_L(a_2) = 1/4$ and $\pi_H(a_2) = 3/4$, so all four joint-action pairs and both response states are observed. Set $r(a_2, b_1) = 1$ and all other rewards to zero. An estimator that pools the opponent regimes then converges to $\mathbb{E}_{\log}[r \mid a_2] = 1/4$. By contrast, the target $\pi_{\mathrm{tar}}(a_2) = 1$ induces b_2 and has $V_{\mathrm{cf}}(\pi_{\mathrm{tar}}) = 0$. A lower confidence bound for the mixture-environment value $V^{\pi,\mu_{\log}}$ therefore converges to $1/4$. Making that bound conservative does not change its estimand and cannot recover $V_{\mathrm{cf}}(\pi_{\mathrm{tar}})$. Both coverage coefficients defined below remain finite in this example: $C_{\mathrm{cf}} = 8/3$ and $C_{\mathrm{sr}} = 2$.

We study when policy-labeled offline trajectories determine (1), how to recover the induced response, and how estimation error should be propagated after recovery. Our contributions are:

- We prove two barriers. Without observable response structure, observationally equivalent adaptive games can disagree by a constant on (1). With structure but missing target support, the minimax error scales at least linearly with the uncovered target mass.

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.

- We prove identification on target-reachable support under a finite response-state model and separate state-joint-action coverage from state-response coverage. Neither condition implies the other.
- Under simultaneous-action timing, the response operator is triangular. The induced path is therefore unique and computable in $O(HS^2AB)$ time per policy.
- We give an end-to-end finite-sample theorem. From policy-labeled trajectories and specified $(\mathcal{Z}, \phi)$, the method reconstructs response labels for the logged policies, recovers each target path by a forward pass, and evaluates that path at the usual pessimistic rate.
- We quantify two failure regimes: incomplete support yields a valid outer interval under response stability, and an ε error in the induced opponent policy adds at most $\min\{H, H(H+1)\varepsilon/4\}$ to value error.

The positive result applies when the response abstraction is part of the system specification, the logging design repeatedly assigns policies from a finite set, and the target-reachable state-joint-action and state-response pairs satisfy the required overlap conditions. Such conditions can be justified in controlled deployments, including posted-price experiments, traffic-routing pilots, and recommendation policies assigned by an experiment platform. Without independent support for the response abstraction and its coverage, the data warrant the negative results or outer intervals, not a point estimate.

2 Related Work

Adaptive opponents and policy regret. Policy regret compares a learner with a policy evaluated against the response that policy would itself induce (Arora et al., 2012). Recent Markov-game results extend this idea to adaptive opponents and function approximation (Nguyen-Tang & Arora, 2024; 2025). Their learner continues to interact and controls regret over time. Their consistent-behavior condition is a Lipschitz restriction on $\pi \mapsto \mu^\pi$; we instead take a designed finite partition of occupancy space, which allows jumps across bins but requires the analyst to specify the bins.

Offline games and opponent modeling. Offline Markov-game methods commonly estimate equilibria or exploitability in a fixed game (Zhang et al., 2023; Cui & Du, 2022; Zhong et al., 2022; Chen et al., 2026b; Lu et al., 2025; Chen et al., 2026a). Offline opponent models learn representations or policies that adapt to an opponent at test time (Jing et al., 2024; 2025; Wu et al., 2023). Our target differs from both: deploying the learner changes the opponent law, so the target environment must be identified before it can be evaluated. Once that environment is identified, we use familiar pessimistic dynamic programming (Jin et al., 2021; Rashidinejad et al., 2021; Xie et al., 2021).

Decision-dependent environments and identification. Performative prediction and reinforcement learning study distributions or dynamics changed by deployed decisions, usually through stability, optimality, or repeated retraining (Perdomo et al., 2020; Mandal et al., 2023; Rank et al., 2024; Mandal & Radanovic, 2025; Sahitaj et al., 2025). Policy-conditioned environment models use the evaluation policy to improve generalization under offline distribution shift (Chen et al., 2024); conditioning alone does not identify a policy-induced opponent response that was absent from the logs. Dynamic-game econometrics supplies complementary non-identification and structural identification results (Kalouptsidi et al., 2017; 2021).

Off-policy and preference-based evaluation. Classical OPE estimates a fixed target in a fixed environment (Precup et al., 2000; Jiang & Li, 2016); causal OPE shows that even that target can require additional assumptions under latent confounding (Bareinboim et al., 2015; Bennett et al., 2021; Kausik et al., 2024). Offline preference learning likewise uses pessimism to handle incomplete comparison coverage (Zhu et al., 2023; Zhan et al., 2024). Our uncertainty is different: changing π changes a dynamic opponent policy before the value can be defined from the log. Appendix A.2 gives a compact comparison, and later appendix sections place the remaining connections beside the corresponding results.

3 Problem and Response Structure

Markov game, adaptive opponent, and offline data. Let $\mathcal{G} = (\mathcal{S}, \mathcal{A}, \mathcal{B}, P, \mathsf{R}, H, \rho)$ be a finite-horizon simultaneous-action Markov game with known initial distribution ρ. Write $S = |\mathcal{S}|$, $A = |\mathcal{A}|$,

and $B = |\mathcal{B}|$. At stage h, the agents draw $a_h \sim \pi_h(\cdot \mid s_h)$ and $b_h \sim \mu_h(\cdot \mid s_h)$ simultaneously, draw a reward $R_h \in [0,1]$ from the common Markov kernel $\mathsf{R}_h(\cdot \mid s_h, a_h, b_h)$ with mean $r_h(s_h, a_h, b_h)$, and draw $s_{h+1} \sim P_h(\cdot \mid s_h, a_h, b_h)$. The reward and next-state draws are conditionally independent given (s_h, a_h, b_h). For a policy pair,

$$V^{\pi,\mu} := \mathbb{E}_{\rho,P,\pi,\mu}\left[\sum_{h=1}^{H} R_h\right].$$

The learner selects a Markov policy π from a finite target class Π; deployment induces an opponent policy μ^π. We observe n independent episodes

$$\mathcal{D} = \{(\pi^{(i)}, \tau^{(i)})\}_{i=1}^{n}, \qquad \pi^{(i)} \sim \nu_{\log},$$

where the exact learner-policy label is recorded and $\tau^{(i)} = ((s_h^{(i)}, a_h^{(i)}, b_h^{(i)}, R_h^{(i)})_{h=1}^{H}, s_{H+1}^{(i)})$ is generated under $(\pi^{(i)}, \mu^{\pi^{(i)}}, P)$. The support $\Pi_{\log} := \mathrm{supp}(\nu_{\log})$ is a finite logging policy set of size M. Write $\nu_{\min} := \min_{\pi_0 \in \Pi_{\log}} \nu_{\log}(\pi_0)$ for the minimum logging-policy probability and $\nu_\star := \min_{\pi \in \Pi} \nu_{\log}(\pi)$ for the smallest target-policy probability, which is zero if some target is unlogged. Response-state labels are not recorded, and a target policy need not belong to the logging policy set.

For any policy pair, let $d_h^{\pi,\mu}(s,a,b)$ be the stage-h occupancy and let $d_h^{\pi,\mu}(s,a)$ and $p_h^{\pi,\mu}(s)$ be its learner and state marginals. The logging occupancy is $d_h^{\log}(s,a,b) = \mathbb{E}_{\pi_0 \sim \nu_{\log}} d_h^{\pi_0,\mu^{\pi_0}}(s,a,b)$.

To represent a policy-dependent opponent, we introduce the following finite abstraction.

Definition 3.1 (Finite response-state model)**.** *There is a finite set $\mathcal{Z}$, a stationary opponent-response kernel $g(\cdot \mid s, z) \in \Delta(\mathcal{B})$, and analyst-specified maps ϕ_h on learner-occupancy prefixes such that, for every π,*

$$z_h^\pi = \phi_h(d_{1:h}^\pi), \qquad \mu_h^\pi(b \mid s) = g(b \mid s, z_h^\pi).$$

The analyst specifies $(\mathcal{Z}, \phi_{1:H})$; the opponent-response kernel is learned from data. The arguments of ϕ_h are learner state-action occupancy prefixes $d_h^\pi(s,a) = p_h^\pi(s)\pi_h(a \mid s)$, not joint occupancies that already include the current opponent action. We call z^π the induced response path.

This is a designed abstraction, not a representation theorem for arbitrary opponents. The estimator queries ϕ_h only at occupancy prefixes generated by $\Pi \cup \Pi_{\log}$ (Lemma B.14). Two maps that agree on those prefixes produce the same labels and lower confidence bounds. A coarser map may mix distinct response-kernel rows and recreate the pooling bias in the introduction. A finer map can preserve the model by duplicating response-kernel rows, but the refined target cells must also be covered. Without a defensible map, the negative result below applies.

We assume a stationary opponent-response kernel so that response information can transfer across stages. The learner policy, transition kernel, reward, response map, and induced response state may still vary by stage. Each map ϕ_h acts on a population learner occupancy, not on one realized trajectory. Thus the response label is not a separate latent variable for every episode: episodes generated by the same logging policy share a population response sequence, which can be reconstructed by pooling episodes with that policy label.

Lemma 3.2 (Triangularity and forward recursion)**.** *Under Definition 3.1, the coordinate $(T_\pi(z))_h := \phi_h(d_{1:h}^{\pi,z})$ depends only on $z_{1:h-1}$. Hence T_π has exactly one fixed point, obtained by the forward recursion*

$$\begin{aligned} p_1 &= \rho, & d_h(s,a) &= p_h(s)\pi_h(a \mid s), \\ z_h &= \phi_h(d_{1:h}), & p_{h+1}(s') &= \sum_{s,a,b} d_h(s,a) g(b \mid s, z_h) P_h(s' \mid s, a, b). \end{aligned}$$

The same conclusion holds for any plug-in stochastic kernels $(\widehat{P}, \widehat{g}_h)$.

The timing matters. The current learner marginal is determined before the current opponent action affects the next state. Models in which the response map uses a same-step quantity that already depends on z_h require a different analysis.

Two-stage finite-log example. The following two-stage example illustrates the forward construction. Let $z_1 = \mathbf{1}\{p_\pi \geq 1/2\}$, where the target assigns probability $p_\pi = 0.8$ to a designated first-stage action. Then $z_1 = 1$, with threshold slack 0.3. Under this response, the first-stage kernel reaches a designated state s_G with probability 0.8. At stage two, let $z_2 = \mathbf{1}\{\frac{1}{2}\Pr(s_2 = s_G) + q_\pi \geq 1/2\}$, where the target's known probability of action 1 is $q_\pi = 0.3$. The summary is 0.7, so $z_2 = 1$ with slack 0.2. The empirical forward pass recovers both labels whenever its plug-in summaries remain within these slacks. With $|\mathcal{Z}| = 2$, the procedure evaluates two coordinates rather than enumerating four response sequences. The same induction extends to general H; Appendix H.9.1 specifies the kernels.

Coverage. Define

$$C_{\mathrm{cf}} := \sup_{\pi \in \Pi, h, s, a, b} \frac{d_h^{\pi, \mu^\pi}(s, a, b)}{d_h^{\log}(s, a, b)}, \tag{2}$$

$$C_{\mathrm{sr}} := \sup_{\pi \in \Pi, h, s: p_h^\pi(s) > 0} \frac{p_h^\pi(s)}{d_h^{\log}(s, z_h^\pi)}, \qquad d_h^{\log}(s, z) := \mathbb{P}(s_h = s, z_h^{\pi^{(i)}} = z). \tag{3}$$

Write $\bar{C} := \max\{C_{\mathrm{cf}}, C_{\mathrm{sr}}\}$. We use $0/0 = 0$ and $x/0 = \infty$ for $x > 0$. The coefficient C_{cf} is a joint-occupancy concentrability coefficient over state-joint-action triples. The coefficient C_{sr} is the analogous coverage coefficient for the opponent-response kernel and is specific to our response-state abstraction. Population identification may borrow a stationary (s, z) cell from another stage because g does not depend on h. The finite-sample theorem estimates stage-specific copies $\widehat{g}_h$ of the stationary opponent-response kernel and therefore uses same-stage counts in C_{sr}. Appendix F gives a conservative cross-stage estimator when same-stage support is absent.

The coefficients are logically independent. A logging policy can visit every state-joint-action triple while visiting a state only under the wrong response label, making C_{cf} finite and C_{sr} infinite. Conversely, it can cover the target (s, z) response-kernel rows while omitting a learner-opponent action pair needed to identify a reward or transition. This separation is why one ordinary concentrability coefficient does not characterize OPE with adaptive opponents: when $|\mathcal{Z}| = 1$ the method reduces to ordinary pessimistic evaluation, and even under benign state coverage, response estimation still costs $\Omega((H-1)\min\{1, \sqrt{|\mathcal{Z}|/n}\})$ (Theorem G.4).

Theorem 3.3 (Barriers and identification on target-reachable support)**.** *Without an observable restriction on $\pi \mapsto \mu^\pi$, there are two adaptive games with the same offline law and different $V_{\mathrm{cf}}(\pi)$, giving constant minimax error. Under Definition 3.1, if target mass α reaches an unlogged state-joint-action triple, two compatible games can differ in value by α, so minimax error is at least $\alpha/2$. Conversely, finite C_{cf} and state-response coverage identify the target-reachable rows of P, R, r, g, and hence the induced response path and trajectory law on the target-reachable support.*

The first claim changes only the response rule outside the logging policy set. The second changes an unobserved reward row reached with target mass α. For the positive direction, state-joint-action coverage identifies every target-reachable row of (P, R), including its mean r, and state-response coverage identifies the response-kernel row $g(\cdot \mid s, z_h^\pi)$. Lemma 3.2 then identifies z^π inductively and hence the entire target trajectory law. Appendix B gives explicit paired games and the induction. These constructions are worst-case: without further restrictions, two unidentified games need not have similar values, and uncovered mass controls the spread only after the response path is held fixed.

Passive finite data cannot uniformly certify population support. For an indicator that records whether a designated cell is reached, success probabilities zero and ϵ induce n-sample laws at total-variation distance $1 - (1-\epsilon)^n \leq n\epsilon$. A zero count therefore cannot distinguish absent support from rare support. Known assignment probabilities can still bound the coverage coefficients: if $\nu_\star > 0$, then $C_{\mathrm{cf}} \leq 1/\nu_\star$ and $C_{\mathrm{sr}} \leq 1/\nu_\star$ (Lemma B.15). If the target is undeployed but a logged policy shares its response path, both coefficients reduce to ordinary concentrability in the common environment, scaled by the logging mass of that policy. The corresponding ratio remains infinite if the shared-path policy never reaches a required target cell.

Theorem 3.3 separates structural identification, response recovery, and statistical evaluation. Additional data improve recovery and evaluation only after the target is identified. Pessimism cannot distinguish observationally equivalent response rules or reveal a zero-probability row. The one-step

example in the introduction has full support and a known map, so it isolates pooling bias. By contrast, the first claim of Theorem 3.3 is a non-identification result: two games can share every logging law and still disagree on an undeployed target.

4 End-to-End Offline Evaluation

Local margins. For a policy π and stage h, let the margin be the smallest ℓ_1 prefix perturbation, among policy-factorized occupancy prefixes, that changes $\phi_h(d^{\pi}_{1:h})$. Let γ_{tar} be the minimum margin over Π and $\gamma_{\log}$ the minimum over $\Pi_{\log}$. A positive margin is necessary for uniform exact label recovery: at a response boundary, arbitrarily close occupancy estimates can produce different labels with non-vanishing probability.

Proposition 4.1 (Reconstructing unobserved logging labels)**.** *For every $\pi_0 \in \Pi_{\log}$, estimate its state marginal $\widetilde{p}^{\pi_0}_h$ among episodes carrying label π_0, then form the policy-factorized occupancy $\widetilde{d}^{\pi_0}_h(s,a) = \widetilde{p}^{\pi_0}_h(s)\pi_{0,h}(a \mid s)$. If*

$$n \gtrsim \frac{H^2\{SA + \log(MH/\delta)\}}{\nu_{\min}\gamma^2_{\log}} + \frac{\log(MH/\delta)}{\nu_{\min}},$$

then, with probability at least $1-\delta$, applying ϕ_h to these empirical prefixes recovers every logging response label exactly.

The proposition uses repeated policy labels, not a smooth extrapolation across a policy parameterization. A continuous logging distribution in which almost every policy appears once would require additional regularity linking nearby policies. Here the finite menu makes each state marginal directly estimable. Factoring it with the known policy keeps the estimate inside the domain controlled by the margin, and conditional multinomial concentration gives an $O(H\sqrt{(S+\log(MH/\delta))/(n\nu_{\min})})$ prefix error. The displayed condition conservatively uses SA.

Estimators. For each logging policy π_0, first estimate its stage-wise state marginals from the episodes bearing label π_0, factor them with π_0, and set $\widehat{z}^{\pi_0}_h = \phi_h(\widetilde{d}^{\pi_0}_{1:h})$. These reconstructed labels define counts $N_h(s,z,b)$ and $N_h(s,z) := \sum_b N_h(s,z,b)$. We estimate $P_h(\cdot \mid s,a,b)$, the mean reward $r_h(s,a,b)$, and $g(\cdot \mid s,z)$ by empirical averages. On zero-count rows, $\widehat{P}_h$ and $\widehat{g}_h$ use fixed probability distributions, so every plug-in forward object remains a probability distribution.

The same trajectories can be used for label reconstruction and response-kernel estimation without sample splitting. To see this, first define hypothetical counts from the true population labels and prove concentration under the original sampling law. Separately, apply Proposition 4.1 to recover those labels. On the intersection of the two events, every reconstructed count equals its hypothetical counterpart. A union bound controls both failures without assuming independence or conditioning concentration on successful recovery. Exact recovery is essential here; approximate labels would contaminate the response-kernel rows and introduce an additional bias term.

For each target policy, apply Lemma 3.2 with $(\widehat{P}, \widehat{g}_h)$ to obtain $\widehat{z}^{\pi}$. The pessimistic recursion uses model and response bonuses

$$\Gamma_h(s,a,b) = cH\sqrt{\frac{S\iota}{N_h(s,a,b)\vee 1}}, \qquad \Lambda_h(s;\pi) = cH\sqrt{\frac{B\iota}{N_h(s,\widehat{z}^{\pi}_h)\vee 1}},$$

and clips each value to $[0, H-h+1]$. With $[x]^u_0 := \min\{u, \max\{0,x\}\}$, its two backups are

$$\overline{Q}^{\pi}_h(s,a,b) = \left[\widehat{r}_h(s,a,b) + (\widehat{P}_h\widehat{V}^{\pi}_{h+1})(s,a,b) - \Gamma_h(s,a,b)\right]^{H-h+1}_0, \tag{4}$$

$$\widehat{V}^{\pi}_h(s) = \left[\sum_{a,b}\pi_h(a \mid s)\widehat{g}_h(b \mid s, \widehat{z}^{\pi}_h)\overline{Q}^{\pi}_h(s,a,b) - \Lambda_h(s;\pi)\right]^{H-h+1}_0. \tag{5}$$

Set $\widehat{V}^{-}(\pi) = \mathbb{E}_{s\sim\rho}\widehat{V}^{\pi}_1(s)$. The model bonus protects reward and transition estimation. The separate response bonus protects the response-kernel row selected by the recovered target response. Algorithm 1 states the full procedure.

Algorithm 1 Response recovery and pessimistic off-policy evaluation

Require: Policy-labeled episodes $\mathcal{D}$, known ρ, target class Π, maps $\phi_{1:H}$
1: **for all** $\pi_0 \in \Pi_{\log}$ **do**
2: Estimate $\widetilde{d}^{\pi_0}_{1:H}$ from episodes labeled π_0 and set $\widehat{z}^{\pi_0}_h = \phi_h(\widetilde{d}^{\pi_0}_{1:h})$
3: **end for**
4: Form $\widehat{P}, \widehat{r}, \widehat{g}_h$ and count-based bonuses
5: **for all** $\pi \in \Pi$ **do**
6: Recover $\widehat{z}^{\pi}$ by the forward recursion
7: Compute $\widehat{V}^{-}(\pi)$ by pessimistic dynamic programming
8: **end for**
9: **return** $\widehat{\pi} \in \arg\max_{\pi\in\Pi} \widehat{V}^{-}(\pi)$ and all one-sided lower confidence bounds

Proposition 4.2 (Exact target response-path recovery)**.** *Suppose all logging labels are correct and, for every target and stage, the stochastic plug-in kernels satisfy*

$$\sum_{s,a,b} d_h^{\pi,\mu^\pi}(s,a,b)\|\widehat{P}_h(\cdot \mid s,a,b) - P_h(\cdot \mid s,a,b)\|_1 \le e_P,$$

$$\sum_{s} p_h^{\pi}(s)\|\widehat{g}_h(\cdot \mid s, z_h^{\pi}) - g(\cdot \mid s, z_h^{\pi})\|_1 \le e_g.$$

If $H^2(e_P + e_g)/2 < \gamma_{\mathrm{tar}}$, the empirical forward sweep recovers $\widehat{z}^{\pi} = z^{\pi}$ for every $\pi \in \Pi$.

To see why only the true target-reachable support is needed, condition on correct labels through stage h. Adding and subtracting one plug-in transition at a time and using stochastic-kernel contraction gives $\Delta_{h+1} \le \Delta_h + e_P + e_g$, with $\Delta_1 = 0$. Thus $\Delta_j \le (j-1)(e_P + e_g)$ and the whole prefix error through stage $h+1$ is at most $h(h+1)(e_P+e_g)/2$. The margin fixes label $h+1$ and closes the induction. Any empirical mass that leaks to an uncovered row is already charged to Δ_h, so a fixed stochastic zero-count fallback is sufficient. Appendix C gives the measure-level calculation.

Theorem 4.3 (End-to-end recovery and evaluation)**.** *Assume Definition 3.1, known ρ, finite Π and $\Pi_{\log}$, $\nu_{\min} > 0$, finite $C_{\mathrm{cf}}, C_{\mathrm{sr}}$, and positive margins $\gamma_{\log}, \gamma_{\mathrm{tar}}$. Define*

$$\kappa = \min\left\{ \min_{d_h^{\log}(s,a,b)>0} d_h^{\log}(s,a,b), \min_{d_h^{\log}(s,z)>0} d_h^{\log}(s,z) \right\}$$

and $\iota = \log(SABH|\mathcal{Z}||\Pi|Mn/\delta)$. There are universal constants such that, if

$$n \gtrsim \frac{H^2(SA+\iota)}{\nu_{\min}\gamma_{\log}^2} + \frac{\iota}{\nu_{\min}} + \frac{\iota}{\kappa}, \tag{6}$$

$$H^2\left(\sqrt{\frac{C_{\mathrm{cf}}S^2AB\iota}{n}} + \sqrt{\frac{C_{\mathrm{sr}}SB\iota}{n}}\right) < c\gamma_{\mathrm{tar}}, \tag{7}$$

then, with probability at least $1-\delta$, all logging and target response states are recovered exactly. Simultaneously, for every $\pi \in \Pi$,

$$0 \le V_{\mathrm{cf}}(\pi) - \widehat{V}^{-}(\pi) \le \widetilde{\mathcal{O}}\left(H^2\sqrt{\frac{C_{\mathrm{cf}}S^2AB}{n}} + H^2\sqrt{\frac{C_{\mathrm{sr}}SB}{n}}\right).$$

Consequently, the returned policy satisfies the same-order suboptimality bound. Dense tabular recovery costs $O(|\Pi|HS^2AB)$ plus $O(nH)$ preprocessing and the cost of evaluating the maps ϕ_h.

What determines the sample requirement. After the logging-label and positive-count requirements in (6) hold, target recovery and value precision impose separate thresholds. Write $D = C_{\mathrm{cf}}S^2AB + C_{\mathrm{sr}}SB$. For exact target recovery and value error at most $\tau > 0$, a sufficient sample-size scale is

$$n \gtrsim H^4\iota D\left(\gamma_{\mathrm{tar}}^{-2} + \tau^{-2}\right).$$

This follows by squaring the two square-root terms and using $(\sqrt{x}+\sqrt{y})^2 \le 2(x+y)$. A small response margin raises the sample threshold for selecting the correct environment; once the labels are exact, it does not appear in the value-error rate itself. Improving $\nu_{\min}$ helps reconstruct rare logging policies, whereas improving C_{sr} helps estimate the response rows used by the targets. These are different design objectives. Table 4 collects the assumptions beside the complete proof.

Corollary 4.4 (Validity after selection)**.** *Let ϵ_n denote the right-hand side of the value bound in Theorem 4.3, and let $\pi^\star \in \arg\max_{\pi\in\Pi} V_{\mathrm{cf}}(\pi)$. On the same event,*

$$V_{\mathrm{cf}}(\pi^\star) - V_{\mathrm{cf}}(\widehat{\pi}) \le \epsilon_n.$$

Moreover, a decision rule that deploys $\widehat{\pi}$ only when $\widehat{V}^-(\widehat{\pi}) \ge v_0$ guarantees $V_{\mathrm{cf}}(\widehat{\pi}) \ge v_0$. Thus both optimization over Π and a data-dependent decision to abstain are covered by the simultaneous theorem, with no additional post-selection correction.

Indeed, $V_{\mathrm{cf}}(\pi^\star) \le \widehat{V}^-(\pi^\star)+\epsilon_n \le \widehat{V}^-(\widehat{\pi})+\epsilon_n \le V_{\mathrm{cf}}(\widehat{\pi})+\epsilon_n$. The threshold claim follows directly from pessimism. This distinction matters in applications: a lower bound can justify a deployment relative to a prespecified floor even when its width is too large to identify the best policy sharply.

Proof sketch. The argument has four steps.

1. Conditional multinomial concentration for each repeated logging policy puts every empirical learner-occupancy prefix inside $\gamma_{\log}$. The reconstructed response labels therefore equal the unobserved population labels.
2. With those labels, state-joint-action counts estimate (P, r) and state-response counts estimate g. Target-weighted errors satisfy
$$e_P \lesssim \sqrt{C_{\mathrm{cf}} S^2 AB\iota/n}, \qquad e_g \lesssim \sqrt{C_{\mathrm{sr}} SB\iota/n}.$$
3. Along the true target forward path, suppose earlier labels are correct and let $\Delta_h = \|\widehat{d}_h^{\pi} - d_h^{\pi}\|_1$. Stochastic-kernel contraction gives $\Delta_{h+1} \le \Delta_h + e_P + e_g$. Hence the occupancy-prefix error is at most $O(H^2(e_P+e_g))$, and (7) prevents the next label from crossing a response boundary.
4. Exact response recovery reduces (4) and (5) to a pessimistic estimator in the correct target environment. Backward induction proves pessimism; unrolling the gap and applying coverage gives the stated rate.

The induction follows the true target path, so the coverage coefficients involve only logged and target-reachable support. Appendix C gives the concentration and backward-recursion details.

5 When Exact Certification Fails

Proposition 5.1 (Outer interval under partial identification)**.** *Fix a target and a set of game completions that agree with the logged law. Suppose all completions share the same target response sequence, either because it is independently known or because a margin dominates every completion-induced occupancy-prefix perturbation. Enclose these completions in a stage-rectangular uncertainty set that holds this response sequence fixed and permits independent stage-wise copies of repeated stationary response-kernel rows. Robust lower and upper Bellman recursions over this relaxation (Iyengar, 2005; Nilim & El Ghaoui, 2005) give the valid outer interval*

$$V^-(\pi) \le V_{\mathrm{cf}}^M(\pi) \le V^+(\pi), \qquad V^+(\pi) - V^-(\pi) \le \sum_{h=1}^{H} (H-h+1) q_h, \tag{8}$$

where q_h is the probability that the first unsupported primitive is needed at stage h. The interval need not equal the sharp identified set when response consistency couples unknown rows across stages, but it contains every compatible value. A finite-sample outer interval replaces supported rows by their confidence regions and computes the resulting width by robust recursion. If only rewards are missing, the sharper population width is the sum of uncovered reward masses. Without a common response or the stated completion-uniform margin, a tiny occupancy change can cross a response boundary and alter all later behavior; $[0, H]$ may then be the only valid interval.

The bound adapts to where support is missing. An unsupported row that the target never reaches contributes nothing, whereas an unsupported early transition can contribute the entire remaining reward range. Missing response stability can be more damaging than a missing reward cell because it may alter every later opponent action.

The interval has an explicit backward computation. With z^π fixed, let $\mathcal{U}_h(s,a,b)$ contain the admissible reward-mean and transition pairs, and let $\mathcal{G}_h(s,z_h^\pi)$ contain the admissible response-kernel rows. Supported population rows are singletons; unsupported rows range over $[0,1]$ and the appropriate probability simplices. Starting from $V_{H+1}^- = 0$, compute

$$Q_h^-(s,a,b) = \inf_{(r,P)\in\mathcal{U}_h(s,a,b)} \{r + PV_{h+1}^-\},$$
$$V_h^-(s) = \inf_{g\in\mathcal{G}_h(s,z_h^\pi)} \sum_{a,b} \pi_h(a \mid s) g(b) Q_h^-(s,a,b).$$

Replace both infima by suprema for V_h^+ and average $V_1^\pm$ under ρ. The response-kernel choice is outside the learner-action average because the opponent does not observe the current action. Each repeated response-kernel row is optimized separately at each stage in this outer relaxation. Consequently, the endpoints can be wider than those attainable by a single stationary completion. The q_h in (8) counts only the first unsupported encounter, so later visits on the same trajectory are not charged twice. For example, an unsupported transition first reached with probability α at stage two contributes at most $(H-1)\alpha$, whereas a missing stage-two reward alone contributes at most α.

Proposition 5.2 (Induced-response robustness)**.** *Let μ^π_{true} be the actual induced response and $\bar{\mu}^\pi$ the response represented by the abstraction. A performance-difference argument gives*

$$\left|V^{\pi,\mu^\pi_{\mathrm{true}}} - V^{\pi,\bar{\mu}^\pi}\right| \le \frac{1}{2}\sum_{h=1}^{H}(H-h+1)\mathbb{E}_{s\sim p_h^{\pi,\mu^\pi_{\mathrm{true}}}} \left\|\mu^\pi_{\mathrm{true},h}(\cdot \mid s) - \bar{\mu}^\pi_h(\cdot \mid s)\right\|_1 . \tag{9}$$

A uniform ℓ_1 error ε therefore gives $B_\varepsilon = \min\{H, H(H+1)\varepsilon/4\}$. Subtracting B_ε from the model-based lower bound preserves validity, and selection regret increases by at most $2B_\varepsilon$. This guarantee requires an externally justified induced-response discrepancy. Passive logs cannot certify behavior under unobserved deployments.

This formulation covers errors in $\mathcal{Z}$, ϕ_h, or g only when they imply a bound on the induced opponent policy or when a margin keeps the response labels unchanged. A small parameter perturbation alone is not enough for a discontinuous threshold map. The method does not estimate ε. A reported lower confidence bound should include the curve $\varepsilon \mapsto \widehat{V}^-(\pi) - B_\varepsilon$; response-kernel residuals on logged (s,z) cells calibrate only observed regimes. Recovery remains stable when the combined statistical and induced-response perturbation stays inside the target margin.

Rates and lower bounds. If the estimator supplies simultaneous Bernstein envelopes, the same proof gives an $H^{3/2}$ leading square-root term under bounded aggregate variance, plus inverse-count remainders. This is a conditional refinement, not a claim that a data-dependent value function automatically satisfies an empirical Bernstein bound. Complementary lower bounds isolate state-joint-action coverage from response-state cardinality: even under benign state coverage, recovering which of $|\mathcal{Z}|$ regimes is active still costs $\Omega((H-1)\min\{1,\sqrt{|\mathcal{Z}|/n}\})$ (Theorem G.4). A Bernstein envelope does not remove that identification layer. Proofs and the conditional function-approximation extension appear in the appendix.

6 Designed Systems and Diagnostics

The guarantees apply to designed data collection, not arbitrary historical logs. The analyst must state the operational meaning of $\mathcal{Z}$ and ϕ_h, and the data-collection record must document repeated policy assignments and minimum exploration probabilities. Counts and sensitivity checks are useful diagnostics, but they do not prove population support or response-model fidelity.

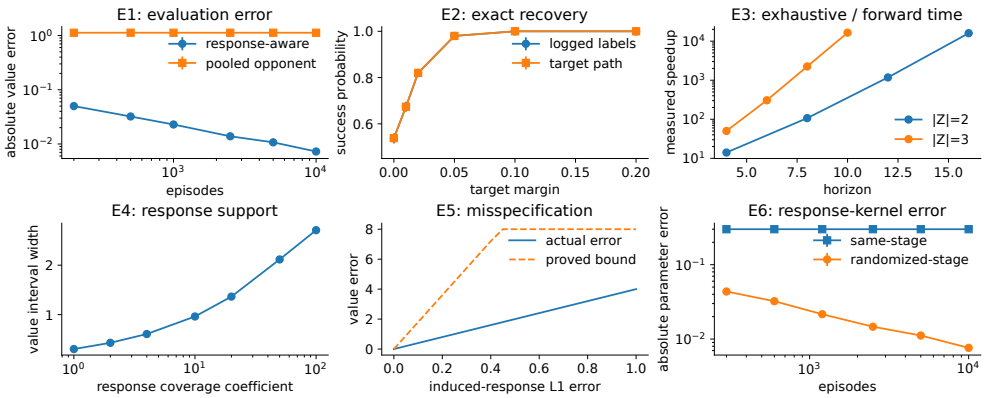


Figure 1. Diagnostics under the specified model: pooling bias (E1), margin-controlled recovery (E2), forward versus enumerate time (E3), C_{sr} versus interval width (E4), performance-difference bound (E5), and cross-stage response-kernel estimation (E6). These check the theorems, not unrestricted opponents. Protocols are in Appendix I.

A designed-system example. Consider two retailers with two inventory states and two prices. In each assigned-policy cohort, the learner announces $z_h^\pi = \mathbf{1}\{q_h^\pi \geq 1/2\}$, where $q_h^\pi = \sum_s p_h^\pi(s)\pi_h(a_{\mathrm{high}} \mid s)$ is its population high-price probability, and then privately draws its price. The rival observes inventory and z, but not the learner's realized price, and chooses its price according to $g(\cdot \mid s, z)$. Logs contain the assigned policy, state, both prices, bounded reward, and next state. Repeated randomized assignments and positive action probabilities provide a route to coverage, subject to a check of target-reachable cells. The rule must be fixed before evaluation; the claimed rival response needs external evidence. Appendix I.3 gives a controlled undeployed-target check, not market validation.

