SAFE TO SWITCH: TARGET-COVERED OFF-POLICY EVALUATION VIA POLICY REPLACEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Offline reinforcement learning enables policy learning from existing data when further interaction is costly or unsafe. However, target occupancy coverage and linear realizability of the target's action values alone do not guarantee sample-efficient off-policy evaluation. We show that target coverage suffices under replacement-policy linear realizability, using safe partial corrections to logged returns. Replacement policies retain a drawn logger action or replace it with a fresh target draw, with arbitrary probabilities across stages, states and drawn actions. At every stage, every replacement policy's action values must be linear on all state–action pairs in known features, with uniform feature and parameter norm bounds. For a fixed target, our estimator achieves error at most ε with probability at least $1-\delta$ using $\tilde{O}(C_e H^6 d/\varepsilon^2)$ iid complete trajectories from one unknown fixed Markov logger, without logger probabilities or model oracles. Here C_e bounds target-to-logger occupancy ratios, H is the horizon and d the feature dimension. The sample bound has logarithmic norm dependence and no state or action cardinality factor. Separate families establish strict representation and coverage relaxations: replacement-policy values admit a two-dimensional feature map while all-policy dimension is unbounded, and finite target coverage coexists with infinite mixture coverage. The finite-sample guarantee and both separations are formalized in Lean 4.

1 INTRODUCTION

Offline reinforcement learning is useful when experience is available but further interaction is costly or unsafe (Duan et al., 2020). Off-policy evaluation (OPE) estimates a target policy's return from experience collected by a behavior policy, or logger, supporting assessment before deployment (Jiang & Li, 2016). If the target always takes a right branch, a lack of data on the left branch need not prevent evaluation. With function approximation, however, coverage is only part of the question: the features must also express the values used to turn logged experience into a target estimate. Jia et al. (2024, Theorem 6) show that target-value realizability and polynomial all-policy concentrability can still require exponentially many complete trajectories in the horizon H to attain expected error of order $1/H$. In contrast, Tkachuk et al. (2026, Theorem 1) obtain polynomial-sample OPE under all-policy linear realizability, all-policy coverage and known logger probabilities. We ask which continuation values must be represented for target occupancy coverage to suffice when the logger is unknown.

Complete trajectories offer a useful starting point. The return following a logged state and action has the logger's action value as its conditional mean, giving a regression problem with an actual policy-value mean. The difficulty is changing the continuation toward the target while keeping this property. Fitted Bellman backups can leave the feature span even when all actual policy values are linear. Bellman completeness guarantees closure of those backups: Duan et al. (2020, Assumptions 1–2 and Theorem 2) give finite-sample guarantees with a function class closed under the target transition operator. We instead construct regression labels whose means retain a policy interpretation.

Suppose we could add to a logged reward the exact advantage of replacing its action with a fresh target action under the same future continuation. In expectation, this would make that decision follow the target. Adding a fraction of the advantage has an equally simple meaning: replace the action with that probability and otherwise keep it. An estimated advantage need not permit this

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.

interpretation, since it can overshoot or have the wrong sign. Our correction takes the point in its confidence interval closest to zero. When the interval is valid, this correction lies between zero and the true advantage and hence corresponds to a partial replacement. An interval containing zero yields no correction. Different observed actions have different prediction uncertainties, so their replacement probabilities can differ. Repeating the operation backward along a trajectory produces a family of logger-to-target replacement policies. We require every replacement policy's action values at every stage to be linear on all state–action pairs in known features, with uniform feature and parameter norm bounds. All-policy realizability implies this condition with the same norm budget. The algorithm adds numerical corrections to observed returns; replacement policies interpret their means. The key distinction is between the policies whose values need representation and the distributions under which errors are controlled. Replacement realizability supplies regression means, while target coverage bounds the uncorrected advantage along the target's path; intermediate policies need not have covered occupancies.

Our contributions are threefold. **1) Statistical and computational guarantee.** Under these conditions, an explicit backward ridge/shrink estimator evaluates a fixed target with error at most ε with probability at least $1-\delta$ using $\tilde{O}(C_e H^6 d/\varepsilon^2)$ iid complete trajectories from one unknown fixed Markov logger (Theorem 3.1 and Corollary 3.2). The sample bound has logarithmic norm dependence and no state or action cardinality factor. The estimator requires no logger probabilities or model oracles and uses polynomially many exact real-arithmetic operations (Section C.5); finite-precision guarantees remain open. **2) Strict relaxations of realizability and coverage.** One family realizes replacement-policy values with two features at fixed horizon, norms and target coverage, while the dimension required for all-policy realization grows without bound (Proposition 5.1). A separate all-policy-realizable family has fixed dimension and norm bounds, finite target coverage and infinite mixture coverage (Proposition 5.2). **3) Formal verification.** The finite-sample guarantee and both separations are formalized in Lean 4 under the complete-trajectory observation law (Section F).

2 From trajectories to replacement policies

Model and observations. Consider a finite layered MDP with horizon $H \geq 1$, state layers $\mathcal{S}_1, \ldots, \mathcal{S}_{H+1}$, finite nonempty action set $\mathcal{A}$, and fixed known initial state s_1. At stage h, the next state has law $P_h(\cdot \mid s, a)$ on $\mathcal{S}_{h+1}$. The reward has an arbitrary Borel law on $[0,1]$ with mean $r_h(s,a)$. Given the complete past through the current state and action, reward and next state have this product law. Terminal value is zero. A policy $\pi = (\pi_h)_{h=1}^H$ is any stochastic memoryless action kernel. Let q_h^π, v_h^π, and $J(\pi) = v_1^\pi(s_1)$ denote its action value, state value, and expected total reward. Write

$$v_h^\pi(s,a) = \mathbb{P}_\pi(S_h = s, A_h = a), \qquad d_h^\pi(s) = \sum_a v_h^\pi(s,a).$$

The target e is known and fixed independently of the data. The dataset contains N iid complete trajectories $Z^j = (S_1^j, A_1^j, R_1^j, \ldots, S_H^j, A_H^j, R_H^j, S_{H+1}^j)$ from one fixed unknown logger b. Dependence within each trajectory is retained. Set $\mu_h = v_h^b$.

2.1 Turning a logged return into a different policy value

Consider two stages. At the last stage, $r_2(s,a) = q_2^\pi(s,a)$ for every policy π. Regressing $R_1 + R_2$ at the first stage would learn q_1^b, since the second action follows the logger. To change this continuation, compare a logged action a at state s with a fresh target action. Write $\Delta_2(s,a) = \sum_{a'} e_2(a' \mid s) r_2(s,a') - r_2(s,a)$. Adding this exact advantage to R_2 would give a target-continuation mean.

Each complete trajectory records the actual second state and reward. The first-stage regression can therefore correct the observed suffix, while averaging over paths retains the MDP's transition law.

A partial correction is useful as well. If $g_2(s,a) = \lambda(s,a)\Delta_2(s,a)$ for $\lambda(s,a) \in [0,1]$, then

$$r_2(s,a) + g_2(s,a) = (1 - \lambda(s,a)) r_2(s,a) + \lambda(s,a) \sum_{a'} e_2(a' \mid s) r_2(s,a'). \tag{1}$$

The right side is the expected reward of keeping the drawn action with probability $1 - \lambda(s,a)$ and redrawing from the target otherwise. Average over the logger's draw of a: the corrected reward now

has the state value of this replacement policy as its mean. Consequently, $R_1 + R_2 + g_2(S_2, A_2)$ has its first-stage action value as the regression mean. The trajectory is still logged data; changing the label changes which policy its conditional mean represents.

The advantage is unknown, so the estimator must obtain a partial correction from a confidence interval. Choosing the endpoint closest to zero ensures that, on a valid interval, the correction has the true advantage's sign and no larger magnitude. An interval containing zero yields no correction. For an advantage estimate m with confidence radius c, define

$$\mathsf{s}_H(m, c) = \mathrm{clip}_{[-H,H]}(\mathrm{sign}(m)(|m| - c)_+)\,, \tag{2}$$

where $(z)_+ = \max\{z, 0\}$. The outer clipping restricts the correction to the possible advantage range.

Lemma 2.1 (Safe scalar correction)**.** *If $c \geq 0$, $|m - \Delta| \leq c$ and $|\Delta| \leq H$, then $g = \mathsf{s}_H(m, c)$ satisfies $g = \lambda\Delta$ for some $\lambda \in [0, 1]$ and $|\Delta - g| \leq 2c$.*

The lemma preserves a policy interpretation while controlling the uncorrected advantage through statistical uncertainty. Here, "safe" refers to that interpretation, not a guarantee of policy improvement or safe deployment. Negative advantages follow the same rule, moving the conditional mean downward when replacing the action would reduce its value. Figure 1 illustrates the three interval configurations; the proof is in Section A.1.

The fitted predictions serve to determine how much advantage can be corrected. The regression label still starts with observed rewards. On valid intervals, the added amount has a policy interpretation through Equation (1), even though the fitted function has sampling error. Consequently, the next regression mean is an actual policy value. This is the property that lets policy-value realizability support the recursion without assuming closure under arbitrary fitted Bellman backups. The three-stage example in Section D.4 exhibits bias in both return averaging and population minimum-norm FQE despite all-policy realizability and target coverage. It concerns that specified FQE rule and is outside the Lean certificate.

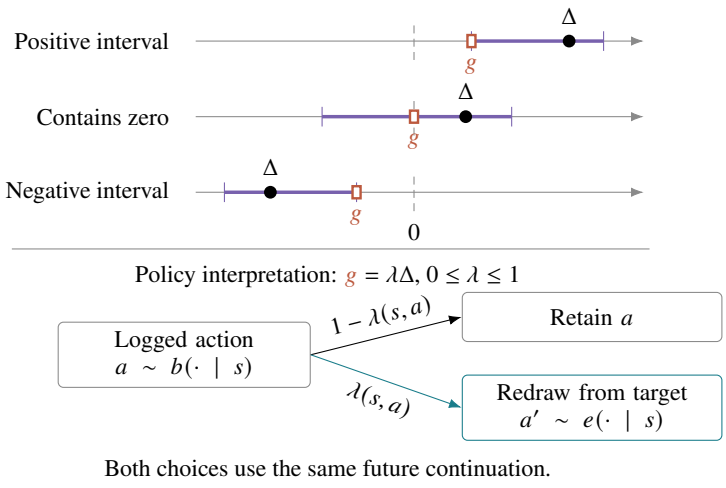

Figure 1: **From a confidence interval to a replacement policy.** The correction g (square) is the interval point closest to zero. For a valid interval, it lies between zero and the true advantage Δ (circle). Thus $g = \lambda\Delta$ for a probability λ and has the conditional-mean interpretation shown below. The estimator adds g to its label; the diagram explains which policy the corrected mean represents.

Why the drawn action matters. At a last-stage state, suppose the three action reward means are $(0, 1/2, 1)$, the logger is uniform and the target chooses action three. For illustration, valid confidence intervals may yield corrections $(1/2, 0, 0)$: the first action's advantage is confidently positive, while the second action's interval contains zero. These corrections correspond to replacing half the draws of action one with action three and retaining action two. The resulting action distribution is $(1/6, 1/3, 1/2)$, with mean reward $2/3$, between the logger value $1/2$ and target value one. A statewise mixture $(1 - t)b + te$ cannot produce this distribution, since it assigns equal probabilities

to actions one and two. Allowing the replacement to depend on the drawn action is therefore part of the correction mechanism.

2.2 The policy class induced by partial replacements

The same interpretation extends backward through any number of stages. Fix a continuation ρ already constructed at later stages. Replace r_2 in Equation (1) by q_h^ρ: the advantage is now $\Delta_h^{\rho,e}(s,a) = \sum_{a'} e_h(a' \mid s) q_h^\rho(s,a') - q_h^\rho(s,a)$. It compares a logger action and a target action with the *same* future policy. A partial correction changes the current action kernel while preserving that continuation. Repeating the step gives the following class.

For any array $\lambda_h(s,a) \in [0,1]$, define

$$T_{b,e}(\lambda)_h(a' \mid s) = \sum_a b_h(a \mid s)\big[(1-\lambda_h(s,a))\mathbf{1}\{a'=a\} + \lambda_h(s,a) e_h(a' \mid s)\big]. \tag{3}$$

Let $\mathcal{R}(b,e) = \{T_{b,e}(\lambda) : 0 \le \lambda \le 1\}$. These are stochastic memoryless policies, with weights varying freely across stages, states and actions. The zero and unit arrays give b and e; action-dependent arrays include the three-action example. The class is closed under splicing stages and is defined before sampling.

With three stages, first fit the last reward and construct g_3. For this interpretation, fix corrections that are fractions of their true continuation advantages. The stage-two label $R_2 + R_3 + g_3(S_3, A_3)$ has the action value of the stage-three replacement as its mean. Comparing stage-two actions under this continuation gives g_2; the first-stage label is the complete return plus $g_2 + g_3$ at the observed future pairs. Each step changes one action kernel, retaining the later replacements within $\mathcal{R}(b,e)$.

Changing the stage-h kernel does not change q_h^ρ, because an action value holds the current action fixed. Thus a stage-h correction enters only *earlier* regressions, whose labels correct the future actions. Conditional expectation along the complete logged suffix gives the invariant.

Lemma 2.2 (Corrected suffix mean)**.** *Fix a suffix policy $\rho = T_{b,e}(\lambda)$ and corrections $g_t = \lambda_t \Delta_t^{\rho,e}$ for every $t > h$, all using this same ρ. Then*

$$\mathbb{E}_b\Bigg[\underbrace{\sum_{t=h}^{H} R_t}_{\text{observed suffix}} + \underbrace{\sum_{t=h+1}^{H} g_t(S_t, A_t)}_{\text{future corrections}} \,\Bigg|\, S_h, A_h\Bigg] = q_h^\rho(S_h, A_h) \quad \textit{almost surely}. \tag{4}$$

For fixed safe corrections, each regression mean is thus a replacement policy's action value in the original MDP. In the estimator, the proposed corrections depend on the data: an estimated advantage may permit a large correction at one pair while uncertainty leaves another close to the logger. Earlier regressions must represent every safe continuation these corrections can induce.

