OFFLINE POLICY LEARNING FROM OUTCOME-ONLY FEEDBACK: IDENTIFICATION AND THE PRICE OF LOGGING MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

We study offline reinforcement learning when each logged episode records the full trajectory but only one noisy scalar outcome, with no per-step rewards. Under history-dependent logging, state–action coverage alone may not determine whether a target policy can be evaluated, because the data reveal rewards only through their sums along logged trajectories. With known dynamics, we show that a policy's value is identified when its occupancy lies in the span of logged trajectory features, that this condition is also necessary for interior rewards, and that the value admits a sharp identified interval when point identification fails. The coefficient $\mathcal{C}_{\mathrm{out}}(\pi;\mu) = d^{\pi\top}\Sigma_\mu^\dagger d^\pi$ is the exact additive change-of-measure constant and measures finite-sample difficulty.

We further show that loggers can visit each state–action pair equally often at every step yet combine those visits into different trajectories. Because each outcome reveals only the sum of rewards along a trajectory, this difference can change both identification and $\mathcal{C}_{\mathrm{out}}$. We quantify how logging memory can inflate $\mathcal{C}_{\mathrm{out}}$ through an order-necessary distortion factor κ_{mem}, together with a target-specific refinement. For policy learning, we propose a pessimistic algorithm based on a projected, box-constrained reward confidence set. It achieves a burn-in-free expected regret bound depending on $\mathcal{C}_{\mathrm{out}}$ of an optimal policy, with a high-probability guarantee obtained through block splitting. A temporal direct-sum family yields the exact relation $\mathcal{C}^\star_{\mathrm{out}} - 1 = H\kappa_{\mathrm{mem}}$ and exposes complementary Gaussian and missing-exposure lower-bound mechanisms. For general even horizons, a parallel-state construction establishes the minimax rate $\Theta(\max\{\sigma, \sqrt{H}\}\sqrt{dC/n})$ for the specified high-dimensional tabular class with $\mathcal{C}^\star_{\mathrm{out}} \le C$ in its non-clipped regime, showing that both the stochastic and low-noise terms are necessary. Matched-marginal experiments on compiler-derived additive geometry and at the Markov logging boundary confirm the predicted changes in outcome geometry and learning difficulty.

1 INTRODUCTION

Offline RL usually assumes that every logged transition carries a reward. In many real-world problems, however, the log records the full sequence of decisions but, instead of a reward at every step, only one final score: a game result, end-to-end latency, or total delivery cost. The learner observes episodes (τ_i, Y_i) with

$$Y_i = \langle \phi(\tau_i), r\rangle + \xi_i,$$

where $\phi(\tau) \in \{0,1\}^d$ records the stage-indexed state–action pairs visited by τ, $r \in [0,1]^d$ is the unknown reward, and ξ_i is conditionally σ-subgaussian noise; individual rewards are never observed.

With per-step reward feedback, state–action coverage provides the observations needed to evaluate a target policy. With outcome-only feedback, visiting the same state–action pairs need not reveal the reward combinations required for its value: the learner observes only trajectory-level sums, and different reward functions may remain indistinguishable from the logged data.

We therefore study the second moment of logged trajectory features, $\Sigma_\mu = \mathbb{E}_\mu[\phi\phi^\top]$, rather than relying only on one-step coverage. Its span determines whether a target value is identified, the coefficient $\mathcal{C}_{\mathrm{out}}$ built from it quantifies finite-sample difficulty, and both depend on cross-time correlations that one-step occupancies do not record.

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.

Contributions.