What can be checked empirically. Policy-label frequencies and cell counts audit the logging mixture and determine the bonuses, but a zero count does not establish zero population support. Boundary slack and adjacent bins assess sensitivity to ϕ; response residuals assess observed regimes. These checks can trigger abstention, but cannot validate ϕ or the opponent response under an undeployed path.

Controlled diagnostics. Figure 1 summarizes diagnostics E1–E6. In E8, the held-out target has a response path absent from every logger. At $n = 10{,}000$ over 200 trials, the non-oracle procedure recovers its path every time; mean absolute value error is 0.0081 ± 0.0005, versus 0.1070 ± 0.0007 for pooling. The lower bound covers in every trial but remains 0.441 below truth on average. Appendix I gives the protocols. These synthetic results do not validate the response map for unrestricted opponents.

7 Discussion

The analysis has a strict order: specify the response abstraction, establish state-joint-action and state-response support, recover the induced response, and only then apply pessimistic evaluation. Pessimism cannot repair an unidentified or misspecified response rule. If an earlier condition fails, report an outer interval, an externally calibrated sensitivity analysis, or abstention.

The theory is finite and tabular, with known ρ, specified response maps, repeated policy assignments, and positive local margins. An unknown ρ adds an initial-state concentration term. Stationarity of g permits cross-stage transfer; Appendix F analyzes that option. Continuous, latent, or history-dependent response states and learned maps require additional structure.

Closing the horizon gap and extending end-to-end recovery to function approximation require estimator-specific envelopes. Reports should include the prespecified map, policy assignments, counts on target-reachable support, boundary slack, the curve $\varepsilon \mapsto \widehat{V}^-(\pi) - B_\varepsilon$, and simultaneous lower confidence bounds for all policies considered. Appendix I.6 gives the complete checklist.

AI USE STATEMENT

Generative AI tools assisted with drafting and copyediting, preparation of numerical verification code, and checks of cross-references and bibliographic metadata. The authors formulated the research questions, model, and results; independently checked every theorem statement and proof; reproduced every numerical constant reported in the paper; and verified each bibliographic entry against the publisher record. The authors take full responsibility for the manuscript and its claims.

ETHICS STATEMENT

This work uses no human-subject or private data. Its motivating domains include markets and recommendation systems, where a misspecified response model could support overconfident or strategically harmful decisions. The proposed bounds should be treated as conditional audit tools, not as deployment authorization when the response abstraction, coverage, or discrepancy assumptions lack independent justification.

REPRODUCIBILITY STATEMENT

The appendix provides complete proofs, explicit estimators and assumptions, and full experimental settings. The supplementary code uses fixed seeds and regenerates every synthetic diagnostic from the documented commands.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, pp. 2312–2320, 2011.

Kenshi Abe and Yusuke Kaneko. Off-policy exploitability-evaluation in two-player zero-sum Markov games. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pp. 78–87, 2021.

Victor Aguirregabiria and Pedro Mira. Sequential estimation of dynamic discrete games. *Econometrica*, 75(1):1–53, 2007.

Raman Arora, Ofer Dekel, and Ambuj Tewari. Online bandit learning against an adaptive adversary: From regret to policy regret. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 1747–1754, 2012.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 263–272, 2017.

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

Patrick Bajari, C. Lanier Benkard, and Jonathan Levin. Estimating dynamic models of imperfect competition. *Econometrica*, 75(5):1331–1370, 2007.

Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*, volume 28, pp. 1342–1350, 2015.

Andrew Bennett, Nathan Kallus, Lihong Li, and Ali Mousavi. Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pp. 1999–2007, 2021.

Claire Chen, Yuheng Zhang, Xinyu Liu, Zixuan Xie, Shuze Daniel Liu, and Nan Jiang. Offline two-player zero-sum Markov games with KL regularization. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026a. URL `https://openreview.net/forum?id=cQD2wxFhFG`.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 1042–1051, 2019.

Jingxiao Chen, Weiji Xie, Weinan Zhang, Yong Yu, and Ying Wen. Offline fictitious self-play for competitive games. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(24): 20118–20126, 2026b. doi: 10.1609/aaai.v40i24.39098.

Ruifeng Chen, Xiong-Hui Chen, Yihao Sun, Siyuan Xiao, Minhui Li, and Yang Yu. Policy-conditioned environment models are more generalizable. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6539–6561, 2024.

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

Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. Near-optimal no-regret learning in general games. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27604–27616, 2021.

Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. Bandits with switching costs: $T^{2/3}$ regret. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pp. 459–467, 2014.

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

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 652–661, 2016.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of the 33rd Conference on Learning Theory*, pp. 2137–2143, 2020.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5084–5096, 2021.

Yuheng Jing, Kai Li, Bingyun Liu, Yifan Zang, Haobo Fu, Qiang Fu, Junliang Xing, and Jian Cheng. Towards offline opponent modeling with in-context learning. In *International Conference on Learning Representations*, 2024.

Yuheng Jing, Kai Li, Bingyun Liu, Ziwen Zhang, Haobo Fu, Qiang Fu, Junliang Xing, and Jian Cheng. Offline opponent modeling with truncated Q-driven instant policy refinement. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 28256–28287, 2025.

Myrto Kalouptsidi, Paul T. Scott, and Eduardo Souza-Rodrigues. On the non-identification of counterfactuals in dynamic discrete games. *International Journal of Industrial Organization*, 50: 362–371, 2017.

Myrto Kalouptsidi, Paul T. Scott, and Eduardo Souza-Rodrigues. Identification of counterfactuals in dynamic discrete choice models. *Quantitative Economics*, 12(2):351–403, 2021.

Chinmaya Kausik, Yangyi Lu, Kevin Tan, Maggie Makar, Yixin Wang, and Ambuj Tewari. Offline policy evaluation and optimization under confounding. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, pp. 1459–1467, 2024.

Samir Khan, Martin Saveski, and Johan Ugander. Off-policy evaluation beyond overlap: Sharp partial identification under smoothness. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23734–23757, 2024.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *The Annals of Statistics*, 52(1):233–260, 2024.

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.

Qinghua Liu, Yuanhao Wang, and Chi Jin. Learning Markov games with adversarial opponents: Efficient algorithms and fundamental limits. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 14036–14053, 2022.

Runyu Lu, Yuanheng Zhu, and Dongbin Zhao. Constrained exploitability descent: An offline reinforcement learning method for finding mixed-strategy Nash equilibrium. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 40977–40994, 2025.

Debmalya Mandal and Goran Radanovic. Performative reinforcement learning with linear Markov decision process. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3232–3240, 2025.

Debmalya Mandal, Stelios Triantafyllou, and Goran Radanovic. Performative reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 23642–23680, 2023.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Richard D. McKelvey and Thomas R. Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(27):815–857, 2008.

Thanh Nguyen-Tang and Raman Arora. Learning in Markov games with adaptive adversaries: Policy regret, fundamental barriers, and efficient algorithms. In *Advances in Neural Information Processing Systems*, volume 37, pp. 56268–56303, 2024.

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.

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*, pp. 1243–1251, 2024.

Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 7599–7609, 2020.

Doina Precup, Richard S. Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 759–766, 2000.

Ben Rank, Stelios Triantafyllou, Debmalya Mandal, and Goran Radanovic. Performative reinforcement learning in gradually shifting environments. In *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence*, pp. 3041–3075, 2024.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. In *Advances in Neural Information Processing Systems*, volume 34, pp. 11702–11716, 2021.

Rilind Sahitaj, Paulius Sasnauskas, Yiğit Yalın, Debmalya Mandal, and Goran Radanovic. Independent learning in performative Markov potential games. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3304–3312, 2025.

Jianzhun Shao, Yun Qu, Chen Chen, Hongchang Zhang, and Xiangyang Ji. Counterfactual conservative Q learning for offline multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 77290–77312, 2023.

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

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 19967–20025, 2022.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdú, and Marcelo J. Weinberger. Inequalities for the L_1 deviation of the empirical distribution. Technical Report HPL-2003-97 (R.1), Hewlett-Packard Laboratories, Palo Alto, CA, June 2003.

Chengjie Wu, Pingzhong Tang, Jun Yang, Yujing Hu, Tangjie Lv, Changjie Fan, and Chongjie Zhang. Conservative offline policy adaptation in multi-agent games. In *Advances in Neural Information Processing Systems*, volume 36, pp. 51881–51903, 2023.

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 the 33rd Conference on Learning Theory*, pp. 3674–3682, 2020.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6683–6694, 2021.

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

Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4065–4078, 2021.

Andrea Zanette, Martin J. Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13626–13640, 2021.

Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D. Lee, and Wen Sun. Provable offline preference-based reinforcement learning. In *International Conference on Learning Representations*, 2024.

Wenhao Zhan, Scott Fujimoto, Zheqing Zhu, Jason D. Lee, Daniel Jiang, and Yonathan Efroni. Exploiting structure in offline multi-agent RL: The benefits of low interaction rank. In *International Conference on Learning Representations*, 2025.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. In Kyriakos G. Vamvoudakis, Yan Wan, Frank L. Lewis, and Derya Cansever (eds.), *Handbook of Reinforcement Learning and Control*, volume 325 of *Studies in Systems, Decision and Control*, pp. 321–384. Springer, Cham, 2021.

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*, pp. 40804–40829, 2023.

Hai Zhong, Xun Wang, Zhuoran Li, and Longbo Huang. Offline-to-online multi-agent reinforcement learning with offline value function memory and sequential exploration. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pp. 2373–2381, 2025.

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*, pp. 27117–27142, 2022.

Han Zhong, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Can reinforcement learning find Stackelberg-Nash equilibria in general-sum Markov games with myopically rational followers? *Journal of Machine Learning Research*, 24(35):1–52, 2023.

Banghua Zhu, Michael I. Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or K-wise comparisons. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 43037–43067, 2023.

A GUIDE TO THE APPENDIX

This appendix follows the logical order of the paper. Section B proves the structural barriers and population-identification result. Section C proves logging-label reconstruction, forward target recovery, pessimistic evaluation, and the end-to-end theorem, then compares two stronger information regimes. Sections D and E treat incomplete support and misspecification. Section F gives a cross-stage response estimator. Section G proves evaluation and policy-selection lower bounds. Section H develops conditional estimation extensions, corollaries, applications, and calibrated open questions. Section I gives the complete diagnostic protocol. Section A.2 first locates the problem among adjacent learning settings; later sections discuss individual connections where they become technically relevant. The six-panel figure in the main text (E1–E6) and the pricing panel E7 are the primary diagnostics. Experiments S1–S8 below are complementary tabular checks of the same theory; they are not a second numbering of E1–E8.

Notation. For a response sequence $z = (z_1, \dots, z_H)$, let $p_h^{\pi,z}$ and $d_h^{\pi,z}$ be the state and learner state-action marginals obtained when the opponent uses $g(\cdot \mid s, z_h)$. For a signed measure u, $\|u\|_1$ denotes its total ℓ_1 norm. For occupancy prefixes,

$$\|d_{1:h} - d'_{1:h}\|_{1,\mathrm{pref}} := \sum_{j=1}^{h} \|d_j - d'_j\|_1.$$

All constants denoted by c or C are universal and may change from line to line. Write $\bar{C} = \max\{C_{\mathrm{cf}}, C_{\mathrm{sr}}\}$. Logarithmic factors are displayed when they affect a margin or a union bound and suppressed elsewhere.

Symbol	Meaning
$\mathcal{S}, \mathcal{A}, \mathcal{B}$; S, A, B	State and action spaces; their tabular cardinalities
P_h, R_h, r_h	Transition kernel, reward kernel, and reward mean
H, ρ	Horizon and initial-state distribution
π, μ; $V^{\pi,\mu}$	Learner and opponent policies; expected return
$\Pi, \Pi_{\log}$	Target policy class and finite repeated logging policy set
$\mu^\pi, V_{\mathrm{cf}}(\pi)$	Response induced by deployment and its induced-response value
$\mathcal{Z}, d_R$; z^π	Response-state space, $d_R = \lvert\mathcal{Z}\rvert$, and induced response sequence
ϕ_h, g	Specified prefix-to-state map and learned stationary opponent-response kernel
$p_h^{\pi,z}, d_h^{\pi,z}$	State and learner occupancies under candidate response sequence z
$d_h^{\pi,\mu}$; $d_h^{\log}$	Joint occupancy and logging-mixture occupancy
$C_{\mathrm{cf}}, C_{\mathrm{sr}}, \bar{C}$	Target-reachable joint-occupancy and state-response coverage coefficients; their maximum
$C_{\mathrm{cf}}^{\mathrm{cand}}, C_{\mathrm{sr}}^{\mathrm{cand}}$	Uniform candidate-path counterparts used only for the stronger benchmark
$\gamma_{\log}, \gamma_{\mathrm{tar}}, \gamma_\phi$	Logging, target-path, and uniform candidate-path response margins
$\kappa_{\mathrm{model}}, \kappa_{\mathrm{resp}}$	Minimum positive state-joint-action and state-response logging masses
$\widehat{P}, \widehat{r}, \widehat{g}, \widehat{\mu}$	Plug-in transition, reward, response-kernel, and opponent estimates
$\Gamma_h^\pi, \Lambda_h^\pi$	Model and response confidence envelopes
$\widehat{V}^-(\pi), \widehat{\pi}$	One-sided lower confidence bound and its maximizing policy

Table 1. Notation used in the main paper and appendix. Symbols local to one construction are defined where they appear.

Scope and model boundaries. Several boundaries are deliberate. First, ϕ_h is a population response rule. Dependence on past actions can be represented when the relevant history is encoded in s or z, but observing the learner's current realized action before choosing b_h falls outside our simultaneous-action model. A response map that uses a same-step joint occupancy already determined by z_h is circular and need not be triangular. Second, multiple deployment responses without a selection rule make $V_{\mathrm{cf}}(\pi)$ set-valued, so a single forward value is inappropriate. Third, stationarity of g is used for cross-stage transfer. The main theorem avoids that transfer by estimating stage-specific copies, while Appendix F gives a conservative randomized-stage estimator when cross-stage support is essential.

Regime	Conclusion	Reference
No observable response structure	V_{cf} is not identifiable from offline data.	Theorem B.1
Target response state absent from the logging support	No estimator can distinguish completions that change the target response.	Theorem B.3
Observable response structure, but $C_{\mathrm{cf}} = \infty$	Uniform offline estimation is impossible.	Theorem B.2
Finite response-state model, $C_{\mathrm{cf}} < \infty$, and joint (s, z) coverage	The target response-state sequence and induced-response rollout are identified on target-reachable support.	Proposition B.4
Identified or supplied target response plus high-probability envelopes	Pessimistic plug-in evaluation and policy selection reduce to controlling two error terms.	Theorem C.13
Finite repeated logging policy set, local support, and response margins	Forward recovery reconstructs logged labels and target responses from policy-labeled trajectories.	Theorem 4.3; Lemmas C.1 and C.3
Observed logged response states, uniform candidate-path coverage, and a uniform margin	The empirical and population operators agree on every candidate path; the algorithm remains a single forward pass.	Theorem C.12
Observed logged response states, a given target path, $C_{\mathrm{sr}} < \infty$, and minimum-count control	The same pessimistic evaluation rate holds when logged labels and the target path are given.	Corollary C.10
Covered subclasses with trivial or adaptive opponents	Capped joint-occupancy and response-state cardinality lower bounds are separately unavoidable.	Theorems G.3–G.4

Table 2. Summary of the theorem arc.

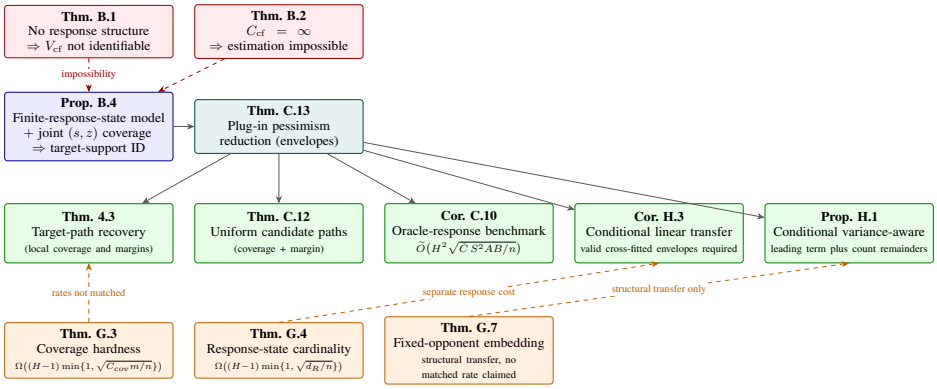


Figure 2. Dependency structure: barriers, identification on target-reachable support, pessimistic reduction, forward response recovery, conditional extensions, and separate capped converses.

A.1 Dependency map for the theorem arc

Table 2 lists each regime with its qualitative conclusion and a formal reference. Figure 2 visualizes the same logical arc. The green target-path theorem is the primary data-driven route. Uniform candidate-path stability and the oracle-response corollary are benchmarks. Linear and variance-aware statements are conditional transfers. The orange row gives separate capped converses and a structural fixed-opponent embedding; no matched $H^{3/2}\sqrt{SAB/n}$ lower bound is claimed.

Line of work	Setting	Target	Coverage or structure	Information	Computation	Role relative to this paper
Cui & Du (2022)	Offline zero-sum MG	Nash gap	Unilateral coverage	Data-driven	Efficient	Pessimistic equilibrium learning in a fixed game
Zhong et al. (2022); Zhang et al. (2023)	Offline MG	Nash gap	Concentrability or Bellman completeness	Data-driven	Efficient	Fixed-game equilibrium learning with tabular or general approximation
Yan et al. (2024)	Offline zero-sum MG	Nash gap	Uniform state-joint-action coverage	Data-driven	Efficient	Model-based fixed-game benchmark
Nguyen-Tang & Arora (2024; 2025)	Online MG, adaptive opponent	Policy regret	Online sampling and response regularity	Continued interaction	Efficient	Comparator is evaluated against an induced response, but new data are collected
End-to-end result	**Offline MG, adaptive opponent**	V_{cf} **and its maximizer**	finite repeated policy set, target-reachable support, local margins	**Policy-labeled fixed logs**	**Forward recursion**	Learns logging labels and the undeployed target response before pessimistic evaluation
Oracle-response benchmark	**Same model**	V_{cf}	$C_{\mathrm{cf}}, C_{\mathrm{sr}} < \infty$	**Target response supplied**	**Efficient**	Isolates value estimation after response identification
Conditional extensions	**MG with adaptive opponent and extra estimator structure**	V_{cf}	valid simultaneous variance or function-approximation envelopes	**Recovered or supplied response**	**Efficient given envelopes**	Transfer statements only; no generic estimator-specific dimension rate is claimed

Table 3. Comparison with the closest learning settings. The target, information regime, and computational status are separated because a rate for a fixed game is not directly comparable with recovery of a policy-induced environment from fixed logs.

A.2 Positioning the problem

The closest literatures differ along two axes that matter here: whether data collection continues after the target is chosen, and whether the target policy changes the opponent law being evaluated. Table 3 makes those distinctions explicit before the technical appendix begins.

Offline Markov-game methods most often target equilibria or exploitability in a fixed game (Cui & Du, 2022; Zhong et al., 2022; Zhang et al., 2023; Yan et al., 2024; Abe & Kaneko, 2021; Chen et al., 2026b; Lu et al., 2025; Chen et al., 2026a; Shao et al., 2023; Zhan et al., 2025; Nika et al., 2024; Zhong et al., 2025). Policy-regret work instead lets a learner continue interacting with an adaptive opponent (Arora et al., 2012; Dekel et al., 2014; Nguyen-Tang & Arora, 2024; 2025). Our setting combines these two features: the opponent changes with the deployed learner policy, but evaluation must use only a fixed archive of earlier policy-labeled deployments. The sections below trace the consequences: identification requires a response model, estimation requires both state-joint-action and state-response support, and pessimism is applied only after the induced target environment has been recovered.

B Barriers and Population Identification

B.1 Non-identifiability without observable response structure

Theorem B.1 (Non-identifiability)**.** *There are two one-step adaptive games with full state-joint-action logging support that generate the same law for every sample size under the logging distribution but whose induced-response values for an undeployed target differ by one. Consequently, for every estimator $\widehat{V}$,*

$$\inf_{\widehat{V}} \sup_{G} \mathbb{P}_G \left(\left| \widehat{V} - V_{\mathrm{cf}}^{G}(\pi_{\mathrm{tar}}) \right| \geq \frac{1}{2} \right) \geq \frac{1}{2}.$$

Proof. Let $H = 1$ and use one state, learner actions a_0, a_1, and opponent actions b_0, b_1. The logging policy set contains $\pi_L(a_1) = 1/4$ and $\pi_H(a_1) = 3/4$ with positive mixing weights. In both games, π_L induces deterministic b_0 and π_H induces deterministic b_1. Thus every action pair (a, b) has positive logging probability. Let the undeployed target satisfy $\pi_{\mathrm{tar}}(a_1) = 1/2$. In game G_0 its response is b_0, while in game G_1 its response is b_1. Use the common reward $r(a, b) = \mathbf{1}\{b = b_1\}$. The two games induce exactly the same policy-labeled logged trajectory law and share the same fully identified reward table, but their target values are zero and one. Since the estimator has the same distribution under both games, the two events

$$|\widehat{V} - 0| < 1/2, \qquad |\widehat{V} - 1| < 1/2$$

are disjoint. At least one game assigns probability at least $1/2$ to error at least $1/2$. □

The construction changes only the unobserved response map. No confidence penalty applied to the common logged law can distinguish the two games.

B.2 Missing state-joint-action support

Theorem B.2 (Uncovered-mass lower bound)**.** *Fix a target policy in a response-state model. Suppose its stage-h occupancy assigns mass $\alpha > 0$ to a set of triples that have zero logging probability. There are two games in the same response model with identical logged laws and target values separated by α. Every estimator therefore has error at least $\alpha/2$ with probability at least $1/2$ on one of the games.*

Proof. Keep P, g, $\mathcal{Z}$, and $\phi_{1:H}$ fixed. Let U be the unsupported set and define two reward functions that agree everywhere except at stage h on U, where $r_0 = 0$ and $r_1 = 1$. Since the logging occupancy of U is zero, every finite or infinite logged-data law is identical. The target dynamics and response sequence are also identical because only the rewards changed. The occupancy representation of value gives

$$V_1(\pi) - V_0(\pi) = \sum_{(s,a,b)\in U} d_h^{\pi,\mu^\pi}(s,a,b) = \alpha.$$

The two-point argument from Theorem B.1 completes the proof. □

The lower bound is linear in missing mass. It is constant only when the target places constant mass on the unsupported region.

B.3 Missing state-response support

Theorem B.3 (Unobserved response-row lower bound)**.** *For every $\alpha \in (0, 1]$, the finite response-state model class contains two games with common P, r, $\mathcal{Z}$, and $\phi_{1:H}$ and identical logged laws such that a target reaches a final-stage pair (s, z_H^π) with probability α, the logging law assigns zero probability to that pair, and the target values differ by α. Consequently, every estimator has error at least $\alpha/2$ with probability at least $1/2$ on one of the games.*

Proof. Use two opponent actions b_0, b_1 and let the common final-stage reward be $r_H(s, a, b_0) = 0$ and $r_H(s, a, b_1) = 1$ for the target action a. Choose common earlier dynamics, policies, and response maps so that the target reaches s with probability α and has final response state $z_H^\pi = z$, while no logging policy reaches the pair (s, z). Let the two opponent-response kernels agree on every logged row and differ only at (s, z): the first selects b_0 and the second selects b_1. Because the row has zero logging probability, the complete logged laws coincide. The response-kernel difference occurs at the final stage, so the target state and response-state occupancies are common to both games. Their expected final rewards, and hence their values, differ by α. The usual two-point argument completes the proof. □

This theorem is the response-conditioned analogue of Theorem B.2. It establishes generic necessity without claiming that every missing response-kernel row changes every value.

B.4 Causality and uniqueness

Proof of Lemma 3.2. For any candidate sequence z,

$$p_1^{\pi,z} = \rho, \qquad d_h^{\pi,z}(s,a) = p_h^{\pi,z}(s)\pi_h(a \mid s),$$

and

$$p_{h+1}^{\pi,z}(s') = \sum_{s,a,b} d_h^{\pi,z}(s,a) g(b \mid s, z_h) P_h(s' \mid s, a, b). \tag{10}$$

The base marginal $d_1^{\pi,z}(s,a) = \rho(s)\pi_1(a \mid s)$ is independent of z. Inductively, if $p_h^{\pi,z}$ depends only on $z_{1:h-1}$, then (10) depends only on $z_{1:h}$. Thus $d_j^{\pi,z}$ depends only on $z_{1:j-1}$, and the full prefix $d_{1:h}^{\pi,z}$ depends only on $z_{1:h-1}$. This proves the first claim.

Now construct z^π sequentially. At stage one, set $z_1^\pi = \phi_1(\rho \otimes \pi_1)$. Given $z_{1:h-1}^\pi$, propagate the dynamics to obtain $d_{1:h}^\pi$ and set $z_h^\pi = \phi_h(d_{1:h}^\pi)$. The constructed sequence satisfies $T_\pi(z^\pi) = z^\pi$ coordinate by coordinate. If z^* is any fixed point, its first coordinate must equal z_1^π. Assuming $z_{1:h-1}^* = z_{1:h-1}^\pi$, causality forces $(T_\pi(z^*))_h = z_h^\pi$, and the fixed-point property gives $z_h^* = z_h^\pi$. Induction proves uniqueness. The argument uses only that P and g are stochastic kernels, so it applies verbatim to stochastic plug-in kernels. □

B.5 Identification on target-reachable support

Proposition B.4 (Population identification on target-reachable support)**.** *Assume Definition 3.1. Fix π. Suppose the logged law identifies P_h and the reward kernel R_h on every triple with positive target occupancy and identifies $g(\cdot \mid s, z_h^\pi)$ on every target-visited state. Then it identifies the response sequence, target trajectory law, value, and every bounded functional of that trajectory law. Finite C_{cf} and joint state-response coverage are sufficient for these local conditions.*

Proof. Conditional next-state and reward laws identify P_h and R_h, and hence the reward mean r_h, on every state-joint-action triple with positive logging mass. Finite C_{cf} therefore identifies them on all target-supported triples. For a logging policy π_0, the population learner occupancy is part of its labeled trajectory law, so the known map ϕ_h identifies $z_h^{\pi_0}$. Whenever a logged deployment reaches s under response state z, the conditional opponent-action law identifies the stationary opponent-response kernel $g(\cdot \mid s, z)$. Joint state-response coverage therefore identifies the response-kernel rows needed on target-reachable support.

Proceed forward. The initial learner marginal and z_1^π are identified from (ρ, π, ϕ_1). Given the identified prefix through stage h, the identified opponent-response and game kernels determine p_{h+1}^π, and the known map ϕ_{h+1} determines z_{h+1}^π. Lemma 3.2 ensures that no global fixed-point selection is needed. This induction identifies the full induced-response trajectory law and hence all of its bounded functionals. □

B.6 Expanded constructions and identification details

The concise arguments above isolate the logical assumptions. This subsection gives the complete constructions, conditional-law calculations, and forward-identification induction used by those results.

Supplementary support-failure construction. Theorem B.1 already gives a full state-joint-action-support witness. For comparison, this construction also drops state-joint-action support and isolates the simpler failure that occurs when both the target action and its induced response are absent from the log. We construct two adaptive games (Stage I), verify that they produce identical data distributions (Stage II), and apply Le Cam's two-point method (Stage III).

Stage I. Construction of two indistinguishable games.

Define a minimal episodic Markov game $\mathcal{G}$ with the following components:

- State space $\mathcal{S} = \{s_0\}$, so $S = 1$.
- Learner action space $\mathcal{A} = \{a_1, a_2\}$, opponent action space $\mathcal{B} = \{b_1, b_2\}$.
- Horizon $H = 1$ (single-step episodes).
- Initial distribution $\rho = \delta_{s_0}$.

Since $H = 1$, no transition kernel is needed. Both games share the common reward function $r_1 \colon \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to [0, 1]$ given by

$$r_1(s_0, a_1, b_1) = r_1(s_0, a_1, b_2) = \tfrac{1}{2}, \qquad r_1(s_0, a_2, b_1) = 1, \qquad r_1(s_0, a_2, b_2) = 0. \tag{11}$$

Logging distribution. Let $\nu_{\log} = \delta_{\pi_{\log}}$ where $\pi_{\log}$ is the deterministic policy $\pi_{\log,1}(a_1 \mid s_0) = 1$. Under $\pi_{\log}$, both games specify the *same* opponent response:

$$\mu^{\pi_{\log}}(b_1 \mid s_0) = \mu^{\pi_{\log}}(b_2 \mid s_0) = \tfrac{1}{2} \quad \text{in both Game 1 and Game 2.} \tag{12}$$

Target policy and divergent responses. Define π_{tar} by $\pi_{\mathrm{tar},1}(a_2 \mid s_0) = 1$. The two games assign different opponent responses to π_{tar}:

$$\text{Game 1:} \quad R_1(\pi_{\mathrm{tar}}): \quad \mu_1^{\pi_{\mathrm{tar}}}(b_1 \mid s_0) = 1, \qquad \mu_1^{\pi_{\mathrm{tar}}}(b_2 \mid s_0) = 0, \tag{13}$$

$$\text{Game 2:} \quad R_2(\pi_{\mathrm{tar}}): \quad \mu_2^{\pi_{\mathrm{tar}}}(b_1 \mid s_0) = 0, \qquad \mu_2^{\pi_{\mathrm{tar}}}(b_2 \mid s_0) = 1. \tag{14}$$

For all other policies $\pi \notin \{\pi_{\log}, \pi_{\mathrm{tar}}\}$, both games assign the same arbitrary opponent response.

Stage II. Verification that the data distributions are identical.

Under the logging process, each episode $i \in [n]$ deploys $\pi^{(i)} = \pi_{\log}$ (since $\nu_{\log} = \delta_{\pi_{\log}}$) and the opponent responds with $\mu^{\pi_{\log}}$. A single trajectory is $\tau^{(i)} = (s_0,\, a_1^{(i)},\, b_1^{(i)},\, r_1^{(i)})$, where:

1. $s_1^{(i)} = s_0$ deterministically (from $\rho = \delta_{s_0}$).
2. $a_1^{(i)} = a_1$ deterministically (from $\pi_{\log,1}(a_1 \mid s_0) = 1$).
3. $b_1^{(i)} \sim \mathrm{Uniform}(\{b_1, b_2\})$ (from (12), identical in both games).
4. $r_1^{(i)} = r_1(s_0, a_1, b_1^{(i)}) = \frac{1}{2}$ regardless of $b_1^{(i)}$ (by (11), since $r_1(s_0, a_1, b) = \frac{1}{2}$ for all b).

Denote by $\mathbb{P}_j^{(n)}$ the joint distribution of the full dataset $\mathcal{D} = \{\tau^{(i)}\}_{i=1}^n$ under Game j, for $j \in \{1, 2\}$. The single-trajectory distribution under Game j is

$$\begin{aligned}\mathbb{P}_j\big(\tau = (s_0, a_1, b, \tfrac{1}{2})\big) &= \rho(s_0)\ \pi_{\log,1}(a_1 \mid s_0)\ \mu^{\pi_{\log}}(b \mid s_0)\ \mathbf{1}\big[r_1(s_0, a_1, b) = \tfrac{1}{2}\big] \\ &= 1 \cdot 1 \cdot \tfrac{1}{2} \cdot 1 \;=\; \tfrac{1}{2}, \qquad \forall\, b \in \{b_1, b_2\},\ j \in \{1, 2\}. \end{aligned} \tag{15}$$

Since the single-trajectory marginals are identical and the n episodes are drawn independently:

$$\mathbb{P}_1^{(n)} = \prod_{i=1}^n \mathbb{P}_1(\tau^{(i)}) = \prod_{i=1}^n \mathbb{P}_2(\tau^{(i)}) = \mathbb{P}_2^{(n)}. \tag{16}$$

In particular, the total variation distance satisfies

$$\mathrm{TV}\big(\mathbb{P}_1^{(n)},\ \mathbb{P}_2^{(n)}\big) \;=\; 0. \tag{17}$$

Stage III. Value separation and Le Cam's two-point method.

Counterfactual values. Under Game 1, the target policy π_{tar} plays a_2 deterministically, and the opponent responds with b_1 deterministically (by (13)):

$$V_{\mathrm{cf}}^{(1)}(\pi_{\mathrm{tar}}) = \mathbb{E}_{s_1 \sim \rho}\Big[\mathbb{E}_{\substack{a \sim \pi_{\mathrm{tar}}(\cdot \mid s_1) \\ b \sim \mu_1^{\pi_{\mathrm{tar}}}(\cdot \mid s_1)}}\big[r_1(s_1, a, b)\big]\Big] = r_1(s_0, a_2, b_1) = 1. \tag{18}$$

Under Game 2, the opponent responds with b_2 deterministically (by (14)):

$$V_{\mathrm{cf}}^{(2)}(\pi_{\mathrm{tar}}) = r_1(s_0, a_2, b_2) = 0. \tag{19}$$

The value separation is

$$\Delta := \big|V_{\mathrm{cf}}^{(1)}(\pi_{\mathrm{tar}}) - V_{\mathrm{cf}}^{(2)}(\pi_{\mathrm{tar}})\big| = |1 - 0| = 1. \tag{20}$$

Le Cam's two-point method (Tsybakov, 2009, Sec. 2.2). We invoke the following classical result, stated here with proof for completeness.