A common feature map and parameter budget must handle all these choices. Requiring every replacement value to be linear supplies a witness for each fixed safe continuation. The witness may vary, while the features and norm budget stay fixed. This motivates the following population assumption; Section 4 handles learned corrections and invalid intervals through an auxiliary recursion before applying the fixed-policy identity.

Assumption 1 (Replacement-policy linear realizability)**.** A known feature map $\phi_h : \mathcal{S}_h \times \mathcal{A} \to \mathbb{R}^d$, $d \ge 1$, has $\|\phi_h(s,a)\|_2 \le L_\phi$ for public bounds $L_\phi, L_\theta \ge 0$. For every $\rho \in \mathcal{R}(b,e)$ and stage $h \le H$, there exists θ_h^ρ such that

$$q_h^\rho(s,a) = \phi_h(s,a)^\top \theta_h^\rho \quad \text{for every } (s,a) \in \mathcal{S}_h \times \mathcal{A}, \qquad \|\theta_h^\rho\|_2 \le L_\theta.$$

Only the continuation policy is restricted; the representation still holds at every valid current state–action pair, including unlogged actions. All-policy realizability with the same norm budget implies this assumption. Realizing only the logger and target does not imply it (Section E.6).

At the root, averaging $q_1^\rho(s_1, \cdot)$ under the known target chooses the target's first action exactly. Only the future decisions remain partially replaced. This explains why the estimator constructs corrections for stages two through H, but outputs a target-action average at stage one.

The learner computes the corrections, not the probabilities λ or the policies ρ. The logger probabilities in Equation (3) enter through the distribution of the observed actions, so computing the labels requires no knowledge of b.

Assumption 2 (Target coverage)**.** A public finite bound $C_e \geq 1$ satisfies $\nu_h^e(s,a) \leq C_e \mu_h(s,a)$ for every decision-layer pair.

Equivalently, the maximum target occupancy ratio is at most C_e.[1] Summing over actions gives $d_h^e \leq C_e d_h^b$. Only the target occupancy is required to be covered.

The learner receives trajectories, s_1, features, target probabilities and bounds $(H, d, L_\phi, L_\theta, C_e)$. It can sum over the finite action set at observed states. It receives no dynamics, reward law, behavior probabilities, occupancy ratios, policy witness or optimization oracle. Computation is in exact real arithmetic and may depend on $|\mathcal{A}|$ even though the sample count does not.

3 Estimator and guarantee

Algorithm 1 implements this construction using ridge regression and feature-dependent confidence widths. Each fit estimates the continuation formed by the later corrections; a target-action average then determines the next correction or, at the root, the final output.

Table 1: **Complete-trajectory OPE guarantees.** The second row isolates the improvement under all-policy realizability; the third also relaxes the representation assumption.
Our sample counts have logarithmic norm dependence.

Result	Policies realized	Coverage	Count, up to logs	Knows b
Tkachuk et al. (2026), Thm. 1	All policies	C_{all}	$C_{\text{all}}^5 H^7 d^3/\varepsilon^2$	Yes
Corollary 3.3	All policies	C_e	$C_e H^6 d/\varepsilon^2$	No
Corollary 3.2	Replacements	C_e	$C_e H^6 d/\varepsilon^2$	No

Setting $C_e \leq C_{\text{all}}$ gives a direct comparison under the baseline assumptions. The main theorem also admits coverage distributions and representations outside that model class, as Section 5 shows.

To keep norm dependence explicit, use the normalized features and bound

$$F = \max\{1, L_\phi\}, \qquad B = \max\{1, FL_\theta\}, \qquad \psi_h = B\phi_h/F. \tag{5}$$

Then $\|\psi_h\|_2 \leq B$ and every replacement policy has a parameter of norm at most one. The normalization lets us use a common regression rule across norm budgets; its effect on the sample count is logarithmic.[2]

Divide a fixed prefix of the data into H independent blocks $D_H, D_{H-1}, \ldots, D_1$, each containing n complete trajectories. Stage h uses only block D_h for fitting. The split makes each regression independent of the fits used to construct its labels. For confidence $0 < \delta < 1$, the radius is

$$\beta = H + 3H\sqrt{2H \log(8nH^2/\delta)}. \tag{6}$$

For a positive definite V, write $\|x\|_{V^{-1}} = \sqrt{x^\top V^{-1} x}$.

Corrections are stored as functions of their fitted vector and design. They need only be evaluated at states occurring in earlier-stage blocks; there is no enumeration of the state space. Empty sums at $h = H$ are zero, so the first regression uses terminal rewards. Actionwise clipping keeps predictions and the output in $[0, H]$ on every dataset. The confidence widths reflect what each block has learned about the queried feature directions. Wide intervals produce no correction; narrow intervals permit a

[1] We use the conventions $0/0 = 0$ and $x/0 = \infty$ for $x > 0$.

[2] Details, including zero norm bounds, are in Section C.4.

Algorithm 1: **Safe to Switch: backward regression with safe corrections**

Input: Independent blocks $D_h = (Z^{h,j})_{j=1}^n$ of complete trajectories; normalized features ψ; target e; initial state s_1; radius β from Equation (6).

Output: Estimate $\widehat{J}$ of the target return.

for $h = H, H-1, \dots, 1$ **do**

  $\psi_h^j \leftarrow \psi_h(S_h^{h,j}, A_h^{h,j})$ for $j = 1, \dots, n$ in block D_h ▷ *Build labels*

  $Y_h^j \leftarrow \sum_{t=h}^{H} R_t^{h,j} + \sum_{t=h+1}^{H} g_t(S_t^{h,j}, A_t^{h,j})$ for $j = 1, \dots, n$

  $V_h \leftarrow I + \sum_{j=1}^{n} \psi_h^j (\psi_h^j)^\top$ ▷ *Fit values and widths*

  $\widehat{\theta}_h \leftarrow V_h^{-1} \sum_{j=1}^{n} \psi_h^j Y_h^j$

  Define $\widehat{q}_h(s,a) \leftarrow \mathrm{clip}_{[0,H]}(\psi_h(s,a)^\top \widehat{\theta}_h)$

  Define $w_h(s,a) \leftarrow \min\{1, \|\psi_h(s,a)\|_{V_h^{-1}}\}$

  Define $\widehat{q}_h^e(s) \leftarrow \sum_a e_h(a \mid s) \widehat{q}_h(s,a)$ ▷ *Target averages and corrections*

  Define $\bar{w}_h^e(s) \leftarrow \sum_a e_h(a \mid s) w_h(s,a)$

  if $h \geq 2$ **then**

    Define $g_h(s,a) \leftarrow \mathsf{s}_H\big(\widehat{q}_h^e(s) - \widehat{q}_h(s,a), \beta[\bar{w}_h^e(s) + w_h(s,a)]\big)$

  end if

end for

return $\widehat{q}_1^e(s_1)$

larger adjustment toward the target continuation. At fixed n, the algorithm does not use C_e; coverage determines the accuracy guarantee and the sample budget.

The two terms in a correction's width reflect its two predictions: the value of the observed action and the target-weighted action value at the same state. Averaging the latter widths with the target probabilities lets rare target actions contribute in proportion to their probability. Both predictions use the same fitted continuation, which incorporates all later corrections. The safe-correction lemma bounds the advantage left uncorrected, the local source of target-value error.

Theorem 3.1 (Trajectory evaluation under target coverage)**.** *Under Assumptions 1 and 2, for any $n \geq 1$, Algorithm 1 uses Hn complete trajectories and satisfies, with probability at least $1-\delta$,*

$$|\widehat{J} - J(e)| \leq (4H-3)\beta\sqrt{\frac{4C_e}{n}\left[d\log(1+nB^2/d) + \log(2H/\delta)\right]}. \tag{7}$$

For fixed public inputs, the estimator is a total measurable function of observations, uniform over compatible MDPs, bounded Borel reward laws, and loggers.

Corollary 3.2 (Explicit public sample count)**.** *For $0 < \varepsilon \leq H$, the explicit prescription in Equation (23) defines $N_\star = Hn_\star$ using only $(H, d, B, C_e, \varepsilon, \delta)$. For every supplied count $N \geq N_\star$, applying Algorithm 1 to the fixed $N_\star$-trajectory prefix gives $|\widehat{J} - J(e)| \leq \varepsilon$ with probability at least $1-\delta$. In particular,*

$$N_\star = \tilde{O}\left(\frac{C_e H^6 d}{\varepsilon^2}\right). \tag{8}$$

The tilde suppresses only logarithms in the displayed public parameters and $1/\delta$. All constants and rounding are specified in Section C.

Corollary 3.3 (Familiar sufficient conditions)**.** *The same guarantee holds if every stochastic policy supported on $U_h(s) = \mathrm{supp}\, b_h(\cdot \mid s) \cup \mathrm{supp}\, e_h(\cdot \mid s)$ has the global representation and norm bound of Assumption 1. In particular, all-policy realizability suffices. No support set or behavior probability is an estimator input.*

A replacement never produces an action outside U, so the support condition suffices and is convenient for verifying model examples.

4 WHY ONLY TARGET COVERAGE IS NEEDED

The proof uses the policy interpretation to identify each regression mean, then bounds the advantage left uncorrected along the target's path. An auxiliary recursion makes the first step valid before confidence events are invoked. Full proofs are in Sections A and B.

4.1 A REALIZABLE REGRESSION ON EVERY DATASET

A regression confidence bound requires a realizable conditional mean. Yet later corrections help form that mean, and their interpretation so far has relied on valid confidence intervals. Simply conditioning on all intervals being valid would change the sampling law used by the next regression. We resolve this dependence by giving the construction a policy interpretation even on datasets where a proposed interval fails.

The public rule computes g_h. On the same blocks, an auxiliary recursion builds labels from its own later corrections, fits the same ridge regression and proposes $\widetilde{g}_h$. It forms the analysis-only correction g_h^* by projecting the proposal onto the segment from zero to the exact continuation advantage:

$$g_h^*(s,a) = \mathrm{proj}_{[\min(0,\Delta_h(s,a)),\max(0,\Delta_h(s,a))]}\,\widetilde{g}_h(s,a). \tag{9}$$

Here $\rho^{>h}$ is the policy after processing the later blocks and $\Delta_h = \Delta_h^{\rho^{>h},e}$. Projection ensures a partial replacement on every dataset.

Writing $\mathcal{F}_{>h} = \sigma(D_{h+1},\dots,D_H)$, Lemma 2.2 gives

$$\mathbb{E}_b[Y_h^* \mid S_h, A_h, \mathcal{F}_{>h}] = q_h^{\rho^{>h}}(S_h, A_h).$$

Thus fixing the later blocks fixes a realizable regression mean, before any confidence event. Unknown advantages enter only this proof construction.

Subtracting continuation values decomposes each centered corrected suffix into at most H martingale increments, each bounded by $3H$, giving variance proxy $9H^3$ (Section A.3). Conditioning on the trajectory's past handles within-trajectory dependence; independence is used between trajectories and between blocks.

4.2 AGREEMENT WHERE THE ESTIMATOR USES A CORRECTION

To connect the auxiliary process to the estimator, observe that a stage-h correction enters earlier regressions only at their observed stage-h pairs. Those queries are independent of the stage-h block and its auxiliary suffix. We can therefore control the advantage interval at each such query using a scalar prediction bound. On a valid interval, Lemma 2.1 says the proposal already lies in the safe segment, so the projection changes nothing.

Across all earlier blocks there are at most $nH(H-1)/2$ queries. A union bound over them costs only a logarithm of the sample size and horizon. Backward induction then equates the two outputs: their terminal labels are identical; agreement of later corrections at queried pairs equates each preceding block's labels, fits and eventual root prediction. No agreement is needed at states that never enter these labels.

Standard ellipsoidal confidence controls every prediction direction at once, adding a dimension factor to the squared radius. Scalar confidence at independent queries avoids this extra factor. A width-weighted exponential bound controls target-action averages directly (Section A.4), avoiding a union bound over target actions.

4.3 TRANSFER UNDER TARGET-WEIGHTED DISTRIBUTIONS

Return to the two-stage example and suppose for the moment that the root regression is exact. At a second-stage state, the gap between the target value and the replacement value is the logger average

of $\Delta_2(s,a) - g_2(s,a)$. The root target action reaches that state according to d_2^e, so the return error averages this gap over target states. It never averages over the states visited by the replacement policy itself. For the auxiliary replacement policy ρ with $\rho_1 = e_1$, the same decomposition gives

$$J(e) - J(\rho) = \sum_{h=2}^{H} \mathbb{E}_{S_h \sim d_h^e,\, A_h \sim b_h(\cdot|S_h)} [\Delta_h(S_h, A_h) - g_h^*(S_h, A_h)]. \tag{10}$$

Section B.2 derives this identity by telescoping the Bellman recursion. After coupling, the estimator's error consists of this policy gap and the root prediction error. An advantage compares a target action with a logger action at the same target state, so only two joint distributions enter the bound. Both are dominated by the logging occupancy:

$$d_h^e(s) e_h(a \mid s) \le C_e \mu_h(s,a), \qquad d_h^e(s) b_h(a \mid s) \le C_e \mu_h(s,a). \tag{11}$$

The first inequality is target coverage itself; the second follows by summing that assumption over actions before multiplying by the logger probability. A replacement may visit poorly covered pairs, yet corrected logged suffixes still provide its action-value means. Error transfer uses only the two target-weighted distributions above, separating the policies that need representation from those that need coverage.

To complete the bound, the expected squared capped leverage under μ_h is of order d/n, up to logarithms. Coverage and Cauchy–Schwarz turn this into a target-weighted width of order $\sqrt{C_e d/n}$. This argument requires no inverse of a population covariance matrix and no positive lower bound on its eigenvalues. Combining the widths with scalar confidence bounds the uncorrected advantages and proves Theorem 3.1.

The H^6 factor comes from the H^3 noise proxy, squaring a sum over H stages, and allocating H independent blocks. Linear dimension dependence comes from combining integrated leverage's d/n with a scalar radius that has no dimension factor; query confidence adds only logarithms. These are sufficient dependencies; their joint optimality remains open.

5 What the assumptions allow

The replacement condition follows from familiar model structure. Two separate families then show that restricting representation and coverage can each change which problems the theorem covers.