- **Policy-value identification and outcome coverage.** We characterize when a policy value is identified from logged outcomes: the target occupancy must lie in the span of logged trajectory features, with necessity holding for interior rewards. When point identification fails, we provide a sharp identified interval (Theorem 3.1). We also characterize $\mathcal{C}_{\mathrm{out}}$ as the exact additive change-of-measure constant, finite precisely on the logged span (Proposition 2.1).
- **Logging-memory distortion.** We derive an order-necessary bound that quantifies how logging memory can inflate $\mathcal{C}_{\mathrm{out}}$ relative to an occupancy-equivalent Markov logger, despite identical one-step occupancies. The bound uses a global distortion factor κ_{mem} and admits a target-specific refinement κ_{rep} (Theorem 3.3, Proposition C.6).
- **Burn-in-free policy learning.** With known dynamics, we propose a projected, box-constrained pessimistic algorithm with polynomial-time optimization and a one-shot expected regret bound depending on $\mathcal{C}_{\mathrm{out}}$ of an optimal policy, without a global minimum-eigenvalue condition. Block splitting gives the corresponding high-probability guarantee (Theorems 4.2 and D.9).
- **Minimax limits of policy learning.** We establish lower bounds from Gaussian testing and missing exposure for every even horizon. On a temporal direct-sum family, we obtain the exact relation $\mathcal{C}^{\star}_{\mathrm{out}} - 1 = H\kappa_{\mathrm{mem}}$. For the specified tabular class with an optimal policy satisfying $\mathcal{C}^{\star}_{\mathrm{out}} \le C$, we establish the minimax rate $\Theta(\max\{\sigma, \sqrt{H}\}\sqrt{dC/n})$ when $d \ge \log n$ and the stated non-clipping conditions hold (Theorem 5.1).

Relation to prior work. Classical linear-design theory characterizes estimable contrasts through span and pseudoinverse geometry. Recent trajectory-supervision work studies behavior- or Markov-policy logging (Xu et al., 2024; Jia et al., 2025; Yuan et al., 2025; Ren and Xie, 2026), for which one-step coverage is sufficient to control the trajectory-level information relevant to learning. We study history-dependent logging caused by budgets, quotas, or random choices made earlier in an episode. Even with the same one-step occupancies, such dependence can change trajectory spans and second moments, affecting identification, learning, and minimax rates.

Table 1: Closest statistical context: logging regime and coverage object in the cited guarantees.

Work	setting	logging / information object	relevant result
PARTED (Xu et al., 2024)	offline	trajectory-feature coverage	reward redistribution + pessimistic upper bounds
Jia et al. (2025)	RL theory	behavior-policy / standard coverage	change-of-trajectory-measure principle
Yuan et al. (2025)	trajectory feedback	Markov / marginal χ^2	sharpened marginal-χ^2 bound
Ren and Xie (2026)	offline	behavior-policy / state–action coverage	pessimistic algorithm + matching lower bound
This paper	offline	history-dependent logger / trajectory span and second moment	sharp partial identification; fixed-marginal memory distortion; projected learning without spectral burn-in; minimax rate on the stated even-H class

2 Outcome-Only Model and Coverage

Consider a finite-horizon MDP with known (P, ρ), horizon H, and tabular reward $r \in [0,1]^d$, $d = HSA$. A trajectory has one-hot visitation vector $\phi(\tau) \in \{0,1\}^d$ with $\|\phi(\tau)\|_2^2 = H$. The dataset contains n independent episodes generated by the same environment and a fixed logging rule μ that may depend on the full within-episode history. The logger may correlate actions across time but cannot alter the transition law. Outcomes satisfy

$$Y = \langle \phi(\tau), r \rangle + \xi, \qquad \mathbb{E}[\xi \mid \tau] = 0,$$

with conditionally σ-subgaussian noise. Each episode is therefore one noisy linear measurement of the reward vector, taken along its trajectory feature. For a Markov target policy π, its occupancy $d^{\pi} = \mathbb{E}_{\pi}[\phi]$ and value are $V^{\pi} = \langle d^{\pi}, r \rangle$. For policy evaluation, only this target functional must be identified. Appendix B gives the full model and Appendix M gives an estimated-dynamics ridge guarantee.