Lemma B.5 (Le Cam's two-point method)**.** *Let $\mathbb{P}_1, \mathbb{P}_2$ be probability measures on a measurable space $(\Omega, \mathcal{F})$, and let $\theta_1, \theta_2 \in \mathbb{R}$ satisfy $|\theta_1 - \theta_2| \geq 2\delta > 0$. Then for every measurable estimator $\hat{\theta}\colon \Omega \to \mathbb{R}$,*

$$\max_{j \in \{1,2\}}\ \mathbb{P}_j\big(|\hat{\theta} - \theta_j| \geq \delta\big) \;\geq\; \frac{1 - \mathrm{TV}(\mathbb{P}_1, \mathbb{P}_2)}{2}. \tag{21}$$

Proof of Lemma B.5. Define $E := \{\omega \in \Omega : |\hat{\theta}(\omega) - \theta_1| < \delta\}$. For any $\omega \in E$, the triangle inequality gives

$$|\hat{\theta}(\omega) - \theta_2| \;\geq\; |\theta_1 - \theta_2| - |\hat{\theta}(\omega) - \theta_1| \;>\; 2\delta - \delta = \delta,$$

so that $E \subseteq \{\omega : |\hat{\theta}(\omega) - \theta_2| \geq \delta\}$. Therefore,

$$\begin{aligned}
\mathbb{P}_1\big(|\hat{\theta} - \theta_1| \geq \delta\big) + \mathbb{P}_2\big(|\hat{\theta} - \theta_2| \geq \delta\big) \;&\geq\; \mathbb{P}_1(E^c) + \mathbb{P}_2(E) \\
&=\; 1 - \mathbb{P}_1(E) + \mathbb{P}_2(E) \\
&=\; 1 - \big(\mathbb{P}_1(E) - \mathbb{P}_2(E)\big) \\
&\geq\; 1 - \sup_{F \in \mathcal{F}} \big|\mathbb{P}_1(F) - \mathbb{P}_2(F)\big| \\
&=\; 1 - \mathrm{TV}(\mathbb{P}_1, \mathbb{P}_2). \qquad (22)
\end{aligned}$$

Since the maximum of two nonnegative quantities is at least half their sum, inequality (21) follows. □

We apply Lemma B.5 with the following identifications:

- $(\Omega, \mathcal{F})$ is the sample space of $\mathcal{D}$, equipped with $\mathbb{P}_j = \mathbb{P}_j^{(n)}$ for $j \in \{1, 2\}$.
- $\theta_j = V_{\mathrm{cf}}^{(j)}(\pi_{\mathrm{tar}})$ for $j \in \{1, 2\}$, with $\theta_1 = 1$ and $\theta_2 = 0$.
- $\delta = \Delta/2 = 1/2$.

We verify the hypotheses: $|\theta_1 - \theta_2| = 1 = 2\delta$, and $\mathrm{TV}(\mathbb{P}_1^{(n)}, \mathbb{P}_2^{(n)}) = 0$ by (17). Applying (21):

$$\max_{j \in \{1,2\}} \; \mathbb{P}_j^{(n)}\Big(\big|\hat{V}(\pi_{\mathrm{tar}}) - V_{\mathrm{cf}}^{(j)}(\pi_{\mathrm{tar}})\big| \geq \tfrac{1}{2}\Big) \;\geq\; \frac{1-0}{2} = \frac{1}{2}. \qquad (23)$$

Taking the infimum over all measurable estimators $\hat{V}$ and observing that the supremum over (M, R) includes both constructed games:

$$\begin{aligned}
\inf_{\hat{V}} \sup_{(M,R)} \Pr\Big(\sup_{\pi \in \Pi} \big|\hat{V}(\pi) - V_{\mathrm{cf}}(\pi)\big| \geq \tfrac{1}{2}\Big) \;&\geq\; \inf_{\hat{V}} \max_{j \in \{1,2\}} \mathbb{P}_j^{(n)}\Big(\big|\hat{V}(\pi_{\mathrm{tar}}) - V_{\mathrm{cf}}^{(j)}(\pi_{\mathrm{tar}})\big| \geq \tfrac{1}{2}\Big) \\
&\geq\; \frac{1}{2}, \qquad (24)
\end{aligned}$$

where the first inequality uses $\sup_\pi |\cdot| \geq |\cdot|$ evaluated at $\pi = \pi_{\mathrm{tar}}$. This proves the companion support-failure barrier with constants $c = c' = 1/2$. It is not used for the stronger full-support claim of Theorem B.1, whose two-logger construction appears in its proof above.

Remark B.6. *The lower bound* (24) *holds for* every *sample size $n \geq 1$, since* $\mathrm{TV}(\mathbb{P}_1^{(n)}, \mathbb{P}_2^{(n)}) = 0$ *regardless of n. This reflects an identifiability barrier: without structural assumptions linking μ^π to observables, no amount of logging data can disambiguate the opponent's counterfactual response to undeployed policies. The impossibility is not statistical but information-theoretic: the logging data are entirely uninformative about the opponent's behavior under π_{tar}.*

□

Expanded proof of Theorem B.2. The proof has three parts: identifying the uncovered region (Part A), constructing a pair of indistinguishable games (Part B), and applying Le Cam's method for the quantitative bound (Part C).

Part A. Identifying the uncovered region.

The counterfactual concentrability coefficient is (cf. the convention $0/0 := 0$, $p/0 := +\infty$ for $p > 0$):

$$C_{\mathrm{cf}} = \sup_{\pi \in \Pi} \sup_{h \in [H]} \max_{(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}} \frac{d_h^{\pi, \mu^\pi}(s, a, b)}{d_h^{\log}(s, a, b)}.$$

If $C_{\mathrm{cf}} = \infty$, there exist $\pi \in \Pi$, $h \in [H]$, and (s^*, a^*, b^*) such that $d_h^{\pi,\mu^\pi}(s^*, a^*, b^*) > 0$ while $d_h^{\log}(s^*, a^*, b^*) = 0$. (If every triple with positive target occupancy also had positive logging occupancy, the ratio would be finite for each triple, and the maximum over the finite set $\mathcal{S} \times \mathcal{A} \times \mathcal{B}$ would be finite.)

Define the uncovered set at step h:

$$\mathcal{U}_h := \big\{(s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} : d_h^{\log}(s, a, b) = 0\big\}, \tag{25}$$

and note that $\alpha = \sum_{(s,a,b)\in\mathcal{U}_h} d_h^{\pi,\mu^\pi}(s, a, b) \geq d_h^{\pi,\mu^\pi}(s^*, a^*, b^*) > 0$.

Part B. Construction of two indistinguishable games.

Fix the transition kernels P, opponent-response kernel g, response space $\mathcal{Z}$, and maps $\phi_{1:H}$ to be identical in both games. Let r be any base reward function satisfying $r_{h'} : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to [0, 1]$ for all $h' \in [H]$. Define two modified reward functions:

$$r_{h'}^{(j)}(s, a, b) = \begin{cases} r_{h'}(s, a, b) & \text{if } h' \neq h \text{ or } (s, a, b) \notin \mathcal{U}_h, \\ 1 & \text{if } h' = h,\ (s, a, b) \in \mathcal{U}_h,\ j = 1, \\ 0 & \text{if } h' = h,\ (s, a, b) \in \mathcal{U}_h,\ j = 2. \end{cases} \tag{26}$$

Both $r^{(1)}$ and $r^{(2)}$ map into $[0, 1]$ as required. In each game take the reward kernel to be deterministic, $\mathsf{R}_{h'}^{(j)} = \delta_{r_{h'}^{(j)}}$.

The data distributions are identical. Let $\mathbb{P}_j^{(n)}$ denote the distribution of $\mathcal{D} = \{\tau^{(i)}\}_{i=1}^n$ under Game j. Since P, g, and $\phi_{1:H}$ are shared, for each logging policy $\pi_0 \in \mathrm{supp}(\nu_{\log})$, the trajectory's state-action-opponent sequence $(s_1, a_1, b_1, \ldots, s_H, a_H, b_H)$ has the same distribution in both games (determined solely by π_0, μ^{π_0}, and P, all of which are identical).

It remains to verify that the observed rewards coincide almost surely. Define

$$\Omega_0 := \big\{\tau = (s_1, a_1, b_1, r_1, \ldots, s_H, a_H, b_H, r_H, s_{H+1}) : (s_h, a_h, b_h) \notin \mathcal{U}_h\big\}.$$

For any $\tau \in \Omega_0$, we have $r_h^{(1)}(s_h, a_h, b_h) = r_h^{(2)}(s_h, a_h, b_h)$ by (26), and $r_{h'}^{(1)} = r_{h'}^{(2)}$ trivially for $h' \neq h$. Hence the trajectory distributions agree on Ω_0.

Under the logging process, the probability of visiting the uncovered region at step h is

$$\Pr_{\log}\big((s_h, a_h, b_h) \in \mathcal{U}_h\big) = \sum_{(s,a,b)\in\mathcal{U}_h} d_h^{\log}(s, a, b) = 0,$$

since every element of $\mathcal{U}_h$ has zero logging occupancy by definition (25). Thus $\Pr_{\log}(\Omega_0) = 1$, and the single-episode distributions agree: $\mathbb{P}_1(\tau^{(i)} \in \cdot\,) = \mathbb{P}_2(\tau^{(i)} \in \cdot\,)$ almost surely. Since episodes are drawn independently, the joint distributions satisfy

$$\mathbb{P}_1^{(n)} = \mathbb{P}_2^{(n)}, \qquad \mathrm{TV}\big(\mathbb{P}_1^{(n)}, \mathbb{P}_2^{(n)}\big) = 0. \tag{27}$$

Part C. Value separation and Le Cam's method.

Since P, g, and $\phi_{1:H}$ are identical in both games and rewards do not affect the dynamics, the occupancy measures $d_{h'}^{\pi,\mu^\pi}(s, a, b)$ are the same across games for every $\pi \in \Pi$ and $h' \in [H]$. Using the occupancy-measure representation of the value function,

$$V_{\mathrm{cf}}^{(j)}(\pi) = \sum_{h'=1}^{H} \sum_{(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}} d_{h'}^{\pi,\mu^\pi}(s, a, b)\ r_{h'}^{(j)}(s, a, b),$$

we compute the value difference:

$$\begin{aligned} V_{\mathrm{cf}}^{(1)}(\pi) - V_{\mathrm{cf}}^{(2)}(\pi) &= \sum_{h'=1}^{H} \sum_{(s,a,b)} d_{h'}^{\pi,\mu^\pi}(s, a, b)\big[r_{h'}^{(1)}(s, a, b) - r_{h'}^{(2)}(s, a, b)\big] \\ &= \sum_{(s,a,b)\in\mathcal{U}_h} d_h^{\pi,\mu^\pi}(s, a, b) \cdot (1 - 0) \\ &= \alpha, \end{aligned} \tag{28}$$

where the second equality uses the fact that $r^{(1)}$ and $r^{(2)}$ agree outside $\mathcal{U}_h$ at step h and agree at all other steps $h' \neq h$.

Setting $\theta_j = V_{\mathrm{cf}}^{(j)}(\pi)$ for $j \in \{1,2\}$ and $\delta = \alpha/2$, we have $|\theta_1 - \theta_2| = \alpha = 2\delta$. Applying Lemma B.5 with (27):

$$\max_{j\in\{1,2\}} \mathbb{P}_j^{(n)}\Big(\big|\hat{V}(\pi) - V_{\mathrm{cf}}^{(j)}(\pi)\big| \geq \frac{\alpha}{2}\Big) \;\geq\; \frac{1-0}{2} = \frac{1}{2}. \tag{29}$$

Taking the infimum over all estimators $\hat{V}$ establishes the claim in Theorem B.2.

Remark B.7. *The lower claim in Theorem B.2 has several notable features:*

1. No dependence on n. *The bound holds for every sample size $n \geq 1$, since* $\mathrm{TV}(\mathbb{P}_1^{(n)}, \mathbb{P}_2^{(n)}) = 0$ *regardless of n. The dataset is completely uninformative about rewards at uncovered triples.*
2. Scaling with uncovered mass. *When $\alpha = \Omega(1)$ (that is, a constant fraction of the target's occupancy lies in the uncovered region), the minimax error is $\Omega(1)$. More generally, the error scales linearly with the total uncovered occupancy mass α.*
3. Necessity of coverage. *This result establishes $C_{\mathrm{cf}} < \infty$ as a* necessary condition *for consistent offline estimation of V_{cf}. Unlike in the standard (non-adaptive) offline RL setting, the coverage requirement here extends to opponent actions, reflecting the additional challenge of adaptive opponents.*

□

Expanded proof of Proposition B.4. We prove the result in three steps. Step 1 identifies the transition kernels and rewards on all target-relevant triples using $C_{\mathrm{cf}} < \infty$. Step 2 identifies the response model on the target-supported state-response pairs using the response-state structure and joint state-response coverage. Step 3 uses causal triangularity to identify the counterfactual response coordinates in chronological order, then identifies the induced opponent response on target-visited states and hence $V_{\mathrm{cf}}(\pi)$.

Throughout, *identifiability* is relative to the structured model class of Definition 3.1, with the known functionals $\{\phi_h\}_{h=1}^H$ fixed. Write $\mathcal{L}$ for the population-level distribution of the observed pair (Π_0, τ), where $\Pi_0 \sim \nu_{\log}$ is the deployed logging-policy label and $\tau = (s_1, a_1, b_1, r_1, \ldots, s_H, a_H, b_H, r_H, s_{H+1})$ is the resulting trajectory generated under (Π_0, μ^{Π_0}, P). Thus, any two models $\mathfrak{m}_1 = (P^{(1)}, \mathsf{R}^{(1)}, g^{(1)})$ and $\mathfrak{m}_2 = (P^{(2)}, \mathsf{R}^{(2)}, g^{(2)})$ in this class that generate the same $\mathcal{L}$ must satisfy $V_{\mathrm{cf}}^{(1)}(\pi) = V_{\mathrm{cf}}^{(2)}(\pi)$ for all $\pi \in \Pi$.

Step 1. Identification of transition and reward kernels. *Assumptions used: $C_{\mathrm{cf}} < \infty$.*

The coverage condition $C_{\mathrm{cf}} < \infty$ implies that for every $\pi \in \Pi$, $h \in [H]$, and $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$:

$$d_h^{\pi,\mu^\pi}(s,a,b) > 0 \quad \Longrightarrow \quad d_h^{\log}(s,a,b) > 0. \tag{30}$$

This follows because if $d_h^{\log}(s,a,b) = 0$ while $d_h^{\pi,\mu^\pi}(s,a,b) > 0$, the concentrability ratio at (s,a,b) would be infinite, contradicting $C_{\mathrm{cf}} < \infty$.

At every triple (s,a,b,h) with $d_h^{\log}(s,a,b) > 0$, the data-generating law $\mathcal{L}$ determines the conditional distributions and reward mean:

$$\mathsf{R}_h(E \mid s,a,b) = \Pr_{\mathcal{L}}\big(R_h \in E \mid s_h = s, a_h = a, b_h = b\big), \quad E \subseteq [0,1], \tag{31}$$

$$r_h(s,a,b) = \mathbb{E}_{\mathcal{L}}\big[R_h \mid s_h = s,\, a_h = a,\, b_h = b\big], \tag{32}$$

$$P_h(s' \mid s,a,b) = \Pr_{\mathcal{L}}\big(s_{h+1} = s' \mid s_h = s,\, a_h = a,\, b_h = b\big), \quad \forall\, s' \in \mathcal{S}. \tag{33}$$

These conditional quantities are identified because the conditioning event $\{s_h = s, a_h = a, b_h = b\}$ has positive probability under the logging measure (since $d_h^{\log}(s,a,b) > 0$).

Remark B.8. *By* (30)*, identification of P_h and R_h, including r_h, at* all *triples that matter for* any *target policy $\pi \in \Pi$ is guaranteed. Triples (s, a, b) with $d_h^{\pi,\mu^\pi}(s, a, b) = 0$ for all $\pi \in \Pi$ do not contribute to any counterfactual value and need not be identified.*

Step 2. Identification of the response function g. *Assumptions used: the finite response-state representation of Definition 3.1, in particular the time-invariant opponent-response kernel g, the response-kernel identity $\mu_h^\pi(\cdot \mid s) = g(\cdot \mid s, z_h^\pi)$, the known update $z_h^\pi = \phi_h(d_{1:h}^\pi)$, and $C_{\mathrm{cf}} < \infty$.*

For each logging policy $\pi_0 \in \mathrm{supp}(\nu_{\log})$, the data-generating law $\mathcal{L}$ reveals the complete trajectory distribution under (π_0, μ^{π_0}, P). We extract the opponent's conditional response at each step h and state s with $d_h^{\pi_0,\mu^{\pi_0}}(s) := \sum_{a,b} d_h^{\pi_0,\mu^{\pi_0}}(s, a, b) > 0$:

$$\mu_h^{\pi_0}(b \mid s) = \Pr_{\mathcal{L}}\big(b_h = b \mid s_h = s,\ \Pi_0 = \pi_0\big), \qquad \forall\, b \in \mathcal{B}. \tag{34}$$

This is identified because the deployed label Π_0 is observed as part of the data and the conditioning event $\{s_h = s, \Pi_0 = \pi_0\}$ has positive probability.

Computing response states from observables. Since the transition kernels P are identified (Step 1) and π_0 is known, we can compute the learner's marginal state-action occupancy measures

$$d_j^{\pi_0}(s, a) := \sum_{b \in \mathcal{B}} d_j^{\pi_0,\mu^{\pi_0}}(s, a, b), \qquad j \in [h], \tag{35}$$

via the forward recursion:

$$\begin{aligned} d_1^{\pi_0}(s, a) &= \rho(s)\, \pi_{0,1}(a \mid s), \\ d_{j+1}^{\pi_0}(s', a') &= \pi_{0,j+1}(a' \mid s') \sum_{s,a,b} d_j^{\pi_0}(s, a)\, \mu_j^{\pi_0}(b \mid s)\, P_j(s' \mid s, a, b), \qquad j \geq 1. \end{aligned} \tag{36}$$

Here $\mu_j^{\pi_0}(b \mid s)$ is available from (34), and P_j is identified from Step 1 at all triples with positive logging occupancy.

By the known update in Definition 3.1, the response state at step h is a deterministic functional of the learner-side marginal occupancies:

$$z_h^{\pi_0} = \phi_h\big(d_{1:h}^{\pi_0}\big), \tag{37}$$

where $d_{1:h}^{\pi_0} = \{d_j^{\pi_0}(s, a)\}_{j \leq h}$. Since ϕ_h is a known function and $d_{1:h}^{\pi_0}$ is identified from the data-generating law, each $z_h^{\pi_0}$ is identified.

Identifying g from observed response states. By the response-kernel identity in Definition 3.1, the observed opponent behavior at state s under response state $z_h^{\pi_0}$ satisfies

$$g(b \mid s,\ z_h^{\pi_0}) = \mu_h^{\pi_0}(b \mid s), \qquad \forall\, b \in \mathcal{B}, \tag{38}$$

at every state s visited under π_0 at step h. Define the set of observed response states:

$$\mathcal{Z}_{\mathrm{obs}} := \big\{z_h^{\pi_0} : \pi_0 \in \mathrm{supp}(\nu_{\log}),\ h \in [H]\big\} \ \subseteq\ \mathcal{Z}. \tag{39}$$

The data-generating law identifies $g(\cdot \mid s, z)$ at every (s, z) where $z \in \mathcal{Z}_{\mathrm{obs}}$ and s is visited under some logging policy at a step with response state z.

We now argue that g is identified on the full domain required for counterfactual evaluation.

Lemma B.9 (Response coverage on target-supported state-response pairs)**.** *Under the hypotheses of Proposition B.4 (including the joint state-response coverage condition), for every $\pi \in \Pi$ and $h \in [H]$, $g(\cdot \mid s, z_h^\pi)$ is identified for every state s with $d_h^{\pi,\mu^\pi}(s) > 0$.*

Proof of Lemma B.9. Fix $\pi \in \Pi$, $h \in [H]$, and $s \in \mathcal{S}$ with $d_h^{\pi,\mu^\pi}(s) > 0$. The joint state-response coverage hypothesis ensures there exists $\pi_0 \in \Pi_{\log}$ with $\nu_{\log}(\pi_0) > 0$ and $h_0 \in [H]$ such that $z_{h_0}^{\pi_0} = z_h^\pi$ and $d_{h_0}^{\pi_0,\mu^{\pi_0}}(s) > 0$. Because $d_{h_0}^{\pi_0,\mu^{\pi_0}}(s) > 0$, the logging distribution visits state s at step h_0 under policy π_0 with positive probability. Since the response state for this policy and step is exactly z_h^π, the data-generating law reveals the opponent's behavior $g(\cdot \mid s, z_h^\pi)$ via equation (38). This identification holds for all relevant states. □

Remark B.10. *The joint state-response coverage condition is not implied by $C_{\mathrm{cf}} < \infty$ alone. The concentrability condition guarantees (s, a, b)-level marginal support but does not guarantee that the logging policy visiting state s will simultaneously induce the required target response state z_h^π. A logging distribution that deploys a mixture of policies may provide adequate action-level coverage marginally while never visiting s* while *in response state z_h^π. The joint coverage condition closes this gap.*

Step 3. Causal identification of the target response and counterfactual value. *Assumptions used: simultaneous-action timing, $C_{\mathrm{cf}} < \infty$, joint response support, and the identified quantities from Steps 1–2.*

For a candidate sequence z, retain the operator notation

$$(T_\pi(z))_h = \phi_h(d_{1:h}^{\pi,z}),$$

where the learner marginals obey

$$\begin{aligned} d_1^{\pi,z}(s,a) &= \rho(s)\pi_1(a \mid s), \\ d_{j+1}^{\pi,z}(s',a') &= \pi_{j+1}(a' \mid s') \sum_{s,a,b} d_j^{\pi,z}(s,a) g(b \mid s, z_j) P_j(s' \mid s, a, b). \end{aligned}$$

The first marginal is independent of z, and coordinate h of T_π depends only on $z_{1:h-1}$. Hence the self-consistent response sequence is constructed forward.

We prove identification across all models consistent with the logged law. At stage one, every such model has the same $d_1^\pi = \rho \otimes \pi_1$ and therefore the same $z_1^\pi = \phi_1(d_1^\pi)$. Suppose the response prefix $z_{1:h-1}^\pi$ and the learner occupancies $d_{1:h-1}^\pi$ have been identified. Lemma B.9 identifies the response-kernel rows used at stage $h-1$, while $C_{\mathrm{cf}} < \infty$ identifies the corresponding transition rows. The forward equation therefore identifies d_h^π, and the known map gives $z_h^\pi = \phi_h(d_{1:h}^\pi)$. Induction identifies the complete sequence. The same argument proves uniqueness, because any self-consistent sequence must agree coordinate by coordinate. This proves claim (a) without a global fixed-point assumption or evaluation of unsupported false candidates.

Identification of the opponent response on target-visited states and of $V_{\mathrm{cf}}(\pi)$. With z^π identified, the counterfactual opponent response is determined by the response-kernel identity at every state that is actually reached under the target process:

$$\mu_h^\pi(b \mid s) = g(b \mid s, z_h^\pi), \qquad \forall\, h \in [H],\ b \in \mathcal{B},\ \text{and all } s \text{ with } d_h^{\pi,\mu^\pi}(s) > 0. \tag{40}$$

This proves claim (b).

For claim (c), it is enough to identify the target occupancy recursion on the reachable region. The target occupancy sequence satisfies

$$\begin{aligned} d_1^{\pi,\mu^\pi}(s,a,b) &= \rho(s)\,\pi_1(a \mid s)\,\mu_1^\pi(b \mid s), \\ d_{h+1}^{\pi,\mu^\pi}(s',a',b') &= \pi_{h+1}(a' \mid s')\,\mu_{h+1}^\pi(b' \mid s') \sum_{s,a,b} d_h^{\pi,\mu^\pi}(s,a,b)\,P_h(s' \mid s,a,b). \end{aligned} \tag{41}$$

Every factor on the right-hand side is identified on the target-supported region:

- $P_h(s' \mid s, a, b)$ and $\mathsf{R}_h(\cdot \mid s, a, b)$, including $r_h(s, a, b)$, are identified by Step 1 whenever $d_h^{\pi,\mu^\pi}(s, a, b) > 0$;
- $\mu_h^\pi(b \mid s)$ is identified by (40) whenever $d_h^{\pi,\mu^\pi}(s) > 0$;
- π and ρ are known primitives.

Thus the target occupancy sequence is uniquely determined by $\mathcal{L}$, establishing the occupancy part of claim (c).

To identify the full counterfactual physical trajectory law, fix any trajectory

$$\tau = (s_1, a_1, b_1, s_2, \ldots, s_H, a_H, b_H, s_{H+1})$$

with positive probability under (π, μ^π, P). Its probability is

$$\overset{\pi}{\Pr}(\tau) = \rho(s_1) \prod_{h=1}^{H} \pi_h(a_h \mid s_h)\, \mu_h^\pi(b_h \mid s_h)\, P_h(s_{h+1} \mid s_h, a_h, b_h). \tag{42}$$

Every factor in (42) is identified on the target-supported region by the preceding steps, so the induced physical trajectory law is uniquely determined by $\mathcal{L}$. The identified Markov reward kernels then determine the joint law after the realized rewards are included. This proves claim (c).

For claim (d), the counterfactual value is

$$V_{\mathrm{cf}}(\pi) = \sum_{h=1}^{H} \sum_{s,a,b} d_h^{\pi,\mu^\pi}(s,a,b)\, r_h(s,a,b). \tag{43}$$

Therefore, $V_{\mathrm{cf}}(\pi)$ is uniquely determined by the data-generating law $\mathcal{L}$. More generally, every bounded functional of the identified counterfactual trajectory law is uniquely determined as well. This completes the proof of claim (d).

Remark B.11 (Proof dependencies)**.** *The proof uses four ingredients. The finite response-state model and the time-invariant response map reduce adaptive behavior to a common tabular object g. The known update and response-kernel identity turn each logged deployment into an identified response state and hence into identified values of g on visited state-response pairs. The support condition $C_{\mathrm{cf}} < \infty$ identifies transitions and rewards on all target-relevant triples. Finally, joint state-response coverage identifies g on the target-supported state-response pairs. Simultaneous-action timing then makes the response map triangular: stage h depends only on the already constructed occupancy prefix. Causal induction identifies z_h^π, the next reachable occupancy, and ultimately the value.*

□

Proposition B.12 (Local support sufficiency and generic necessity)**.** *Under the hypotheses of Proposition B.4, fix a target policy $\pi \in \Pi$. To identify the full rewarded counterfactual trajectory of π, including s_{H+1}, it suffices that the data-generating law identify:*

(i) *P_h and R_h, including r_h, on every triple (s,a,b) with $d_h^{\pi,\mu^\pi}(s,a,b) > 0$;*

(ii) *$g(\cdot \mid s, z_h^\pi)$ on every state s with $d_h^{\pi,\mu^\pi}(s) > 0$.*

No identification of P, R (including its mean r), or g away from this target-supported slice is required. For the expected return alone, the weaker collection consisting of P_h for $h < H$, the means r_h for all h, and the response-kernel rows on target-reachable support is sufficient. Reward means and response-kernel rows on that support are generically value-separating, as are transitions at stages $h < H$ when later rewards depend on the next state. The terminal transition P_H and features of R_h beyond its mean can affect the full trajectory law but not the expected H-stage return. These are generic, completion-relative claims, not universal necessity statements for a fixed instance with possible cancellations or value-irrelevant rows.

Proof. We first prove sufficiency and then record the qualified converse.

Sufficiency. Suppose conditions (i) and (ii) hold. The counterfactual rollout under (π, μ^π, P) is the sequence of joint occupancies $(d_h^{\pi,\mu^\pi})_{h=1}^H$ and the joint trajectory law over $\bigotimes_h (\mathcal{S} \times \mathcal{A} \times \mathcal{B} \times [0,1])$. The forward recursion

$$d_{h+1}^{\pi,\mu^\pi}(s') \;=\; \sum_{(s,a,b)} d_h^{\pi,\mu^\pi}(s,a,b)\, P_h(s' \mid s,a,b),$$

shows that the step-$(h{+}1)$ state occupancy depends only on P_h evaluated at triples with $d_h^{\pi,\mu^\pi}(s,a,b) > 0$, which are identified by (i). The step-h joint occupancy $d_h^{\pi,\mu^\pi}(s,a,b) = d_h^{\pi,\mu^\pi}(s)\, \pi_h(a \mid s)\, g(b \mid s, z_h^\pi)$ depends only on $g(\cdot \mid s, z_h^\pi)$ evaluated at s with $d_h^{\pi,\mu^\pi}(s) > 0$, which is identified by (ii). The response coordinate z_h^π is determined by the deterministic functional ϕ_h applied to the causal learner occupancy prefix, and is therefore identified once that prefix is.

The realized reward sequence is identified from the kernels R_h at triples covered by (i), while its expectation satisfies $\mathbb{E}[R_h] = \sum_{(s,a,b)} d_h^{\pi,\mu^\pi}(s,a,b)\, r_h(s,a,b)$. Hence the entire counterfactual rollout and the value $V_{\mathrm{cf}}(\pi)$ are uniquely determined.

Generic value separation under compatible completions. If the reward mean in condition (i) is unidentified at some triple (s_0, a_0, b_0, h_0) with $d_{h_0}^{\pi,\mu^\pi}(s_0, a_0, b_0) > 0$, then by the construction in Theorem B.2, two reward functions agreeing elsewhere but disagreeing on $r_{h_0}(s_0, a_0, b_0)$ produce two adaptive games with the same logged law but distinct $V_{\mathrm{cf}}(\pi)$. Hence identification at this triple is necessary for a completion class that permits that separating reward perturbation. If an on-slice transition P_{h_0} is unidentified for $h_0 < H$, choose two compatible transition rows that send the process to states with different known downstream rewards; the resulting target values differ. An analogous construction applies to condition (ii): if $g(\cdot \mid s_0, z_{h_0}^\pi)$ is unidentified at some s_0 with $d_{h_0}^{\pi,\mu^\pi}(s_0) > 0$, two opponent-response kernels agreeing on the logged (s, z) pairs but disagreeing at $(s_0, z_{h_0}^\pi)$ produce distinct counterfactual values, by the same $0/1$-reward construction restricted to $(s_0, z_{h_0}^\pi)$.

The joint-occupancy and response-state lower bounds formalize two such separating families. Neither the shape of a reward distribution at fixed mean nor P_H can alter the expected return, although both remain part of the full-law identification claim. For a particular fixed value problem, cancellation or a value-irrelevant unidentified row can also defeat universal necessity. Thus the proposition gives exact full-law sufficiency and a precisely scoped, completion-relative account of value separation; no off-slice values of P, r, or g are invoked. □

Relation to structural counterfactual identification. The barriers above echo the econometric fact that dynamic-game counterfactuals require behavioral and equilibrium restrictions (Kalouptsidi et al., 2017; Aguirregabiria & Mira, 2007; Bajari et al., 2007; Kalouptsidi et al., 2021). Here the finite response state makes the deployment response itself an explicit estimand and separates state-joint-action support from state-response support. This yields a finite-log recovery theorem, but it does not make the abstraction testable from arbitrary passive data. The same distinction separates our problem from confounded OPE: latent-confounding methods study identification of a fixed target law (Bareinboim et al., 2015; Bennett et al., 2021; Kausik et al., 2024), whereas changing the policy here changes the opponent law that defines the target.

B.7 Why support cannot be certified from passive counts

Lemma B.13 (Zero support versus rare support for a cell indicator)**.** *For the Bernoulli indicator that a potentially relevant cell is hit, the total-variation distance between n observations generated with success probability zero and with success probability ϵ is*

$$1 - (1-\epsilon)^n \le n\epsilon.$$

No finite-sample procedure can therefore certify positivity uniformly without a known lower-mass condition.

Proof. Under zero probability the indicator sequence is identically zero. Under probability ϵ, that same sequence occurs with probability $(1-\epsilon)^n$. The overlap of the two Bernoulli product laws therefore has mass $(1-\epsilon)^n$, giving the equality. Bernoulli's inequality gives the upper bound. The same calculation applies to paired categorical laws that are identical conditional on no hit of the designated cell. □

Lemma B.14 (Maps that agree on the menu)**.** *The reconstructed logging labels, recovered target paths, counts, and lower confidence bounds of Algorithm 1 depend on $\phi_{1:H}$ only through its values on occupancy prefixes generated by $\Pi \cup \Pi_{\log}$. Any two maps that agree on that finite set produce identical outputs.*

Proof. For each logging policy π_0, the reconstructed label $\widehat{z}_h^{\pi_0} = \phi_h(\widetilde{d}_{1:h}^{\pi_0})$ is evaluated at a policy-factorized occupancy prefix of π_0. Target recovery evaluates ϕ_h only at plug-in prefixes of policies in Π. Counts $N_h(s, z, b)$ are indexed by those reconstructed labels, and the pessimistic backups use

only the recovered target path. Therefore every output is a function of ϕ restricted to prefixes of $\Pi \cup \Pi_{\log}$. □

Lemma B.15 (Design-based coverage bounds)**.** *Let $\nu_\star := \min_{\pi\in\Pi} \nu_{\log}(\pi)$. If $\nu_\star > 0$, equivalently if every target is logged, then $C_{\mathrm{cf}} \le 1/\nu_\star$ and $C_{\mathrm{sr}} \le 1/\nu_\star$. If a target π with $\nu_{\log}(\pi) = 0$ shares its response path with some $\pi_z \in \Pi_{\log}$, then that target's contribution to either supremum is at most ordinary concentrability between π and π_z in the common environment $(P, g(\cdot \mid s, z^\pi))$, scaled by $1/\nu_{\log}(\pi_z)$. Explicitly,*

$$\sup_{h,s,a,b} \frac{d_h^{\pi,\mu^\pi}(s,a,b)}{d_h^{\log}(s,a,b)} \le \frac{1}{\nu_{\log}(\pi_z)} \sup_{h,s,a,b} \frac{d_h^{\pi,\mu^\pi}(s,a,b)}{d_h^{\pi_z,\mu^{\pi_z}}(s,a,b)},$$
$$\sup_{h,s:p_h^\pi(s)>0} \frac{p_h^\pi(s)}{d_h^{\log}(s,z_h^\pi)} \le \frac{1}{\nu_{\log}(\pi_z)} \sup_{h,s:p_h^\pi(s)>0} \frac{p_h^\pi(s)}{p_h^{\pi_z}(s)},$$

with the convention $0/0 = 0$ and $x/0 = \infty$ for $x > 0$.

Proof. If $\pi \in \Pi$ and $\nu_{\log}(\pi) \ge \nu_\star > 0$, then

$$d_h^{\log}(s,a,b) \ge \nu_{\log}(\pi)\, d_h^{\pi,\mu^\pi}(s,a,b) \ge \nu_\star\, d_h^{\pi,\mu^\pi}(s,a,b)$$

and

$$d_h^{\log}(s,z_h^\pi) \ge \nu_{\log}(\pi)\, p_h^\pi(s) \ge \nu_\star\, p_h^\pi(s).$$

The two coverage ratios are therefore at most $1/\nu_\star$ for every such target, which proves the first claim.