A structural sufficient condition. Linear MDPs in the supplied features satisfy all-policy realizability. State aggregation gives a further sufficient condition: suppose rewards and next-block probabilities factor linearly through those features, with bounded reward and transition factors. Within each block containing multiple states, suppose all actions used by the logger or target anywhere in that block have the same reward mean and next-block distribution. Their values then agree for every supported continuation, even when replacement probabilities vary across the concrete states. This gives support and replacement realizability; singleton blocks require no action equivalence. Transitions within a destination block may still differ, allowing concrete transition rank to exceed the feature dimension. Proposition E.1 states the precise condition and norm bound. It applies to the two-feature family below. Relations to low-rank MDPs, Bellman completeness and Bellman rank are developed in Section E.3.

5.1 Two features can suffice when all-policy dimension diverges

Proposition 5.1 (Strict representation separation)**.** *For every $m \ge 3$ and $u \in [1/4, 1/2]$, there is an instance satisfying the support condition in Corollary 3.3 with $H = d = 2$, $L_\phi = 1$, $L_\theta = 2$, $C_e = 4$ and $J(e) = u$. The features and target are independent of u. Every feature map representing all policies' action values has dimension at least m. For the supplied map, failure occurs on logged current pairs, for every parameter norm bound.*

Positive root action i leads to branch x_i. Every logger- or target-supported continuation has reward mean u at each branch, so two features represent their values and the known root reward offsets. An unrestricted policy can instead take an unlogged action at any one branch, changing just the corresponding root value by $1 - u$. The m independent changes force all-policy dimension at least

m. The logger return is $u + 1/8$, so merely averaging logged returns does not recover $J(e)$. The complete construction and rank proof are in Section D.1.

The representation gap uses missing logger support. With full support, replacement realizability recovers all-policy representation in the same features, potentially at a larger parameter norm budget (Proposition E.2 and Section E.5).

5.2 Target coverage can coexist with infinite mixture coverage

Let C_{mix} be the largest occupancy ratio over statewise mixtures of the logger and target. Always $C_e \le C_{\text{mix}} \le C_{\text{all}}$. A mixture can follow the logger into a branch the target never visits, then take the target's actions there. Target coverage need not cover these combinations.

Proposition 5.2 (Coverage separation without Bellman completeness)**.** *For every $m \ge 2$, $p \in [0, 1/2]$ and $u \in [1/4, 1/2]$, there is a finite MDP with $H = d = 3$, feature and all-policy parameter norms at most one, and $J(e) = u$, such that $C_e = 4$, while $C_{\text{mix}} = 2/p$ for $p > 0$ and $C_{\text{mix}} = \infty$ for $p = 0$. Bellman completeness fails for the target operator in the supplied feature class. The sample count in Corollary 3.2 is independent of m, p, u.*

In this family the target takes a right branch. The logger also enters a left branch, where it takes a particular action with probability p; a mixture can take that action always. Its ratio grows as $2/p$ while the target's ratio stays four. The features represent all actual policy values, but arbitrary linear continuations can distinguish successors with the same true value and thereby break Bellman closure. Section D.2 gives the model and proof.

6 Related work and discussion

Trajectories and policy-value realizability. Tkachuk et al. (2024; 2026) use complete trajectories and all-policy realizability, building on the skipping constructions of Weisz et al. (2023). Our partial replacements preserve policy-value means through corrected returns; Table 1 gives the direct OPE comparison. The lower bounds of Jia et al. (2024) use target-only realizability, and those of Foster et al. (2022) use non-trajectory data. These settings leave open the combination studied here.

Feature-dynamics coverage. Amortila et al. (2026) analyze discounted LSTDQ under target-only realizability using iid transition tuples. Their feature-dynamics coefficient measures propagation through a compressed model. Under target Bellman completeness, their Proposition 5 identifies it with true feature-occupancy energy, and target domination gives $C_{\text{FD}} \le C_e$. Without completeness, our separating family has fixed dimension, all-policy norm bounds and $C_e = 4$, yet C_{FD} diverges with both population matrices invertible (Section H). Compression propagates mass into a rare branch the target never visits. Duan et al. (2020) use stronger closure structure; Perdomo et al. (2023) characterize linear estimators, whereas our corrections depend nonlinearly on the data.

Weight learning and corrected returns. Single-policy coverage also appears in weight-learning and primal-dual methods with additional weight or value-class conditions (Uehara et al., 2020; Zhan et al., 2022; Ozdaglar et al., 2023). Q-MMR (Li & Jiang, 2026, Theorem 1) gives a trajectory-based certificate under target realizability using empirical matching losses and learned weight norms. Its population bound (Corollary 10) assumes bounded generalized leverage and exact minimization, and depends on projected dynamics and population weight norms. Our rate follows from replacement realizability and target occupancy coverage through actual-policy regression means. Sequential doubly robust OPE corrects trajectory returns using importance ratios (Jiang & Li, 2016, Eq. (10)). Our corrections instead interpolate between policy continuations, with target coverage controlling the advantage left uncorrected.

Scope and further questions. Implementation costs and sample calibration are in Sections C.2 and C.5. The closed-form counts certify statistical learnability under the stated assumptions; they do not establish the estimator's practical sample efficiency. Sharp horizon dependence and a matching joint rate remain open (Section G). Reusing trajectories across stages may save the factor H paid by independent blocks, but requires a new dependence analysis for fitted corrections and regression labels.

REPRODUCIBILITY STATEMENT

Sections 2 and 3 specify the observation model, assumptions and public estimator. Complete proofs and boundary constructions appear in Sections A to E, G and H. The Lean code includes verification scripts and numerical diagnostics; the code documentation records the dependency versions and verification procedure. Section F gives the formal correspondence and verification scope.

AI USE STATEMENT

We used AI tools to assist with drafting and revising the manuscript, including improving the clarity, organization, and consistency of the exposition. GPT6-Astra, Fable5.1, and GLM5.3 assisted with developing and refining the Lean formalization of the proofs. The authors take responsibility for the final manuscript and the Lean code.

REFERENCES

Philip Amortila, Audrey Huang, Akshay Krishnamurthy, and Nan Jiang. A unifying view of coverage in linear off-policy evaluation. In *International Conference on Learning Representations*, 2026. arXiv:2601.19030v1.

Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pp. 2701–2709. PMLR, 2020.

Dylan J. Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. In *Conference on Learning Theory*, pp. 3489–3489. PMLR, 2022.

Zeyu Jia, Alexander Rakhlin, Ayush Sekhari, and Chen-Yu Wei. Offline reinforcement learning: Role of state aggregation and trajectory data. In *Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 2644–2719. PMLR, 2024.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 652–661. PMLR, 2016.

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

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

Xiang Li and Nan Jiang. Q-MMR: Off-policy evaluation via recursive reweighting and moment matching. arXiv:2605.06474v2, 2026.

Cong Ma, Banghua Zhu, Jiantao Jiao, and Martin J. Wainwright. Minimax off-policy evaluation for multi-armed bandits. *IEEE Transactions on Information Theory*, 68(8):5314–5339, 2022. doi: 10.1109/TIT.2022.3162335. URL https://arxiv.org/abs/2101.07781v1. Theorem numbering refers to arXiv:2101.07781v1 (2021).

Asuman E. Ozdaglar, Sarath Pattathil, Jiawei Zhang, and Kaiqing Zhang. Revisiting the linear-programming framework for offline RL with general function approximation. In *International Conference on Machine Learning*, pp. 26769–26791. PMLR, 2023.

Juan C. Perdomo, Akshay Krishnamurthy, Peter Bartlett, and Sham Kakade. A complete characterization of linear estimators for offline policy evaluation. *Journal of Machine Learning Research*, 24(284):1–50, 2023.

Volodymyr Tkachuk, Gellért Weisz, and Csaba Szepesvári. Trajectory data suffices for statistically efficient learning in offline RL with linear q^π-realizability and concentrability. In *Advances in Neural Information Processing Systems*, volume 37, pp. 83268–83313, 2024. arXiv:2405.16809.

Volodymyr Tkachuk, Csaba Szepesvári, and Xiaoqi Tan. Trajectory data suffices for statistically efficient policy evaluation in fixed-horizon offline RL with linear q^π-realizability and concentrability. In *Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 6372–6405. PMLR, 2026. arXiv:2510.03494v2.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and Q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pp. 9659–9668. PMLR, 2020.

Gellért Weisz, András György, and Csaba Szepesvári. Online RL in linearly q^π-realizable MDPs is as easy as in linear MDPs if you learn what to ignore. In *Advances in Neural Information Processing Systems*, volume 36, pp. 59172–59205, 2023.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pp. 2730–2775. PMLR, 2022.

APPENDIX CONTENTS

Read A–C for the main theorem. D groups the two core separation proofs and further examples; E gives structural conditions and representation boundaries. F records formal verification, G statistical benchmarks, and H feature-dynamics comparisons.

A Trajectory law and regression concentration . . . 13

A.1 Safe interpolation and replacement identities . . . 13

A.2 The observation law and conditional means . . . 13

A.3 Exponential concentration along the observed suffix . . . 14

A.4 Scalar and weighted ridge confidence . . . 14

A.5 Integrated leverage without covariance invertibility . . . 15

B Auxiliary recursion, coupling and the error bound . . . 16

B.1 Measurable recursion and finite-query agreement . . . 16

B.2 Target-only transfer and the root fit . . . 16

C Explicit count and public estimator . . . 17

C.1 Solving the logarithmic sample threshold . . . 17

C.2 Numerical scale of the sufficient count . . . 18

C.3 A deterministic alternative to the closed count . . . 18

C.4 Totality, quantifiers, and sample extraction . . . 18

C.5 Evaluating the correction functions . . . 18

D Core separations and legal examples . . . 19

D.1 A strict separation with growing dimension . . . 19

D.2 F3: infinite mixture coverage and failed Bellman completeness . . . 19

D.3 F2: an obstruction to covered skipping constructions . . . 20

D.4 Population failure of minimum-norm FQE . . . 21

E Realizability conditions and boundary proofs . . . 22

E.1 Replacement and support policy classes . . . 22

E.2 A structural abstraction criterion . . . 22

E.3 Relations to standard structural assumptions . . . 23

E.4 Full-support extrapolation . . . 24

E.5 A norm separation despite full support . . . 25

E.6 Endpoint realization is insufficient . . . 25

E.7 Deterministic and stochastic policy quantifiers . . . 25

F Formal correspondence and reproducibility . . . 26

F.1 Paper-to-Lean map . . . 26

F.2 Verification scope . . . 27

G Statistical benchmarks and the remaining gap . . . 27

H Comparison with feature-dynamics coverage . . . 28

A TRAJECTORY LAW AND REGRESSION CONCENTRATION

This appendix proves the two single-block probability bounds used by Theorem 3.1. The next appendix assembles them along the auxiliary backward recursion. All probabilities are over complete reward-inclusive trajectories.

For the probability calculations, use $\ell = \log(2H/\delta)$, $\alpha = \delta/(4nH^2)$ and $\sigma_H = 3H\sqrt{H}$. Thus the radius in Equation (6) equals $H + \sigma_H\sqrt{2\log(2/\alpha)}$.

A.1 SAFE INTERPOLATION AND REPLACEMENT IDENTITIES

Proof of Lemma 2.1. If $|m| \le c$, then $g = 0$ and $|\Delta| \le |\Delta - m| + |m| \le 2c$. If $m > c$, then $0 < m - c \le \Delta \le m + c$, so $g = m - c$ lies between zero and Δ and leaves error at most $2c$. For $m < -c$ the signs reverse. In the latter two cases $|g| \le |\Delta| \le H$, so clipping has no effect. □

Fix an already constructed suffix policy ρ after stage h, and put

$$\Delta_h^{\rho,e}(s,a) = \sum_{a'} e_h(a' \mid s) q_h^\rho(s,a') - q_h^\rho(s,a). \tag{12}$$

Write $\Delta_h = \Delta_h^{\rho,e}$ when the continuation is fixed. If $g_h = \lambda_h \Delta_h$ with weights in $[0,1]$, define the current kernel by

$$\rho_h(a' \mid s) = \sum_a b_h(a \mid s)\left[\lambda_h(s,a) e_h(a' \mid s) + (1 - \lambda_h(s,a))\mathbf{1}\{a' = a\}\right]. \tag{13}$$

Summing its action values gives

$$v_h^\rho(s) = \sum_a b_h(a \mid s)[q_h^\rho(s,a) + g_h(s,a)]. \tag{14}$$

The current update changes no current action value or later value. Stage-splicing closure keeps the resulting policy in $\mathcal{R}(b,e)$. These identities justify the partial-replacement premise of Lemma 2.2.

A.2 THE OBSERVATION LAW AND CONDITIONAL MEANS

Let $\mathcal{R}_h(s,a)$ be the Borel reward probability measure on $[0,1]$. For a finite state–action path $z = (s_1, a_1, \ldots, s_H, a_H, s_{H+1})$, its probability under b is

$$p_b(z) = \mathbf{1}\{s_1 \text{ is the prescribed initial state}\} \prod_{h=1}^{H} b_h(a_h \mid s_h) P_h(s_{h+1} \mid s_h, a_h).$$

The trajectory law is the finite mixture, with weights $p_b(z)$, of the product reward measures $\bigotimes_{h=1}^{H} \mathcal{R}_h(s_h, a_h)$ attached to those paths. It is a probability measure on the finite path space times $[0,1]^H$. Its stage-h pair marginal is μ_h, and integrating the observed return gives $J(b)$. The data law is its N-fold product. This construction retains arbitrary Borel reward laws and the stated conditional product convention.

In this subsection and the next, write g_t for the auxiliary corrections g_t^*. For fixed bounded corrections and a policy ρ satisfying Equation (14) on layers $t > h$, put $U_h = \sum_{t=h}^{H} R_t + \sum_{t=h+1}^{H} g_t(S_t, A_t)$. Then $U_H = R_H$ and $U_h = R_h + g_{h+1}(S_{h+1}, A_{h+1}) + U_{h+1}$. For a logger-supported pair (s,a), backward conditioning gives

$$\begin{aligned}
&\mathbb{E}_b[U_h \mid S_h = s, A_h = a] \\
&\quad = r_h(s,a) + \sum_{s'} P_h(s' \mid s,a) \sum_{a'} b_{h+1}(a' \mid s')[g_{h+1}(s',a') + q_{h+1}^\rho(s',a')] \\
&\quad = r_h(s,a) + \sum_{s'} P_h(s' \mid s,a) v_{h+1}^\rho(s') = q_h^\rho(s,a).
\end{aligned}$$

The first equality uses induction, the second the correction identity, and the third the original Bellman equation. At $h = H$ the continuation is zero. Pairs of zero logger mass impose no conditional-expectation constraint. Layer closure ensures that only valid-layer identities are used. Thus Lemma 2.2 holds under the full law of the observed rewards.