Let

$$\mathcal{U}_\mu = \operatorname{span}\{\phi(\tau) : q_\mu(\tau) > 0\}, \qquad \Sigma_\mu = \mathbb{E}_\mu[\phi\phi^\top],$$

so $\operatorname{range}(\Sigma_\mu) = \mathcal{U}_\mu$. The span $\mathcal{U}_\mu$ describes which reward combinations the log can reveal, while Σ_μ describes how strongly those combinations are represented: $v^\top \Sigma_\mu v = \mathbb{E}_\mu[\langle \phi(\tau), v\rangle^2]$. For a target policy, both enter the outcome coefficient

$$\mathcal{C}_{\mathrm{out}}(\pi;\mu) = \begin{cases} d^{\pi\top}\Sigma_\mu^\dagger d^\pi, & d^\pi \in \mathcal{U}_\mu, \\ +\infty, & \text{otherwise.} \end{cases}$$

It is the smallest C satisfying, for every additive trajectory functional $F_v(\tau) = \langle \phi(\tau), v\rangle$,

$$\big(\mathbb{E}_\pi F_v\big)^2 \le C\, \mathbb{E}_\mu[F_v^2]. \tag{1}$$

This is an exact target-direction quantity: it is finite precisely when the logged design contains the target occupancy direction. By contrast, the standard coefficient $C_{\mathrm{std}}(\pi;\mu) = \max_{d^\pi(h,s,a)>0} d^\pi(h,s,a)/d^\mu(h,s,a)$, with value $+\infty$ if such a denominator vanishes, only compares one-step marginals. In short, the logged span characterizes identification for interior rewards and provides a sufficient condition more generally (Section 3), while $\mathcal{C}_{\mathrm{out}}$ measures how hard the target is to learn from finite data (Sections 4–5).

Proposition 2.1 (Equivalent views of outcome coverage)**.** *Whenever $d^\pi \in \mathcal{U}_\mu$,*

$$\mathcal{C}_{\mathrm{out}}(\pi;\mu) = \sup_{v\neq 0} \frac{\langle d^\pi, v\rangle^2}{v^\top \Sigma_\mu v} = \min_{g:\, \mathbb{E}_\mu[g(\tau)\phi(\tau)] = d^\pi} \mathbb{E}_\mu[g(\tau)^2].$$

The minimum is attained by the importance-like score $g_\pi(\tau) = \phi(\tau)^\top \Sigma_\mu^\dagger d^\pi$. Thus $\mathcal{C}_{\mathrm{out}}$ is simultaneously a change-of-measure constant, a target leverage, and the minimum second moment of any linear unbiased trajectory reweighting that reconstructs the target occupancy.

The first equality is Lemma B.4; for the second, g_π is feasible and Cauchy–Schwarz against $\Sigma_\mu^\dagger d^\pi$ rules out a smaller second moment. The same score drives the burn-in-free proof of Section 4.

3 Policy-Value Identification and Logging Memory

This section addresses two questions: when logged outcomes determine a target policy's value, and how logging memory changes the information in the log even when one-step coverage is unchanged. We first show that identifying a policy value does not require recovering the entire reward vector.

Theorem 3.1 (Identification and sharp partial identification; main form)**.** *If $d^\pi \in \mathcal{U}_\mu$, then V^π is identified from the population law of (τ, Y). For every interior reward $r \in (0,1)^d$, the converse also holds. In general the sharp identified set is*

$$\left[\min_{u\in\mathcal{R}_0} \langle d^\pi, u\rangle, \max_{u\in\mathcal{R}_0} \langle d^\pi, u\rangle\right], \quad \mathcal{R}_0 = \{u \in [0,1]^d : \langle \phi(\tau), u\rangle = m(\tau)\ \forall \tau \in \operatorname{supp}(q_\mu)\}.$$

Consequently, reward-vector recovery is unnecessary: only the target functional must be spanned. The exact statement, including boundary degeneracies, is Theorem C.1 and Proposition C.2.