If $z^{\pi_z} = z^\pi$, then $\mu^{\pi_z} = \mu^\pi$ because both policies use the same response-kernel rows $g(\cdot \mid s, z_h^\pi)$. Hence

$$d_h^{\log}(s,a,b) \ge \nu_{\log}(\pi_z)\, d_h^{\pi_z,\mu^{\pi_z}}(s,a,b), \qquad d_h^{\log}(s,z_h^\pi) \ge \nu_{\log}(\pi_z)\, p_h^{\pi_z}(s).$$

Dividing the target occupancies by these lower bounds yields the displayed inequalities. A zero occupancy of π_z on a target-supported cell still makes the corresponding ratio infinite. □

C Proof of the End-to-End Guarantee

Once the response path is recovered, our backward recursion deliberately uses the pessimistic machinery developed for offline RL (Jin et al., 2021; Yin & Wang, 2021; Rashidinejad et al., 2021; Xie et al., 2021; Zanette et al., 2021; Shi et al., 2022; Li et al., 2024); see Levine et al. (2020) for a broader overview. Concentrability and batch learning lower bounds explain the state-joint-action support term (Munos & Szepesvári, 2008; Chen & Jiang, 2019). The proof below adds the layer that those fixed-environment analyses do not need: reconstructing response labels, estimating a separate opponent-response kernel, and verifying that the forward target path remains inside its response cells before the Bellman argument begins.

C.1 Margins and empirical conventions

For any policy π and stage h, define

$$\gamma_h(\pi) := \inf \left\{ \|\widetilde{d}_{1:h} - d_{1:h}^\pi\|_{1,\mathrm{pref}} : \begin{array}{l} \widetilde{d}_j(s,a) = \widetilde{p}_j(s)\pi_j(a \mid s), \quad \widetilde{p}_j \in \Delta(\mathcal{S}), \\ \phi_h(\widetilde{d}_{1:h}) \ne \phi_h(d_{1:h}^\pi) \end{array} \right\}, \tag{44}$$

with the infimum of the empty set equal to $+\infty$. Then

$$\gamma_{\mathrm{tar}} = \min_{\pi\in\Pi,h} \gamma_h(\pi), \qquad \gamma_{\log} = \min_{\pi_0\in\Pi_{\log},h} \gamma_h(\pi_0).$$

At a zero-count state-joint-action row, define $\widehat{P}_h(\cdot \mid s,a,b)$ to be a fixed distribution, such as the uniform distribution on $\mathcal{S}$, and set $\widehat{r}_h(s,a,b) = 0$. At a zero-count response row, define $\widehat{g}_h(\cdot \mid s,z)$ to be uniform on $\mathcal{B}$. These choices guarantee stochastic forward updates. The proof never requires them to approximate the corresponding population row; their influence is controlled through occupancy leakage from the covered target-reachable support.

Component	What Theorem 4.3 assumes or learns
Response structure	Finite $\mathcal{Z}$, maps $\phi_{1:H}$, and initial law ρ are specified; the opponent-response kernel g is learned.
Logged information	Learner-policy labels are observed; response labels and target responses are not.
Logging design	A finite menu with probability floor $\nu_{\min}$ is repeated across independent episodes.
Recoverability	Logging and target occupancy prefixes have positive local margins from response boundaries.
Support	Target state-joint-action and same-stage (s, z) cells are covered. Logging every target at mass $\nu_\star > 0$ yields $C_{\mathrm{cf}}, C_{\mathrm{sr}} \le 1/\nu_\star$.
Sample size	Equations (6) and (7) protect logging labels, positive-count rows, and target labels.
Computation	One forward pass per target; no enumeration of response sequences or coverage of false candidate paths is needed.

Table 4. Assumptions and learned objects in the end-to-end guarantee. Population support and response-model fidelity require design or domain knowledge; counts alone do not prove them.

Algorithm 2 Response recovery and pessimistic off-policy evaluation

Require: Policy-labeled dataset $\mathcal{D}$, finite logging policy set $\Pi_{\log}$, target class Π, confidence δ, mode in $\{\text{end-to-end}, \text{observed-label}, \text{oracle-response}\}$, and supplied labels or target responses only in the corresponding benchmark modes.

if mode $=$ end-to-end **then**
  Reconstruct every logger response sequence with the policy-factorized occupancy induction of Lemma C.1.
end if
Form $\widehat{P}_h, \widehat{r}_h, \widehat{g}_h$ and their game-kernel and response-side confidence envelopes from the recovered or observed labels.
for each target policy $\pi \in \Pi$ **do**
  if mode $=$ oracle-response **then**
    Use the supplied target sequence z^π.
    Set $\widehat{\mu}_h^\pi(\cdot \mid s) \leftarrow \widehat{g}_h(\cdot \mid s, z_h^\pi)$ for every h, s.
  else
    Set $\widehat{d}_1^\pi = \rho \otimes \pi_1$.
    for $h = 1, \dots, H$ **do**
      $\widehat{z}_h^\pi \leftarrow \phi_h(\widehat{d}_{1:h}^\pi)$ and $\widehat{\mu}_h^\pi(\cdot \mid s) \leftarrow \widehat{g}_h(\cdot \mid s, \widehat{z}_h^\pi)$.
      Propagate $\widehat{d}_{h+1}^\pi$ with the stochastic zero-count fallback for $\widehat{P}_h$.
    end for
  end if
  $\widehat{V}_{H+1}^\pi(s) \leftarrow 0$ for all s.
  for $h = H, H-1, \dots, 1$ **do**
    $\overline{Q}_h^\pi(s, a, b) \leftarrow [\widehat{r}_h + (\widehat{P}_h \widehat{V}_{h+1}^\pi) - \Gamma_h]_0^{H-h+1}$.
    $\widetilde{V}_h^\pi(s) \leftarrow \sum_a \pi_h(a \mid s) \sum_b \widehat{\mu}_h^\pi(b \mid s) \overline{Q}_h^\pi(s, a, b) - \Lambda_h(s; \pi)$.
    $\widehat{V}_h^\pi(s) \leftarrow [\widetilde{V}_h^\pi(s)]_0^{H-h+1}$.
  end for
  $\widehat{V}^-(\pi) \leftarrow \mathbb{E}_{s \sim \rho}[\widehat{V}_1^\pi(s)]$.
end for
return all one-sided lower confidence bounds and $\widehat{\pi} \in \arg\max_{\pi \in \Pi} \widehat{V}^-(\pi)$.

C.2 Mode-specific implementation details

For concreteness, Algorithm 2 states the unified end-to-end, observed-label, and oracle-response variants. The end-to-end mode uses only the policy-labeled dataset. The other modes are retained as intermediate benchmarks.

The response-recovery phase is a forward pass and the value phase is a backward pass. The total runtime is polynomial in $|\Pi_{\log}|, |\Pi|, S, A, B, H$, and d_R; there is no d_R^H enumeration term. The oracle-response mode remains useful for isolating the inherited value-estimation component.

The pessimistic recursion (lines 10–14) is identical to the standard offline RL backward induction (Jin et al., 2021), with two modifications: the bonus Γ_h now also accounts for the opponent action coordinate, and the additional response-side bonus $\Lambda_h(s;\pi)$ subtracts a one-sided correction for plug-in error in $\widehat{\mu}_h^{\pi}$. Both modifications are required for the plug-in pessimism reduction to deliver the uniform evaluation guarantee of Theorem C.13.

C.3 Reconstructing the logged response labels

Lemma C.1 (Logging-label reconstruction)**.** *Suppose $\Pi_{\log}$ is finite, $\nu_{\min} > 0$, and $\gamma_{\log} > 0$. For each π_0, let n_0 be its episode count, let $\widetilde{p}_h^{\pi_0}$ be the empirical state frequency at stage h among these episodes, and set*

$$\widetilde{d}_h^{\pi_0}(s,a) = \widetilde{p}_h^{\pi_0}(s)\pi_{0,h}(a \mid s).$$

There is a universal c such that, if

$$n \geq c\left[\frac{H^2(S+\iota)}{\nu_{\min}\gamma_{\log}^2} + \frac{\iota}{\nu_{\min}}\right], \qquad \iota = \log\frac{SAHMn}{\delta},$$

then, with probability at least $1-\delta$, $\phi_h(\widetilde{d}_{1:h}^{\pi_0}) = z_h^{\pi_0}$ for every logging policy and stage.

Proposition 4.1 is this statement with the conservative cardinality SA in place of S.

Proof of Proposition 4.1. A multiplicative Chernoff bound and a union bound give $n_0 \geq n\nu_{\min}/2$ for every π_0 with probability at least $1-\delta/3$, after increasing the universal constant. Conditional on the policy labels, the trajectories bearing a fixed label π_0 are independent draws from the deployment law of π_0. At every fixed stage, their states are therefore i.i.d. categorical draws with distribution $p_h^{\pi_0}$. Standard multinomial ℓ_1 concentration gives, simultaneously over π_0 and h,

$$\|\widetilde{p}_h^{\pi_0} - p_h^{\pi_0}\|_1 \leq c\sqrt{\frac{S+\iota}{n\nu_{\min}}}.$$

Because both occupancies are factored with the same known policy, $\|\widetilde{d}_h^{\pi_0} - d_h^{\pi_0}\|_1 = \|\widetilde{p}_h^{\pi_0} - p_h^{\pi_0}\|_1$. Summing through stage h gives prefix error at most $cH\sqrt{(S+\iota)/(n\nu_{\min})} < \gamma_{\log}$. Each $\widetilde{d}_h^{\pi_0}$ is policy-factorized, so the margin definition now applies and yields the asserted label equality. No estimate of P or g is used in this reconstruction. □

On the event in Lemma C.1 (Proposition 4.1), every reconstructed state-response count is exactly the count that would have been formed from observed response labels. The same trajectories may therefore be used for reconstruction and primitive estimation without sample splitting.

C.4 Primitive concentration on target-reachable support

Let $N_h(s,a,b)$ and $N_h(s,z)$ be the state-joint-action and reconstructed state-response counts. Define

$$\kappa_{\text{model}} := \min\{d_h^{\log}(s,a,b) : d_h^{\log}(s,a,b) > 0\},$$

$$\kappa_{\text{resp}} := \min\{d_h^{\log}(s,z) : d_h^{\log}(s,z) > 0\}, \qquad \kappa := \min\{\kappa_{\text{model}}, \kappa_{\text{resp}}\}.$$

Lemma C.2 (Target-weighted primitive errors)**.** *On an event of probability at least $1-\delta$, simultaneously for all targets and stages,*

$$\sum_{s,a,b} d_h^{\pi,\mu^\pi}(s,a,b)\|\widehat{P}_h(\cdot \mid s,a,b) - P_h(\cdot \mid s,a,b)\|_1 \leq e_P, \tag{45}$$

$$\sum_{s} p_h^{\pi}(s)\|\widehat{g}_h(\cdot \mid s, z_h^{\pi}) - g(\cdot \mid s, z_h^{\pi})\|_1 \leq e_g, \tag{46}$$

$$\sum_{s,a,b} d_h^{\pi,\mu^\pi}(s,a,b)|\widehat{r}_h(s,a,b) - r_h(s,a,b)| \leq e_r, \tag{47}$$

where

$$e_P = c\sqrt{\frac{C_{\mathrm{cf}} S^2 AB\iota}{n}}, \qquad e_g = c\sqrt{\frac{C_{\mathrm{sr}} SB\iota}{n}},$$
$$e_r = c\sqrt{\frac{C_{\mathrm{cf}} SAB\iota}{n}}, \qquad \iota = \log \frac{SABH|\mathcal{Z}||\Pi|Mn}{\delta}.$$

provided $n \geq c\iota/\kappa$ and the logging-label event holds.

Proof. Conditional multinomial concentration gives, on every positive-count population cell,

$$\|\widehat{P}_h - P_h\|_1 \leq c\sqrt{\frac{S\iota}{N_h(s,a,b)}}, \qquad \|\widehat{g}_h - g\|_1 \leq c\sqrt{\frac{B\iota}{N_h(s,z)}}, \qquad |\widehat{r}_h - r_h| \leq c\sqrt{\frac{\iota}{N_h(s,a,b)}}.$$

The minimum-mass condition and Bernstein concentration also give $N_h(s,a,b) \geq n d_h^{\log}(s,a,b)/2$ and $N_h(s,z) \geq n d_h^{\log}(s,z)/2$ for every positive-mass cell.

For the transition term, finite C_{cf} implies $d_h^{\pi,\mu^\pi}(s,a,b) \leq C_{\mathrm{cf}} d_h^{\log}(s,a,b)$. Hence

$$\sum_{s,a,b} d_h^{\pi,\mu^\pi}(s,a,b)\sqrt{\frac{S\iota}{n d_h^{\log}(s,a,b)}}$$
$$\leq \sqrt{\frac{C_{\mathrm{cf}} S\iota}{n}} \sum_{s,a,b} \sqrt{d_h^{\pi,\mu^\pi}(s,a,b)} \leq \sqrt{\frac{C_{\mathrm{cf}} S^2 AB\iota}{n}},$$

where the final step is Cauchy-Schwarz. The response bound is identical with $p_h^\pi(s) \leq C_{\mathrm{sr}} d_h^{\log}(s, z_h^\pi)$ and at most S states. Applying the transition calculation without its leading $\sqrt{S}$ factor proves the reward bound. Since $e_r \leq e_P$, the reward contribution is absorbed by the theorem's displayed rate. □

C.5 Forward recovery of the target response

Lemma C.3 (Target-path recovery)**.** *Assume the event of Lemma C.2. If*

$$\frac{H^2}{2}(e_P + e_g) < \gamma_{\mathrm{tar}},$$

then the empirical forward pass returns $\widehat{z}^\pi = z^\pi$ for every $\pi \in \Pi$.

Proposition 4.2 is this statement.

Proof of Proposition 4.2. Fix π and write $\Delta_h = \|\widehat{d}_h^\pi - d_h^\pi\|_1$. The first marginal is exact because ρ is known, so $\Delta_1 = 0$ and $\widehat{z}_1^\pi = z_1^\pi$.

Suppose $\widehat{z}_j^\pi = z_j^\pi$ for $j \leq h$, where $h < H$. Using the correct labels through stage h, decompose the next-state difference as

$$\widehat{p}_{h+1} - p_{h+1} = \sum_{s,a,b} (\widehat{d}_h - d_h)(s,a)\widehat{g}_h(b \mid s, z_h^\pi)\widehat{P}_h(\cdot \mid s,a,b)$$
$$+ \sum_{s,a,b} d_h(s,a)[\widehat{g}_h - g](b \mid s, z_h^\pi)\widehat{P}_h(\cdot \mid s,a,b)$$
$$+ \sum_{s,a,b} d_h(s,a) g(b \mid s, z_h^\pi)[\widehat{P}_h - P_h](\cdot \mid s,a,b).$$

Because all plug-in rows are stochastic, the three terms have ℓ_1 norm at most Δ_h, e_g, and e_P, respectively. Since multiplication by the fixed learner policy preserves the state-marginal ℓ_1 distance,

$$\Delta_{h+1} \leq \Delta_h + e_P + e_g.$$

Therefore $\Delta_j \le (j-1)(e_P + e_g)$ and

$$\|\widehat{d}^{\pi}_{1:h+1} - d^{\pi}_{1:h+1}\|_{1,\mathrm{pref}} \le \sum_{j=1}^{h+1} \Delta_j \le \frac{H^2}{2}(e_P + e_g) < \gamma_{\mathrm{tar}}.$$

The margin definition forces $\widehat{z}^{\pi}_{h+1} = z^{\pi}_{h+1}$, closing the induction. The concentration event is simultaneous, so no additional union bound over response sequences is used. □

C.6 Pessimistic evaluation after recovery

For completeness, define $\widehat{V}^{\pi}_{H+1} = 0$ and

$$\overline{Q}^{\pi}_h(s,a,b) := \left[\widehat{r}_h(s,a,b) + (\widehat{P}_h \widehat{V}^{\pi}_{h+1})(s,a,b) - \Gamma_h(s,a,b)\right]_0^{H-h+1}, \tag{48}$$

$$\widehat{V}^{\pi}_h(s) := \left[\sum_a \pi_h(a \mid s) \sum_b \widehat{g}_h(b \mid s, \widehat{z}^{\pi}_h) \overline{Q}^{\pi}_h(s,a,b) - \Lambda_h(s;\pi)\right]_0^{H-h+1}, \tag{49}$$

$$\widehat{V}^{-}(\pi) := \mathbb{E}_{s\sim\rho} \widehat{V}^{\pi}_1(s). \tag{50}$$

Lemma C.4 (Pessimistic plug-in bound)**.** *Suppose $\widehat{z}^{\pi} = z^{\pi}$ for every target and the transition, response, and mean-reward concentration bounds hold. With bonuses chosen as above, simultaneously for all π,*

$$0 \le V_{\mathrm{cf}}(\pi) - \widehat{V}^{-}(\pi) \le C \sum_{h=1}^{H} \left(\mathbb{E}_{d^{\pi,\mu^{\pi}}_h} \Gamma_h(s,a,b) + \mathbb{E}_{p^{\pi}_h} \Lambda_h(s;\pi)\right).$$

The right side is at most

$$\widetilde{\mathcal{O}}\left(H^2 \sqrt{\frac{C_{\mathrm{cf}} S^2 AB}{n}} + H^2 \sqrt{\frac{C_{\mathrm{sr}} SB}{n}}\right).$$

Proof. The componentwise transition and reward events imply, for every function $v \in [0,H]^S$,

$$|(\widehat{P}_h - P_h)v|(s,a,b) \le H \|\widehat{P}_h(\cdot \mid s,a,b) - P_h(\cdot \mid s,a,b)\|_1 \le \Gamma_h(s,a,b),$$

after increasing the constant; the scalar mean-reward error is absorbed in the same envelope. Similarly, for every $q \in [0,H]^B$,

$$|(\widehat{g}_h - g)q|(s, z^{\pi}_h) \le H \|\widehat{g}_h(\cdot \mid s, z^{\pi}_h) - g(\cdot \mid s, z^{\pi}_h)\|_1 \le \Lambda_h(s;\pi).$$

Backward induction in (48) and (49) shows $\widehat{V}^{\pi}_h(s) \le V^{\pi,\mu^{\pi}}_h(s)$. For the gap, add and subtract the plug-in Bellman terms, use the two envelopes, and propagate the remaining continuation gap under the true target dynamics. Unrolling gives the displayed occupancy-weighted sum. Clipping cannot increase the distance to a target in $[0, H-h+1]$. Applying the same Cauchy-Schwarz calculation as in Lemma C.2 to the count-based bonuses yields the rate. □

Proof of Theorem 4.3. Allocate failure probability among Lemma C.1 (Proposition 4.1), Lemma C.2, and Lemma C.3 (Proposition 4.2). Condition (6) gives exact logged labels and enough counts. Condition (7), with a sufficiently small universal c, gives exact target labels. Lemma C.4 then gives the uniform evaluation bound.

Let $\widehat{\pi}$ maximize $\widehat{V}^{-}$. Since $\widehat{V}^{-}(\pi) \le V_{\mathrm{cf}}(\pi)$ for every π,

$$\begin{aligned} V_{\mathrm{cf}}(\pi^*) - V_{\mathrm{cf}}(\widehat{\pi}) &\le V_{\mathrm{cf}}(\pi^*) - \widehat{V}^{-}(\widehat{\pi}) \\ &\le V_{\mathrm{cf}}(\pi^*) - \widehat{V}^{-}(\pi^*), \end{aligned}$$

so the same uniform rate controls selection. Computing all empirical counts costs $O(nH)$. A dense transition update costs $O(S^2AB)$ per stage and target, proving the runtime statement. □

C.6.1 EQUIVALENT FORMULATION ON TARGET-REACHABLE SUPPORT

Definition C.5 (Logged-supported response coordinates)**.** *For each stage h, let*

$$\mathcal{Z}_{h,\log} := \{z \in \mathcal{Z} : \exists \pi_0 \in \Pi_{\log} \text{ with } \nu_{\log}(\pi_0) > 0 \text{ and } z_h^{\pi_0} = z\}.$$

The product $\prod_h \mathcal{Z}_{h,\log}$ records which response coordinates occur in the finite logging policy set. It is a reporting device, not a set on which coverage of target-reachable support controls every candidate rollout.

Assumption C.6 (True-path local response margin)**.** *For each logger and target policy, every response map ϕ_h is constant on a neighborhood of the corresponding population learner-occupancy prefix. The radius may depend on the policy and stage, and the minimum radius over the finite declared classes is positive.*

Remark C.7 (Equivalent formulation of Theorem 4.3)**.** *Under the finite repeated logging policy set, simultaneous-action timing, target-reachable state-joint-action and same-stage state-response coverage, and the local margins above, the count, concentration, and margin inequalities of Theorem 4.3 already imply exact logged-label reconstruction and sequential recovery of z_h^π for every $\pi \in \Pi$. The runtime remains polynomial in the tabular dimensions and H. No union bound or coverage condition is taken over false response paths. The argument is the one given for Theorem 4.3: Lemma C.1 reconstructs labels, Lemma C.2 controls primitive errors on target-reachable support, and Lemma C.3 closes the forward induction.*

Remark C.8 (Relationship to the uniform candidate-path benchmark)**.** *The uniform candidate-path theorem remains a valid stronger-assumption benchmark when its uniform coverage and margin conditions hold. Exhaustive enumeration is not needed in the simultaneous-action model. The logged-supported set in Definition C.5 remains useful for reporting which response regimes were actually observed, not as a substitute for the causal proof.*

C.7 STRONGER INFORMATION REGIMES

The end-to-end theorem reconstructs the logged labels and follows the true target path. The next results show how the guarantees change when response labels are supplied or when coverage is imposed on every candidate path.

Corollary C.9 (Observed logged response states)**.** *If the response-state sequence is recorded for every logged episode, Lemma C.1 and the label-reconstruction terms involving $\gamma_{\log}$ and $\nu_{\min}$ may be removed. The positive-cell requirement $n \gtrsim \iota/\kappa$, the target-recovery condition, and the evaluation and selection rates are unchanged.*

Proof. The recorded labels define the same counts that Lemma C.1 reconstructs. The remaining proof is identical. □

Corollary C.10 (Evaluation with a given target path)**.** *Suppose the response-state labels of logged episodes are observed and the target sequence z^π is supplied for every $\pi \in \Pi$. Under finite $C_{\mathrm{cf}}, C_{\mathrm{sr}}$ and the minimum-positive-mass condition $n \gtrsim \iota/\min\{\kappa_{\mathrm{model}}, \kappa_{\mathrm{resp}}\}$, the recursion in* (48)–(49) *satisfies, with probability at least $1-\delta$, simultaneously for every target,*

$$0 \le V_{\mathrm{cf}}(\pi) - \widehat{V}^-(\pi) \le \widetilde{\mathcal{O}}\left(H^2 \sqrt{\frac{C_{\mathrm{cf}} S^2 AB}{n}} + H^2 \sqrt{\frac{C_{\mathrm{sr}} SB}{n}} \right).$$

The same bound controls the regret of the maximizing policy. No logging or target response margin is needed in this supplied-information regime.

Proof. Observed logged labels define the state-joint-action and state-response counts directly, and the given target sequence selects the relevant response-kernel rows. The simultaneous multinomial and reward concentration event in Lemma C.2 therefore holds without Lemmas C.1 and C.3. Lemma C.4 gives the displayed evaluation bound, and the selection argument in the proof of Theorem 4.3 gives the regret claim. This comparison shows that the end-to-end result matches the oracle-response value rate once its response-recovery conditions are met. □

For comparison, consider the stronger requirement that every candidate response path is covered. Let $\mu_h^z(\cdot \mid s) = g(\cdot \mid s, z_h)$ for every sequence $z \in \mathcal{Z}^H$, and let $p_h^{\pi,z}$ and $d_h^{\pi,z}$ be its state and joint occupancies. Define

$$C_{\mathrm{cf}}^{\mathrm{cand}} := \sup_{\pi\in\Pi, z\in\mathcal{Z}^H, h, s, a, b} \frac{d_h^{\pi,z}(s,a,b)}{d_h^{\log}(s,a,b)}, \tag{51}$$

$$C_{\mathrm{sr}}^{\mathrm{cand}} := \sup_{\pi\in\Pi, z\in\mathcal{Z}^H, h, s: p_h^{\pi,z}(s)>0} \frac{p_h^{\pi,z}(s)}{d_h^{\log}(s,z_h)}. \tag{52}$$

Assumption C.11 (Uniform candidate-path response margin)**.** *There is $\gamma_\phi > 0$ such that for every $\pi \in \Pi$, $z \in \mathcal{Z}^H$, and $h \in [H]$,*

$$\|\widetilde{d}_{1:h} - d_{1:h}^{\pi,z}\|_{1,\mathrm{pref}} < \gamma_\phi \quad \Longrightarrow \quad \phi_h(\widetilde{d}_{1:h}) = \phi_h(d_{1:h}^{\pi,z}).$$

Theorem C.12 (Uniform stability over candidate paths)**.** *Suppose logged response labels are observed or have been reconstructed correctly, $C_{\mathrm{cf}}^{\mathrm{cand}}, C_{\mathrm{sr}}^{\mathrm{cand}} < \infty$, and Assumption C.11 holds. Let*

$$\iota_{\mathrm{cand}} = \log(SABH|\mathcal{Z}|^H|\Pi|n/\delta).$$

If $n \gtrsim \iota_{\mathrm{cand}} / \min\{\kappa_{\mathrm{model}}, \kappa_{\mathrm{resp}}\}$ and

$$H^2\left(\sqrt{\frac{C_{\mathrm{cf}}^{\mathrm{cand}} S^2 AB\iota_{\mathrm{cand}}}{n}} + \sqrt{\frac{C_{\mathrm{sr}}^{\mathrm{cand}} SB\iota_{\mathrm{cand}}}{n}}\right) < c\gamma_\phi,$$

then the empirical and population response operators agree on every $\pi \in \Pi$ and $z \in \mathcal{Z}^H$. The empirical forward recursion consequently recovers every z^π, and the evaluation and selection conclusions of Corollary C.10 hold. The $|\mathcal{Z}|^H$ factor appears only inside a logarithm; the algorithm still performs one forward pass and does not enumerate candidate sequences.

Proof. On the simultaneous cell-wise concentration event, uniform candidate-path coverage and Cauchy–Schwarz give, uniformly over π, z, h,

$$\sum_{s,a,b} d_h^{\pi,z}(s,a,b)\|\widehat{P}_h - P_h\|_1 \le c\sqrt{\frac{C_{\mathrm{cf}}^{\mathrm{cand}} S^2 AB\iota_{\mathrm{cand}}}{n}},$$

$$\sum_{s} p_h^{\pi,z}(s)\|\widehat{g}_h(\cdot \mid s, z_h) - g(\cdot \mid s, z_h)\|_1 \le c\sqrt{\frac{C_{\mathrm{sr}}^{\mathrm{cand}} SB\iota_{\mathrm{cand}}}{n}}.$$

For a fixed candidate rollout, stochastic-kernel contraction gives the same one-step occupancy recursion as Lemma C.3. Summing it through stage h bounds every candidate prefix perturbation by the left side of the margin condition. Assumption C.11 therefore makes the empirical and population maps agree coordinatewise on the whole candidate domain. Applying the common operator to the forward recursion gives exact target-path recovery. Corollary C.10 then applies on the recovered path. □

Uniform stability over every candidate path requires coverage and a margin on those paths. We call this the uniform candidate-path benchmark; it is a stronger comparison regime defined in this paper, not a standard OPE assumption. Theorem 4.3 needs those conditions only along logging-policy and target-reachable support.

C.7.1 Generic plug-in reduction and expanded benchmark proofs

The end-to-end theorem supplies one concrete way to construct the required response and model envelopes. The following reduction separates that construction from the pessimistic transfer argument, after which we give the full uniform candidate-path and oracle-response calculations.

Theorem C.13 (Plug-in pessimism under identified opponent responses)**.** *Assume that for each target policy $\pi \in \Pi$, the induced opponent policy μ^π and value $V_{\mathrm{cf}}(\pi) = V^{\pi,\mu^\pi}$ are well-defined. In settings satisfying Proposition B.4, these objects are identified on target-reachable support. Suppose that for each target policy $\pi \in \Pi$ the learner is given data-dependent plug-in objects $(\widehat{P}_h, \widehat{r}_h, \widehat{\mu}^\pi)$*

and nonnegative envelopes $\Gamma_h^\pi(s,a,b)$ and $\Lambda_h^\pi(s)$ such that, on an event of probability at least $1-\delta$, the model envelope holds simultaneously for all π, h, s, a, b, and the response envelope holds for every s with $p_h^{\pi,\mu^\pi}(s) > 0$,

$$\left|(\widehat{r}_h - r_h)(s,a,b) + [(\widehat{P}_h - P_h)\widehat{V}_{h+1}^\pi](s,a,b)\right| \le \Gamma_h^\pi(s,a,b), \tag{53}$$

$$H\,\|\widehat{\mu}_h^\pi(\cdot \mid s) - \mu_h^\pi(\cdot \mid s)\|_1 \le \Lambda_h^\pi(s). \tag{54}$$

Construct the pessimistic recursion

$$\widehat{V}_{H+1}^\pi(s) := 0, \tag{55}$$

$$\overline{Q}_h^\pi(s,a,b) := \left[\widehat{r}_h(s,a,b) + [\widehat{P}_h\widehat{V}_{h+1}^\pi](s,a,b) - \Gamma_h^\pi(s,a,b)\right]_0^{H-h+1}, \tag{56}$$

$$\widetilde{V}_h^\pi(s) := \sum_a \pi_h(a \mid s) \sum_b \widehat{\mu}_h^\pi(b \mid s)\,\overline{Q}_h^\pi(s,a,b) - \Lambda_h^\pi(s), \tag{57}$$

$$\widehat{V}_h^\pi(s) := [\widetilde{V}_h^\pi(s)]_0^{H-h+1}, \tag{58}$$

and let

$$\widehat{V}^-(\pi) := \mathbb{E}_{s\sim\rho}[\widehat{V}_1^\pi(s)], \qquad \widehat{\pi} \in \arg\max_{\pi\in\Pi} \widehat{V}^-(\pi).$$

Then, on the same high-probability event,

$$\begin{aligned} 0 &\le V_{\mathrm{cf}}(\pi) - \widehat{V}^-(\pi) \\ &\le 2\sum_{h=1}^H \mathbb{E}_{d_h^{\pi,\mu^\pi}}[\Gamma_h^\pi] \\ &\quad + 2\sum_{h=1}^H \mathbb{E}_{s\sim p_h^{\pi,\mu^\pi}}[\Lambda_h^\pi(s)], \qquad \forall \pi \in \Pi, \end{aligned} \tag{59}$$

and consequently

$$\begin{aligned} 0 &\le V_{\mathrm{cf}}(\pi^*) - V_{\mathrm{cf}}(\widehat{\pi}) \\ &\le 2\sup_{\pi\in\Pi}\left\{\sum_{h=1}^H \mathbb{E}_{d_h^{\pi,\mu^\pi}}[\Gamma_h^\pi] + \sum_{h=1}^H \mathbb{E}_{s\sim p_h^{\pi,\mu^\pi}}[\Lambda_h^\pi(s)]\right\}. \end{aligned} \tag{60}$$

Proof. Fix a target policy π and write $V_h^* = V_h^{\pi,\mu^\pi}$ and $Q_h^* = Q_h^{\pi,\mu^\pi}$. The proof proceeds in three steps: establish pessimism, derive the one-step recursion with the abstract envelopes, and then pass to policy optimization.

Pessimism. By backward induction on h, suppose $\widehat{V}_{h+1}^\pi \le V_{h+1}^*$. Then the model envelope (53) gives

$$\overline{Q}_h^\pi(s,a,b) \le Q_h^*(s,a,b)$$

for every (s,a,b). The response envelope (54) then gives

$$\widetilde{V}_h^\pi(s) \le \sum_a \pi_h(a \mid s) \sum_b \mu_h^\pi(b \mid s)\, Q_h^*(s,a,b) = V_h^*(s),$$

and clipping preserves the inequality $\widehat{V}_h^\pi(s) \le V_h^*(s)$. Hence $\widehat{V}^-(\pi) \le V_{\mathrm{cf}}(\pi)$.

One-sided error bound. Define $e_h(s) := V_h^*(s) - \widehat{V}_h^\pi(s)$. Keeping the envelopes explicit in the same one-step decomposition yields

$$\begin{aligned} e_h(s) \le\ & \sum_{a,b} \pi_h(a \mid s)\mu_h^\pi(b \mid s)[P_h e_{h+1}](s,a,b) \\ &+ 2\sum_{a,b} \pi_h(a \mid s)\mu_h^\pi(b \mid s)\Gamma_h^\pi(s,a,b) \\ &+ 2\Lambda_h^\pi(s). \end{aligned} \tag{61}$$

Taking expectations under the target occupancy and telescoping the $P_h e_{h+1}$ term gives eq. (59).

Policy optimization. Since $\widehat{\pi}$ maximizes $\widehat{V}^-$ and the estimator is pessimistic for every policy,

$$V_{\mathrm{cf}}(\pi^*) - V_{\mathrm{cf}}(\widehat{\pi}) \le \sup_{\pi\in\Pi}\big(V_{\mathrm{cf}}(\pi) - \widehat{V}^-(\pi)\big).$$

Combining this with eq. (59) proves eq. (60). The final displayed implication is immediate. □

Expanded proof of Theorem C.12. We prove exact recovery of the target response-state sequences and then invoke the oracle-state analysis.

Step 1: uniform concentration on candidate-relevant cells. Using the observed response-state labels, define the same tabular counts and estimators as in the proof of Corollary C.10. By Chernoff bounds together with Weissman's multinomial ℓ_1 inequality (Weissman et al., 2003), there is an event $\mathcal{E}_{\mathrm{cand}}$ with probability at least $1-\delta$ on which, simultaneously for every relevant step and cell,

$$N_h(s,a,b) \ge \frac{n}{2} d_h^{\log}(s,a,b), \qquad N_h(s,z) \ge \frac{n}{2} d_h^{\log}(s,z), \tag{62}$$

$$\|\widehat{P}_h(\cdot \mid s,a,b) - P_h(\cdot \mid s,a,b)\|_1 \le \Xi_h(s,a,b) := c\sqrt{\frac{S\iota_{\mathrm{cand}}}{N_h(s,a,b)\vee 1}}, \tag{63}$$

$$\|\widehat{g}_h(\cdot \mid s,z) - g(\cdot \mid s,z)\|_1 \le \Upsilon_h(s,z) := c\sqrt{\frac{B\iota_{\mathrm{cand}}}{N_h(s,z)\vee 1}}. \tag{64}$$

Because ι_{cand} already includes the union bound over $\Pi \times \mathcal{Z}^H$, these inequalities hold uniformly for all candidate evaluations that appear below.