A.3 Exponential concentration along the observed suffix

Centered increments. Fix any processed auxiliary history, hence its corrections and actual policy ρ. With the current pair fixed, define

$$\xi_h = R_h + v^\rho_{h+1}(S_{h+1}) - q^\rho_h(S_h, A_h),$$
$$\xi_t = R_t + g_t(S_t, A_t) + v^\rho_{t+1}(S_{t+1}) - v^\rho_t(S_t), \quad t > h. \tag{15}$$

These increments satisfy

$$Y_h - q^\rho_h(S_h, A_h) = \xi_h + \sum_{t=h+1}^{H} \xi_t.$$

The equality is pathwise: the state potentials telescope and $v^\rho_{H+1} = 0$. The first increment has conditional mean zero given the history through (S_h, A_h), by the Bellman equation. For $t > h$, condition on the history through S_t, before drawing A_t. Its conditional mean is

$$\sum_a b_t(a \mid S_t)\left[r_t(S_t, a) + g_t(S_t, a) + P_t v^\rho_{t+1}(S_t, a)\right] - v^\rho_t(S_t) = 0,$$

by Equation (14). The logger is fixed and memoryless, so this identity also holds conditional on earlier observed rewards and states. **Exponential bound.** Since $R_t \in [0, 1]$, $|g_t| \le H$, and $q^\rho_t, v^\rho_t \in [0, H]$, every increment has magnitude at most $3H$. Conditional Hoeffding gives

$$\mathbb{E}[\exp(z\xi_t) \mid \text{preceding history}] \le \exp(9H^2 z^2/2).$$

Repeated conditioning multiplies these bounds over the $H - h + 1$ remaining steps. The resulting conditional variance proxy is $9H^2(H - h + 1) \le 9H^3$. Thus the auxiliary label satisfies

$$\mathbb{E}_b[\exp\{z(Y^*_h - q^\rho_h)\} \mid S_h, A_h] \le \exp(9H^3 z^2/2), \qquad z \in \mathbb{R}. \tag{16}$$

The multiplier z may be any fixed function of (S_h, A_h) after that pair has been observed. This version supplies the scalar tests in the regression argument below. Independence is used between sampled trajectories; within each trajectory the exponential bounds are conditional.

The formal proof integrates the original Borel reward laws at each step and then the finite Markov path weights. It proves the same scalar exponential test with a pair-dependent multiplier. No replacement of the reward law by its mean, additional trajectory observation, or independent-transition model is used in this concentration step.

A.4 Scalar and weighted ridge confidence

Scalar confidence. Fix the already processed blocks and their auxiliary suffix. Write X_j for the stage-h feature vector of trajectory j in the fresh block, and $\eta_j = Y^*_j - X_j^\top \theta$ for its centered label noise, where $\|\theta\|_2 \le 1$. The current state–action pairs take finitely many values. Conditioning their independent trajectories on these pairs preserves a product law: for a design fiber of positive mass, its joint conditional measure is the product of the individual conditional trajectory measures. Zero-mass fibers can be omitted. Equation (16) therefore gives independent, conditionally σ_H-sub-Gaussian noises on each design fiber. Conditioning only on feature vectors gives the same conclusion by grouping finite pair fibers, or one may retain the pair design throughout.

Let $V = I + \sum_j X_j X_j^\top$ and $\widehat{\theta} = V^{-1} \sum_j X_j Y^*_j$. For a valid stage-$h$ pair (s, a), put $x = \psi_h(s, a)$, $q(x) = x^\top\theta = q^\rho_h(s, a) \in [0, H]$, $t = \sqrt{x^\top V^{-1} x}$ and $w(x) = \min\{1, t\}$. The prediction decomposition is

$$x^\top(\widehat{\theta} - \theta) = \sum_j x^\top V^{-1} X_j \eta_j - x^\top V^{-1}\theta.$$

Since $V \succeq I$, Cauchy–Schwarz gives $|x^\top V^{-1}\theta| \le t\|\theta\|_{V^{-1}} \le t$. Moreover,

$$\sum_j (x^\top V^{-1} X_j)^2 = x^\top V^{-1}(V - I)V^{-1}x \le t^2.$$

When $t > 0$, the first term divided by t has conditional moment-generating function at most $\exp(\sigma_H^2 z^2/2)$, for every real z. Taking the sum of the two signed exponential bounds controls its

absolute value. For $0 < t \le 1$, clipping cannot increase the prediction error. For $t > 1$, the clipped error is at most H because both values lie in $[0, H]$. It follows, for every $z \ge 0$, that

$$\mathbb{E}\left[\exp\left(z\frac{|\widehat{q}(x) - q(x)|}{w(x)}\right)\middle| X\right] \le 2\exp(Hz + \sigma_H^2 z^2/2). \tag{17}$$

If $t = 0$, positive definiteness gives $x = 0$, both predictions are zero, and we define the ratio to be zero. Thus the same bound holds in that case.

Weighted confidence. Let ξ be any finite nonnegative measure on query pairs, fixed conditional on the design independently of the current noise. Set $W = \sum_x \xi(x)w(x)$. If $W = 0$, the weighted error vanishes. Otherwise $p_x = \xi(x)w(x)/W$ are probability weights. Jensen's inequality gives

$$\exp\left(z\frac{\sum_x \xi(x)|\widehat{q}(x) - q(x)|}{W}\right) \le \sum_{x:w(x)>0} p_x \exp\left(z\frac{|\widehat{q}(x) - q(x)|}{w(x)}\right).$$

Taking conditional expectations and using Equation (17) gives the same exponential bound for the weighted error. No independence of different query predictions has been used. Markov's inequality, optimized at $z = \sqrt{2\log(2/\alpha)}/\sigma_H$, implies

$$\mathbb{P}\left\{\sum_x \xi(x)|\widehat{q}(x) - q(x)| > [H + \sigma_H\sqrt{2\log(2/\alpha)}]W \,\middle|\, X\right\} \le \alpha. \tag{18}$$

The bound also covers a random query measure by conditioning on it when it is independent of the regression noise given the design. It does not cover arbitrary measures selected after fitting.

For a fixed independent query (s, a), take the sum of a point mass at (s, a) and the target action law at s. The left side of the corresponding weighted error bounds $|\widehat{q}_h^{*,e}(s) - \widehat{q}_h^*(s, a) - \Delta_h(s, a)|$ by the triangle inequality, and its weighted width is $\bar{w}_h^e(s) + w_h(s, a)$. This yields a valid advantage interval with failure at most α, without a factor depending on the number of target actions.

Why clipping precedes averaging. Rescaling removes polynomial norm dependence from the ridge bias but permits large leverage in unobserved directions. Actionwise caps retain the probability weight of a rare target action. If an action has probability p and leverage norm $1/p$, with zero widths elsewhere, capping the average gives one, whereas averaging the capped widths gives p. The latter quantity, together with clipped action predictions, is what the weighted confidence argument controls.

A.5 Integrated leverage without covariance invertibility

For iid $X_1, \ldots, X_n$ with $\|X_i\|_2 \le B$, let $V_i = I + \sum_{j=1}^i X_j X_j^\top$ and $\mathcal{L}_n = \mathbb{E}_X \min\{1, X^\top V_n^{-1} X\}$, using an independent vector X of the same distribution. With probability at least $1 - \alpha$,

$$\mathcal{L}_n \le \frac{4}{n}[d\log(1 + nB^2/d) + \log(1/\alpha)]. \tag{19}$$

Proof. Put $u_i = X_i^\top V_{i-1}^{-1} X_i$, $f_i = \min\{1, u_i\} \in [0, 1]$, and $m_i = \mathbb{E}[f_i \mid X_1, \ldots, X_{i-1}]$. For $z \in [0, 1]$, $e^{-z} \le 1 - z/2$, hence

$$\mathbb{E}[e^{m_i/2 - f_i} \mid X_1, \ldots, X_{i-1}] \le e^{m_i/2}(1 - m_i/2) \le 1.$$

Iteration gives $\mathbb{E}\exp(\sum_i (m_i/2 - f_i)) \le 1$. Since $V_n \succeq V_{i-1}$, inverse monotonicity gives $m_i \ge \mathcal{L}_n$ for every realized sample. The determinant lemma and $\min\{1, z\} \le 2\log(1 + z)$ for $z \ge 0$ give

$$\sum_i f_i \le 2\sum_i \log(1 + u_i) = 2\log\det V_n.$$

Thus $\mathbb{E}\exp(n\mathcal{L}_n/2 - 2\log\det V_n) \le 1$. Markov's inequality and the determinant bound imply

$$\mathcal{L}_n \le [4\log\det V_n + 2\log(1/\alpha)]/n \le 4[d\log(1 + nB^2/d) + \log(1/\alpha)]/n.$$

□

We retain the looser factor four on the confidence logarithm to simplify tuning. The leverage cap is active in general and is retained throughout the proof. No population covariance inverse is used. This is a standard elliptical-potential argument; the trajectory version also appears as Lemma 26 of Tkachuk et al. (2026).

B Auxiliary recursion, coupling and the error bound

B.1 Measurable recursion and finite-query agreement

Initialize the auxiliary policy with the logger and its correction list empty. At stage h, use the already defined g_t^*, $t > h$, in the labels and compute the same ridge fit, predictions and proposed correction as the public rule. Project that proposal by Equation (9). For a nonzero exact advantage, set $\lambda_h = g_h^*/\Delta_h$, and set it to zero otherwise. Projection ensures $0 \le \lambda_h \le 1$ and $g_h^* = \lambda_h \Delta_h$, including at a zero advantage. Splice Equation (13) into the auxiliary policy at stage h.

Unconditional structural invariant. The initial logger is in $\mathcal{R}(b, e)$. Stage-splicing closure implies membership of every auxiliary policy on every history. The corrected-value identity holds at the updated stage because $g_h^* = \lambda_h \Delta_h$. Its action values depend only on future kernels, which are unchanged; earlier updates cannot affect this or any later identity. Induction therefore proves all processed suffix identities without a confidence hypothesis. The policy $\rho^{>h}$ and corrections used by the stage-h fit are $\mathcal{F}_{>h}$-measurable. Conditional on these later blocks, the fresh block D_h retains its original trajectory law. The invariant therefore gives its label mean by Lemma 2.2; fixing its pair design then gives the independent centered noises of Section A.4. Realizability and the bounded-reward suffix argument justify Equation (18) on each such fiber.

Measurability. The auxiliary recursion uses finite Bellman sums, positive-definite matrix inversion, clipping, and division on a nonzero denominator with a specified zero fallback. All are Borel operations. The exact advantages and policy kernels are therefore measurable functions of the processed blocks. The public rule is measurable by the same operations without projection. All events below are finite sums or intersections of inequalities involving these functions. Realizing parameters are chosen only to prove each fixed fiber's bound; no parameter selection is used to define an event or recursion.

Independent queries. For $k < h$ and $j \in [n]$, let $Z_{khj} = (S_h^j, A_h^j)$ in block D_k. This pair is independent of D_h and of the processed blocks $D_{h+1}, \dots, D_H$. Fixing those blocks, the current pair design, and Z_{khj} leaves the regression noise with the product conditional law used in Section A.4. The query version of Equation (18) bounds the auxiliary advantage interval's failure by α. Integrating over the design, query and suffix history proves the same unconditional bound. Layer validity holds almost surely under the trajectory law, so every realization is used only on its specified domain. No later observation from a trajectory of D_h is being conditioned upon as an independent query.

Let $\mathcal{A}$ be the intersection of the valid-interval events for all these queries. A union bound gives

$$\mathbb{P}(\mathcal{A}^{\mathsf{c}}) \le nH(H-1)\alpha/2 \le \delta/4. \tag{20}$$

On $\mathcal{A}$, Lemma 2.1 gives $\widetilde{g}_h(Z_{khj}) = g_h^*(Z_{khj})$ for every query. Backward induction now equates the public and auxiliary fits. At H both labels are the observed rewards. If later fits agree, the proposed corrections agree as functions; the queried equality with g_t^* then equates each stage-h label and fit. Thus both root outputs agree. Corrections at pairs absent from all earlier blocks may differ, and no pointwise confidence claim is made there.

B.2 Target-only transfer and the root fit

Let $\xi_1 = \nu_1^e$ and $\xi_h(s,a) = d_h^e(s)[e_h(a \mid s) + b_h(a \mid s)]$ for $h \ge 2$. These measures are fixed by the population model, independently of the current block noise and all data. Conditional on any later-block history, Equation (18) therefore controls their auxiliary prediction error. Integrating these conditional bounds and taking a union over stages yields an event $\mathcal{P}$ with $\mathbb{P}(\mathcal{P}^{\mathsf{c}}) \le H\alpha \le \delta/4$. Define $W_h = \sum \xi_h w_h$ and let $\mathcal{L}$ be simultaneous leverage success from Equation (19), with per-stage budget $\delta/(2H)$. Then $\mathbb{P}(\mathcal{L}^{\mathsf{c}}) \le \delta/2$. On this event, simultaneously over stages,

$$\mathcal{L}_h := \mathbb{E}_{\mu_h} w_h^2 \le u_n := \frac{4}{n}[d \log(1 + nB^2/d) + \ell]. \tag{21}$$

These events need not be independent.

On $\mathcal{P}$, the triangle inequality and the definition of ξ_h give

$$\mathbb{E}_{d_h^e b_h}|\Delta_h - m_h^*| \le \sum_{s,a} d_h^e(s)[e_h(a \mid s) + b_h(a \mid s)]|q_h^{\rho}(s,a) - \widehat{q}_h^*(s,a)| \le \beta W_h.$$

Also $\mathbb{E}_{d_h^e b_h} c_h = \beta W_h$. Actionwise clipping gives $|m_h^*| \le H$. Soft-thresholding does not increase this magnitude, so its outer $[-H, H]$ clipping is inactive and $|m_h^* - \widetilde{g}_h| \le c_h$. Since projection cannot increase distance to Δ_h,

$$\mathbb{E}_{d_h^e b_h} |\Delta_h - g_h^*| \le 2\beta W_h.$$

For either distribution η in Equation (11), $\mathbb{E}_\eta w_h \le \sqrt{\mathbb{E}_\eta w_h^2} \le \sqrt{C_e \mathcal{L}_h}$. Expanding W_h therefore gives

$$W_1 \le \sqrt{C_e \mathcal{L}_1}, \qquad W_h \le 2\sqrt{C_e \mathcal{L}_h} \quad (h \ge 2). \tag{22}$$

If a logger state has zero mass, target coverage forces its target mass to zero; no occupancy division is required.

Set $\rho_1 = e_1$. For $\bar{\Delta}_h(s) = \sum_a e_h(a \mid s) q_h^\rho(s,a) - v_h^\rho(s)$, Bellman recursion gives

$$v_h^e(s) - v_h^\rho(s) = \bar{\Delta}_h(s) + \sum_{a,s'} e_h(a \mid s) P_h(s' \mid s, a) [v_{h+1}^e(s') - v_{h+1}^\rho(s')].$$

Multiplying by $d_h^e(s)$ and summing telescopes the future terms. The terminal values and $\bar{\Delta}_1$ vanish, proving Equation (10). On $\mathcal{P} \cap \mathcal{L}$, this bounds the continuation error by $4\beta \sum_{h=2}^H \sqrt{C_e \mathcal{L}_h}$ and the auxiliary root error by $\beta\sqrt{C_e \mathcal{L}_1}$. On $\mathcal{A}$ its output is the public one. The combined failure probability is at most $Q\alpha + H\alpha + \delta/2 \le \delta$, yielding Equation (7). At $H = 1$ the coupling and continuation sums are empty, so the argument reduces to root regression. No occupancy of a replacement policy requires coverage.

C Explicit count and public estimator

C.1 Solving the logarithmic sample threshold

For $0 < \varepsilon \le H$, define the public quantities

$$\begin{aligned} D &= d + \ell + d\log(1 + B^2), & E &= 1 + \log(8H^2/\delta), \\ K &= 128 C_e H^4/\varepsilon^2, & L &= 2304 C_e H^5/\varepsilon^2, \\ M_0 &= 1 + KD + LDE, & n_\star &= \lceil 64 M_0 \log^2(64 M_0) \rceil, \qquad N_\star = H n_\star. \end{aligned} \tag{23}$$

We verify the accuracy guarantee at $n = n_\star$. Put $T = d\log(1+n) + \ell + d\log(1+B^2)$ and $u = \log(8nH^2/\delta)$. Since $d \ge 1$, $d\log(1 + nB^2/d) + \ell \le T$. The scalar radius obeys

$$\beta^2 \le 2H^2 + 36H^3 u, \qquad C_e u_n \le 4C_e T/n.$$

Using $4H - 3 \le 4H$, the square of the right side of Equation (7) is at most

$$\frac{128 C_e H^4 T + 2304 C_e H^5 u T}{n} = \varepsilon^2 \frac{KT + LuT}{n}. \tag{24}$$

Let $z = \log(64 M_0) > 4$. Since $M_0 \ge 1$, rounding gives $64 M_0 z^2 \le n \le 65 M_0 z^2$, and $z \le 64 M_0$ gives

$$8n \le 520 M_0 z^2 \le 520 M_0 (64 M_0)^2 \le (64 M_0)^4.$$

Hence $\log(8n) \le 4z$. For $n \ge 1$, $T \le D\log(8n)$ and $u \le E\log(8n)$. Therefore

$$KT + LuT \le 4KDz + 16LDEz^2 \le 16(KD + LDE)z^2 \le 16 M_0 z^2 \le n.$$

Equation (24) proves the guarantee. Also $N_\star \le 65 H M_0 \log^2(64 M_0)$, which gives Equation (8): D is linear in d up to logarithms, while E carries no polynomial dimension factor. The additive H term is absorbed for $0 < \varepsilon \le H$ and $C_e, B, d \ge 1$. The prescription uses only public inputs, with no sample count inside its defining logarithms and no state or action cardinality.

C.2 Numerical scale of the sufficient count

At $H = d = 3$, $B = 1$, $C_e = 4$ and $\delta = 0.05$, Equation (23) gives $N_\star \approx 1.89 \times 10^{13}$ for $\varepsilon = 1$ and 2.72×10^{15} for $\varepsilon = 0.1$. These conservative constants certify a sufficient threshold; they are not practical sample recommendations. Directly checking Equation (7) gives sufficient total counts 3 595 926 927 and 519 319 701 162, respectively, about 5200 times smaller. The public-parameter procedure below verifies these inequalities using rational upper bounds; even these counts are large. Numerical diagnostics run below this threshold and check the mechanism and implementation on finite examples. They do not verify the high-probability guarantee at those smaller counts. For $\varepsilon \geq H$, a constant output already meets the error tolerance without data.

C.3 A deterministic alternative to the closed count

Theorem 3.1 holds at every fixed positive integer block size. Thus a count chosen from public parameters alone is sufficient whenever its right side is at most ε. This is not a data-dependent stopping rule. The Lean theorem `estimateLinearAtCount_guarantee` establishes this contract under the actual observation law; the closed count is a corollary.

For rational public inputs, a rational-arithmetic procedure checks an upper bound on the square of the right side. Write $x = 2^k y$, $1 \leq y < 2$, $t = (y-1)/(y+1)$. The identity and positive remainder bound

$$\log y = 2 \sum_{j=0}^{K-1} \frac{t^{2j+1}}{2j+1} + R_K, \qquad 0 \leq R_K \leq \frac{2t^{2K+1}}{(2K+1)(1-t^2)}$$

follow by integrating the geometric series. Apply the same bound to $\log 2$, combine with $k \log 2$, and round outward to dyadic rationals. Integer square root with upward rounding bounds $\sqrt{18H^3 \log(8nH^2/\delta)}$ in β. All remaining operations use exact fractions. The program starts from an upward-rounded closed prescription and retains a checked sufficient upper endpoint while bisecting; it reports failure if the initial enclosure is too coarse. It claims sufficiency, not the smallest integer. Neither the enclosure program nor floating-point regression is a Lean-verified executable.

C.4 Totality, quantifiers, and sample extraction

Fix the finite public spaces, horizon, initial state, target, features, bounds, accuracy, confidence, and available count N. Normalize by Equation (5) and compute $n_\star$ from Equation (23). If $N < Hn_\star$, define the estimate to be zero. Otherwise take a fixed prefix of length $Hn_\star$, reshape it into H blocks, and apply Algorithm 1. This defines the rule also on off-law paths. Reward coordinates lie in $[0, 1]$ by the observation type.

Prefix projection sends the N-fold trajectory product to its $Hn_\star$-fold product; finite reindexing and currying then give H independent blocks of $n_\star$ trajectories, establishing the required sampling law for every $N \geq N_\star$. Finite feature evaluations, positive-definite matrix inversion, quadratic forms, square roots, and clipping are measurable. Their backward composition makes the public rule measurable.

The rescaled realizing parameter is $(F/B)\theta_h^\pi$, of norm at most $FL_\theta/B \leq 1$; $F, B \geq 1$ avoid division by zero. This works even if an original norm bound is zero. The rule uses no unknown model information, so quantifying over compatible MDPs, reward laws, and loggers after constructing it proves the uniform endpoint. Lean allows all $\varepsilon > 0$ and $0 < \delta \leq 1$; the paper states the customary nontrivial range.

C.5 Evaluating the correction functions

Store $(V_h^{-1}, \widehat{\theta}_h)$ at each stage. Evaluating $g_h(s, a)$ requires the clipped predictions and quadratic forms for every target action, followed by two finite action averages. This takes $O(|\mathcal{A}|d^2)$ arithmetic operations with direct evaluation. There are at most $nH(H-1)/2$ such evaluations in all labels. Including ridge fitting, the total real-arithmetic operation count is

$$O\left(Hd^3 + N_\star d^2 + N_\star H|\mathcal{A}|d^2\right).$$

This excludes feature/target query costs and public logarithmic tuning. No state-space enumeration or policy search is involved. By comparison, no efficient implementation is known for the evaluation baseline of Tkachuk et al. (2026, Section 4.4, footnote 17). Our count is a paper-level calculation, not a Lean-verified bit-complexity result. The effect of finite precision near shrink thresholds remains unproved.

D Core separations and legal examples

D.1 A strict separation with growing dimension

We prove Proposition 5.1.

Proof. Use actions $\{0, 1, \dots, m\}$, one root, second-stage states $x_1, \dots, x_m$, and a terminal state. Root action $i \geq 1$ leads to x_i; action zero leads to x_1. The root reward means are $c_0 = c_1 = 0$ and $c_i = 1/4$ for $i \geq 2$. At every x_j, action zero has reward one and every positive action has reward mean u. Rewards are independent Bernoulli draws with these means, conditional on the visited pairs; transitions are deterministic and the terminal state is observed. At every state the target chooses action one and the logger uses

$$b(0) = 0, \qquad b(1) = \tfrac{1}{2}, \qquad b(i) = \frac{1}{2(m-1)} \quad (2 \leq i \leq m).$$

Choose the public features

$$\phi_1(\text{root}, i) = (1/2, 2c_i), \qquad \phi_2(x_j, i) = \begin{cases} (0,1), & i = 0, \\ (1,0), & i \geq 1. \end{cases}$$

Their norms are at most one. Every policy supported on the positive actions has second-stage value u at every x_j. Its action values, including the current action zero, are represented by $\theta_1 = (2u, 1/2)$ and $\theta_2 = (u, 1)$. Both squared norms are at most $5/4 < 4$. This proves the support condition uniformly over all such stochastic policies.

The target visits $(\text{root}, 1)$ and $(x_1, 1)$ with probability one. Their logger occupancies are $1/2$ and $1/4$. All other target occupancies vanish, so $C_e = 4$ exactly. The target return is u; the logger return is $u + 1/8$.

Let π^0 choose action one everywhere. For $1 \leq i \leq m$, let π^i choose action zero only at x_i. On root actions $j = 1, \dots, m$,

$$q_1^{\pi^i}(\text{root}, j) - q_1^{\pi^0}(\text{root}, j) = (1-u)\mathbf{1}\{j = i\}. \tag{25}$$

These m vectors span $\mathbb{R}^m$. If an all-policy feature matrix on the root pairs had d' columns, every difference would lie in its column space, forcing $d' \geq m$. This rank argument imposes no parameter norm constraint. For the supplied features, root actions two and three have identical features but, under π^2, values $5/4$ and $u + 1/4$. Both root pairs have positive logger probability. Thus their distinct values cannot be represented by any common parameter, regardless of its norm. □

D.2 F3: infinite mixture coverage and failed Bellman completeness

Fix $m \geq 2$, $p \in [0, 1/2]$, and $u \in [1/4, 1/2]$. The decision layers are $\{s_1\}$, $\{l, r\}$, and $\{x, y_1, \dots, y_m, z\}$, followed by a terminal state. Actions are $0, 1$. Let $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ denote standard basis vectors; these feature symbols are distinct from the target policy e. Table 2 specifies the features, transitions, and rewards. Every unlisted final transition goes to the terminal state. All rewards are deterministic except the Bernoulli reward at z.

The target takes action 1 at s_1, l and action 0 elsewhere. The logger is uniform except that $b(1 \mid l) = p$. This model has $m + 5$ nonterminal states. It is legal under the conditional product convention: the sole random reward is paired with a deterministic terminal transition.

All-policy realizability and norms. For every stochastic memoryless policy π, use the parameters

$$\begin{aligned} \theta_3^\pi &= (1/4, 1/4, u), & \theta_2^\pi &= (1/4, u, u + 1/4), \\ \theta_1^\pi &= (1/4, u + \pi(1 \mid r)/4, 0). \end{aligned} \tag{26}$$

Table 2: **The F3 model.** Features and target are independent of u.

State	Action	Feature	Next state	Reward
s_1	0, 1	$\mathbf{e}_1, \mathbf{e}_2$	l, r, respectively	0
l	0	$\mathbf{e}_1$	x	0
l	1	$\mathbf{e}_1$	Uniform on $y_1, \ldots, y_m$	0
r	0, 1	$\mathbf{e}_2, \mathbf{e}_3$	z	$0, 1/4$, respectively
x	either	$\mathbf{e}_1$	Terminal	$1/4$
y_i	either	$\mathbf{e}_2$	Terminal	$1/4$
z	either	$\mathbf{e}_3$	Terminal	Bernoulli(u)

Substitution into the Bellman equations verifies every state–action value. Their squared norms are at most $3/8$, $7/8$, and $5/8$, respectively. Every feature has norm one, so Assumption 1 holds with both original bounds one. These witnesses cover arbitrary stochastic policies, including policies with different action probabilities at each y_i.

Target return and coverage. The target takes the right branch and receives no reward at r, so $J(e) = u$. Its occupied pairs have target/logger ratios 2 at $(s_1, 1)$, 4 at $(r, 0)$, and 4 at $(z, 0)$; all other target masses are zero. Thus $C_e = 4$. The unknown parameter u changes the observed reward law and the target return but not the public features or policy.

Mixture coverage. A statewise mixture has root probability of l in $[0, 1/2]$, action-1 probability at l in $[p, 1]$, and action-0 probability at each last-layer state in $[1/2, 1]$. At $(y_i, 0)$ its occupancy is therefore at most $1/(2m)$, while the logger mass is $p/(4m)$. The ratio $2/p$ is attained by using the logger at the root and the target at l, y_i. At $p = 0$ the same pair witnesses infinite coverage. At $(l, 1)$ the ratio is at most $1/p$; at $(l, 0)$ it is at most one. At x the mixture state mass is at most the logger state mass, so its pair ratios are at most two. Right-branch ratios are at most four, and root ratios at most two. Because $p \le 1/2$, all are bounded by $2/p$. The maximum ratio is therefore exactly the coefficient in Proposition 5.2.