Step 2: forward occupancy perturbation for candidate rollouts. Fix $\pi \in \Pi$ and $z \in \mathcal{Z}^H$. Let $d_h^{\pi,z}(s,a)$ denote the true learner-side occupancy under (π, μ^z, P), and define the empirical forward recursion

$$\begin{aligned} \widehat{d}_1^{\pi,z}(s,a) &= \rho(s)\,\pi_1(a \mid s), \\ \widehat{d}_{h+1}^{\pi,z}(s',a') &= \pi_{h+1}(a' \mid s') \sum_{s,a,b} \widehat{d}_h^{\pi,z}(s,a)\,\widehat{\mu}_h^z(b \mid s)\,\widehat{P}_h(s' \mid s,a,b). \end{aligned} \tag{65}$$

Write $\Delta_h^{\pi,z} := \|\widehat{d}_h^{\pi,z} - d_h^{\pi,z}\|_1$. Using the true occupancy as the reference measure and the fact that stochastic kernels contract total variation, we obtain

$$\begin{aligned} \Delta_{h+1}^{\pi,z} \le \Delta_h^{\pi,z} &\\ &+ \sum_s d_h^{\pi,z}(s)\,\|\widehat{\mu}_h^z(\cdot \mid s) - \mu_h^z(\cdot \mid s)\|_1 \\ &+ \sum_{s,a,b} d_h^{\pi,z}(s,a,b)\,\|\widehat{P}_h(\cdot \mid s,a,b) - P_h(\cdot \mid s,a,b)\|_1. \end{aligned} \tag{66}$$

On $\mathcal{E}_{\mathrm{cand}}$, uniform candidate-path coverage gives

$$\sum_s d_h^{\pi,z}(s)\,\Upsilon_h(s,z_h) \le c\sqrt{\frac{C_{\mathrm{sr}}^{\mathrm{cand}}\,SB\,\iota_{\mathrm{cand}}}{n}}, \tag{67}$$

$$\sum_{s,a,b} d_h^{\pi,z}(s,a,b)\,\Xi_h(s,a,b) \le c\sqrt{\frac{C_{\mathrm{cf}}^{\mathrm{cand}}\,S^2AB\,\iota_{\mathrm{cand}}}{n}}. \tag{68}$$

Indeed, the first inequality uses $d_h^{\pi,z}(s) \le C_{\mathrm{sr}}^{\mathrm{cand}} d_h^{\log}(s,z_h)$, eqs. (62) and (64), and $\sum_s \sqrt{d_h^{\pi,z}(s)} \le \sqrt{S}$; the second uses $d_h^{\pi,z}(s,a,b) \le C_{\mathrm{cf}}^{\mathrm{cand}} d_h^{\log}(s,a,b)$, eqs. (62) and (63), and $\sum_{s,a,b}\sqrt{d_h^{\pi,z}(s,a,b)} \le \sqrt{SAB}$. Starting from $\Delta_1^{\pi,z} = 0$ and iterating eqs. (66)–(68), we conclude

$$\Delta_h^{\pi,z} \le cH\left(\sqrt{\frac{C_{\mathrm{cf}}^{\mathrm{cand}}\,S^2AB\,\iota_{\mathrm{cand}}}{n}} + \sqrt{\frac{C_{\mathrm{sr}}^{\mathrm{cand}}\,SB\,\iota_{\mathrm{cand}}}{n}}\right) \tag{69}$$

for every π, z, h.

Step 3: exact recovery by causal recursion. Summing (69) over $j \le h$ yields

$$\|\widehat{d}^{\pi,z}_{1:h} - d^{\pi,z}_{1:h}\|_{1,\mathrm{pref}} \le cH^2 \left(\sqrt{\frac{C^{\mathrm{cand}}_{\mathrm{cf}}\, S^2AB\, \iota_{\mathrm{cand}}}{n}} + \sqrt{\frac{C^{\mathrm{cand}}_{\mathrm{sr}}\, SB\, \iota_{\mathrm{cand}}}{n}} \right). \tag{70}$$

By the assumed sample-size condition, the right-hand side is at most γ_ϕ. Assumption C.11 therefore implies

$$\phi_h(\widehat{d}^{\pi,z}_{1:h}) = \phi_h(d^{\pi,z}_{1:h}) \qquad \forall \pi \in \Pi,\ z \in \mathcal{Z}^H,\ h \in [H].$$

Hence $\widehat{T}_\pi(z) = T_\pi(z)$ for all $\pi \in \Pi$ and $z \in \mathcal{Z}^H$. In particular, causal forward recursion through the empirical operator agrees coordinate by coordinate with the population recursion, so $\widehat{z}^\pi = z^\pi$ for every π.

Step 4: reduction to the oracle-state corollaries. On the exact-recovery event,

$$\widehat{\mu}^\pi_h(\cdot \mid s) = \widehat{g}_h(\cdot \mid s, \widehat{z}^\pi_h) = \widehat{g}_h(\cdot \mid s, z^\pi_h),$$

so the pessimistic estimator is exactly the oracle-state estimator analyzed in Corollary C.10. Since the event $\mathcal{E}_{\mathrm{cand}}$ already supplies the concentration inequalities needed there, Corollary C.10 applies verbatim. This completes the proof. □

Expanded proof of Corollary C.10. Fix a target policy π and an arbitrary fixed fallback distribution $q_0 \in \Delta(\mathcal{S})$. Using the observed response-state labels, define the tabular counts $N_h(s,a,b)$, $N_h(s,a,b,s')$, $N_h(s,z)$, and $N_h(s,z,b)$, and let

$$\widehat{P}_h(s' \mid s,a,b) := \begin{cases} N_h(s,a,b,s')/N_h(s,a,b), & N_h(s,a,b) > 0, \\ q_0(s'), & N_h(s,a,b) = 0, \end{cases}$$

$$\widehat{r}_h(s,a,b) := \frac{\sum_i r^{(i)}_h \mathbf{1}\{(s^{(i)}_h, a^{(i)}_h, b^{(i)}_h) = (s,a,b)\}}{N_h(s,a,b) \vee 1},$$

$$\widehat{\mu}^\pi_h(\cdot \mid s) := \widehat{g}_h(\cdot \mid s, z^\pi_h),$$

where

$$\widehat{g}_h(b \mid s,z) := \begin{cases} N_h(s,z,b)/N_h(s,z), & N_h(s,z) > 0, \\ 1/B, & N_h(s,z) = 0. \end{cases}$$

Choose bonuses

$$\Gamma_h(s,a,b) := c_1 H \sqrt{\frac{S\iota}{N_h(s,a,b) \vee 1}}, \qquad \Lambda_h(s;\pi) := c_2 H \sqrt{\frac{B\iota}{N_h(s,z^\pi_h) \vee 1}}. \tag{71}$$

Chernoff and Hoeffding bounds, together with the multinomial ℓ_1 concentration result of Weissman et al. (2003), imply that there is an event $\mathcal{E}$ with probability at least $1-\delta$ on which the following hold simultaneously for all relevant cells:

$$N_h(s,a,b) \ge \frac{n}{2} d^{\log}_h(s,a,b), \qquad N_h(s,z) \ge \frac{n}{2} d^{\log}_h(s,z), \tag{72}$$

$$\left|(\widehat{r}_h - r_h)(s,a,b) + [(\widehat{P}_h - P_h)V](s,a,b)\right| \le \Gamma_h(s,a,b) \quad \forall V : \mathcal{S} \to [0,H], \tag{73}$$

$$\|\widehat{\mu}^\pi_h(\cdot \mid s) - \mu^\pi_h(\cdot \mid s)\|_1 \le \frac{\Lambda_h(s;\pi)}{H} \quad \forall s \text{ with } d^{\pi,\mu^\pi}_h(s) > 0. \tag{74}$$

Therefore the hypotheses of Theorem C.13 hold on $\mathcal{E}$ with

$$\Gamma^\pi_h = \Gamma_h, \qquad \Lambda^\pi_h(\cdot) = \Lambda_h(\cdot;\pi).$$

Applying Theorem C.13 yields

$$\begin{aligned} 0 &\le V_{\mathrm{cf}}(\pi) - \widehat{V}^-(\pi) \\ &\le 2 \sum_{h=1}^H \mathbb{E}_{d^{\pi,\mu^\pi}_h}[\Gamma_h] \\ &\quad + 2 \sum_{h=1}^H \mathbb{E}_{s \sim p^{\pi,\mu^\pi}_h}[\Lambda_h(s;\pi)]. \end{aligned} \tag{75}$$

It remains to convert the two sums into an explicit n-rate. By eq. (72),

$$\Gamma_h(s,a,b) \le cH\sqrt{\frac{S\iota}{n\, d_h^{\log}(s,a,b)}}, \qquad \Lambda_h(s;\pi) \le cH\sqrt{\frac{B\iota}{n\, d_h^{\log}(s,z_h^{\pi})}}.$$

Hence

$$\begin{aligned}\sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi,\mu^\pi}}[\Gamma_h] &\le cH\sqrt{\frac{S\iota}{n}} \sum_{h=1}^{H}\sum_{s,a,b} \frac{d_h^{\pi,\mu^\pi}(s,a,b)}{\sqrt{d_h^{\log}(s,a,b)}} \\ &\le cH^2\sqrt{\frac{C_{\mathrm{cf}}\, S^2AB\,\iota}{n}}, \qquad (76)\\ \sum_{h=1}^{H} \mathbb{E}_{s\sim p_h^{\pi,\mu^\pi}}[\Lambda_h(s;\pi)] &\le cH\sqrt{\frac{B\iota}{n}} \sum_{h=1}^{H}\sum_{s} \frac{d_h^{\pi,\mu^\pi}(s)}{\sqrt{d_h^{\log}(s,z_h^{\pi})}} \\ &\le cH^2\sqrt{\frac{C_{\mathrm{sr}}\, SB\,\iota}{n}}. \qquad (77)\end{aligned}$$

The first inequality uses $d_h^{\pi,\mu^\pi}(s,a,b) \le C_{\mathrm{cf}} d_h^{\log}(s,a,b)$ and $\sum_{s,a,b}\sqrt{d_h^{\pi,\mu^\pi}(s,a,b)} \le \sqrt{SAB}$; the second uses $d_h^{\pi,\mu^\pi}(s) \le C_{\mathrm{sr}} d_h^{\log}(s,z_h^{\pi})$ and $\sum_s \sqrt{d_h^{\pi,\mu^\pi}(s)} \le \sqrt{S}$. Combining eqs. (75)–(77) gives the decomposed rate in Corollary C.10. The displayed simplification follows from $SB \le S^2AB$ and $\bar{C} = \max\{C_{\mathrm{cf}}, C_{\mathrm{sr}}\}$. Because the event $\mathcal{E}$ holds uniformly over all target policies, the bound is uniform in π. □

Corollary C.14 (Pessimistic policy selection with supplied target response states)**.** *Under the hypotheses of Corollary C.10, let*

$$\widehat{\pi} \in \arg\max_{\pi\in\Pi} \widehat{V}^{-}(\pi).$$

Then, with probability at least $1-\delta$*,*

$$\begin{aligned}0 &\le V_{\mathrm{cf}}(\pi^*) - V_{\mathrm{cf}}(\widehat{\pi}) \\ &\le \widetilde{O}\left(H^2\sqrt{\frac{C_{\mathrm{cf}}\, S^2AB}{n}} + H^2\sqrt{\frac{C_{\mathrm{sr}}\, SB}{n}}\right) \qquad (78)\\ &\le \widetilde{O}\left(H^2\sqrt{\frac{\bar{C}\, S^2AB}{n}}\right),\end{aligned}$$

where $\pi^* \in \arg\max_{\pi\in\Pi} V_{\mathrm{cf}}(\pi)$*. Equivalently,*

$$n \gtrsim \frac{H^4\,\bar{C}\,S^2AB}{\varepsilon^2} + \frac{1}{\min\{\kappa_{\mathrm{model}}, \kappa_{\mathrm{resp}}\}} \qquad (79)$$

and, up to logarithms and universal constants, this suffices to guarantee $V_{\mathrm{cf}}(\pi^*) - V_{\mathrm{cf}}(\widehat{\pi}) \le \varepsilon$*.*

Proof. On the high-probability event of Corollary C.10,

$$\widehat{V}^{-}(\pi) \le V_{\mathrm{cf}}(\pi) \qquad \forall \pi \in \Pi.$$

Since $\widehat{\pi}$ maximizes $\widehat{V}^{-}$,

$$V_{\mathrm{cf}}(\pi^*) - V_{\mathrm{cf}}(\widehat{\pi}) \le V_{\mathrm{cf}}(\pi^*) - \widehat{V}^{-}(\pi^*) \le \sup_{\pi\in\Pi}\big(V_{\mathrm{cf}}(\pi) - \widehat{V}^{-}(\pi)\big).$$

Applying Corollary C.10 proves eq. (78), and solving for n gives eq. (79). □

Remark C.15 (Scope of the oracle-response corollary)**.** *Corollary C.10 is a* one-sided *lower-confidence instantiation of Theorem C.13. The additional coefficient* C_{sr} *is the quantitative finite-sample analogue of the joint state-response coverage condition used in Proposition B.4. The displayed rate first separates the model and response contributions and only then collapses them into the simpler* $\bar{C}$ *form. The result is conservative, but it is obtained from a fully explicit tabular-count argument with no pointwise pessimism gap left unresolved.*

C.8 Necessity of a local margin for exact labels

Proposition C.16 (Boundary instability)**.** *There is a two-step response-state model in which the target response label cannot be recovered with probability tending uniformly to one over instances whose occupancy lies arbitrarily close to a response boundary.*

Proof. Let z_1 be constant. From the initial state, the first opponent action moves the process to state s_1 with probability θ and to s_0 otherwise. Define $\phi_2(d_{1:2}) = \mathbf{1}\{p_2(s_1) \geq 1/2\}$. Compare two games with $\theta_- = 1/2 - \eta$ and $\theta_+ = 1/2 + \eta$; all learner policies and response maps are known and identical. Each logged episode supplies one Bernoulli transition observation. Distinguishing the two instances requires $n\eta^2$ to grow. Taking $\eta = c/\sqrt{n}$ keeps their n-sample total-variation distance bounded away from one, so every estimator of the target's second response label has constant error on one instance. Exact uniform recovery therefore requires a separation condition or an equivalent restriction. □

D Partial Identification Under Incomplete Support

When state-joint-action or state-response cells are missing, a point estimate may no longer be justified. Partial-identification methods use sets or intervals when overlap is insufficient for point identification (Khan et al., 2024). Here the adaptive opponent creates an additional requirement: the induced response must remain stable across compatible completions.

Fix a target policy π. Let $\mathcal{M}_{\log}$ be the set of completions that agree with the population logged law on every supported state-joint-action triple and state-response cell, and that use the same specified $(\mathcal{Z}, \phi_{1:H})$. We first consider a subset $\mathcal{M}_{\mathrm{stab}} \subseteq \mathcal{M}_{\log}$ on which the target response sequence z^π is common. This occurs when the response sequence is independently known, already identified, or protected by a margin against every completion-induced prefix perturbation. The restriction to a common response can couple unknown rows across stages, so $\mathcal{M}_{\mathrm{stab}}$ need not be rectangular.

Proposition D.1 (Partial-identification outer interval)**.** *Hold the common target response sequence fixed and let $\mathcal{R}_{\mathrm{stab}}$ be the stage-rectangular row-wise hull that fixes every logged primitive row, permits independent copies $g_h(\cdot \mid s, z)$ across stages, and allows the full reward range or probability simplex on each unsupported row. It contains every rollout law induced by $\mathcal{M}_{\mathrm{stab}}$, including those generated by a stationary opponent-response kernel. Define*

$$V^-(\pi) = \inf_{M \in \mathcal{R}_{\mathrm{stab}}} V^M(\pi), \qquad V^+(\pi) = \sup_{M \in \mathcal{R}_{\mathrm{stab}}} V^M(\pi).$$

Robust backward recursion (Iyengar, 2005; Nilim & El Ghaoui, 2005) computes these outer endpoints and, for every $M \in \mathcal{M}_{\mathrm{stab}}$,

$$V^-(\pi) \leq V^M_{\mathrm{cf}}(\pi) \leq V^+(\pi).$$

If q_h is the probability, under the common pre-hit law, that stage h is the first stage at which an unsupported primitive is required, then

$$V^+(\pi) - V^-(\pi) \leq \sum_{h=1}^{H} (H - h + 1) q_h. \tag{80}$$

If only rewards are unidentified and α_h is the target mass on unsupported reward triples at stage h, then

$$V^+(\pi) - V^-(\pi) \leq \sum_{h=1}^{H} \alpha_h. \tag{81}$$

The dependence in (81) *is sharp by Theorem B.2. The outer interval equals the sharp identified interval when $\mathcal{M}_{\mathrm{stab}}$ coincides with $\mathcal{R}_{\mathrm{stab}}$; equality is not claimed in general.*

Proof. After fixing z^π, the stage-indexed row choices in $\mathcal{R}_{\mathrm{stab}}$ are rectangular by construction. On a supported state-joint-action triple, the reward and transition law are fixed. On a supported (h, s, z^π_h) cell, the opponent-action law is fixed. Unsupported stage-indexed rows range independently over their natural sets. Standard robust backward induction therefore computes the infimum and supremum over $\mathcal{R}_{\mathrm{stab}}$. A stationary stable completion embeds in this relaxation by assigning the

same response-kernel row to every copy, which proves coverage of every value in $\mathcal{M}_{\mathrm{stab}}$ without assuming that the smaller stationary class is rectangular.

For the width, couple two arbitrary models in the rectangular hull to use the same randomness while their common supported primitives are used. Their trajectories agree until the first unsupported primitive is encountered. Conditional on the first encounter occurring at stage h, both remaining returns lie in $[0, H-h+1]$, so their difference is at most $H-h+1$. The first-hit events are disjoint; averaging proves (80). If only rewards differ, the trajectory law is common and the occupancy representation gives

$$|V^{M_1}(\pi) - V^{M_2}(\pi)| \le \sum_h \sum_{x \in U_h} d_h^\pi(x) = \sum_h \alpha_h,$$

which proves (81). □

Why uncovered mass is not enough without stability. Suppose an unsupported transition changes an occupancy prefix by an arbitrarily small amount, but the original prefix lies equally close to a threshold of ϕ_h. Two compatible completions can then assign different z_h, after which their opponent policies may differ on every later target state. The resulting value difference can be of order H even as the initial uncovered mass tends to zero. Thus Proposition D.1 cannot hold over all of $\mathcal{M}_{\log}$ without a common response, a local margin, or a direct induced-response discrepancy bound.

D.1 Finite-sample intervals

Let $\mathfrak{C}_n$ be the stage-rectangular outer confidence set obtained by placing two-sided confidence regions around every positive-count reward and transition row and every stage-indexed response row, and the full simplex or reward range on every zero-count row. Repeated copies of a stationary opponent-response kernel are allowed to vary independently in this relaxation. When target recovery succeeds, hold the recovered z^π fixed in the target rollout. This set can include nonstationary row combinations or combinations that do not jointly reproduce that label through $\phi_{1:H}$, which is harmless for coverage and may make the interval conservative.

Theorem D.2 (Finite-sample outer confidence interval)**.** *On the intersection of the simultaneous primitive-confidence event and the event that $\widehat{z}^\pi = z^\pi$ for every target, the robust dynamic programs over the rectangular outer set*

$$\widehat{V}^-_{\mathrm{rob}}(\pi) := \inf_{M \in \mathfrak{C}_n} V^M(\pi), \qquad \widehat{V}^+_{\mathrm{rob}}(\pi) := \sup_{M \in \mathfrak{C}_n} V^M(\pi)$$

satisfy

$$\widehat{V}^-_{\mathrm{rob}}(\pi) \le V_{\mathrm{cf}}(\pi) \le \widehat{V}^+_{\mathrm{rob}}(\pi)$$

simultaneously over the target class. The robust recursion computes their finite-sample width exactly for the stated outer set. With complete target support, that width contracts at the same order as Theorem 4.3. These are valid outer confidence intervals; exact equality with a nonrectangular confidence region is not asserted.

Proof. Exact recovery fixes the true target response sequence in the outer rollout. The true primitive rows belong to their confidence regions, so that rollout belongs to the outer set. The two inequalities follow immediately from the infimum and supremum definitions, and rectangularity permits backward optimization over primitive rows at each cell. Under full support, the same occupancy-weighted concentration and gap-unrolling calculation as Lemma C.4, applied to both signs, gives the stated contraction rate. For incomplete support, Proposition D.1 explains the population first-hit contribution; the finite-sample recursion itself is the precise width computation. □

An equivalent implementation mirrors (48) and (49) with plus signs and upper clipping. The robust-set definition is preferable for zero-count cells because it makes the coverage contribution and interval validity explicit.

E RESPONSE-MODEL MISSPECIFICATION

The following result measures misspecification at the induced-policy level. Errors in $\mathcal{Z}$, ϕ_h, or the opponent-response kernel are covered only when they imply such a discrepancy or when a margin ensures that the response labels do not change.

Proposition E.1 (Value stability under induced-response error)**.** *Fix a learner policy π, physical game (P, r, ρ), and two opponent policies μ and $\bar{\mu}$. With rewards in $[0, 1]$,*

$$|V^{\pi,\mu} - V^{\pi,\bar{\mu}}| \le \frac{1}{2} \sum_{h=1}^{H} (H - h + 1) \mathbb{E}_{s \sim p_h^{\pi,\mu}} \|\mu_h(\cdot \mid s) - \bar{\mu}_h(\cdot \mid s)\|_1. \tag{82}$$

If the ℓ_1 discrepancy is uniformly at most ε, then

$$|V^{\pi,\mu} - V^{\pi,\bar{\mu}}| \le B_\varepsilon := \min\left\{ H, \frac{H(H+1)}{4}\varepsilon \right\}.$$

If this bound holds uniformly over Π, a policy with suboptimality η for the modeled values has suboptimality at most $\eta + 2B_\varepsilon$ for the true values.

Proof. The finite-horizon performance-difference identity, changing only the opponent policy, gives

$$V^{\pi,\mu} - V^{\pi,\bar{\mu}} = \sum_{h=1}^{H} \mathbb{E}_{s \sim p_h^{\pi,\mu}, a \sim \pi_h} \sum_b [\mu_h - \bar{\mu}_h](b \mid s) Q_h^{\pi,\bar{\mu}}(s, a, b).$$

The span of $Q_h^{\pi,\bar{\mu}}(s, a, \cdot)$ is at most $H - h + 1$. For any signed vector u with zero sum,

$$\left| \sum_b u_b q_b \right| \le \frac{1}{2} \operatorname{span}(q) \|u\|_1.$$

This proves (82). Summing the remaining horizons gives $H(H+1)\varepsilon/4$, while the return range gives the additional cap at H. For selection, add and subtract the modeled value at the true and returned optimizers; each comparison contributes at most B_ε. □

Corollary E.2 (Misspecification-adjusted lower confidence bound)**.** *If $\widehat{V}^-(\pi)$ is a valid lower bound for the modeled value and the discrepancy in Proposition E.1 is at most B_ε, then $\widehat{V}^-(\pi) - B_\varepsilon$ is a valid lower bound for the true value.*

Proof. The claim follows from $V_{\mathrm{true}}(\pi) \ge V_{\mathrm{model}}(\pi) - B_\varepsilon \ge \widehat{V}^-(\pi) - B_\varepsilon$. □

Proposition E.3 (Recovery stability)**.** *Let $(\bar{P}, \bar{g})$ define the nominal model and let $\bar{d}_{1:H}^{\pi}$ and $\bar{p}_{1:H}^{\pi}$ be its target learner-occupancy and state-marginal paths. The true, nominal, and empirical paths use the same initial law ρ, target policy π, and specified maps $\phi_{1:H}$. Compute γ_{tar} around the nominal path. Suppose the true physical transition is $\bar{P}$ and, whenever the true and nominal labels agree through stage h, the true induced opponent row at each visited state is within ε in ℓ_1 of $\bar{g}(\cdot \mid s, \bar{z}_h^{\pi})$. Assume the empirical primitives satisfy, for every stage,*

$$\sum_{s,a,b} \bar{d}_h^{\pi}(s, a, b) \|\widehat{P}_h(\cdot \mid s, a, b) - \bar{P}_h(\cdot \mid s, a, b)\|_1 \le e_P,$$

$$\sum_s \bar{p}_h^{\pi}(s) \|\widehat{g}_h(\cdot \mid s, \bar{z}_h^{\pi}) - \bar{g}(\cdot \mid s, \bar{z}_h^{\pi})\|_1 \le e_g.$$

The true learner-occupancy prefix drift from the nominal path is at most $H^2\varepsilon/2$, while the empirical prefix error is at most $H^2(e_P + e_g)/2$. Exact target-label recovery remains valid whenever

$$\frac{H^2}{2}(e_P + e_g + \varepsilon) < \gamma_{\mathrm{tar}}.$$

Proof. Run two inductions around the nominal path. Agreement of the corresponding labels through stage h activates the discrepancy premise for the true path and the primitive-error bounds for the empirical path. Stochastic-kernel contraction gives

$$\|d_{h+1}^{\mathrm{true}} - \bar{d}_{h+1}\|_1 \le \|d_h^{\mathrm{true}} - \bar{d}_h\|_1 + \varepsilon,$$
$$\|\widehat{d}_{h+1}^{\pi} - \bar{d}_{h+1}^{\pi}\|_1 \le \|\widehat{d}_h^{\pi} - \bar{d}_h^{\pi}\|_1 + e_P + e_g.$$

Thus the two prefix errors are at most $H^2\varepsilon/2$ and $H^2(e_P + e_g)/2$. The combined condition puts both strictly inside the nominal margin, preserving the next labels and proving $z^{\mathrm{true}} = \bar{z} = \widehat{z}$. □

The premise of Proposition E.3 is stronger than a small parameter error. A small perturbation of a discontinuous map ϕ_h need not imply a small response-policy error. This is why the paper states robustness in terms of the induced response and treats the discrepancy bound as external information.

F Cross-Stage Response Estimation

The main theorem estimates a separate $\widehat{g}_h$ at each stage. Each is an estimator of the same stationary opponent-response kernel, so there is no model mismatch, but the proof requires same-stage state-response coverage. A simple randomized-stage construction aligns finite-sample support with the cross-stage population condition without invoking within-episode independence.

Before inspecting an episode, draw J_i uniformly from $[H]$, independently across episodes and of the trajectories. Retain only $(s_{J_i}^{(i)}, z_{J_i}^{(i)}, b_{J_i}^{(i)})$ for estimating g. Define

$$\bar{d}^{\log}(s,z) := \frac{1}{H}\sum_{j=1}^{H} d_j^{\log}(s,z), \qquad C_{\mathrm{sr}}^{\mathrm{cross}} := \sup_{\pi,h,s:p_h^{\pi}(s)>0} \frac{p_h^{\pi}(s)}{\bar{d}^{\log}(s, z_h^{\pi})}.$$

Proposition F.1 (Cross-stage response-kernel estimation)**.** *Define an oracle version of the retained observations using the population response label attached to each logging policy. These oracle triples are independent and satisfy*

$$\mathbb{P}(b_{J_i} = b \mid s_{J_i} = s, z_{J_i} = z) = g(b \mid s, z).$$

The empirical opponent-response kernel based on these observations satisfies the response-envelope and target-weighted error bounds of Lemma C.2 with C_{sr} replaced by $C_{\mathrm{sr}}^{\mathrm{cross}}$. On the logging-label recovery event, the implementable estimator formed from reconstructed labels equals this oracle estimator exactly. Consequently, Theorem 4.3 remains valid with this replacement and the corresponding minimum positive mass of $\bar{d}^{\log}$.

Proof. For each episode,

$$\mathbb{P}(s_{J_i} = s, z_{J_i} = z, b_{J_i} = b)$$
$$= \frac{1}{H}\sum_{j=1}^{H} d_j^{\log}(s,z) g(b \mid s,z) = \bar{d}^{\log}(s,z) g(b \mid s,z).$$

Stationarity permits factoring out g. The selected triples are independent across episodes, so ordinary multinomial concentration applies. Repeating the Cauchy-Schwarz argument in (46) with denominator $\bar{d}^{\log}$ proves the target-weighted bound. The recovery and evaluation proofs then proceed unchanged. □

This construction may discard useful visits and can lose a factor of at most H relative to an ideally pooled estimator. Using all visits requires a trajectory-martingale or episode-block analysis. We do not claim a sharper all-visit rate without that additional proof.

G Lower Bounds and Rate Comparisons

The negative results in Section B concern identification and support. We next isolate finite-sample costs within an identified class.

Policy regret also evaluates a comparator against the response it induces (Arora et al., 2012; Dekel et al., 2014), and online Markov-game methods learn against stationary or adaptive opponents through continued interaction (Shapley, 1953; Littman, 1994; Bai & Jin, 2020; Xie et al., 2020; Liu et al., 2022; Daskalakis et al., 2021; Zhang et al., 2021; Nguyen-Tang & Arora, 2024; 2025). Our lower bounds ask a different information question: after interaction has ended, how much fixed logged data is needed to identify the induced response and select among undeployed policies? This is why state-joint-action coverage and response-state cardinality appear as separate costs.

Theorem G.1 (Response-state cardinality lower bound)**.** *For every $H \geq 2$, there is a family satisfying Definition 3.1 with three game states, finite coverage coefficients, and $d_R = |\mathcal{Z}|$ target policies such that, for $n \gtrsim d_R$, every uniform value estimator has expected error*

$$\sup_{G,\pi} \mathbb{E}_G |\widehat{V}(\pi) - V_{\mathrm{cf}}^G(\pi)| \geq c(H-1)\sqrt{\frac{d_R}{n}}.$$

Proof. Let the learner have two actions and choose d_R distinct first-stage action probabilities in $[1/4, 3/4]$. A known quantization map ϕ_1 assigns these policies to the d_R response states, and the logging policy set chooses them uniformly. For every $h \geq 2$, set $\phi_h(d_{1:h}) = \phi_1(d_1)$, so the response label persists. The process starts in a common state with zero reward. Conditional on z, the first opponent action is Bernoulli with mean $\theta_z \in [1/4, 3/4]$ and moves the process to a good or bad absorbing state. At later stages, give both players singleton, outcome-irrelevant actions and define the remaining response-kernel rows deterministically. Stages 2 through H give reward one in the good state and zero in the bad state. A target associated with z therefore has value exactly $(H-1)\theta_z$. The response sequence is determined by the known policy and response maps, all state-joint-action occupancies have constant coverage, and the target state-response coefficient is of order d_R.

Fix one response state z_0. Under the two games G_- and G_+, set $\theta_{z_0} = 1/2 \mp \epsilon$ and leave every other row unchanged. The KL divergence between the two laws of one logged episode is $d_R^{-1}\mathrm{kl}(1/2 - \epsilon, 1/2 + \epsilon) \leq c_0\epsilon^2/d_R$. Tensorization gives $D_{\mathrm{KL}}(G_-^n \| G_+^n) \leq c_0 n\epsilon^2/d_R$. Choose $\epsilon = c_1\sqrt{d_R/n}$ with c_1 small enough that Pinsker's inequality keeps total variation below $1/2$. Le Cam's lemma then gives expected error at least $(H-1)\epsilon/4$ in one of the two games for the target associated with z_0. □

Theorem G.2 (Joint-occupancy coverage lower bound)**.** *For every $H \geq 2$ and feasible coverage level $1 \leq C \leq n$, there is a covered subclass satisfying Definition 3.1 on which value estimation has minimax error at least*

$$c(H-1)\sqrt{\frac{C}{n}}$$

up to the natural cap at $H-1$.

Proof. Set $|\mathcal{Z}| = 1$ and let the opponent have one deterministic action, so response estimation is vacuous. The learner starts in state s_0 and has an informative action a_1 and a safe action a_0. The target always selects a_1, while the logging policy selects it with probability $1/C$. Under parameter $\theta \in [1/4, 3/4]$, action a_1 moves to a good absorbing state with probability θ and a bad absorbing state otherwise. Action a_0 moves to the bad state. The first reward is zero; at stages 2 through H, the good state gives reward one and the bad state gives reward zero. Thus the target value is $(H-1)\theta$, and the maximum target-to-logging occupancy ratio is exactly C.

Compare $\theta_- = 1/2 - \epsilon$ and $\theta_+ = 1/2 + \epsilon$. A logged episode carries information about the parameter only when a_1 is selected, so its KL divergence is

$$D_{\mathrm{KL}}(P_- \| P_+) = \frac{1}{C}\mathrm{kl}(\theta_-, \theta_+) \leq \frac{c_0\epsilon^2}{C}.$$

After n episodes the KL divergence is at most $c_0 n\epsilon^2/C$. Taking $\epsilon = c_1 \min\{1, \sqrt{C/n}\}$ with a sufficiently small universal c_1, Pinsker's inequality gives total variation at most $1/2$. The two target

values differ by $2(H-1)\epsilon$. Le Cam's two-point lemma therefore gives minimax absolute error at least $c_2(H-1)\epsilon$, which is the claimed bound after adjusting constants. □

The two lower bounds separate independent obstacles. The joint-occupancy construction has no adaptive response, while the response construction keeps the game state space fixed. We next show that both obstacles persist for policy selection, not only uniform value estimation.

Theorem G.3 (Joint-occupancy coverage lower bound for policy selection)**.** *For every $C_{\mathrm{cov}} \geq 1$ and integer $m \geq 1$, there is a subclass satisfying Definition 3.1 with $m+2$ states, $|\mathcal{Z}| = 1$, and $C_{\mathrm{cf}} \leq 2C_{\mathrm{cov}}$ such that every offline selector obeys*

$$\inf_{\widehat{\pi}} \sup_{G} \mathbb{E}_G[V_{\mathrm{cf}}(\pi^*) - V_{\mathrm{cf}}(\widehat{\pi})] \geq c(H-1)\min\left\{1, \sqrt{\frac{C_{\mathrm{cov}} m}{n}}\right\}.$$

Proof. Let the initial state be uniform on $s_1, \ldots, s_m$, use learner actions a_+, a_-, a_0, and give the opponent one action. For a sign vector $\theta \in \{-1,+1\}^m$, action a_+ at s_j reaches a good absorbing tail with probability $1/2 + \epsilon\theta_j$, while a_- reaches it with probability $1/2 - \epsilon\theta_j$; a_0 reaches a zero-reward tail. The good tail pays one at stages $2, \ldots, H$. The target class contains every deterministic sign policy, and the optimum chooses the action indexed by θ_j at each initial state. Logging chooses a_+ and a_- with probability $1/(2C_{\mathrm{cov}})$ each and a_0 otherwise. Taking a singleton response state places the family inside our model and gives $C_{\mathrm{cf}} \leq 2C_{\mathrm{cov}}$.