Failure of target Bellman completeness. Here the bounded linear class is $\mathcal{F}_h = \{(s, a) \mapsto \phi_h(s, a)^\top \theta : \|\theta\|_2 \le 1,\ 0 \le \phi_h(s, a)^\top \theta \le H\}$ on the valid layer. For a fixed policy π, closure would require $\mathcal{T}_h^\pi f(s, a) = r_h(s, a) + \sum_{s'} P_h(s' \mid s, a) \sum_{a'} \pi_{h+1}(a' \mid s') f(s', a')$ to belong to the current linear span for every $f \in \mathcal{F}_{h+1}$. Choose the linear final value $f(s, a) = \phi_3(s, a)^\top \mathbf{e}_2$. It is bounded in $[0, 1]$, equals zero at x and one at every y_i, and is action-independent. Its one-step backups at $(l, 0)$ and $(l, 1)$ are zero and one. Both current features equal $\mathbf{e}_1$, so no single linear parameter can represent that backup. This is an explicit closure failure, even though all actual policy values are realizable. Applying Corollary 3.2 with $(H, d, B, C_e) = (3, 3, 1, 4)$ proves the final claim of Proposition 5.2, uniformly over m, p, u.

D.3 F2: AN OBSTRUCTION TO COVERED SKIPPING CONSTRUCTIONS

The following condition describes a covered skipping construction. We show that target coverage need not imply this additional premise. A skipping construction chooses a statewise continuation probability $\omega_G(s)$. Its skippy policy is $\omega_G b + (1 - \omega_G) e$. For threshold $\alpha > 0$, the usual range-based rule has $\omega_G(s) = 0$ whenever $\mathrm{range}_G(s) \ge \alpha / \sqrt{2d}$. A good construction satisfies

$$\sup_\pi \max_{a, a'} [q_h^\pi(s, a) - q_h^\pi(s, a')] \le \sqrt{2d}\ \mathrm{range}_G(s).$$

Its rollout coverage bounds the occupancies of every policy following the logger through a stage and the skippy policy afterward. Other requirements of a good construction, including stopped-backup completeness, can only restrict this class further; the obstruction below uses range domination alone.

In F2, $H = 3$, $d = 2$, and the layers are $\{s_1\}$, $\{l, r\}$, and $\{x, y_1, \ldots, y_m\}$ for $m \ge 2$. Root actions lead to l, r with zero reward and features $\mathbf{e}_1, \mathbf{e}_2$. At l, action 0 leads to x with reward $1/2$ and feature

$\mathbf{e}_1$; action 1 leads uniformly to the y_i with reward zero and feature $\mathbf{e}_2$. At r, both actions lead to x with reward $1/2$ and feature $\mathbf{e}_1$. At the last layer, action rewards are $1, 1/3$ and features $\mathbf{e}_1, \mathbf{e}_2$. All rewards are deterministic. The target takes action 1 at s_1, l and action 0 elsewhere. The logger is uniform except that it always takes action 0 at l.

For any policy π, let $v_x = v_3^\pi(x)$ and $\bar{v}_y = m^{-1} \sum_i v_3^\pi(y_i)$. Valid realizing parameters are

$$\theta_3^\pi = (1, 1/3), \qquad \theta_2^\pi = (1/2 + v_x, \bar{v}_y), \qquad \theta_1^\pi = (v_2^\pi(l), v_2^\pi(r)).$$

Since $v_3^\pi \in [1/3, 1]$ and $v_2^\pi \in [1/3, 3/2]$, each parameter has norm at most three; all features have norm one. The target return is $3/2$, and its maximum occupancy ratio is four, attained at $(r, 0)$. The logger never visits any y_i or the pair $(l, 1)$.

Under the target, $q_2^e(l, 0) = 3/2$ and $q_2^e(l, 1) = 1$. Range domination therefore forces every good construction to have $\text{range}_G(l) \geq 1/(2\sqrt{2d})$. For $0 < \alpha \leq 1/2$, its skip rule satisfies $\omega_G(l) = 0$. A logger prefix of length one followed by the skippy policy consequently puts mass $1/2$ on $(l, 1)$ at stage two, while that pair has logger mass zero. No finite rollout coverage bound exists for any such construction.

This is also a feature-direction obstruction to that proof route. Every logger-supported stage-two feature equals $\mathbf{e}_1$, so any ridge design from n observations is $\text{diag}(1 + n, 1)$. The pair $(l, 1)$ has feature $\mathbf{e}_2$ and squared leverage one, giving expected squared leverage at least $1/2$ under that rollout for every sample size. Safe advantage OPE does not need this quantity to vanish: the target never visits l, so its state-weighted residual there contributes zero. Our target-coverage theorem applies to F2 without the impossible covered-construction promise. This example is an exact linear MDP and establishes no statistical hardness.

D.4 Population failure of minimum-norm FQE

Fix $u \in [1/4, 1/2]$. The two root actions lead deterministically to l, r, respectively, with zero root reward and the same feature $(1, 0)$. Both actions at l, r have deterministic reward u and then terminate. At $(l, 1)$ the feature is $(0, 1)$; at the other second-stage pairs it is $(1, 0)$. The target chooses action one at the root and at l, and zero at r. The logger chooses both root actions equally, always chooses zero at l, and is uniform at r. All policies have root action values u and terminal action values u, realized by $(u, 0)$ and (u, u); both norm budgets are one. Target coverage is four, attained at $(r, 0)$, and $J(e) = u$.

Define population FQE to minimize the logger-weighted squared Bellman error backward, selecting the minimum Euclidean norm among minimizers at each stage. The terminal loss is $(\theta_{2,1} - u)^2$, so the selected vector is $(u, 0)$. Its fitted target continuation is zero at l and u at r. Consequently the root loss is

$$\tfrac{1}{2}\theta_{1,1}^2 + \tfrac{1}{2}(\theta_{1,1} - u)^2 = (\theta_{1,1} - u/2)^2 + u^2/4.$$

Its selected vector is $(u/2, 0)$ and its target estimate is $u/2$. These losses, minimum-norm minimizers and legal-model claims are formalized in `MinimumNormFQE`. The corresponding isotropic ridge limit, with penalty tending to zero in the population loss, has the same coefficients. This does not cover other tie-breaking rules or a nonvanishing penalty. The example is not statistical hardness: even averaging logged returns works here. Its role is to exhibit a concrete failure of ordinary fitted backups.

A three-stage variant with a nonzero correction. Retain the root, second-stage features, rewards and policies above, but send every second-stage action to a common state w. At stage three, actions zero and one pay zero and one, with features $(0, 0)$ and $(1, 0)$, respectively, then terminate. The logger is uniform at w and the target chooses one. All rewards are deterministic. For any stochastic Markov policy π, put $p = \pi_3(1 \mid w)$. Its global action values are realized by

$$\theta_1^\pi = (u + p, 0), \qquad \theta_2^\pi = (u + p, u + p), \qquad \theta_3^\pi = (1, 0).$$

Thus $H = 3$, $d = 2$, $L_\phi = 1$ and $L_\theta = 3$ suffice uniformly in u, p; features and target do not depend on u. The occupancy ratios remain at most four, attained at $(r, 0)$; the new target pair has ratio two. The resulting returns and population minimum-norm FQE estimate are

$$J(e) = u + 1, \qquad J(b) = u + \tfrac{1}{2}, \qquad \widehat{J}_{\text{FQE,pop}} = \tfrac{u+1}{2}.$$

Indeed, terminal regression gives $(1, 0)$, and the second-stage fit gives $(u + 1, 0)$ because all logged second-stage features are $(1, 0)$. The fitted target values at l, r are therefore zero and $u+1$; the root fit averages them. Both FQE and return averaging have nonvanishing bias. Here $\Delta_3(w, 0) = 1$: adding its full correction to logged zero actions makes the terminal contribution one on every path. The preceding regression then has the target-continuation mean $u + 1$. Omitting all corrections instead leaves the root fit at $u+1/2$ in the population limit. This paper calculation illustrates the role of suffix correction; it is outside the Lean certificate and concerns the specified FQE rule, not all estimators.

E Realizability conditions and boundary proofs

E.1 Replacement and support policy classes

The main assumption is already restricted to $\mathcal{R}(b, e)$, as defined in Equation (3). We record the structural properties used by the proof. Every kernel is nonnegative, and summing over a' yields $\sum_a b_h(a \mid s)[1 - \lambda_h(s, a) + \lambda_h(s, a)] = 1$. The zero and unit arrays give b and e, respectively. Choosing separate arrays at separate stages shows closure under stagewise splicing. Outside $U_h(s) = \operatorname{supp} b_h(\cdot \mid s) \cup \operatorname{supp} e_h(\cdot \mid s)$, both terms vanish. Thus support-restricted realizability implies Assumption 1, proving Corollary 3.3; no support or replacement-class oracle is introduced.

The auxiliary policy constructed in Section B.1 belongs to this class on every history. This membership, together with Assumption 1, supplies the single realizing witness used on each conditional regression fiber. No realizability outside the class is used elsewhere: conditional means come from the original MDP, concentration from bounded rewards, and error transfer from target coverage. Normalization rescales each such witness by F/B. This explains why the same public estimator specializes directly to support and all-policy realizability.

E.2 A structural abstraction criterion

We state and prove the abstraction criterion described in Section 5. All conditions concern the population MDP and the supplied feature map; no additional input is supplied to the learner. Let $\zeta_h : \mathcal{S}_h \to \mathcal{Z}_h$ be surjective maps, including at the terminal layer, with $\phi_h(s, a) = \bar{\phi}_h(\zeta_h(s), a)$ and $\|\phi_h(s, a)\|_2 \le L_\phi$. For every valid current pair, including actions outside $U_h(s)$, assume

$$r_h(s, a) = \bar{\phi}_h(\zeta_h(s), a)^\top \eta_h, \tag{27}$$

$$\sum_{s' : \zeta_{h+1}(s') = z'} P_h(s' \mid s, a) = \bar{\phi}_h(\zeta_h(s), a)^\top m_h(z') \quad (z' \in \mathcal{Z}_{h+1}). \tag{28}$$

The vector factors may be signed; the displayed transition probabilities are those of the given legal MDP. For fixed $R, M \ge 0$, require

$$\|\eta_h\|_2 \le R, \qquad \left\| \sum_{z'} m_h(z') f(z') \right\|_2 \le M \|f\|_\infty \quad \text{for every } f : \mathcal{Z}_{h+1} \to \mathbb{R}. \tag{29}$$

For a block z, set $U_h(z) = \bigcup_{s : \zeta_h(s) = z} U_h(s)$. If $|\zeta_h^{-1}(z)| > 1$, assume that at this block all actions in $U_h(z)$ have identical reward means and identical distributions of the next block. Equations (27)–(28) already make these quantities independent of the representative state within a block. The extra assumption requires independence of the supported action as well. No such extra assumption is imposed at singleton blocks.

Proposition E.1 (A sufficient abstraction condition)**.** *Under (27)–(29) and the supported-action equivalence condition above, every stochastic memoryless policy supported on $U_h(s)$ has linear action values, globally in the given features, with parameter norm at most $R + (H - h)M$ at stage h. In particular, Assumption 1 holds with $L_\theta = R + (H - 1)M$.*

Proof. Fix any stochastic memoryless policy ρ with $\operatorname{supp} \rho_h(\cdot \mid s) \subseteq U_h(s)$ at every pair (h, s). We prove backward that v_h^ρ is constant on each block and q_h^ρ has a global linear representation. At layer $H+1$ the value is zero, hence blockwise constant. Suppose $v_{h+1}^\rho(s') = \bar{v}_{h+1}^\rho(\zeta_{h+1}(s'))$. The Bellman identity and the two factorizations give, for every (s, a),

$$q_h^\rho(s, a) = \phi_h(s, a)^\top \theta_h^\rho, \qquad \theta_h^\rho = \eta_h + \sum_{z'} m_h(z') \bar{v}_{h+1}^\rho(z'). \tag{30}$$

At a nonsingleton block, the common reward and next-block law make this q-value the same for every action in $U_h(z)$ and every state in the block. Each state may mix these actions differently, but its state value is still that common number. At a singleton block, blockwise constancy is automatic. This completes the induction. Since rewards lie in $[0, 1]$, $\|\bar{v}^{\rho}_{h+1}\|_\infty \le H - h$. Equation (29) therefore gives $\|\theta^{\rho}_h\|_2 \le R + (H-h)M$. Every replacement policy is supported on U, as proved in Section E.1, so the final assertion follows. □

The criterion restricts only aggregate transitions, not the distribution within a destination block. It can therefore apply when the concrete transition matrix has much larger rank than the feature dimension. It also imposes no occupancy domination: Assumption 2 remains a separate condition. Bounds L_ϕ and $R + (H-1)M$ can be used as the public norm promises without revealing the factors, abstraction, logger or support sets.

For Proposition 5.1, retain the root as a singleton and merge all second-stage states x_i into one block. With the features of Section D.1, take $\eta_1 = (0, 1/2)$, $m_1(\text{second}) = (2, 0)$, $\eta_2 = (u, 1)$ and $m_2(\text{terminal}) = (1, 1)$. Thus $R = \sqrt{5}/2$ and $M = 2$ work uniformly in u and m. All supported terminal actions have reward mean u and the same terminal transition; the unlogged action zero has a different reward and is not subject to this action-equivalence requirement. The criterion proves support realizability in dimension two. The direct witnesses in Section D.1 give the sharper radius $L_\theta = 2$.

E.3 Relations to standard structural assumptions

Linear and low-rank MDPs. Taking ζ_h to be the identity reduces the factorization above to a linear MDP in the given features (Jin et al., 2020). Every block is a singleton, so the action restriction disappears. The same induction then applies to every policy, with the explicit radius $R + (H-1)M$. A low-rank transition factorization gives this implication only when its left factor is the supplied feature map, rewards lie in that span, and the factor norms are controlled. Rank alone gives none of these requirements. For example, at $H = 1$ a constant scalar feature and two different action reward means fail even target realizability, although both actions have the same terminal transition and the transition rank is one. Conversely, with zero rewards and the zero scalar feature every policy value is realized with radius zero, while transition matrices between layers may have arbitrarily high rank. Thus replacement realizability places no dimension-only bound on transition rank.

Bellman completeness. For a next-stage function f, write $\mathcal{T}^{\rho}_h f(s, a) = r_h(s, a) + \sum_{s'} P_h(s' \mid s, a) \sum_{a'} \rho_{h+1}(a' \mid s') f(s', a')$. If prescribed bounded linear classes contain terminal zero and are closed under these operators for every replacement continuation, backward induction gives Assumption 1 with those class radii. Closure of the unbounded spans alone establishes representation, without a specified common norm budget.

Closure for the target alone is insufficient. In Section E.6, root features are one, terminal features equal action $a \in \{0, 1\}$, and the target chooses zero at both terminal states. For any terminal linear function $f(s, a) = a\theta$, its target backup is zero at both root actions; the terminal reward is itself linear. Hence the target Bellman operator preserves the feature spans (also the unit-radius classes), yet the replacement continuation used there has root values $(0, 1/2)$, which are not representable. In the opposite direction, Proposition 5.2 satisfies all-policy realizability but violates even target Bellman completeness: the distinct successor features have no action dependence, so the nonlinear-in-current-features backup in Section D.2 is a target backup. Thus neither implication holds between target completeness and replacement realizability in the supplied features.

State aggregation. An exact reward/transition abstraction preserves blockwise values for policies whose action kernels are constant within each block. Replacement weights are allowed to vary across concrete states even when b and e are blockwise constant. The endpoint example just used makes this issue explicit: merge its two terminal states, whose rewards and abstract transitions coincide action by action. Both endpoints respect this abstraction, but switching to the target at only one terminal state produces the unequal root values. Ordinary aggregation therefore does not guarantee the needed representation. The action-equivalence condition in Proposition E.1 ensures that these state-dependent choices cannot change the blockwise value.

Bellman rank. Bellman rank factors averaged Bellman errors over a chosen function class and roll-in policy class (Jiang et al., 2017, Definition 5). This differs from representing each replacement policy's action values at every pair. To make the distinction precise, let π_f be the greedy policy of a bounded candidate f, and let $g_h^f(s)$ be its expected Bellman residual at state s when actions at stages $h, h+1$ follow π_f. Then

$$\mathcal{E}_h(f, f') = \sum_{s \in \mathcal{S}_h} d_h^{\pi_{f'}}(s) g_h^f(s).$$

The endpoint counterexample has at most two states per layer. This factors every stagewise error matrix through two coordinates, so its Bellman rank is at most two. For candidates in $[0, 1]$, the state-distribution factor has norm at most one and the residual factor at most $2\sqrt{2}$, satisfying the norm requirement as well. This is the statewise argument underlying Jiang et al. (2017, Proposition 1). It applies, in particular, to the bounded scalar-feature candidate class in the example, yet those features fail replacement realizability. Small Bellman rank alone thus does not imply Assumption 1. We assert no reverse rank bound without fixing the candidate and roll-in classes.

E.4 Full-support extrapolation

Proposition E.2 (Full-support extrapolation)**.** *If $b_h(a \mid s) \geq p > 0$ at every valid decision pair, Assumption 1 implies that every stochastic memoryless policy has a realizing parameter at stage h of norm at most $L_\theta(4/p - 1)^{H-h}$.*

We prove Proposition E.2. Suppose $b_h(a \mid s) \geq p > 0$ at every valid decision pair. Normalization of any logger action kernel implies $p \leq 1$. Let $t = p/2$, $\rho^0 = (b + e)/2$ and $\rho^1 = (1 - t)\rho^0 + t\pi$ for an arbitrary stochastic policy π. The midpoint is a replacement policy with constant weight $1/2$. For the tilted policy use

$$\lambda_h(s, a) = \frac{1}{2} - t \frac{\pi_h(a \mid s) - \rho_h^0(a \mid s)}{b_h(a \mid s)}.$$

At valid pairs, $|\pi_h - \rho_h^0| \leq 1$ and $t/b_h \leq 1/2$, so these weights lie in $[0, 1]$. Summing gives $\sum_a b_h(a \mid s)\lambda_h(s, a) = 1/2$. Equation (3) therefore equals

$$b_h(a' \mid s)/2 + t[\pi_h(a' \mid s) - \rho_h^0(a' \mid s)] + e_h(a' \mid s)/2 = (1 - t)\rho_h^0(a' \mid s) + t\pi_h(a' \mid s).$$

Weights outside valid decision layers can be set to zero. The resulting replacement kernels agree with the tilted kernels wherever an MDP trajectory can use them, which suffices for every valid-layer action value.

At each future stage, $\pi = t^{-1}\rho^1 - (1 - t)t^{-1}\rho^0$ with coefficients summing to one. Action values at stage h are separately affine in each of the later stage kernels: Bellman expansion visits each time index once, and its path weights are products of these kernels. This identity also follows by induction on the remaining horizon, distributing a signed combination through the action sum and retaining the reward term because the coefficients sum to one. Expanding all $H - h$ future stages expresses q_h^π as a signed sum of 2^{H-h} action values of stage splices of ρ^0, ρ^1. Each splice belongs to $\mathcal{R}(b, e)$. The sum of absolute coefficients is

$$\left(t^{-1} + (1 - t)t^{-1}\right)^{H-h} = (4/p - 1)^{H-h}.$$

Take the same signed sum of their realizing parameters and apply the triangle inequality. This proves the stagewise radius and the uniform bound $L_\theta(4/p - 1)^{H-1}$, with no state-count factor. Conversely, all-policy realizability implies replacement realizability. Thus with finite spaces and full support, the two exact representation requirements agree when no fixed parameter budget is stipulated. The displayed conversion bound need not preserve the original budget and is not claimed sharp.

For the support condition of Corollary 3.3, full logger support makes U equal to the entire action set on valid layers. Its realizability requirement therefore equals all-policy realizability with the *same* norm budget. Kernels outside valid layers can be replaced by b without changing a queried action value, so no positivity assumption is needed at invalid indices.

E.5 A NORM SEPARATION DESPITE FULL SUPPORT

For $0 < \gamma \le 1/2$, take $H = 2$, actions zero and one, and root actions leading respectively to second-stage states x, y. Root reward is zero. Set $c = 1/4$ and let every terminal reward mean be c except action one at x, which pays one. The target always chooses zero. The logger is uniform at the root and chooses terminal action one with probability γ at both states. It has full support, with every valid action probability at least γ. The target ratios are two at the root and $2/(1-\gamma)$ at $(x, 0)$, so $C_e = 2/(1-\gamma) \le 4$ exactly.

Choose

$$\phi_1(s_1, 0) = (c, \gamma(1-c)), \quad \phi_1(s_1, 1) = (c, 0), \qquad \phi_2(s, a) = (r_2(s, a), 0).$$

Their norms are at most one. A replacement policy's probability z of action one at x lies in $[0, \gamma]$, since replacing a logger action by the target can only remove that action. Its witnesses are $\theta_1 = (1, z/\gamma)$ and $\theta_2 = (1, 0)$, with norms at most $\sqrt{2}$. In contrast, a policy choosing action one at x has root values $(1, c)$. The second root action forces the first parameter coordinate to be one; the first then forces the second coordinate to be $1/\gamma$. Every realizing parameter has norm at least $1/\gamma$, unbounded as $\gamma \downarrow 0$. For example, $\gamma = 1/4$ already separates the common budget $\sqrt{2}$. This example separates norm requirements. Since the sample upper bound depends only logarithmically on the parameter budget, this norm gap does not imply a comparable sample-complexity gap or a statistical lower bound.

E.6 ENDPOINT REALIZATION IS INSUFFICIENT

Proposition E.3 (Two policy-quantifier boundary tests)**.** *Realizing only the logger and target does not imply replacement realizability, even with $H = 2$, $d = 1$, unit norms and $C_e = 2$. In contrast, realizing every deterministic memoryless policy with a common norm bound is equivalent to realizing every stochastic memoryless policy with that same bound. The latter equivalence also holds for policies supported on U.*

For the first assertion of Proposition E.3, use the same binary root transitions, zero root reward, and terminal reward equal to the action in $\{0, 1\}$. Let the scalar feature be one at the root and equal to the reward at the terminal stage. Both logger and target are uniform at the root. At each terminal state the logger is uniform, while the target chooses zero. The logger's root action value is $1/2$ and the target's is zero, independent of the current action; their terminal action values equal the common feature. Both are realized with parameter norm at most one, and the feature norm is at most one. The target/logger ratio is one at the root and two at each occupied terminal pair, so $C_e = 2$.

The replacement policy using the target at x and the logger at y has root action values $(0, 1/2)$. Identical scalar root features cannot represent them with any common parameter. Thus the endpoint assumption fails to imply the one used in our proof. This is a counterexample to an implication, not a claim of statistical impossibility under endpoint realizability alone.

E.7 DETERMINISTIC AND STOCHASTIC POLICY QUANTIFIERS

For a stochastic policy π, independently sample an action at every stage/state according to $\pi_h(\cdot \mid s)$. This gives a distribution on finitely many deterministic decision policies. The weight of a stage rule $f : \mathcal{S}_h \to \mathcal{A}$ is $\prod_{s \in \mathcal{S}_h} \pi_h(f(s) \mid s)$; these weights are nonnegative, sum to one, and have action marginal $\pi_h(a \mid s)$. A path in the layered model visits each stage once, so averaging its probability under the sampled deterministic policy gives its probability under π. The current action in q_h is fixed, and the same identity applies to every future path and reward. Hence q_h^π is the corresponding convex combination of the deterministic-policy action values, simultaneously at every valid current pair. The Lean proof establishes this identity stage by stage using single-stage affinity, rather than assuming a value-mixture representation.

Take the same convex combination of the realizing parameters. Linearity is preserved and its norm is at most the common budget L_θ by convexity. The reverse implication follows because deterministic policies are stochastic policies. For policies supported on U, every deterministic component of nonzero weight is also supported on U; zero-weight components require no witness. The argument therefore proves both equivalences in Proposition E.3 without increasing the norm bound.

F FORMAL CORRESPONDENCE AND REPRODUCIBILITY

F.1 PAPER-TO-LEAN MAP

The development uses Lean 4.33.1 and Mathlib `v4.33.1`. Names in Table 3 have prefix `CoverageFrontier`; most statistical declarations are in its `SafeAdvantage` namespace.

Table 3: **Formal proof interfaces for the linear-dimension theorem.** All probability and policy invariants are derived inside the final endpoint.

Paper component	Module and declaration
Population policy classes	`RestrictedRealizability`: `InReplacement`, `MDP.ReplacementRealizable`
Public rule	`ScalarPublic`: `estimateLinearN`, `measurable_estimateLinearN`
Deterministic count selection	`ScalarPublic`: `estimateLinearAtCount_guarantee`
Finite-design conditioning	`FiniteDesignConcentration`: `cond_pi`, `scalar_fiber_mgf`
Weighted scalar confidence	`ScalarPrediction`: `ScalarRidge.fixed_design_tail`
Actual-law regression	`ScalarTrajectoryRegression`: `scalar_regression_tail`
Always-valid auxiliary suffix	`ProjectedAdvantage`: `projectedRun_suffix`, `projectedRun_inReplacement`
Queried output agreement	`ObservedQueryCoupling`: `coupled_estimate`
Query-event probability	`ProjectedQueryProbability`: `queryCoupling_failure`
Target transfer and Theorem 3.1	`ProjectedError`: `scalar_block_estimate_tail`
Exact count	`ScalarSampleRate`: `linearSampleCount_sufficient`
Ordinary-data guarantee	`ScalarPublic`: `estimateLinearN_guarantee_of_replacement`

The uniform endpoint is

`exists_uniform_linear_replacement_trajectory_estimator`.

It instantiates one total measurable function with `SafeAdvantage.estimateLinearN`, before the compatible unknown MDP, reward law and logger. Replacing `replacement` by `support` or `target_only` gives the support and all-policy specializations. The observation type has $H+1$ states, H actions and H bounded reward coordinates. A public default action fills unused coordinates in an internal infinite-path extension; no observed decision depends on it.

The formal map is defined at every stage/state/action index; extending the paper's features by zero outside valid layers meets this interface. Realizing parameters are required only on valid layers. The formal recursion computes an unused root correction, omitted in Algorithm 1; both output the same root target average. The return proof can retain the interpolated root kernel, since its target-root action-value average is unchanged by replacing that kernel with e_1. The formal union bound uses the looser query count nH^2, which still fits the allocated $\delta/4$ budget. These interface choices add no statistical premise or learner input.

`RestrictedSeparation` establishes Proposition 5.1, with branch count m three above the Lean parameter; `FamilyF3` and the all-policy endpoint give Proposition 5.2. `MinimumNormFQE` proves the two-stage population fits in Section D.4. The full-support, norm and endpoint results are in `FullSupportExtrapolation`, `FullSupportNormBoundary` and `EndpointRealizability Boundary`, respectively. `DeterministicRealizability` proves the policy-quantifier equivalences.

F.2 Verification scope

The environment audit checks all project declarations, including private names, against the axiom allowlist `propext`, `Classical.choice`, `Quot.sound`. The theorem manifest lists hypotheses and public inputs. The code documentation gives dependency versions and verification commands.

The certificate covers the exact-real statistical endpoint and the original structural examples, including the two-stage FQE calculation in Section D.4. The three-stage variant, abstraction criterion, statistical benchmarks and discounted comparisons are paper proofs. Numerical diagnostics check finite examples below the sufficient sample threshold; they certify neither the probability theorem nor floating-point correctness.

G Statistical benchmarks and the remaining gap

This appendix supplies paper proofs and a version-specific literature reduction; these statistical lower bounds are outside the Lean certificate. All examples retain iid complete reward-inclusive trajectories from one fixed Markov logger. They are all-policy realizable, hence also satisfy replacement realizability. Separate hard subclasses provide separate lower bounds, not their product.

G.1 Coverage, accuracy and confidence are necessary

Put $G = \max\{1, H-1\}$. For every $C \geq 1$, $0 < \varepsilon \leq G/8$ and $0 < \delta \leq 1/4$, consider common-public-input instances with $d = 1$, $L_\phi = 1$, $L_\theta = G$ and target coverage exactly C. Any estimator that is ε-accurate with failure at most δ on every instance in this class requires

$$N \geq \frac{CG^2}{256\varepsilon^2} \log \frac{1}{\delta}. \tag{31}$$

Since $G \geq H/2$, this is $\Omega(CH^2\varepsilon^{-2}\log(1/\delta))$.