For a candidate sign vector u,

$$V_\theta(\pi^\theta) - V_\theta(\pi^u) = \frac{2(H-1)\epsilon}{m}\mathrm{Ham}(u,\theta).$$

Flipping coordinate j changes the data law only when the episode starts at s_j and logging selects an informative action, so the n-sample KL divergence is at most $16n\epsilon^2/(mC_{\mathrm{cov}})$. With $\epsilon = \min\{1/4, c_1\sqrt{C_{\mathrm{cov}} m/n}\}$, Assouad's lemma gives expected Hamming error at least $c_2 m$ for one model (Tsybakov, 2009, Theorem 2.12). Substitution proves the claim. □

Theorem G.4 (Response-state cardinality lower bound for policy selection)**.** *For every $d_R = |\mathcal{Z}| \geq 2$, there is a subclass satisfying Definition 3.1 with $C_{\mathrm{cf}} \leq 2$ such that every offline selector obeys*

$$\inf_{\widehat{\pi}} \sup_{G} \mathbb{E}_G[V_{\mathrm{cf}}(\pi^*) - V_{\mathrm{cf}}(\widehat{\pi})] \geq c(H-1)\min\left\{1, \sqrt{\frac{d_R}{n}}\right\}.$$

Proof. Use policies $\pi_1, \ldots, \pi_{d_R}$ whose first-stage probability of action a_1 is a distinct value $p_j \in [7/16, 9/16]$. A known quantizer maps a small interval around p_j to response state z_j at every stage, so the induced response sequence of π_j is the constant sequence z_j. In model k, the first opponent action is Bernoulli with mean $1/2+\epsilon$ under z_k and $1/2$ under every other response state. Action one moves to a good absorbing tail worth $H-1$, while action zero moves to a zero tail. Logging is uniform over the d_R policies. The learner action probabilities are bounded away from zero and one, and $\epsilon \leq 1/16$, so direct comparison of the four first-stage cells and the two tail occupancies gives $C_{\mathrm{cf}} \leq 2$.

Policy π_k is uniquely optimal and every wrong selection loses $(H-1)\epsilon$. Two models differ only on episodes labeled by one of two policies, which gives pairwise n-sample KL divergence at most $cn\epsilon^2/d_R$. Choose $\epsilon = \min\{1/16, c_1\sqrt{d_R/n}\}$. For $d_R = 2$, Le Cam's two-point method gives a constant probability of error. For $d_R \geq 3$, Fano's inequality gives the same conclusion on one model (Tsybakov, 2009, Theorem 2.5). The value separation proves the result. □

G.1 Expanded information-theoretic derivations

The short proofs above display the hard families. Here we retain the full reductions to Hamming and finite-hypothesis testing, including the cell-wise divergence calculations. For the response family with $d_R = 2$, the two-point argument in the concise proof applies; the expanded Fano calculation below treats $d_R \geq 3$.

Expanded proof of Theorem G.3. We use a standard hypercube construction together with Assouad's lemma.

Lemma G.5 (Assouad in Hamming form)**.** *Let $\Theta = \{-1, +1\}^m$, and let $\{P_\theta^{(n)} : \theta \in \Theta\}$ be a family of data laws. If for every coordinate $j \in [m]$ and every pair $\theta, \theta^{(j)}$ differing only in coordinate j,*

$$\mathrm{KL}\Big(P_\theta^{(n)} \,\|\, P_{\theta^{(j)}}^{(n)}\Big) \le \alpha,$$

then any estimator $\widehat{\theta} \in \{-1, +1\}^m$ satisfies

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta\Big[\mathrm{Ham}(\widehat{\theta}, \theta)\Big] \ge \frac{m}{2}\Big(1 - \sqrt{\alpha/2}\Big).$$

Proof. This is the standard Assouad bound for the Hamming loss; see, e.g., Tsybakov (2009, Theorem 2.12). We use only the displayed corollary. □

The hard family. Let $\Theta = \{-1, +1\}^m$ and let $\theta = (\theta_1, \dots, \theta_m)$. For each θ, define a horizon-H game with:

- initial state set $\{s_1, \dots, s_m\}$ and $\rho(s_j) = 1/m$;
- learner actions $\{a_+, a_-, a_0\}$;
- a single opponent action b_0;
- two absorbing tail states g and z, where g yields reward 1 at every remaining step and z yields reward 0.

To place this family inside the finite response-state model class, take $\mathcal{Z} = \{z_1\}$, define $\phi_h \equiv z_1$ for all h, and let $g(\cdot \mid s, z_1)$ be the degenerate distribution on b_0. At stage 1, if the initial state is s_j:

$$P_\theta(g \mid s_j, a_+, b_0) = \tfrac{1}{2} + \varepsilon\theta_j, \qquad P_\theta(g \mid s_j, a_-, b_0) = \tfrac{1}{2} - \varepsilon\theta_j, \qquad (83)$$
$$P_\theta(g \mid s_j, a_0, b_0) = 0.$$

The remaining transition probability goes to z. All rewards are 0 at stage 1; from stages $2, \dots, H$, state g gives reward 1 and state z gives reward 0.

The policy class contains all deterministic maps $u = (u_1, \dots, u_m) \in \{-1, +1\}^m$, where policy π^u chooses a_+ at state s_j if $u_j = +1$ and a_- if $u_j = -1$. The optimal policy is π^θ.

The logging policy chooses a_+ and a_- each with probability $1/(2C_{\mathrm{cov}})$ at every initial state and chooses the dummy action a_0 with probability $1 - 1/C_{\mathrm{cov}}$.

Coverage calibration. For the target policy π^θ, each visited triple has logging probability at least $1/(2C_{\mathrm{cov}}m)$. Hence

$$C_{\mathrm{cf}} \le 2C_{\mathrm{cov}} \qquad (84)$$

for every game in this family.

Value separation. If policy π^u is deployed in model θ, then

$$V_\theta(\pi^\theta) - V_\theta(\pi^u) = \frac{2(H-1)}{m}\,\varepsilon\,\mathrm{Ham}(u, \theta). \qquad (85)$$

Indeed, at each state s_j, choosing the wrong informative action changes the probability of entering the good tail by exactly 2ε, and the good tail contributes $H - 1$ future rewards.

KL bound. Fix a coordinate j and let $\theta^{(j)}$ denote the sign vector obtained by flipping θ_j. The two models differ only when the episode starts in state s_j and the logging policy chooses one of the informative actions. This occurs with probability $1/(mC_{\mathrm{cov}})$. Conditional on that event, the data law differs only in one Bernoulli transition parameter, from $1/2 + \varepsilon$ to $1/2 - \varepsilon$, whose KL divergence is at most $16\varepsilon^2$ for $\varepsilon \le 1/4$. Therefore

$$\mathrm{KL}\Big(P_\theta^{(n)} \,\|\, P_{\theta^{(j)}}^{(n)}\Big) \le \frac{16n\varepsilon^2}{mC_{\mathrm{cov}}}. \qquad (86)$$

Reduction to policy optimization. Given any offline policy-selection algorithm $\widehat{\pi}$, define $\widehat{\theta} \in \{-1,+1\}^m$ by the actions chosen by $\widehat{\pi}$ on the states $s_1,\dots,s_m$. Then (85) yields

$$V_\theta(\pi^\theta) - V_\theta(\widehat{\pi}) = \frac{2(H-1)}{m}\,\varepsilon\,\mathrm{Ham}(\widehat{\theta},\theta).$$

Choose

$$\varepsilon := \min\left\{\frac{1}{4},\ c_1\sqrt{\frac{C_{\mathrm{cov}}m}{n}}\right\}$$

with $c_1 > 0$ small enough so that the KL bound in (86) is at most a universal constant below 2. Lemma G.5 then gives

$$\sup_{\theta\in\Theta}\mathbb{E}_\theta\Big[\mathrm{Ham}(\widehat{\theta},\theta)\Big] \ge c_2 m$$

for a universal constant $c_2 > 0$. Consequently,

$$\inf_{\widehat{\pi}}\sup_{\theta\in\Theta}\mathbb{E}_\theta\big[V_\theta(\pi^\theta) - V_\theta(\widehat{\pi})\big] \ge c_3(H-1)\min\left\{1,\sqrt{\frac{C_{\mathrm{cov}}m}{n}}\right\}, \tag{87}$$

which proves Theorem G.3. □

Expanded proof of Theorem G.4 for $d_R \ge 3$. We construct a hard subclass in which the transition model is fixed and the difficulty lies entirely in identifying the best response state.

Lemma G.6 (Fano for finite hypotheses)**.** *Let $\{P_j^{(n)}\}_{j=1}^M$ be $M \ge 2$ data laws. If*

$$\max_{j\neq k}\mathrm{KL}\Big(P_j^{(n)} \,\|\, P_k^{(n)}\Big) \le \alpha\log M$$

for some $\alpha < 1/8$, then any estimator $\widehat{J}$ of the index $J \in [M]$ satisfies

$$\sup_{j\in[M]}\Pr_j(\widehat{J}\neq j) \ge c_4$$

for a universal constant $c_4 > 0$.

Proof. This is the standard Fano inequality; see, e.g., Tsybakov (2009, Theorem 2.5). □

The hard family. Fix one nonabsorbing physical state s_0, learner actions $\{a_0,a_1\}$, opponent actions $\{0,1\}$, and absorbing tail states g and z. For each $j \in [d_R]$, let

$$p_j := \frac{1}{2} + \frac{2j - d_R - 1}{16 d_R}.$$

Then the p_j are distinct, lie in $[7/16, 9/16]$, and satisfy

$$\frac{1}{d_R}\sum_{j=1}^{d_R} p_j = \frac{1}{2}.$$

The construction assigns each target policy a distinct first-step occupancy signature and therefore a distinct response state. The hard instance hides which response state receives the slight reward advantage, so identifying the best policy amounts to identifying the advantaged response state from offline data.

Let $\Pi = \{\pi_1,\dots,\pi_{d_R}\}$, where π_j plays a_1 at stage 1 with probability p_j and a_0 with probability $1-p_j$; for stages $2,\dots,H$, all policies play the common action a_0. Let $\mathcal{Z} = \{z_1,\dots,z_{d_R}\}$ and define disjoint intervals

$$I_j := \left[p_j - \frac{1}{32 d_R}, p_j + \frac{1}{32 d_R}\right), \qquad j \in [d_R].$$

For every $h \in [H]$, define the known functional

$$\phi_h(d_{1:h}) := \begin{cases} z_j, & d_1(s_0, a_1) \in I_j \text{ for some } j \in [d_R], \\ z_1, & \text{otherwise.} \end{cases}$$

For each distinguished index $j^\star \in [d_R]$, define the response model

$$g_{j^\star}(1 \mid s_0, z_j) = \begin{cases} \frac{1}{2} + \varepsilon, & j = j^\star, \\ \frac{1}{2}, & j \neq j^\star, \end{cases} \qquad g_{j^\star}(0 \mid s_0, z_j) = 1 - g_{j^\star}(1 \mid s_0, z_j). \tag{88}$$

Extend $g_{j^\star}$ arbitrarily on the absorbing states g and z, where the opponent action has no effect on rewards or transitions. At stage 1 the reward is 0; if the opponent action is 1, the process enters the good absorbing tail g and receives reward 1 on stages $2, \ldots, H$, while if the opponent action is 0, it enters the zero tail z and receives reward 0 thereafter. The logging distribution is uniform over Π.

This family satisfies the finite response-state model in Definition 3.1. Indeed, for every policy π_j and every candidate response-state sequence u, the first-step learner-side occupancy satisfies

$$d_1^{\pi_j, u}(s_0, a_1) = p_j,$$

so $\phi_h(d_{1:h}^{\pi_j, u}) = z_j$ for every h. Hence $T_{\pi_j}(u)$ is the constant sequence $(z_j, \ldots, z_j)$, and $z^{\pi_j} = (z_j, \ldots, z_j)$ is the unique fixed point.

Moreover, for $\varepsilon \leq 1/16$, the counterfactual concentrability is bounded by an absolute constant. At stage 1, every target triple has probability at most

$$\frac{9}{16}\Big(\frac{1}{2} + \varepsilon\Big) \leq \frac{81}{256},$$

while under uniform logging,

$$d_1^{\log}(s_0, a_1, 1) = \frac{1}{4} + \frac{\varepsilon p_{j^\star}}{d_R}, \tag{89}$$

$$d_1^{\log}(s_0, a_1, 0) = \frac{1}{4} - \frac{\varepsilon p_{j^\star}}{d_R}, \tag{90}$$

$$d_1^{\log}(s_0, a_0, 1) = \frac{1}{4} + \frac{\varepsilon(1 - p_{j^\star})}{d_R}, \tag{91}$$

$$d_1^{\log}(s_0, a_0, 0) = \frac{1}{4} - \frac{\varepsilon(1 - p_{j^\star})}{d_R}. \tag{92}$$

Because $p_{j^\star} \in [7/16, 9/16]$, $d_R \geq 2$, and $\varepsilon \leq 1/16$, every positive stage-1 logging cell is at least $119/512$. For $h \geq 2$, the process is already in one of the absorbing tails and both players use common actions there. The target mass on the good tail is at most $1/2 + \varepsilon$, while under uniform logging it is $1/2 + \varepsilon/d_R$; similarly, the target mass on the zero tail is at most $1/2$, while under uniform logging it is at least $1/2 - \varepsilon/d_R$. Hence the ratio between target and logging masses is at most $9/8$. Therefore

$$C_{\mathrm{cf}} \leq 2. \tag{93}$$

Value separation. Write

$$q_j := g_{j^\star}(1 \mid s_0, z_j) = \frac{1}{2} + \varepsilon \mathbf{1}\{j = j^\star\}.$$

Since the learner action affects the value only through the induced response state, we have

$$V_{j^\star}(\pi_j) = (H - 1) q_j.$$

For the distinguished index $j^\star$,

$$V_{j^\star}(\pi_{j^\star}) - V_{j^\star}(\pi_j) = (H - 1)\varepsilon, \qquad \forall j \neq j^\star. \tag{94}$$

Hence a policy-selection error immediately incurs regret $(H - 1)\varepsilon$.

KL bound. For two indices $j^\star \neq k^\star$, the models differ only on episodes whose logged policy label is $\pi_{j^\star}$ or $\pi_{k^\star}$, each occurring with probability $1/d_R$ under the uniform logging distribution. Conditional

on the logged label, the learner-action distribution is identical in the two models; only the Bernoulli opponent-action parameter changes from $1/2 + \varepsilon$ to $1/2$. Therefore

$$\max_{j^\star \neq k^\star} \mathrm{KL}\Big(P_{j^\star}^{(n)} \,\|\, P_{k^\star}^{(n)}\Big) \leq c_5 \frac{n\varepsilon^2}{d_R} \tag{95}$$

for a universal constant $c_5 > 0$.

Choose

$$\varepsilon := \min\left\{\frac{1}{16},\ c_6\sqrt{\frac{d_R}{n}}\right\}$$

with $c_6 > 0$ small enough so that the right-hand side of eq. (95) is at most $\frac{1}{16}\log d_R$ and $\varepsilon \leq 1/16$. Lemma G.6 then implies that every algorithm misidentifies the best policy with probability at least c_4 on some instance. Combining this with (94) gives

$$\inf_{\widehat{\pi}} \sup_{j^\star \in [d_R]} \mathbb{E}_{j^\star}[V_{j^\star}(\pi_{j^\star}) - V_{j^\star}(\widehat{\pi})] \geq c_7(H-1)\min\left\{1, \sqrt{\frac{d_R}{n}}\right\}, \tag{96}$$

which proves Theorem G.4. □

Theorem G.7 (Embedding of fixed-opponent offline lower bounds)**.** *Every information-theoretic lower bound for finite-horizon episodic offline MDP policy optimization under the same policy-labeled i.i.d. logging protocol, bounded-reward model, and compatible policy class transfers to the present class by taking $|\mathcal{Z}| = 1$ and a deterministic opponent action. The logged laws, target values, and policy regrets are unchanged.*

Proof. Introduce one response state z_0, set every ϕ_h equal to z_0, and let $g(\cdot \mid s, z_0)$ select a fixed opponent action. On that action, copy the MDP transition and reward kernels. Marginalizing the deterministic opponent reproduces the original MDP experiment exactly. □

These lower bounds establish separate dependence on joint-occupancy coverage and response-state cardinality. Their horizon scaling is H, leaving a $\sqrt{H}$ gap to the conditional variance-aware upper bound below. The singleton-response embedding does not create a $\sqrt{B}$ game difficulty, so it cannot by itself close that gap. We do not claim a matched converse without a direct construction.

G.2 Synthesis of the full theory

Proposition G.8 (Synthesis of the theory)**.** *For off-policy evaluation and policy selection in Markov games with adaptive opponents:*

(i) *without observable response structure, the induced-response value is not identified; missing state-joint-action or state-response support yields a lower bound proportional to target mass on an unsupported row;*
(ii) *under simultaneous-action timing, the finite response-state operator is triangular and its unique response sequence is obtained by a forward recursion rather than an assumed or searched-for fixed point;*
(iii) *a finite positive-mass logging policy set, target state-joint-action support, and state-response support identify the reachable target trajectory;*
(iv) *under local logging and target margins, policy-factorized occupancy estimation reconstructs unobserved logged response labels and recovers each target response on its true path;*
(v) *conditional on recovery, the generic pessimistic recursion gives simultaneous one-sided lower confidence bounds and valid policy selection;*
(vi) *incomplete support, response-model discrepancy, or a failed margin test changes the output to a valid outer interval, a discrepancy-adjusted lower confidence bound, or abstention rather than an unjustified point estimate;*
(vii) *even on covered structured subclasses, joint-occupancy coverage and response-state cardinality yield the separate capped lower bounds in Theorems G.3 and G.4.*

Proof. The first item follows from Theorems B.1, B.2, and B.3. The causal lemma and its forward proof give item (ii). The identification argument on target-reachable support and the finite-policy-set correction give item (iii). Lemmas C.1, C.2, and C.3 give item (iv). The generic plug-in theorem and Lemma C.4 give item (v). The outer-interval and misspecification results in Appendices D and E give item (vi). The two explicit Assouad and Fano constructions prove item (vii). □

Remark G.9 (Remaining theoretical gaps)**.** *Causal target-path recovery resolves the computational question about exhaustive response search. The remaining gaps concern sharper direct lower bounds for conditional variance rates, fully specified function approximation, and broader response models with continuous latent states, history dependence, or same-step feedback.*

H Extensions and Further Consequences

H.1 A conditional variance-aware refinement

The count-based theorem uses range envelopes, which contribute one factor of H per stage. A Bernstein improvement is valid whenever the estimator supplies simultaneous, potentially data-dependent envelopes of the form

$$\Gamma_h^{\mathrm{B}}(s,a,b) \le c\left[\sqrt{\frac{\sigma_{P,h}^2(s,a,b)\iota}{N_h(s,a,b)\vee 1}} + \frac{H\iota}{N_h(s,a,b)\vee 1}\right],$$

$$\Lambda_h^{\mathrm{B}}(s;\pi) \le c\left[\sqrt{\frac{\sigma_{g,h}^2(s,z_h^\pi)\iota}{N_h(s,z_h^\pi)\vee 1}} + \frac{H\iota}{N_h(s,z_h^\pi)\vee 1}\right].$$

The statement below is deliberately conditional on valid envelopes. It does not treat an empirical Bernstein inequality for a value function computed from the same samples as if the value function were fixed.

Proposition H.1 (Variance-aware transfer)**.** *Suppose the hypotheses and recovery event of Theorem 4.3 hold and the displayed Bernstein envelopes are valid simultaneously. If*

$$\sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi,\mu^\pi}} \sigma_{P,h}^2 \le H^2, \qquad \sum_{h=1}^{H} \mathbb{E}_{p_h^\pi} \sigma_{g,h}^2 \le H^2$$

uniformly over Π*, then the evaluation and selection gaps are*

$$\widetilde{\mathcal{O}}\left(H^{3/2}\sqrt{\frac{C_{\mathrm{cf}}S^2AB}{n}} + H^{3/2}\sqrt{\frac{C_{\mathrm{sr}}SB}{n}} + \frac{H^2C_{\mathrm{cf}}SAB}{n} + \frac{H^2C_{\mathrm{sr}}S}{n}\right)$$

up to logarithmic factors already absorbed by $\widetilde{\mathcal{O}}$*.*

Proof. Pessimistic transfer is identical to Lemma C.4. For the transition term, use coverage and Cauchy-Schwarz first across cells and then across stages:

$$\begin{aligned}\sum_{h=1}^{H} \mathbb{E}_{d_h}\sqrt{\frac{\sigma_{P,h}^2\iota}{N_h}} &\le c\sqrt{\frac{C_{\mathrm{cf}}S^2AB\iota}{n}}\sum_{h=1}^{H}\sqrt{\mathbb{E}_{d_h}\sigma_{P,h}^2} \\ &\le cH^{3/2}\sqrt{\frac{C_{\mathrm{cf}}S^2AB\iota}{n}}.\end{aligned}$$

The response square-root term is identical. Coverage converts the transition inverse-count remainders to $\widetilde{\mathcal{O}}(H^2C_{\mathrm{cf}}SAB/n)$ and the response remainders to $\widetilde{\mathcal{O}}(H^2C_{\mathrm{sr}}S/n)$, which gives the two explicit lower-order terms. □

Obtaining such envelopes without sample splitting or a uniform value-function argument is a separate estimator-design question. The main theorem does not rely on Proposition H.1.

H.2 Conditional function approximation

The structural results do not require tabularity. If a function-approximation method returns valid model and response envelopes on target-reachable support, the same pessimistic reduction applies after a target response is known or recovered.

Proposition H.2 (Envelope-based function approximation)**.** *Suppose an estimator returns* $(\widehat{P}, \widehat{r}, \widehat{\mu}^\pi)$ *and nonnegative envelopes* $(\Gamma_h^\pi, \Lambda_h^\pi)$ *satisfying, simultaneously,*

$$|(\widehat{r}_h - r_h) + (\widehat{P}_h - P_h)\widehat{V}_{h+1}^\pi|(s,a,b) \le \Gamma_h^\pi(s,a,b),$$
$$H\|\widehat{\mu}_h^\pi(\cdot \mid s) - \mu_h^\pi(\cdot \mid s)\|_1 \le \Lambda_h^\pi(s).$$

Then the pessimistic recursion satisfies the occupancy-weighted conclusion of Lemma C.4. This includes linear-MDP estimators with elliptical confidence radii, conditional on supplying or separately recovering z^π*.*

Proof. The backward pessimism and gap-unrolling proof uses only the two displayed envelopes and bounded values, not the tabular construction of those envelopes. □

The current forward-recovery rate is tabular because Lemma C.2 controls occupancy propagation by tabular conditional-distribution error. Establishing an analogous target-path perturbation theorem for general function approximation remains open.

H.3 A dimensionally valid linear specialization

Let $\varphi : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}^{d_P}$ and $\psi : \mathcal{S} \times \mathcal{Z} \to \mathbb{R}^{d_g}$ be known bounded features, and consider

$$P_h(s' \mid s,a,b) = \langle \varphi(s,a,b), \theta_h(s') \rangle, \qquad r_h(s,a,b) = \langle \varphi(s,a,b), \xi_h \rangle,$$
$$g(b \mid s,z) = \langle \psi(s,z), \omega_b \rangle, \qquad \Omega = [\omega_1, \dots, \omega_B] \in \mathbb{R}^{d_g \times B},$$

with parameter sets restricted so that the displayed rows are probability distributions. The matrix Ω, rather than one scalar-output coefficient vector, is required for a B-class response distribution. Stage-wise estimators may fit separate copies of this stationary matrix; Proposition F.1 applies when cross-stage pooling is needed.

A valid procedure may fit transition, reward, and response regressions on one fold and evaluate the pessimistic recursion on an independent fold. Without cross-fitting, it must instead prove a simultaneous empirical-process bound over the data-dependent value class. Population feature coverage alone is insufficient: a finite-sample result must also control empirical Gram matrices, target leverage, and the simplex projection used for the estimated response rows. Self-normalized concentration is one possible ingredient (Abbasi-Yadkori et al., 2011).

Corollary H.3 (Conditional linear-function guarantee)**.** *Under the linear model above, suppose the target response is supplied or certified, and suppose a cross-fitted or otherwise justified estimator supplies the simultaneous envelopes in Proposition H.2. Then, uniformly over* Π*,*

$$0 \le V_{\mathrm{cf}}(\pi) - \widehat{V}^-(\pi) \le 2\sum_{h=1}^H \mathbb{E}_{d_h^{\pi,\mu^\pi}}[\Gamma_h^\pi] + 2\sum_{h=1}^H \mathbb{E}_{p_h^\pi}[\Lambda_h^\pi],$$

with the same post-selection conclusion as Corollary 4.4.

Proof. The linear parameterization is used to construct the envelopes. Once they hold, Proposition H.2 and the selection argument apply directly. □

Population diagnostics such as

$$\sup_{\pi,h} \mathrm{tr}(\Sigma_h^\pi(\Sigma_h^{\log})^{-1}), \qquad \sup_{\pi,h} \mathrm{tr}(\Psi_h^\pi(\Psi_h^{\log,z_h^\pi})^{-1})$$

describe game-model and response-feature leverage. They do not by themselves prove empirical conditioning. We therefore make no generic dimension-only rate claim: such a rate requires an estimator-specific proof of cross-fitting or uniformity, empirical Gram control, leverage bounds, and matrix-valued multiclass response estimation.

H.4 Detailed conditional estimation extensions

The generic plug-in reduction is not intrinsically tabular. This section develops the linear model with a dimensionally valid response parameterization and states the rate at the level justified by the available envelopes. Let $\varphi : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}^{d_P}$ and $\psi : \mathcal{S} \times \mathcal{Z} \to \mathbb{R}^{d_g}$ be known bounded features. Consider

$$P_h(s' \mid s, a, b) = \langle \varphi(s, a, b), \theta_h(s') \rangle, \qquad r_h(s, a, b) = \langle \varphi(s, a, b), \xi_h \rangle,$$
$$g(b \mid s, z) = \langle \psi(s, z), \omega_b \rangle, \qquad \Omega = [\omega_1, \dots, \omega_B] \in \mathbb{R}^{d_g \times B},$$

with parameters restricted so the displayed rows are probability distributions. The matrix Ω is required for the vector-valued response regression. Stage-wise estimators may fit separate copies of this stationary matrix; the randomized-stage construction in Proposition F.1 is used when cross-stage transfer is needed.

Regression and independence requirements. A concrete procedure may fit ridge reward, transition, and response regressions on a nuisance fold and evaluate its pessimistic recursion on an independent fold. Alternatively, it must prove a simultaneous empirical-process bound over the data-dependent value class. Population trace coverage alone is not enough: the theorem must also establish high-probability empirical Gram-matrix control and leverage bounds on the target occupancy. Simplex projection must be included in the response estimator.

Conditional envelopes. Let $\widehat{P}, \widehat{r}, \widehat{g}$ be any such cross-fitted or otherwise rigorously controlled estimates. Suppose that, simultaneously over the declared target class, they provide nonnegative envelopes $\Gamma_h^\pi(s, a, b)$ and $\Lambda_h^\pi(s)$ satisfying

$$|(\widehat{r}_h - r_h)(s, a, b) + [(\widehat{P}_h - P_h)\widehat{V}_{h+1}^\pi](s, a, b)| \le \Gamma_h^\pi(s, a, b),$$
$$H \|\widehat{\mu}_h^\pi(\cdot \mid s) - \mu_h^\pi(\cdot \mid s)\|_1 \le \Lambda_h^\pi(s).$$

These are the same assumptions used by Proposition H.2; their proof may use self-normalized concentration (Abbasi-Yadkori et al., 2011), sample splitting, and empirical leverage control.

Linear coverage diagnostics. Define the physical Gram matrices

$$\Sigma_h^\pi := \mathbb{E}_{d_h^{\pi,\mu^\pi}}[\varphi\varphi^\top], \qquad \Sigma_h^{\log} := \mathbb{E}_{d_h^{\log}}[\varphi\varphi^\top],$$

and the response Gram matrices

$$\Psi_h^\pi := \sum_s p_h^\pi(s)\psi(s, z_h^\pi)\psi(s, z_h^\pi)^\top, \qquad \Psi_h^{\log,z} := \sum_s d_h^{\log}(s, z)\psi(s, z)\psi(s, z)^\top.$$

Below, $M^\dagger$ denotes the Moore–Penrose pseudoinverse, and each trace quantity is defined as $+\infty$ unless the target Gram matrix is supported on the range of its logging counterpart. The population quantities

$$C_{\mathrm{cf}}^{\mathrm{lin}} := \sup_{\pi,h} \mathrm{tr}(\Sigma_h^\pi(\Sigma_h^{\log})^\dagger), \qquad C_{\mathrm{sr}}^{\mathrm{lin}} := \sup_{\pi,h} \mathrm{tr}(\Psi_h^\pi(\Psi_h^{\log,z_h^\pi})^\dagger)$$

measure target leverage in the physical and response feature spaces. They are useful population descriptors, but a finite-sample theorem must separately control the corresponding empirical matrices.

Corollary H.4 (Conditional linear-function-approximation guarantee)**.** *Under the linear model above, exact or certified target-response recovery, and the simultaneous envelopes just stated, the pessimistic estimator satisfies*

$$0 \le V_{\mathrm{cf}}(\pi) - \widehat{V}^-(\pi) \le 2\sum_{h=1}^H \mathbb{E}_{d_h^{\pi,\mu^\pi}}[\Gamma_h^\pi] + 2\sum_{h=1}^H \mathbb{E}_{s\sim p_h^{\pi,\mu^\pi}}[\Lambda_h^\pi] \tag{97}$$

uniformly over $\pi \in \Pi$, with the same post-selection conclusion as the generic plug-in theorem.

Proof. The linear parameterization is used only to construct the envelopes. Once they hold, Theorem C.13, equivalently Proposition H.2, gives the displayed result directly. □

What the conditional transfer does not imply. The population trace quantities above do not by themselves control empirical conditioning, and the value function entering a regression target is data-dependent unless it is cross-fitted or controlled uniformly. An explicit dimension-only rate therefore requires an estimator-specific proof of the matrix-valued response regression, empirical leverage bounds, and every independence step. We state only the envelope transfer that the current assumptions justify.

H.4.1 VARIANCE-AWARE TRANSFER DETAILS

The range-based tabular theorem pays one factor of H in each stage-wise model bonus. A Bernstein refinement can save a factor of $\sqrt{H}$ only when the estimator supplies valid simultaneous variance envelopes and the relevant occupancy-weighted conditional variances aggregate. This qualification is essential because a value function computed from the same samples cannot be treated as fixed inside an empirical Bernstein inequality. Classical variance-sensitive RL analyses and empirical Bernstein inequalities provide useful technical templates (Azar et al., 2017; Jin et al., 2020; Li et al., 2024; Shi et al., 2022; Maurer & Pontil, 2009), but their martingale or sample-splitting arguments must be rebuilt for the response-augmented occupancy law rather than imported by analogy.

Suppose the model and response envelopes have the form

$$\Gamma_h^{\mathrm{B}}(s,a,b) \le c\left[\sqrt{\frac{\sigma_{P,h}^2(s,a,b)\iota}{N_h(s,a,b)\vee 1}} + \frac{H\iota}{N_h(s,a,b)\vee 1}\right], \tag{98}$$

$$\Lambda_h^{\mathrm{B}}(s;\pi) \le c\left[\sqrt{\frac{\sigma_{g,h}^2(s,z_h^\pi)\iota}{N_h(s,z_h^\pi)\vee 1}} + \frac{H\iota}{N_h(s,z_h^\pi)\vee 1}\right].$$

These envelopes may be obtained by sample splitting or by a proved uniform empirical-process argument; their validity is an assumption of the transfer result, not a consequence of the display itself.

Proposition H.5 (Conditional variance-aware tabular transfer)**.** *Suppose the hypotheses and exact-recovery event of Theorem 4.3 hold, the displayed Bernstein envelopes are valid simultaneously, and*

$$\sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi,\mu^\pi}} \sigma_{P,h}^2 \le H^2, \qquad \sum_{h=1}^{H} \mathbb{E}_{p_h^\pi} \sigma_{g,h}^2 \le H^2$$

uniformly over $\pi \in \Pi$. Then the leading square-root terms in the uniform evaluation and selection gaps are

$$\widetilde{O}\left(H^{3/2}\sqrt{\frac{C_{\mathrm{cf}} S^2 AB}{n}} + H^{3/2}\sqrt{\frac{C_{\mathrm{sr}} SB}{n}}\right) + \widetilde{O}\left(\frac{H^2 C_{\mathrm{cf}} SAB}{n} + \frac{H^2 C_{\mathrm{sr}} S}{n}\right), \tag{99}$$

where the second term is the explicit contribution of the inverse-count parts of (98)*.*

Proof. Apply the generic pessimism bound to the displayed envelopes. Cauchy–Schwarz under target occupancy combines the square-root variance terms with the coverage-weighted reciprocal counts. The two assumed aggregate variance bounds contribute the $H^{3/2}$ leading order. Coverage and the positive-cell count event bound the linear $H\iota/(N\vee 1)$ components by $\widetilde{O}(H^2 C_{\mathrm{cf}} SAB/n + H^2 C_{\mathrm{sr}} S/n)$. □

Why the variance statement is conditional. A same-sample proof cannot simply reuse the data to construct both the value function and its empirical variance. It must also justify any total-variance argument for stage-wise suprema and define the response-side variance term precisely. Proposition H.1 states the rate comparison only under valid envelopes. The matched converse remains open; Theorem G.7 records only the valid structural transfer of fixed-opponent offline-MDP lower bounds.