Legal complete-trajectory experiment. The target chooses an informative root action; the logger chooses it with probability $1/C$ and otherwise chooses a dummy action. At $H = 1$, the informative reward is Bernoulli(p) and the dummy reward is zero, with the same deterministic terminal transition. At $H \geq 2$, the root reward is zero. The informative action draws a next-state bit with probability p of one; the dummy action produces bit zero. At later layers the bit persists, every action has the same effect, and the reward is the current bit. Both target and logger choose a fixed action after the root. This is a finite layered MDP under the conditional product convention: every random reward has a deterministic next state, and every random transition has deterministic reward.

Use feature one at the informative root action, zero at the dummy, and the state bit at later layers, independently of action. Parameters Gp at the root and $H - h + 1$ at layer $h \geq 2$ realize every policy with radius G. The target return is Gp. Root coverage is exactly C; subsequent common transitions preserve occupancy domination. A complete informative trajectory contains only one random bit. Its repeated states and rewards are copies, not further independent observations. Thus for the two data laws

$$D_{\mathrm{KL}}(P_{p_-}^{\otimes N} \| P_{p_+}^{\otimes N}) = \frac{N}{C} \mathrm{kl}(p_- \| p_+).$$

Testing bound. Set $a = 2\varepsilon/G \leq 1/4$ and $p_\pm = 1/2 \pm a$, so the returns differ by 4ε. Since $\log(1+x) \leq x$,

$$\mathrm{kl}(p_- \| p_+) = 2a \log \frac{1+2a}{1-2a} \leq 16a^2 = 64\varepsilon^2/G^2.$$

Thresholding an accurate estimate at $G/2$ gives a test with both errors at most δ. The log-sum inequality applied to its two decision regions (equivalently, data processing for relative entropy) gives

$$\frac{64N\varepsilon^2}{CG^2} \geq \mathrm{kl}(\delta \| 1-\delta) = (1-2\delta)\log\frac{1-\delta}{\delta} \geq \tfrac{1}{4}\log(1/\delta).$$

For the last inequality, use $1 - 2\delta \geq 1/2$ and $1 - \delta \geq \sqrt{\delta}$ when $\delta \leq 1/4$. This proves Equation (31). Randomized estimators are covered by appending an independent common seed to both experiments.

G.2 The on-policy boundary

At $C = 1$, domination and equal total occupancy mass give $\nu_h^e = \nu_h^b$ at each stage, hence $J(e) = J(b)$. Averaging the logged returns, which lie in $[0, H]$, gives failure at most $2\exp(-2N\varepsilon^2/H^2)$ by Hoeffding's inequality. Together with Equation (31), this matches $\Theta(H^2\varepsilon^{-2}\log(1/\delta))$ for the class containing the norm-G examples above. No dimension factor is necessary at this boundary. More generally, occupancy domination gives $|J(e) - J(b)| \le H(1 - 1/C)$, so a dimension lower bound must restrict accuracy below this near-on-policy bias scale or keep C away from one.

G.3 Why linear dimension is a meaningful benchmark

For fixed H, sufficiently large d, $C \ge 2$, $L_\phi = 1$ and $L_\theta = \sqrt{d}$, a subclass requires $\Omega(Cd/\log d)$ observations at sufficiently small fixed accuracy and confidence $3/4$. This follows from Ma et al. (2022, Theorem 5, Eq. (28)).[3]

At $H = 1$, take d actions, one-hot features, and the uniform target. Target coverage is equivalent to $b(a) \ge \nu := 1/(Cd)$ for every action. Every bounded reward vector is all-policy realizable with parameter radius $\sqrt{d}$. The source considers an unknown fixed logger and fixed sample count. Its theorem gives a positive universal constant MSE lower bound at $N_0 = \lceil ACd/\log d \rceil$ for a sufficiently large fixed A: the conditions $\nu < \log d/N_0$, $\nu \le 1/(2d)$ and $\nu \ge c/(N_0 \log d)$ hold for sufficiently large d, and its term $\exp[-200\sqrt{N_0\nu\log d}]$ stays bounded below by a positive constant r_0. Monotonicity extends that lower bound to smaller sample counts.

To convert MSE to fixed confidence, suppose an n-sample estimator succeeds with probability $3/4$ at accuracy ε_0, where $\varepsilon_0^2 \le r_0/4$. Clip it to $[0, 1]$, run it on t independent blocks, and take the median. For a fixed sufficiently large odd t, the median fails with probability at most $\exp(-t/8) \le r_0/4$, so its MSE is at most $r_0/2$. If $tn \le N_0$, this contradicts the source lower bound. Therefore $n = \Omega(Cd/\log d)$. Appending deterministic zero-reward steps embeds the same observation experiment at any fixed horizon, without revealing additional information or changing the norm budget.

The lower bound uses a growing parameter radius, which enters our upper bound only logarithmically. It does not establish dimension hardness with both original norm budgets fixed at one, nor a joint Cd/ε^2 lower bound. Together with the linear-d upper rate, it matches the polynomial dimension dependence up to logarithms at fixed H, ε, δ in the stated norm regime.

H Comparison with feature-dynamics coverage

This appendix explicitly changes to the infinite-horizon discounted model of Amortila et al. (2026), arXiv:2601.19030v1. Its data are iid tuples (S, A, R, S', A'), with $(S, A) \sim \mu$, the reward and transition drawn from the MDP, and $A' \sim e(\cdot \mid S')$. They are not our complete-trajectory observations. Write

$$\nu_\gamma^e = (1-\gamma)\sum_{t\ge 0}\gamma^t \nu_t^e, \quad \phi_0 = \mathbb{E}_{S_0, A_0\sim e}\phi(S_0, A_0), \quad \Sigma = \mathbb{E}_\mu[\phi\phi^\top], \quad \Sigma_{\mathrm{cr}} = \mathbb{E}_\mu[\phi(S, A)\phi(S', A')^\top].$$

Their Assumptions 1–2 require target Q-realizability, bounded features, and invertibility of Σ and $A = \Sigma - \gamma\Sigma_{\mathrm{cr}}$. Theorem 1 and Eq. (11) use the population coefficient

$$C_{\mathrm{FD}} = (1-\gamma)^2 \phi_0^\top A^{-1}\Sigma A^{-\top}\phi_0. \tag{32}$$

The leading directional term in their bound is dimension-free; its burn-in and lower-order terms may depend on d and $1/\sigma_{\min}(A)$. Their Theorem 2 instead uses an empirical coefficient and an explicit dimension factor. Stability of the compressed dynamics is a condition for the infinite-series interpretation in their Proposition 1, not an additional blanket assumption of Theorem 1. Their Proposition 4 recovers aggregated concentrability for state-action abstractions; the resulting occupancy is that of the compressed model, which need not equal the true occupancy. All numbered references in this paragraph refer to the pinned version.

[3] Theorem numbering is taken from *arXiv:2101.07781v1*.

H.1 A POSITIVE BRIDGE UNDER TARGET COMPLETENESS

Proposition H.1 (Occupancy coverage controls feature-dynamics coverage under completeness)**.** *In the discounted model above, suppose Σ, A are invertible and the linear feature span is Bellman complete for e, as in Assumption 3 of Amortila et al. (2026). If $\nu_\gamma^e \le C_e \mu$ pointwise, then*

$$C_{\mathrm{FD}} = \left\| \mathbb{E}_{\nu_\gamma^e} \phi \right\|_{\Sigma^{-1}}^2 \le \mathbb{E}_\mu \left[\left(\frac{\nu_\gamma^e}{\mu} \right)^2 \right] \le C_e. \tag{33}$$

Proof. The equality is Proposition 5 of Amortila et al. (2026). Set $w = \nu_\gamma^e / \mu$ on the support of μ, and zero elsewhere; domination gives $0 \le w \le C_e$ and $\mathbb{E}_\mu w = 1$. Let $v = \mathbb{E}_\mu[w\phi]$ and $t = v^\top \Sigma^{-1} v$. Cauchy–Schwarz gives

$$t = \mathbb{E}_\mu[w \phi^\top \Sigma^{-1} v] \le \sqrt{\mathbb{E}_\mu w^2} \sqrt{\mathbb{E}_\mu (\phi^\top \Sigma^{-1} v)^2} = \sqrt{\mathbb{E}_\mu w^2} \sqrt{t}.$$

If $t = 0$ the claim is immediate; otherwise divide by $\sqrt{t}$ and square. Finally $w^2 \le C_e w$ implies $\mathbb{E}_\mu w^2 \le C_e$. □

This connects the two coefficients under an additional structural condition. It does not identify the two papers' observation protocols or transfer their finite-sample counts between those protocols.

H.2 A NONSINGULAR SEPARATION WITHOUT COMPLETENESS

Proposition H.2 (Fixed target coverage, diverging feature-dynamics coverage)**.** *For every $0 < p \le 1/2$ and $u \in [1/4, 1/2]$, there is a discounted MDP with $\gamma = 1/2$, supplied dimension $d = 3$, feature and all-policy parameter norms at most one, and normalized logger occupancy μ, such that $\nu_\gamma^e \le 4\mu$ with smallest constant four, both Σ, A are invertible, and*

$$C_{\mathrm{FD}} = \frac{1}{2} + \frac{1}{4p(2-p)}. \tag{34}$$

The features and target are independent of p, u. The compressed dynamics are nilpotent, so this separation does not rely on spectral instability.

Proof. Use states s, ℓ, r, z, actions $0, 1$, and initial state s. At s, action zero leads to ℓ and action one leads to r, with zero reward. Every action at ℓ, r gives an independent Bernoulli reward of mean u and leads to z. State z is absorbing with zero rewards. The target chooses actions $1, 1, 0, 0$ at s, ℓ, r, z, respectively. The logger is uniform at s, r, has $b(1 \mid \ell) = p$, and chooses zero at z. Let μ be its normalized discounted occupancy. Denote the coordinate vectors of $\mathbb{R}^3$ by e_0, e_1, e_2 and set

$$\phi(s,a) = e_0, \qquad \phi(\ell, 1) = e_2, \qquad \phi(\ell, 0) = \phi(r, 0) = \phi(r, 1) = e_1, \qquad \phi(z, a) = 0.$$

For every stochastic memoryless policy, action values are $u/2$ at s, u at ℓ, r, and zero at z. The common realizing parameter is $(u/2, u, u)$, of norm $3u/2 \le 3/4$. Thus all-policy realizability holds globally, including the unlogged absorbing action, with fixed norm bounds.

The nonzero occupancies are

(s,a)	$(s,0)$	$(s,1)$	$(\ell,0)$	$(\ell,1)$	$(r,0)$	$(r,1)$
$\mu(s,a)$	$1/4$	$1/4$	$(1-p)/8$	$p/8$	$1/16$	$1/16$
$\nu_\gamma^e(s,a)$	0	$1/2$	0	0	$1/4$	0

and both occupancies give mass $1/4$ to $(z, 0)$. The largest target/logger ratio is four at $(r, 0)$, independent of p, u. With target next actions as in their sampling protocol,

$$\Sigma = \begin{pmatrix} 1/2 & 0 & 0 \\ 0 & (2-p)/8 & 0 \\ 0 & 0 & p/8 \end{pmatrix}, \qquad \Sigma_{\mathrm{cr}} = \begin{pmatrix} 0 & 1/4 & 1/4 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \phi_0 = e_0. \tag{35}$$

All second-step transitions have zero next feature. Root transitions produce e_2 after action zero and e_1 after action one, explaining the two off-diagonal cross moments. The matrix $A = \Sigma - \Sigma_{\mathrm{cr}}/2$

is upper triangular with diagonal $1/2, (2-p)/8, p/8$, hence invertible for every $p > 0$. Solving $A^\top x = e_0$ gives

$$x = \left(2, \frac{2}{2-p}, \frac{2}{p}\right)^\top, \qquad C_{\mathrm{FD}} = \tfrac{1}{4} x^\top \Sigma x = \frac{1}{2} + \frac{1}{8(2-p)} + \frac{1}{8p} = \frac{1}{2} + \frac{1}{4p(2-p)}.$$

Moreover $B = (\Sigma^{-1}\Sigma_{\mathrm{cr}})^\top$ has only its first column nonzero, with zero first entry, so $B^2 = 0$. □

For comparison, the true target feature mean is $(1/2, 1/4, 0)^\top$, whose squared Σ^{-1} norm is $1/2 + 1/[2(2-p)]$, uniformly bounded. Compression makes the two root actions share one feature and propagates mass into the rare ℓ direction even though the target takes the right branch. Target completeness fails: a candidate with unequal e_1, e_2 coefficients has different target backups at the two identical root features. All actual policy values agree in those directions, which explains why all-policy realizability does not prevent the coefficient from diverging.

This is a separation of coverage requirements, not a statistical lower bound, an inconsistency claim for LSTDQ, or a counterexample to their theorem. Indeed the unknown reward mean can be estimated by pooling second-step rewards. The finite two-step version has the same target coverage and all-policy feature dimension, so our trajectory theorem has a count uniform in p. The displayed discounted calculation itself uses their tuple model; it makes no direct comparison between tuple and trajectory sample counts.

H.3 Target realizability need not realize replacements

For completeness, a discounted version of Section E.6 shows why their realizability promise does not imply ours. Use the same four-state transition graph as above, $\gamma = 1/2$, zero root and absorbing rewards, and deterministic reward a at either second-step state. Both logger and target are uniform at the root. At ℓ, r, the logger is uniform and the target chooses zero; both choose zero at z. Set $\phi(s,a) = e_0$, $\phi(\ell,a) = \phi(r,a) = ae_1$, and $\phi(z,a) = 0$. The target action value is represented by $(0,1)$, with unit feature and parameter norms. The normalized logger occupancy gives $\Sigma = \mathrm{diag}(1/2, 1/8)$ and $\Sigma_{\mathrm{cr}} = 0$, so $A = \Sigma$ is invertible and $C_{\mathrm{FD}} = 1/2$. Target Bellman completeness also holds: every next target feature is zero, so every target backup equals the reward. But the replacement policy using the target at ℓ and the logger at r has root action values $(0, 1/4)$. Identical root features cannot represent them. All tuples and rewards follow the same discounted sampling convention as above. Thus neither paper's assumptions can be obtained from the other's by merely renaming the coverage coefficient.