Regime	Requirement for ε-optimal selection
Supplied target response	$\widetilde{\mathcal{O}}(H^4\bar{C}S^2AB/\varepsilon^2)$ plus positive-cell count requirements
Uniform candidate-path benchmark	Maximum of the oracle-response requirement, uniform candidate-path margin condition, and positive-cell counts
End-to-end target path	Maximum of logging-label recovery, local target-margin recovery, positive-cell counts, and $\widetilde{\mathcal{O}}(H^4\bar{C}S^2AB/\varepsilon^2)$
Conditional variance transfer	Leading requirement $\widetilde{\mathcal{O}}(H^3\bar{C}S^2AB/\varepsilon^2)$ plus inverse-count terms, only when the stated simultaneous Bernstein envelopes hold
Conditional linear approximation	No generic dimension-only sample complexity without an estimator-specific envelope and empirical-Gram theorem

Table 5. Sample-complexity status. Conditional envelope results are not estimator-specific rates.

H.5 Further policy-selection consequences

Write $\bar{C} = \max\{C_{\mathrm{cf}}, C_{\mathrm{sr}}\}$. Table 5 separates unconditional results from conditional transfers. Every end-to-end requirement must hold simultaneously; a value-rate calculation alone does not certify response recovery.

A scale calibration. Take $H = 10$, $S = A = B = 4$, $|\mathcal{Z}| = 4$, and $\bar{C} = 12$. Ignoring universal constants, logarithms, and the positive-cell floor, setting the range-based upper bound equal to $\varepsilon = 0.1$ gives

$$n \gtrsim \frac{H^4\bar{C}S^2AB}{\varepsilon^2} = \frac{10^4 \cdot 12 \cdot 16 \cdot 4 \cdot 4}{0.01} \approx 3.1 \times 10^9.$$

This is a deliberately conservative rate calibration, not an operational budget. The end-to-end method must additionally meet both margin conditions and every supported-cell count requirement. The calculation makes clear when a theoretically identifiable deployment may still be statistically impractical.

Corollary H.6 (Expectation and simple regret)**.** *Let E be a simultaneous event on which a data-dependent certificate $U(\mathcal{D})$ satisfies $V_{\mathrm{cf}}(\pi^*) - V_{\mathrm{cf}}(\widehat{\pi}) \le U(\mathcal{D})$. Then*

$$\mathbb{E}[V_{\mathrm{cf}}(\pi^*) - V_{\mathrm{cf}}(\widehat{\pi})] \le \mathbb{E}[U(\mathcal{D})\mathbf{1}\{E\}] + H\mathbb{P}(E^c).$$

In particular, if $\mathbb{P}(E^c) \le \delta$ and $U \le u_n$ on E, the expected regret is at most $u_n + H\delta$.

Proof. Decompose the expectation over E and E^c. Regret is at most H because rewards lie in $[0, 1]$. □

Corollary H.7 (Anytime-valid bounds at dyadic checkpoints)**.** *Consider nested prefixes of one i.i.d. episode stream. At checkpoints $n_k = 2^k$, $k \ge 1$, run the end-to-end procedure with $\delta_k = 6\delta/(\pi^2 k^2)$. Under the hypotheses of Theorem 4.3, let $k_0 \ge 2$ be the first index such that every checkpoint $k \ge k_0$ satisfies all fixed-sample count and margin conditions at confidence δ_k. Such an index exists for a fixed problem with positive covered-cell masses and positive margins because $n_k/\log(n_k/\delta_k)$ eventually increases without bound. Between checkpoints retain the most recent lower confidence bounds. With probability at least $1 - \delta$, every frozen bound after n_{k_0} is valid, and its square-root rate differs from the fixed-sample rate only by a universal constant and an additional $O(\log\log(en))$ inside the logarithm.*

Proof. Apply Theorem 4.3 at every $k \ge k_0$ and union bound $\sum_{k \ge k_0} \delta_k \le \delta$. If $n_k \le n < n_{k+1}$, use the estimator from n_k. Then $n_k \ge n/2$, and $\log(1/\delta_k)$ adds only $O(\log\log(en))$. □

Corollary H.8 (Almost-sure consistency for a fixed problem)**.** *Fix a tabular problem satisfying the end-to-end assumptions with positive constants, and use nested data prefixes. Then the pessimistic selector satisfies*

$$V_{\mathrm{cf}}(\pi^*) - V_{\mathrm{cf}}(\widehat{\pi}_n) = O\left(\sqrt{\frac{\log n}{n}}\right) \quad \textit{almost surely}.$$

Proof. Set $\delta_n = n^{-2}$. The fixed positive masses and margins make all sample-size conditions hold eventually. The deviation probabilities are summable, so Borel-Cantelli gives the claimed eventual rate. □

The finite-class logarithm can be written as $\log|\Pi|$ plus the cell-wise terms. For an infinite class, run the method only on a finite ε-net. A full-class comparison is valid only when the net is response stable and V_{cf} is locally Lipschitz: if every π has a net point $\widetilde{\pi}$ with the same response path and $|V_{\mathrm{cf}}(\pi) - V_{\mathrm{cf}}(\widetilde{\pi})| \le L_V\varepsilon$, the union bound uses $\log \mathcal{N}(\Pi, \varepsilon)$ and the approximation contribution is at most $2L_V\varepsilon$. An ordinary policy-space cover is insufficient near a response boundary.

The simultaneous lower confidence bounds also support abstention. One may retain only policies whose recovery radius lies below empirical response-boundary slack and whose robust outer-interval width lies below a prespecified tolerance. This is a decision rule built from the same event, not a new identification assumption.

H.6 Rate comparison and research directions

Table 5 separates proved sample requirements from conditional transfers. The oracle-response row also requires the positive-cell floor $1/\min\{\kappa_{\mathrm{model}}, \kappa_{\mathrm{resp}}\}$. An envelope transfer becomes a sample-complexity theorem only after a concrete estimator is shown to satisfy those envelopes simultaneously.

H.6.1 Doubly-robust variant

Research blueprint, not a theorem. A valid doubly robust extension is attractive, but the one-step pseudo-reward based on a single one-step augmentation is not a sequential influence function and does not justify an either-rate guarantee. A complete result would need to define the longitudinal counterfactual law, use sequential density ratios, cross-fit every nuisance including the recovered response model, and prove that the remaining bias is a sum of products of nuisance errors. It would also need a one-sided confidence construction compatible with pessimistic selection. Standard OPE ideas such as Jiang & Li (2016) motivate this program, but equation-level transfer is not automatic because the target opponent response changes with the deployed policy. We therefore retain this topic as an explicit research direction and make no doubly robust rate claim.

H.6.2 Sharper response-state cardinality lower bound

Open refined response-state cardinality conjecture. It is natural to ask whether there exists a covered subclass for which

$$\inf_{\widehat{\pi}} \sup \mathbb{E}\big[V_{\mathrm{cf}}(\pi^\star) - V_{\mathrm{cf}}(\widehat{\pi})\big] \;\ge\; c\,H\sqrt{\frac{d_R B}{n}} \tag{100}$$

up to a natural cap. Counting $d_R B$ hypotheses while describing only $\log B$ bits per response state and asserting a KL scaling without the required packing would not prove this claim. The rigorously established response lower bound is Theorem G.4, which gives $c(H-1)\min\{1, \sqrt{d_R/n}\}$. Proving or refuting the additional $\sqrt{B}$ dependence remains open.

H.6.3 Detailed expectation conversion

Corollary H.9 (Simple regret bound)**.** *Under the hypotheses of Theorem C.13, the policy selector $\widehat{\pi} = \arg\max_{\pi\in\Pi} \widehat{V}^-(\pi)$ satisfies the following safe expectation conversion. Let E be the theorem's simultaneous event and let $U(D)$ be any nonnegative, data-dependent certificate such that $V_{\mathrm{cf}}(\pi^\star) - V_{\mathrm{cf}}(\widehat{\pi}) \le U(D)$ on E. Then*

$$\mathbb{E}\big[V_{\mathrm{cf}}(\pi^\star) - V_{\mathrm{cf}}(\widehat{\pi})\big] \;\le\; \mathbb{E}[U(D)\mathbf{1}\{E\}] + H\delta. \tag{101}$$

If $U(D) \le u_n$ deterministically on E, this simplifies to $u_n + H\delta$. In the tabular oracle regime, taking $\delta = 1/n$ and the deterministic uniform certificate yields $\widetilde{O}(H^2\sqrt{\bar{C}S^2AB/n}) + H/n$.

Proof. Decompose the expectation over E and E^c. The first term is bounded by $\mathbb{E}[U(D)\mathbf{1}\{E\}]$. Since rewards lie in $[0,1]$, policy regret lies in $[0,H]$, so the second term is at most $H\mathbb{P}(E^c) \le H\delta$. The stated specializations follow immediately. □

H.6.4 ASYMPTOTIC REGRET

Corollary H.10 (Asymptotic regret rate)**.** *Fix one tabular oracle problem satisfying the assumptions with finite constants, and view the datasets as nested prefixes of one infinite i.i.d. stream. Then* $V_{\mathrm{cf}}(\pi^\star) - V_{\mathrm{cf}}(\widehat{\pi}_n) \to 0$ *almost surely and*

$$V_{\mathrm{cf}}(\pi^\star) - V_{\mathrm{cf}}(\widehat{\pi}_n) = O\left(\sqrt{\frac{\log n}{n}}\right) \qquad \textit{almost surely}.$$

The hidden constant depends on the fixed problem parameters. No exact limsup constant is asserted from a bound stated only up to logarithmic factors.

Proof. Choose $\delta_n = 1/n^2$ in Corollary C.10. The deviation event has probability at most $1/n^2$; the Borel-Cantelli lemma gives that the bound holds for all but finitely many n almost surely. Substituting δ_n into ι contributes a $\log n$ factor under the square root. The fixed positive cell probabilities imply that the count condition eventually holds almost surely, giving the stated rate. □

H.6.5 EXPLICIT POLICY-CLASS LOG-CARDINALITY DEPENDENCE

Corollary H.11 (Target-class cardinality refinement)**.** *The bound of Corollary C.10 can be refined to make the policy-class union bound explicit:*

$$V_{\mathrm{cf}}(\pi) - \widehat{V}^{-}(\pi) \;\le\; c\,H^2\sqrt{\frac{\bar{C}\,S^2AB(\log(SABHd_R n/\delta) + \log|\Pi|)}{n}}.$$

For an infinite target class, choose a finite ε*-net* Π_ε*, run the estimator and selector only over that net, and compare the selected net policy with the optimum over the full class. This replacement is valid only if the net is response-stable and the counterfactual value is Lipschitz on each covered neighborhood. Concretely, if every* π *has* $\widetilde{\pi} \in \Pi_\varepsilon$ *with the same recovered response path and* $|V_{\mathrm{cf}}(\pi) - V_{\mathrm{cf}}(\widetilde{\pi})| \le L_V\varepsilon$*, then the finite-net bound uses* $\log \mathcal{N}(\Pi, \varepsilon)$ *and the comparison with the full-class optimum costs at most* $2L_V\varepsilon$*.*

Proof. The proof of Theorem C.13 applies a union bound over the $|\Pi|$ value functions $\widehat{V}^{\pi}_{h+1}$ in the model envelope. The contribution to ι is exactly $\log|\Pi|$. For the conditional infinite-class statement, union bound over the finite response-stable cover and apply the assumed Lipschitz inequality twice, once for the optimal policy and once for the selected policy. Without response stability, an arbitrarily small policy perturbation can cross a response boundary, so ordinary total-variation covering alone is insufficient. □

H.7 APPLICATIONS AND CONCRETE RESPONSE MAPS

The model is most credible when the response state and its update map are part of system design. Four recurring patterns illustrate the intended scope. In repeated pricing or auctions, a_h and b_h can be simultaneous price or bid choices, with z_h encoding a public, prespecified exposure regime. In human-AI systems, z_h can encode a documented load regime that affects an engagement action chosen before the current prompt is seen. In recommendation systems, z_h can encode a prespecified engagement state driven by cumulative delivered content, provided the modeled user action does not depend on the current recommendation. In budgeted markets or security settings, z_h can record a documented resource threshold. Sequential replies, purchases, or clicks that depend on the learner's current action require a different model. These are design templates, not claims that passive data in these domains satisfy the assumptions.

The same domains appear in nearby work for different reasons. Opponent modeling learns representations or adaptation policies for use at deployment (Jing et al., 2024; 2025; Wu et al., 2023), while Stackelberg RL usually specifies a best-response follower and permits interaction (Zhong et al., 2023). Performative learning studies repeated deployment and convergence to stable or optimal policies (Perdomo et al., 2020; Mandal et al., 2023; Rank et al., 2024; Mandal & Radanovic, 2025; Sahitaj et al., 2025). Preference-based offline RL uses static comparison feedback and reward uncertainty (Zhu et al., 2023; Zhan et al., 2024). In contrast, the response maps below are operational specifications

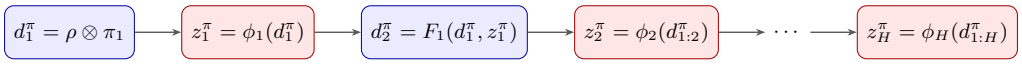


Figure 3. Triangular construction of the response sequence. Each coordinate uses only the already determined prefix, so one forward pass yields the unique response.

for identifying a dynamic opponent law from a fixed archive. They should be adopted only when their states and update rules have meaning before the target values are examined.

Three response maps cover these patterns.

Quantized aggregates. Let $\psi_h(d_{1:h}) \in \mathbb{R}^m$ be a known deployment summary and let $\{\mathcal{C}_z : z \in \mathcal{Z}\}$ be a finite partition. Set $\phi_h(d_{1:h}) = z$ when $\psi_h(d_{1:h}) \in \mathcal{C}_z$ and learn one response-kernel row family $g(\cdot \mid s, z)$ per cell. The margin is the distance from the population summary to the nearest cell boundary.

Quantized logit response. A bounded-rationality model may use a logit quantal-response model with a temperature regime selected by a quantized exposure summary (McKelvey & Palfrey, 1995). Conditional on z, the opponent-response kernel can be estimated nonparametrically or parameterized as

$$g(b \mid s, z) = \frac{\exp(\beta_z u_b(s))}{\sum_{b' \in \mathcal{B}} \exp(\beta_z u_{b'}(s))}.$$

The causal timing concerns the summary that selects z; smoothness of the logit row does not replace separation from a discrete bin boundary.

Threshold responses. For thresholds $0 = t_0 < \cdots < t_{d_R} = 1$ on a scalar summary, set $\phi_h(d_{1:h}) = j$ on $[t_{j-1}, t_j)$ for $j < d_R$ and on $[t_{d_R-1}, 1]$ for $j = d_R$. Let $g(\cdot \mid s, j)$ be a learned stochastic row or a specified pure action. This is appropriate only when the thresholds have an operational interpretation independent of the estimated target values.

Lemma H.12 (Boundary stability for Lipschitz summaries)**.** *Suppose $L_\psi > 0$, the output of ψ_h is equipped with a norm $\|\cdot\|_\psi$, and ψ_h is L_ψ-Lipschitz from prefix ℓ_1 distance to $\|\cdot\|_\psi$. If the population summary is at least $\eta_h^\pi > 0$ from the nearest quantization boundary in the same output norm, every prefix satisfying*

$$\|\widetilde{d}_{1:h} - d_{1:h}^\pi\|_{1,\mathrm{pref}} < \eta_h^\pi / L_\psi$$

has the same response label as the population prefix.

Proof. Lipschitzness keeps the perturbed summary within η_h^π of its population value, so it cannot cross a boundary. □

For any application, credibility rests on the same evidence: the response states and boundaries are prespecified; policy assignments and minimum exploration probabilities are recorded; the target reaches only state-joint-action and state-response rows with adequate support; and the stationary opponent-response kernel is adequate after conditioning on (s, z). If longer history affects behavior, it must enter the game state or response state. Otherwise the misspecification analysis, not the point theorem, is the appropriate output.

H.8 Detailed application mappings and audit conditions

The finite response-state model in Definition 3.1 is restrictive, but it has concrete interpretations. Each response family in Appendix H.8.1 can be specified in a designed system whose response state is recorded directly or defined by a prespecified finite rule. We sketch four domains and identify what would have to determine the map ϕ_h and opponent-response kernel g in each one.

(D1) Repeated pricing and dynamic auctions. A pricing platform and a rival choose prices simultaneously from finite menus. The rival observes the public state and a prespecified bin of aggregate price exposure, not the platform's realized current price. Logit quantal response is an established model of boundedly rational choice (McKelvey & Palfrey, 1995); here its payoffs must

be functions of the shared state and announced bin. The summary ψ_h can be the population average posted price, with z_h^π its bin index. Recovery requires policy labels, both price actions, state and outcome records, and bin boundaries fixed independently of the outcomes. A buyer's response to the current posted price is sequential and is not covered by this simultaneous-action model.

(D2) Human-AI interaction with shared context. An assistant chooses an intervention a_h from a finite menu while a human chooses an engagement action from $|\mathcal{B}|$ possibilities using the shared task state and a previously announced load regime. One designed model lets z_h^π track cumulative intervention load or time pressure. The human does not observe the current intervention before choosing b_h; a reply to the current prompt would require a sequential-action model. The response abstraction and its boundary slack still require external justification.

(D3) Adaptive recommender systems with engagement-state users. One possible recommender model uses an engagement state, such as cold, warm, or churning, that evolves with platform-side decisions (Shao et al., 2023; Jing et al., 2024; 2025). The platform's cumulative delivered-content rate ψ_h partitions into a finite engagement-state z_h^π, and the user's next-session engagement decision conditional on z_h^π is a class-(i) quantized-aggregate response state if made before the current item is shown. Clicks on the current item are outside this timing. Some systems record an engagement label for analytics, but the theorem does not infer a valid response state merely from the presence of such a field.

(D4) Budget-constrained or threshold-switching adversaries. One designed model for ad auctions or security games lets a bidder or attacker switch between a small number of pure strategies when an operator-induced budget crosses a prespecified threshold. This fits class (iii): ψ_h is the cumulative ad-spend (or attack-cost) imposed by the deployed policy, z_h^π is the current threshold cell, and g is a deterministic response kernel. Threshold positions may be externally documented by declared budgets or contractual constraints; otherwise they are not certified by passive trajectories alone.

Verifiability. These domains illustrate how a response state could be specified by system design, direct logging, or externally justified cumulative quantities. They do not establish that any passive dataset satisfies the abstraction. Conditional on a valid response map, the data identify the opponent-response kernel only on covered (s, z) pairs (Proposition B.4). The joint state-response coverage condition (the state-response coverage condition of Proposition B.4) translates concretely to logging across (i) a range of price levels (D1), (ii) prompt menus spanning the relevant load regimes (D2), (iii) recommendation densities covering each engagement state (D3), or (iv) budget allocations bracketing each threshold cell (D4). A deployment must deliberately collect or verify this operational coverage; passive logs need not contain it.

H.8.1 Three explicit response-map families

We give three explicit instantiations of Definition 3.1. In each case we exhibit the response-state space $\mathcal{Z}$, the occupancy-to-response-state functional ϕ_h, the opponent-response kernel g, and the causal timing and boundary conditions needed for forward response recovery.

(i) Quantized aggregate responses. Fix a learner-side summary map $\psi_h(d_{1:h}) \in \mathbb{R}^m$ (e.g., cumulative learner activity in particular state-action cells) and a finite partition $\{\mathcal{C}_1, \ldots, \mathcal{C}_{d_R}\}$ of its range. Set $\phi_h(d_{1:h}) := j$ whenever $\psi_h(d_{1:h}) \in \mathcal{C}_j$, and let $g(b \mid s, z) = q_z(b \mid s)$ for fixed conditional distributions $\{q_z\}_{z\in\mathcal{Z}}$. Under simultaneous-action timing, ψ_h uses only the learner occupancy prefix available before the stage-h opponent action. Thus z_h^π is computed after $z_{1:h-1}^\pi$ with no circular same-stage dependence. If the realized summary is at least η from the union of the partition boundary, any prefix perturbation smaller than η preserves its label. This subclass models *congestion, load shedding, or coarse market response*.

(ii) Quantal-response opponent with quantized temperature. Bounded-rationality opponents are commonly modeled by logit (quantal) response (McKelvey & Palfrey, 1995): given a payoff vector $u_b(s)$, the opponent selects b with probability proportional to $\exp(\beta\, u_b(s))$ for some inverse temperature β. We allow β to depend on cumulative learner activity through a quantized aggregate.

Concretely, fix a learner-side load summary $\psi_h(d_{1:h})$ taking values in $[0,1]$, a finite grid $0 = \beta_1 < \beta_2 < \cdots < \beta_{d_R}$, and partition $[0,1]$ into d_R contiguous bins $\{\mathcal{C}_j\}_{j=1}^{d_R}$. Set $\phi_h(d_{1:h}) = j$ whenever $\psi_h(d_{1:h}) \in \mathcal{C}_j$, and let

$$g(b \mid s, z = j) \;=\; \frac{\exp\big(\beta_j\, u_b(s)\big)}{\sum_{b' \in \mathcal{B}} \exp\big(\beta_j\, u_{b'}(s)\big)}$$

for known $u_b(s)$. The quantized temperature is recovered chronologically when ψ_h respects the causal timing above. Smoothness of the logit response kernel controls how response probabilities vary within a label, while distance from bin boundaries supplies the discrete margin needed for exact label recovery. This subclass captures *behavioral opponents whose rationality level adapts to deployment intensity*.

(iii) Threshold best-response opponent. Suppose the opponent has d_R pure strategies $\{b_1^\star, \ldots, b_{d_R}^\star\}$ and switches between them according to which of d_R disjoint regions of deployment-summary space the learner occupies. Fix thresholds $0 = t_0 < t_1 < \cdots < t_{d_R - 1} < t_{d_R} = 1$ on a learner-side load $\psi_h(d_{1:h}) \in [0,1]$. Set $\phi_h(d_{1:h}) = j$ whenever $\psi_h(d_{1:h}) \in [t_{j-1}, t_j)$ for $j < d_R$, and use the closed final cell $[t_{d_R - 1}, 1]$ for $j = d_R$. Let $g(b \mid s, z = j) = \mathbf{1}\{b = b_j^\star\}$. This can model seller withdrawal: thresholds t_j encode the levels of deployed market activity at which the seller switches between participation strategies. Verification uses boundary slack rather than a separate uniqueness condition: if the true summary is at distance at least η from every threshold, then any prefix perturbation smaller than η preserves the response label.

Common structure. All three examples share the same template: a known scalar or low-dimensional deployment summary ψ_h, a finite quantization producing z_h^π, and an opponent-response kernel g with a specified form and unknown population parameters. Causal timing gives one response sequence. Quantization separation controls stability of that sequence under estimation error, which is exactly the role of the response-state margin.

Lemma H.13 (Causal uniqueness and local perturbation stability)**.** *Under simultaneous-action timing, T_π has the unique forward solution in Lemma 3.2. In addition, suppose $L_\psi > 0$ and ψ_h is L_ψ-Lipschitz from prefix ℓ^1 distance to a specified output norm $\|\cdot\|_{\mathrm{out}}$. Suppose also that the population summary at the true prefix is at least $\eta_h^\pi > 0$ from the nearest quantization boundary in that norm. Then every prefix $\widetilde{d}_{1:h}$ satisfying*

$$\|\widetilde{d}_{1:h} - d_{1:h}^\pi\|_{1,\mathrm{pref}} < \eta_h^\pi / L_\psi$$

has the same response label as the population prefix.

Proof. Uniqueness is the causal induction of Lemma 3.2. Lipschitzness gives $\|\psi_h(\widetilde{d}_{1:h}) - \psi_h(d_{1:h}^\pi)\|_{\mathrm{out}} < \eta_h^\pi$, so the perturbed summary cannot cross a quantization boundary. Hence ϕ_h is unchanged. □

This lemma does not use a derivative with respect to the current response state: under the stated timing, $d_h^{\pi,z}(s,a)$ is independent of z_h. A model in which the response map consumes a same-step quantity already affected by z_h is genuinely noncausal and lies outside the present theorem.

H.8.2 Support, causal timing, and pessimism diagrams

We collect three small diagrams that visualize the central conceptual objects: the coverage requirement on target-reachable support (Figure 4), the causal response recursion (Figure 5), and the pessimistic lower-confidence band underlying the policy selector (Figure 6).

Coverage requirement. Finite C_{cf} requires the logged occupancy to dominate the induced-response target occupancy on every (s,a,b) triple actually visited under the target deployment. Joint state-response coverage (the state-response coverage condition of Proposition B.4) additionally requires logged visits to the same game state s while instantiating the target response state z_h^π.

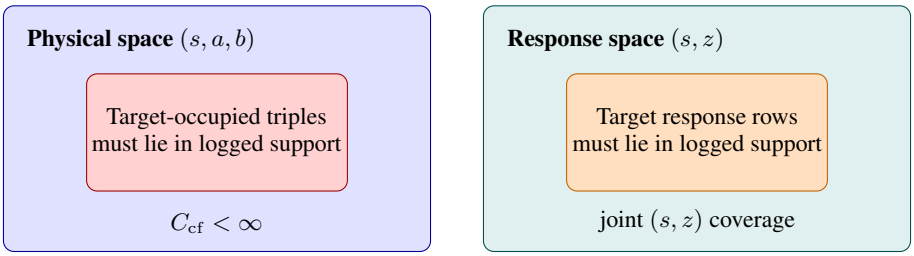


Figure 4. The two support requirements live in different spaces. State-joint-action coverage controls target (s, a, b) triples, while state-response coverage controls target opponent-response-kernel rows (s, z). Neither containment implies the other.

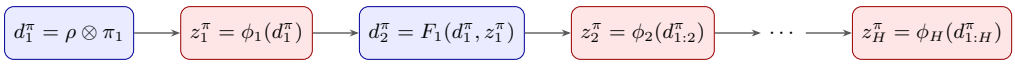


Figure 5. Triangular causal construction of the response sequence. Each coordinate uses only the already determined prefix, so one forward pass yields the unique response and no enumeration over $\mathcal{Z}^H$ is required.

Causal response recursion. For each target policy π, the stage-h response is computed from the learner-occupancy prefix generated by earlier responses. The apparent fixed-point problem is therefore a triangular dependency chain.

Pessimism band. The estimator is intentionally one-sided: $\widehat{V}^-(\pi) \le V_{\mathrm{cf}}(\pi)$ with high probability simultaneously over Π. Selecting $\widehat{\pi} = \arg\max_\pi \widehat{V}^-(\pi)$ is therefore safe because no policy can be selected on the basis of an over-optimistic estimate.

H.9 Worked numerical examples

We give a concrete numerical instance of Theorem B.1 that keeps full physical logging support. Take a one-step game with one state, learner actions a_0, a_1, and opponent actions b_0, b_1. The logging policy set contains $\pi_L(a_1) = 1/4$ and $\pi_H(a_1) = 3/4$ with positive mixing weights. In both games, π_L induces deterministic b_0 and π_H induces deterministic b_1. Thus every physical pair (a, b) has positive logging probability. Use the common reward $r(a, b) = \mathbf{1}\{b = b_1\}$, so the physical reward table is identified. Let the undeployed target satisfy $\pi_{\mathrm{tar}}(a_1) = 1/2$. In game G_0 its response is b_0; in game G_1 it is b_1. The two games induce exactly the same policy-labeled logged trajectory law, but their target values are zero and one. Hence $\mathrm{TV}(\mathbb{P}_0^{(n)}, \mathbb{P}_1^{(n)}) = 0$ for every $n \ge 1$. By Le Cam's two-point inequality (Lemma B.5), every estimator has error at least $1/2$ with probability at least $1/2$ on one of the games. A companion construction in Appendix B.6 drops state-joint-action support as well; that example is a coverage failure, not this full-support non-identifiability theorem.

The instance also separates two requirements that are easy to conflate. A one-bit response state with a specified policy-to-state rule would rule out the two arbitrary off-menu response assignments used above, but it would not by itself identify the target response-kernel row. Identification additionally requires the logs to cover the target (s, z) cell, as Theorem B.3 shows. The physical reward table is already identified here; the missing information is the opponent row induced by the undeployed target.

H.9.1 A two-stage causal-recovery calculation

A concrete two-step example makes the fixed-point map explicit. Let $H = 2$, $\mathcal{S} = \{s_0, s_G, s_B\}$, and $\mathcal{A} = \mathcal{B} = \{0, 1\}$. At $h = 1$ the only active state is s_0, and s_G, s_B are terminal step-2 states. Let a learner policy π choose action 1 at $h = 1$ with probability p_π and, at $h = 2$, choose action 1 with

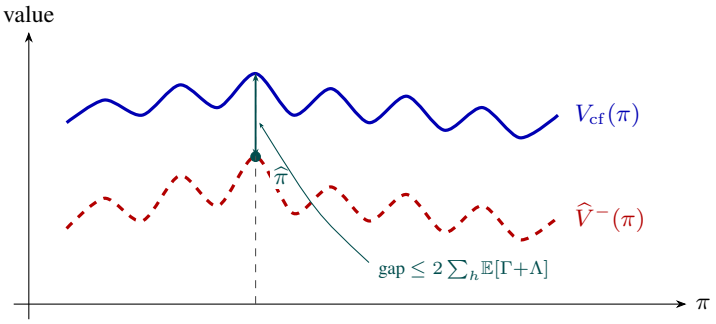


Figure 6. One-sided pessimism band. The lower-confidence value $\widehat{V}^{-}(\pi)$ (red, dashed) lies below $V_{\mathrm{cf}}(\pi)$ (blue) uniformly over $\pi \in \Pi$ with high probability. Selecting $\widehat{\pi} = \arg\max \widehat{V}^{-}$ guarantees $V_{\mathrm{cf}}(\pi^\star) - V_{\mathrm{cf}}(\widehat{\pi}) \leq 2\sup_\pi \sum_h \mathbb{E}[\Gamma_h^\pi + \Lambda_h^\pi]$ (Theorem C.13).

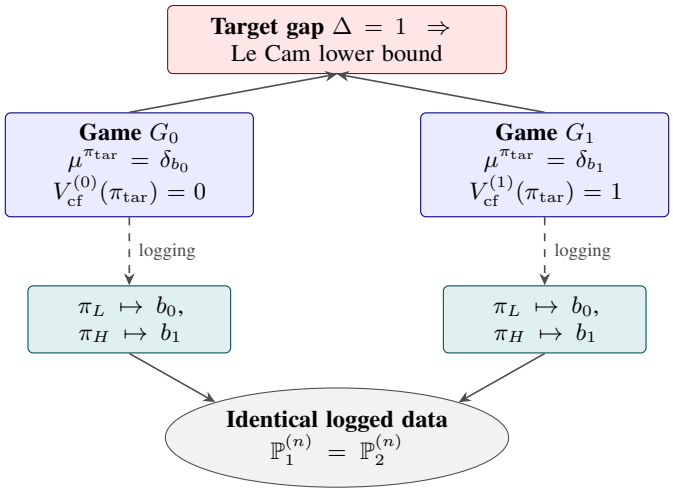


Figure 7. Identifiability barrier (Theorem B.1). Two Markov games with adaptive opponents and identical logged data (TV $= 0$) but divergent counterfactual values on the undeployed target policy.

probability q_π at both s_G and s_B. Define the response states by

$$z_1^\pi = \mathbf{1}\{p_\pi \geq 1/2\}, \qquad z_2^\pi = \mathbf{1}\left\{\frac{1}{2}\Pr_{(\pi,\mu^\pi,P)}(s_2 = s_G) + \sum_{s\in\mathcal{S}} d_2^\pi(s,1) \geq \frac{1}{2}\right\}.$$

Since $\sum_s d_2^\pi(s,1) = q_\pi$ (and $\Pr(s_2 = s_G) = \sum_a d_2^\pi(s_G,a)$), the second response state is again a known functional of learner-side occupancies. Let the opponent-response kernel be deterministic, $g(1 \mid s,1) = 1$ and $g(0 \mid s,0) = 1$ for all $s \in \mathcal{S}$, and let the first-step transition and second-step reward be $P_1(s_G \mid s_0,a,b) = \mathbf{1}\{a = b = 1\}$, $r_2(s_G,a,b) = 1$, and $r_2(s_B,a,b) = 0$ for all a,b.

Now fix a target policy with $(p_\pi, q_\pi) = (0.8, 0.3)$. Forward recovery first gives $z_1^\pi = 1$ from $p_\pi = 0.8$. Propagating that response through P_1 gives $\Pr(s_2 = s_G) = 0.8$, so the second summary is $\frac{1}{2}\cdot 0.8 + 0.3 = 0.7$ and $z_2^\pi = 1$. Thus $z^\pi = (1,1)$ and $V_{\mathrm{cf}}(\pi) = 0.8$, without evaluating any false candidate sequence. The threshold slacks are 0.3 and 0.2; hence any scalar-summary error strictly below 0.2 preserves both response labels. The candidate operator maps $(1,\cdot)$ to $(1,1)$ and $(0,\cdot)$ to $(1,0)$, so enumerating $\mathcal{Z}^H$ is unnecessary.

H.10 Open statistical questions

A sequential doubly robust estimator is an attractive extension, but ordinary one-step augmentation is insufficient because deployment changes the opponent law. A valid result would require sequential density ratios, cross-fitting of every nuisance including the response model, a product-of-errors

ID	Quantity varied	Diagnostic purpose
E1	Sample size	Check the response-aware plug-in rate against a response-agnostic bias floor.
E2	Target response margin	Check exact logged-label and target-label recovery.
E3	Horizon and response alphabet	Compare forward-recursion and exhaustive runtime, and verify agreement where enumeration is feasible.
E4	Logging weight on the target response	Hold state coverage fixed while measuring the state-response coefficient and interval width.
E5	Induced-response discrepancy	Compare actual value error with the bound B_ε.
E6	Stage of response-cell support	Compare stage-wise with cross-stage response-kernel estimation.
E7	Pricing-shaped adaptive game	Compare an oracle-response plug-in estimator with response-agnostic pooling.
E8	Undeployed target in a designed pricing-shaped game	Reconstruct hidden logging labels, recover the target response, and evaluate a lower bound from policy-labeled trajectories.

Table 6. Diagnostic map. Randomized diagnostics average independent trials with standard-error bars: 300 trials for E1, 500 for E2 and E4, and 400 for E6. E2 is run at $n = 1000$. E3 reports measured runtime. E5 is exact with horizon $H = 8$.

remainder, and a one-sided confidence construction compatible with pessimistic selection; standard OPE ideas provide a blueprint, not an automatic theorem (Jiang & Li, 2016).

The response-selection lower bound proves an $H\sqrt{d_R/n}$ dependence up to its natural cap. Whether a covered subclass forces an additional $\sqrt{B}$ factor remains open; no such dependence follows merely by counting $d_R B$ parameters. Other open problems include closing the horizon gap to the conditional variance-aware upper bound and proving end-to-end target-path recovery under general function approximation.

I Controlled Synthetic Diagnostics

The experiments are controlled checks of the theory. They use no external data and are not intended as real-world validation. The implementation uses fixed integer seed schedules and produces vector figures. Table 6 states the role of each primary experiment (E1–E8). Appendix I.4 reports a complementary suite labeled S1–S8.

I.1 Common protocol

All environments have finite S, A, B, H, stationary $g(b \mid s, z)$, and known threshold or quantization maps ϕ_h. Logging episodes carry learner-policy labels. Unless a diagnostic is labeled oracle-response, the estimator reconstructs logging response labels from empirical learner occupancies. E7 and the complementary panels S1, S2, S4, S5, S6, and S8 evaluate at the true target response sequence. Transition zero-count rows use a uniform stochastic fallback. Unless a diagnostic intentionally removes support, the logging policy set contains policies on both sides of every response boundary.

For recovery diagnostics, success means equality of the entire recovered response sequence to the population forward sequence. For runtime comparisons, exhaustive enumeration is stopped before memory or runtime becomes excessive and is used only at horizons where it is a meaningful correctness oracle. For incomplete support, the reported interval is the robust outer confidence interval from Theorem D.2. For misspecification, the game dynamics and learner policy are fixed while the true opponent row moves an ℓ_1 distance ε from the modeled row.

I.2 Results

Figure 1 reports all six diagnostics. In E1, increasing the sample size from 200 to 10,000 reduces mean absolute response-aware value error from 0.0500 ± 0.0023 to 0.0073 ± 0.0003. Pooling the two opponent regimes instead remains between 1.124 and 1.127, which is the predicted structural bias rather than a small-sample effect. Logging-label reconstruction succeeds in every E1 trial.

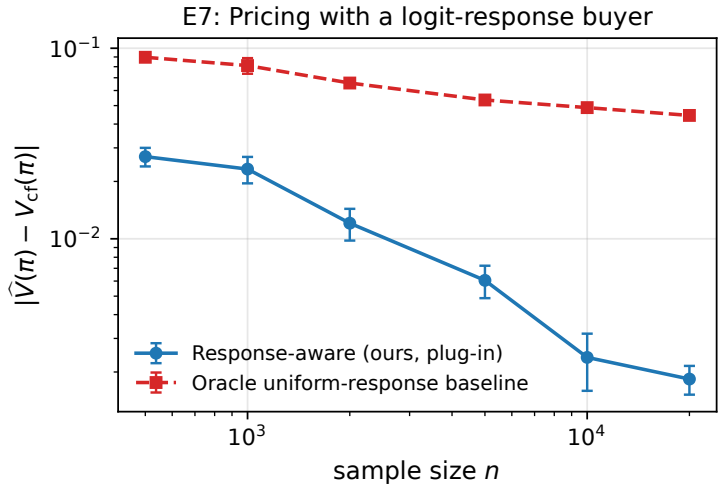


Figure 8. E7: pricing-shaped oracle-response benchmark. The response-aware plug-in estimator uses the true target response sequence. The response-agnostic baseline uses an oracle uniform average of the true response rows, isolating pooling bias rather than testing an estimated opponent model.

E2 makes the margin condition visible at $n = 1000$. At the response boundary, exact logging and target-path recovery succeeds in 0.538 ± 0.022 of trials. Success rises to 0.820 ± 0.017 at margin 0.02, 0.980 ± 0.006 at margin 0.05, and 1.000 at margins 0.10 and 0.20. The same empirical occupancy is used to reconstruct a logging label and to drive the corresponding two-stage target recursion, so the two curves overlap. The experiment illustrates the transition at the boundary; it does not test independence between the two events.

E3 checks every enumerated instance exactly. Causal recovery examines H coordinates, while exhaustive search examines $|\mathcal{Z}|^H$ candidates. On the machine used for this run, the measured speedup at the largest cases was 1.58×10^4 for $(|\mathcal{Z}|, H) = (2, 16)$ and 1.59×10^4 for $(3, 10)$. Wall-clock values are recorded separately from the deterministic scientific summary because they depend on hardware and load.

E4 is the one-cell specialization of the robust outer interval. With no target-response support, its value width is the full range $H = 5$. At $n = 2000$, increasing the logging weight from 0.01 to one decreases C_{sr} from 100 to one and the mean width from 2.708 ± 0.020 to 0.304. Empirical coverage is at least 0.996 in every supported setting. E5 computes the misspecification gap exactly at horizon $H = 8$: the actual gap $H\varepsilon/2$ remains below B_ε for all displayed $\varepsilon \in [0, 1]$. E6 removes same-stage support for the target response cell. The same-stage fallback then has constant response-kernel error 0.3, whereas the randomized-stage estimator falls from 0.0436 ± 0.0018 at $n = 300$ to 0.0076 ± 0.0003 at $n = 10{,}000$.

E7 is a secondary application diagnostic rather than evidence for end-to-end recovery. It uses a four-round game with three inventory states, three price actions, three quantity actions, and three price-anchor response regimes. A logit quantal-response buyer (McKelvey & Palfrey, 1995) selects a quantity from the public anchor and inventory before seeing the learner's realized price; this is a stylized timing assumption, not ordinary point-of-sale demand. Five logging policies cover the response regimes. Averaged over three targets and five trials, the estimator supplied with the true target response path improves by roughly one order of magnitude as n grows from 500 to $20{,}000$. A response-agnostic baseline supplied with the uniform average of the true response rows approaches an absolute-error floor near 0.044. That oracle baseline isolates pooling bias; it is not fitted behavior cloning.

I.3 HELD-OUT TARGET WITH RECONSTRUCTED RESPONSES

E8 exercises the non-oracle pipeline on one controlled instance of the designed-system example. The horizon is three, with two inventory states and two price actions for each retailer. The initial state is $s = 0$. At each stage the prespecified response map sets

$$z_h^\pi = \mathbf{1}\left\{\sum_s p_h^\pi(s)\pi_h(a_{\text{high}} \mid s) \geq \frac{1}{2}\right\}.$$

The rival chooses its price without observing the learner's current price. Its high-price probabilities $g(b = 1 \mid s, z)$ for $(s, z) = (0,0), (0,1), (1,0), (1,1)$ are respectively $0.20, 0.80, 0.35, 0.70$. The transition probability of next-stage state 1 is $P_h(1 \mid s, a, b) = 0.10 + 0.35a + 0.35b$; the reward mean is $r_h(s, a, b) = 0.10 + 0.25\mathbf{1}\{s = 0\} + 0.15a + 0.15b + 0.10\mathbf{1}\{a = b\} + 0.03(h - 1)$. Rewards are sampled as Bernoulli variables with these means. These quantities define the data generator and the ground truth; the estimator is not given P, r, g, or any response label.

Each episode is independently assigned with probability $1/3$ to one of three logged policies. Their high-price probabilities in states $(0, 1)$ are $(0.20, 0.10)$, $(0.85, 0.65)$, and $(0.60, 0.10)$. The target, with probabilities $(0.90, 0.05)$, is excluded from the logging menu. Its population response path is $(1, 0, 1)$, whereas the three logging paths are $(0, 0, 0)$, $(1, 1, 1)$, and $(1, 0, 0)$. Thus the complete target response path is not observed in any logging policy, although its required state-response cells are covered. The target's distances in high-price probability from the threshold are 0.400, 0.191, and 0.132 across the three stages. Because both logged actions and both opponent actions have positive probability and the transition reaches both states, the population coefficients are finite: $C_{\text{cf}} = 3.885$ and $C_{\text{sr}} = 5.223$. These values describe this generator, not estimated guarantees for an arbitrary deployment.

For each logging policy, the estimator first computes its empirical state frequencies and applies the known map to policy-factorized learner occupancies. It uses the resulting labels to form stage-specific estimates of P_h, r_h, and g_h; zero-count probability rows receive uniform stochastic fallbacks. A forward pass then recovers the target labels from the estimated model. The point estimate evaluates that recovered path. The lower bound uses simultaneous row-wise Hoeffding radii for rewards and multinomial ℓ_1 radii for transition and opponent rows, with nominal error level 0.10, followed by the pessimistic recursion. The pooled baseline uses the same estimated physical model but ignores response labels when estimating the opponent policy. The estimator never receives the population target path; the generator computes it separately for scoring. The row-wise radii are calibrated for correctly labeled data; by themselves they do not provide an unconditional 90% guarantee when labels are estimated. We therefore report label recovery and empirical bound coverage separately.

n	Joint path	Aware error	Pooled error	Lower gap	Coverage
250	0.995	0.0624	0.1053	1.632	1.000
500	1.000	0.0411	0.1078	1.631	1.000
1,000	1.000	0.0295	0.1040	1.388	1.000
2,000	1.000	0.0172	0.1050	0.985	1.000
5,000	1.000	0.0121	0.1085	0.624	1.000
10,000	1.000	0.0081	0.1070	0.441	1.000

Table 7. E8, 200 independent trials per sample size. Joint path is the fraction with all logging and target response labels correct. Errors are mean absolute value errors. Lower gap is the true value minus the pessimistic lower value; coverage is its empirical one-sided coverage. The true value is 1.632. Standard errors and unrounded values are in the supplement.

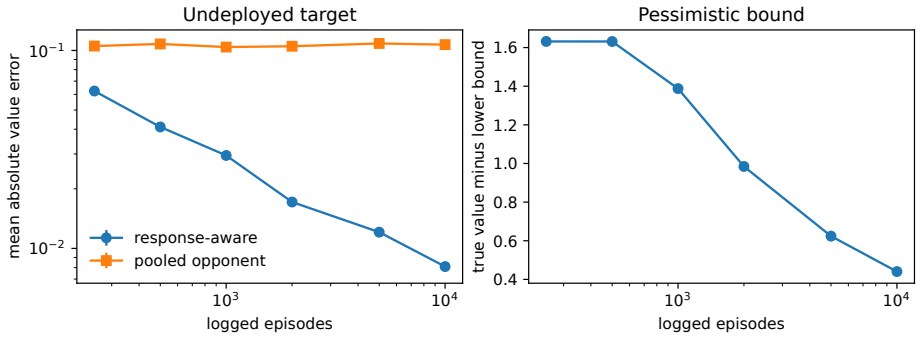


Figure 9. E8: held-out target evaluated without response labels supplied to the estimator. Error bars are one standard error over 200 independent trials. The row-wise pessimistic lower bound is zero on average at $n = 250$ and remains conservative at $n = 10{,}000$; empirical coverage alone is not a validation of the response model outside this controlled generator.

With a correctly specified map, overlap across policies lets the estimator evaluate this undeployed target. Certification remains conservative: at $n = 10{,}000$, the row-wise lower bound averages 0.441 below truth despite small point-estimation error; at $n = 250$, it is zero on average and cannot clear a positive deployment threshold. Neither result validates a real rival's response to the announced category or support in historical data.

I.4 Additional controlled diagnostic suite

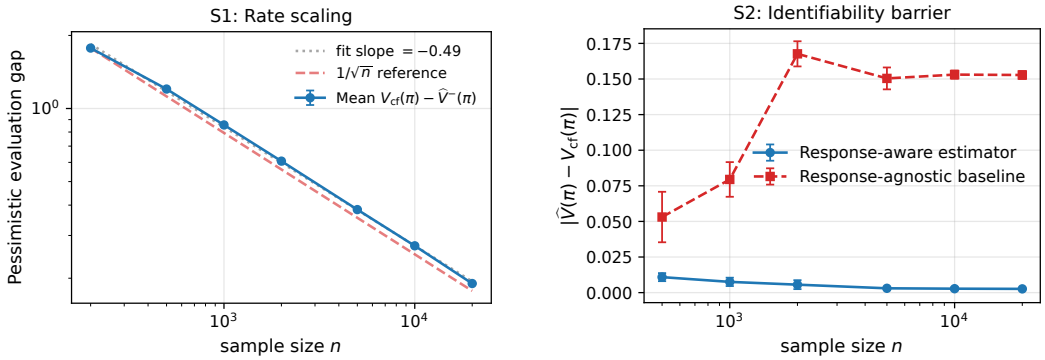
The six-panel figure in the main text (E1–E6) and the pricing panel E7 are the primary diagnostics. The eight experiments below are complementary tabular checks of the same mechanisms and are labeled S1–S8. S1, S2, S4, S5, S6, and S8 evaluate at the true target response sequence and therefore diagnose the common $1/\sqrt{n}$ envelope of Corollary C.10 and of Theorem 4.3 after the response path is given, together with the bias caused by pooling distinct response regimes. They are not tests of label reconstruction from policy-labeled logs. S3 and S7 test recovery of z^π.

The first five experiments (S1–S5) use an abstract tabular game implemented in the supplementary reproduction package. In that generator, $P_h(\cdot \mid s, a, b)$ and ρ are drawn once per seed from symmetric Dirichlet priors, rewards $r_h(s, a, b)$ are drawn uniformly in $[0, 1]$, and the stationary opponent-response kernel $g(\cdot \mid s, z)$ is drawn row-wise from Dirichlet$(0.5\,\mathbf{1}_B)$ and then sharpened to $0.2\,g(\cdot \mid s, z) + 0.8\,e_{z \bmod B}$ so that response state z favors opponent action $b = z \bmod B$. For S1–S5 and S7 the response-state functional ϕ_h quantizes the cumulative learner-side mass on the single state-action cell $(s = 0, a = 0)$ into d_R bins. Logged episodes sample a logging policy uniformly from the stated finite logger set unless an experiment explicitly specifies a non-uniform mixture (S4 and S8). For S5, the Bernstein-style bonus uses the implementation's empirical statewise variance proxy $\mathrm{var}\{\widehat{V}_{h+1}(s') : s' \in \mathcal{S}\}$ inside the bonus term, which is intended only as a qualitative diagnostic for the variance-aware sharpening rather than a separate theorem-level estimator construction. The reproduction package fixes a master seed and uses deterministic per-experiment seed schedules. Plots with error bars report means with ± 1 standard-error bars; when a bar is not visually distinguishable, it is smaller than the marker diameter. S3 and S7 report recovery rates without error bars. The reproduction materials record the software environment and reference hardware. Three further experiments instantiate domain-specific structure: **S6** runs an oracle-response estimator on a pricing-shaped game (Appendix H.7) with an anchor-conditioned logit buyer and contrasts it against an oracle uniform-response baseline; **S7** compares Picard iteration with exhaustive search as a computational cross-check of the same unique fixed point already computed by the forward pass; and **S8** stress-tests target response-state support while keeping the physical state process fixed. The supplementary package contains the reproduction code and the generated PDF figures. Table 8 collects the environment, logging design, and sweep parameters for each experiment.

S1: Rate scaling (oracle-response envelope). We sweep the sample size from 200 to 20,000 in a fixed game with $H = 5$ and 6 candidate target policies and report the mean pessimistic evaluation gap $\frac{1}{|\Pi|}\sum_\pi (V_{\mathrm{cf}}(\pi) - \widehat{V}^-(\pi))$ averaged over 10 trials. This panel uses an oracle-response estimator, so

Exp.	Environment	Policies / logging design	Sweep, trials, and notes
S1	Base tabular game: $S = 4$, $A = B = 2$, $H = 5$, $d_R = 2$	6 random target policies; 8 random logging policies; uniform logger mixture	sample-size sweep from 2×10^2 to 2×10^4; 10 trials
S2	Base tabular game: $S = 4$, $A = B = 2$, $H = 4$, $d_R = 2$	1 random target policy; 4 random logging policies; uniform logger mixture	sample-size sweep from 5×10^2 to 2×10^4; 6 trials; agnostic baseline uses g_{avg}
S3	Controlled-margin instance: $S = 2$, $A = B = 2$, $H = 3$, $d_R = 2$	target action-0 probability $0.5 + \gamma_\phi$; logging set is target policy plus 3 random policies	γ_ϕ sweep from 0.05 to 0.45; $n = 2000$; 30 trials; exact response-path recovery
S4	Base tabular game: $S = 3$, $A = B = 2$, $H = 4$, $d_R = 2$	4 random target policies; 5 random logging policies with mixture weights proportional to $(1 + 100 \cdot \text{skew}, 1, \ldots, 1)$	skew $\in [0, 0.9]$; $n = 4000$; 8 trials; reports empirical $\bar{C}$
S5	Base tabular game: $S = 3$, $A = B = 2$, $d_R = 2$	4 random target policies; 5 random logging policies; uniform logger mixture	horizon sweep $H \in \{3, 5, 7, 10, 14\}$; $n = 4000$; 8 trials; Hoeffding and Bernstein comparison
S6	Pricing-shaped game: $S = 3$, $A = B = 3$, $H = 4$, $d_R = 3$, $\beta = 6$	3 random target policies; 3 near-deterministic low/mid/high price loggers plus 2 random mixers	sample-size sweep from 5×10^2 to 2×10^4; 5 trials
S7	Same family as S3	same target and logging construction as S3	γ_ϕ sweep from 0.05 to 0.45; $n = 2000$; 30 trials; Picard iteration with 3 restarts and 20 steps compared with exhaustive search
S8	Response-coverage stress game: $S = 2$, $A = B = 2$, $H = 4$, $d_R = 2$	target action-0 probability 0.7; logging interpolates between target and opposite response regimes	logging-weight sweep from 0.02 to 0.95; $n = 1000$; 12 trials; transitions are action-independent

Table 8. Experimental settings for S1–S8.



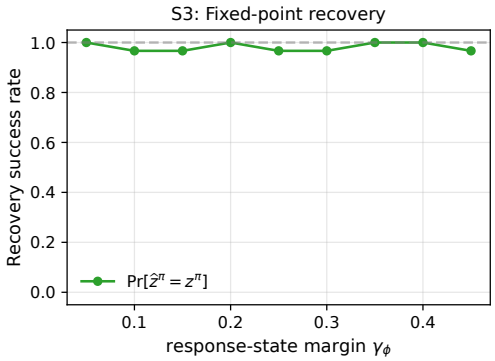
(a) S1: Pessimistic evaluation gap as a function of sample size on log-log axes, using an oracle-response estimator. The empirical curve tracks the predicted $1/\sqrt{n}$ rate.

(b) S2: Oracle-response plug-in versus response-agnostic pooling. The pooled baseline retains an asymptotic bias.

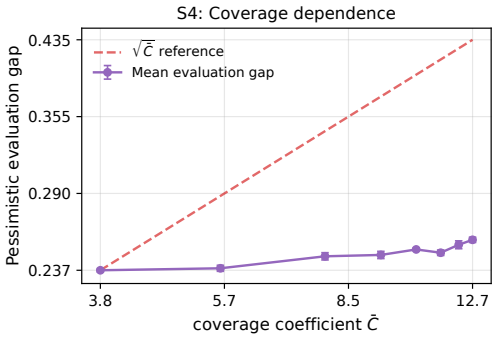
Figure 10. S1 (rate scaling) and S2 (response-agnostic pooling bias).

it diagnoses the common $1/\sqrt{n}$ envelope of Corollary C.10 and of Theorem 4.3 after exact recovery. Figure 10a shows the result on log-log axes alongside a $1/\sqrt{n}$ reference. The empirical curve tracks that rate.

S2: Bias from pooling response regimes. We compare two evaluations on the same logged data: (a) an oracle-response plug-in estimator that uses the true target response sequence z^π, and (b) an oracle response-agnostic baseline that treats the opponent as a static policy obtained by averaging the true response-kernel rows uniformly across z. Figure 10b plots $|\widehat{V}(\pi_{\text{tar}}) - V_{\text{cf}}(\pi_{\text{tar}})|$ as a function of n for both. The oracle-response estimator converges to zero at the predicted rate; the agnostic baseline plateaus at a positive bias of order 0.15 no matter how much data is collected. This experiment demonstrates the bias of this particular response-agnostic baseline; it is not an empirical proof of the non-identifiability theorem, which instead compares observationally equivalent games and quantifies over all estimators. The panel is also not a test of causal recovery.



(a) S3: Empirical fixed-point recovery rate as a function of the response-state margin. Recovery is near 100% above the diagnostic threshold.



(b) S4: Oracle-response evaluation gap as a function of the empirical coverage coefficient $\bar{C}$, with a $\sqrt{\bar{C}}$ reference for orientation.

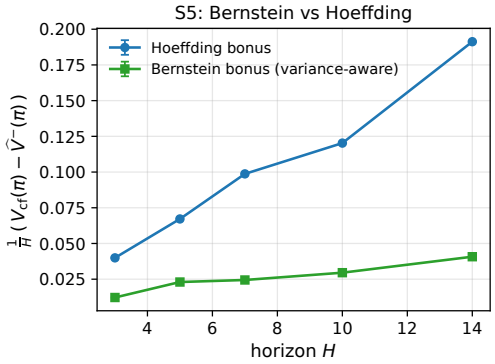
Figure 11. S3 (response-path recovery) and S4 (coverage-coefficient dependence).

S3: Response-path recovery rate (Thm. 4.3). We sweep the response-state margin γ_ϕ from 0.05 to 0.45 and measure the success rate of exact response-sequence recovery $\Pr[\widehat{z}^\pi = z^\pi]$ over 30 trials per margin level at fixed $n = 2000$. Figure 11a shows that recovery succeeds at near 100% across the range above the diagnostic threshold, consistent with the local margin mechanism in Theorem 4.3. Exhaustive enumeration is used only as a correctness oracle on instances small enough to enumerate. Causal forward recursion is the theorem-backed implementation.

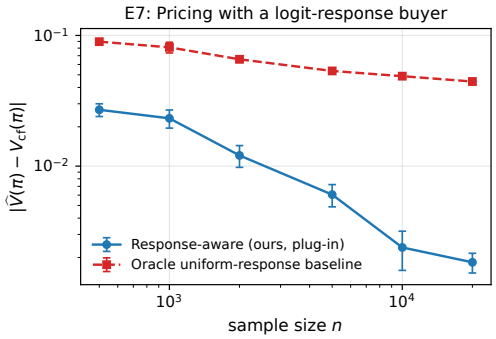
S4: Coverage coefficient dependence (Cor. C.10). We hold n fixed at 4000 and vary the logging-policy mixture skew so that the empirical coverage coefficient $\bar{C}$ varies by a factor of roughly three. Figure 11b plots the resulting oracle-response evaluation gap on log-log axes alongside a $\sqrt{\bar{C}}$ reference anchored at the smallest $\bar{C}$. The evaluation gap generally increases with $\bar{C}$, though more mildly than the worst-case $\sqrt{\bar{C}}$ envelope; at this fixed sample size the bonus term is partially offset by the natural plug-in error and the dependence is upper-bound rather than tight.

S5: Bernstein and Hoeffding bonuses (Prop. H.1). We compare the count-based bonuses in the main paper with the conditional Bernstein envelopes in Proposition H.1 as a function of horizon $H \in \{3, 5, 7, 10, 14\}$, holding $n = 4000$ fixed. We report the horizon-normalized evaluation gap $H^{-1}(V_{\mathrm{cf}} - \widehat{V}^-)$ to make rates comparable. Figure 12a shows the Bernstein bonus consistently yielding a smaller gap, with the gap between the two growing with H. This is a qualitative diagnostic only: Proposition H.1 requires valid simultaneous Bernstein envelopes and inverse-count remainders, which the plotting proxy does not itself establish.

S6: Pricing-shaped game with anchor-conditioned logit buyer (Appendix H.7, D1). We instantiate the framework on a pricing-shaped Markov game with an adaptive opponent and $|\mathcal{S}| = 3$ inventory levels, $|\mathcal{A}| = 3$ price levels, $|\mathcal{B}| = 3$ quantity choices, $H = 4$ rounds, and $d_R = 3$ buyer price-anchor regimes. The opponent chooses quantity with a logit rule of inverse temperature $\beta = 6$ applied to a payoff that depends on inventory and the public price-anchor regime z_h^π, not the learner's current realized price. Logging policies span low-, mid-, and high-price strategies plus two random mixers, ensuring coverage of all (s, z) cells. We compare two estimators of $V_{\mathrm{cf}}(\pi)$ on the same logged data: (a) an oracle-response plug-in estimator using the true target response path z^π, and (b) an oracle response-agnostic baseline that ignores response state and uses the uniform average of the true opponent rows, $g_{\mathrm{avg}}(b \mid s) := \frac{1}{d_R} \sum_z g(b \mid s, z)$. The second baseline is deliberately given population information to isolate the bias from pooling regimes; it is not behavior cloning from logged data. This is the same oracle-response pricing instance as E7 (Figure 8), not an independent confirmation. It is a theorem diagnostic aligned with Corollary C.10, not a claim that the pricing instance is solved without a supplied target response.

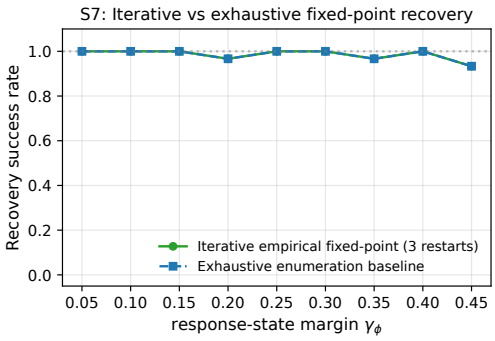


(a) S5: Comparison of Bernstein and Hoeffding bonuses across the horizon. The Bernstein gap is consistently smaller and the gap widens with H.

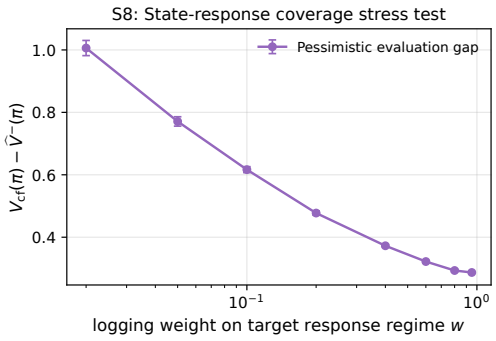


(b) S6: Same oracle-response pricing instance as E7, reprinted here next to S5. It is not a second independent run.

Figure 12. S5 (variance-aware sharpening) and S6 (response pooling in a pricing instance).



(a) S7: Iterative empirical fixed-point recovery (Picard iteration, 3 random restarts) tracks the exhaustive enumeration baseline across the entire margin range. This is empirical evidence only, not a finite-sample guarantee. Causal forward recovery is the method's theorem-backed algorithm.



(b) S8: State-response coverage stress test. As the logging mixture puts less mass w on the target response regime, the pessimistic evaluation gap grows monotonically.

Figure 13. S7 (iterative fixed-point recovery) and S8 (state-response coverage stress test).

Figure 12b reprints the E7 curve: absolute evaluation error $|\widehat{V}(\pi) - V_{\mathrm{cf}}(\pi)|$ averaged over 3 target policies and 5 trials. As in E7, the oracle-response estimator drops by an order of magnitude as n grows from 500 to 20,000, while the uniform-response baseline approaches an irreducible bias floor around 0.044.

S7: Picard-versus-enumeration stress test. This computational check compares two generic solvers for the same unique fixed point already computed by the forward pass. On the same instance family as S3, we run *Picard iteration*: starting from a random $z^{(0)} \in \mathcal{Z}^H$, we iterate $z^{(k+1)} = \widehat{T}_\pi(z^{(k)})$ until a fixed point is reached or 20 iterations elapse, with three random restarts. We compare the recovery rate to the exhaustive enumeration baseline of S3 across the margin sweep $\gamma_\phi \in [0.05, 0.45]$.

Figure 13a shows the two recovery curves overlapping at every margin level. Iterative search reached the exhaustive baseline in these trials within a small number of Picard steps. This comparison is descriptive, not a finite-sample claim. The causal forward theorem makes the solver comparison secondary: under the stated timing, response recovery follows one chronological path and requires no global fixed-point solver.

S8: State-response coverage stress test. We stress the role of target response-state support in a small game in which transitions are independent of (a, b), so the physical state process is fixed across logging mixtures. The logging data are then generated from a mixture of one policy that induces the target response regime and one policy that induces the opposite regime. Let w denote the logging weight on the target regime. In this construction, the state marginals are unchanged while the state-response coverage coefficient scales as $C_{\mathrm{sr}} \approx 1/w$ up to sampling noise. Because the logger mixture also changes learner and opponent action frequencies, this panel does not hold the full joint-occupancy coefficient C_{cf} fixed. Figure 13b plots the pessimistic evaluation gap $V_{\mathrm{cf}}(\pi) - \widehat{V}^{-}(\pi)$ against w at fixed $n = 1000$.

Figure 13b shows a monotone degradation as the target response regime is starved. This is a targeted diagnostic for the distinction emphasized in the state-response coverage condition of Proposition B.4 and Corollary C.10: even when the same physical states remain logged, conservative evaluation becomes markedly harder when the logs realize the wrong response regime.

Summary. The additional suite checks distinct parts of the theory: $1/\sqrt{n}$ scaling (S1), bias from response-agnostic pooling (S2), recovery away from response boundaries (S3), sensitivity to coverage (S4 and S8), a qualitative comparison of Bernstein and Hoeffding bonuses (S5), and agreement between two generic fixed-point solvers on small instances (S7). E7/S6 supplies the pricing-shaped example. These are controlled diagnostics on instances where the relevant quantities can be computed directly, not tests of the model on observational data.

I.5 Reproducibility checks

We validate the release scripts using the following checks:

1. repeated runs with the same seeds produce byte-identical scientific summaries; wall-clock measurements are stored separately;
2. every categorical row constructed by the diagnostics is stochastic, including the Bernoulli zero-count fallbacks;
3. forward recovery agrees with enumeration on every enumerated population and empirical instance;
4. every true value lies inside its reported outer confidence interval on the simultaneous confidence event;
5. the misspecification curve never exceeds B_ε beyond numerical tolerance;
6. all figure labels, sample sizes, trial counts, and error-bar definitions match the manuscript, including the separately labeled oracle-response benchmark.

The released run passed all checks. Repeating the primary run produced the same SHA-256 digest for the scientific summary; the verification value is included with the reproduction materials. The release also contains the underlying tables and vector figures for Figure 1 and the oracle-response pricing benchmark.

I.6 Reporting checklist

A numerical lower confidence bound is interpretable only together with the evidence for its assumptions. Table 9 summarizes the resulting choices. We recommend reporting five items. First, define each response state in domain terms and publish the maps $\phi_{1:H}$ before examining target-policy values. This separates a system specification from a response partition chosen to improve the result. Second, describe the logging policy set, its assignment probabilities, and the number of episodes assigned to every policy, and record the resulting assignment bound from Lemma B.15 when it applies. Policy labels are part of the observation model here; grouping superficially similar trajectories after data collection is not a substitute. Third, report state-joint-action counts $N_h(s, a, b)$ and state-response counts $N_h(s, z)$ on all target-reachable support. Zero counts flag unobserved cells, not necessarily zero population support; positive counts quantify uncertainty but do not prove a population probability floor. Fourth, report the estimated response path, its empirical distance to each decision boundary, and the answer under adjacent defensible coarsenings of $\mathcal{Z}$. If the estimated perturbation radius reaches a boundary, replace the point estimate by a set-valued analysis over compatible response paths or

Condition	Valid conclusion	Output
No defensible response abstraction	The logged law need not identify an undeployed response	Report non-identification; no certified ranking
Incomplete support; common target response across completions	Missing primitives can leave a range of compatible values	Outer interval under partial identification, Proposition 5.1
Incomplete support; response not stable across completions	Small occupancy differences may change later responses	$[0, H]$ unless an additional restriction justifies a narrower interval
Support holds; margin not protected	Exact label recovery is not certified	Abstain, or bound all compatible responses
All end-to-end assumptions and sample requirements hold	Simultaneous recovery and pessimistic evaluation apply	Lower confidence bounds and policy selection
Model-based lower confidence bound valid; induced-response discrepancy bounded	Actual value differs from modeled value by at most B_ε	Subtract B_ε from each lower confidence bound

Table 9. Choosing a valid output from the available assumptions. Failure of structural identification, support, or response stability cannot be repaired by reducing a statistical bonus.

abstain. Fifth, provide the selected policy's lower confidence bound, the curve $\varepsilon \mapsto \widehat{V}^{-}(\pi) - B_\varepsilon$, and the lower confidence bounds for every policy in the declared target class. This makes the simultaneous claim in Corollary 4.4, and any later enlargement of the class, auditable.

The report should keep three kinds of uncertainty separate. Sampling error is represented by the bonuses in Algorithm 1. Missing support is represented by the rectangular outer interval in Proposition 5.1. Disagreement between the response abstraction and deployment behavior is represented, when externally bounded, by Proposition 5.2. Combining them into a single standard error would hide their different meanings. More logged episodes shrink sampling error, but they cannot validate an extrapolated response rule or reveal a row that the logging design assigns zero probability. Opponent-response-kernel residuals on logged (s, z) cells may be reported as a calibration of ε inside observed regimes; they do not bound the discrepancy under an undeployed path.

How the audit changes the output. The first decision is structural. The analyst should record who chose the response states, when the maps $\phi_{1:H}$ were fixed, and what each boundary means operationally. If those choices were made after examining target values, the resulting point estimate is exploratory rather than confirmatory. If no response abstraction has independent support, Theorem 3.3 calls for a non-identification statement. This conclusion should not be softened because an estimator that pools response regimes has a small standard error.

The second decision concerns design and reachability. Assignment records can justify a policy-level probability floor, while mandated action randomization can justify lower bounds on particular state-joint-action cells. These records should be connected explicitly to the target-reachable ratios in (2) and (3). A large positive count is useful for estimating a covered row, but it is not evidence that every row with a zero count has zero population mass, nor does it prove that a rare positive cell has a prespecified lower bound. When the design does not justify a needed row, the analysis should move to Proposition 5.1; when compatible completions can also change the response path, it should report the unrestricted interval $[0, H]$ unless an external stability argument is supplied.

The third decision is whether the recovered path is separated from the response boundaries. For every target, report the perturbation radius used by the recovery theorem beside the smallest empirical boundary slack. A point estimate is appropriate only when the former is smaller than the latter on the simultaneous event. Otherwise, the analysis should range over compatible response paths or abstain. Passing this check justifies the transition to ordinary pessimistic evaluation; it does not retroactively validate the response abstraction or the logging design.

Finally, preserve the target class and every data-dependent choice used in selection. Report all simultaneous lower confidence bounds, not only the winning policy, and state whether any policy, binning rule, or deployment threshold was added after inspecting the results. If the class changes, the

union bound and selection statement must be recomputed for the enlarged class. These records make the analysis reproducible and show exactly which conclusion follows from each layer of evidence.