Proof idea. If $d^\pi \in \mathcal{U}_\mu$, write it as a linear combination of supported trajectory features; applying the same coefficients to the identified conditional means $m(\tau)$ recovers V^π. Conversely, if the target has a component z in $\mathcal{U}_\mu^\perp$, then every supported trajectory is orthogonal to z. For an interior reward, a sufficiently small perturbation $r + \epsilon z$ therefore leaves the entire law of observed (τ, Y) unchanged while shifting $\langle d^\pi, r\rangle$ by $\epsilon\|z\|_2^2$. The argument is exact observational equivalence and does not depend on the estimator or the sample size. The box constraint is why boundary rewards require the small qualification stated above.

Marginal coverage is insufficient. A two-step XOR family (Appendix C.1) makes the gap between one-step and trajectory coverage concrete. The logger draws its first action uniformly from $\{a, b\}$ and repeats it with probability $1-\varepsilon$, so every action marginal is uniform, whereas the optimal target plays a crossing pair. For this deterministic optimal target,

$$C_{\mathrm{std}}(\pi^\star;\mu) = 2, \qquad \mathcal{C}_{\mathrm{out}}(\pi^\star;\mu) = 1 + \frac{1}{\varepsilon} \quad (\varepsilon > 0),$$

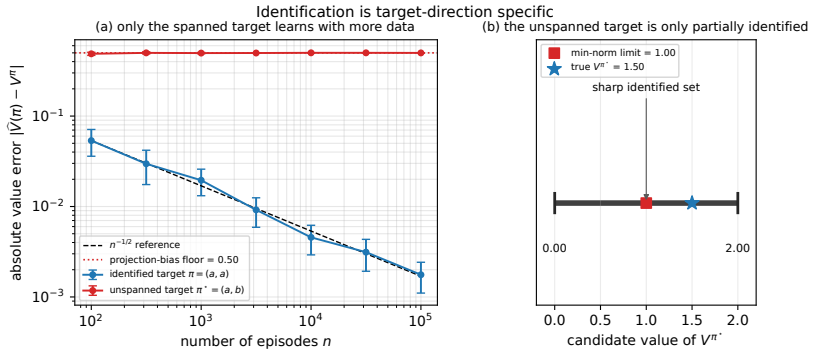

Figure 1: Identification diagnostic. For a spanned target direction, least-squares value error decays with data; for an unspanned target it converges to a nonzero projection bias, while the population identified interval remains nontrivial. The experiment is illustrative; the span statement itself is exact and population-level. Error bars are two-sided Student-t 95% CIs over 20 independent seeds.

while at $\varepsilon = 0$ the target value is not identified at all. Hence perfectly benign one-step coverage can coexist with arbitrarily weak trajectory-level credit information. The mechanism is visible directly in the trajectory law:

trajectory	(a, a)	(a, b)	(b, a)	(b, b)
q_{μ_ε}	$(1-\varepsilon)/2$	$\varepsilon/2$	$\varepsilon/2$	$(1-\varepsilon)/2$
$q_{\bar{\mu}}$	$1/4$	$1/4$	$1/4$	$1/4$

Both laws have the same one-step action marginals. Markovization therefore leaves every occupancy-based coverage statistic unchanged while replacing the cross-time correlation by independence.

At $\varepsilon = 0$, two reward worlds induce the same logged-data law, so the missing contrast becomes an algorithmic impossibility.

Theorem 3.2 (Nonidentification is an algorithmic impossibility; main form)**.** *At the endpoint $\varepsilon = 0$, for every $\Delta \in (0, 2]$ there are two reward vectors with identical logged-data distributions for every sample size n but opposite optimal crossing policies. Every algorithm therefore satisfies*

$$\max_{s \in \{+,-\}} \mathbb{E}_s[\mathrm{SubOpt}_s(\hat{\pi})] \geq \Delta/2,$$

even though $C_{\mathrm{std}}(\pi; \mu) \leq 2$ for every policy. Theorem E.1 gives the exact construction and shows that the two suboptimalities sum to Δ policy-by-policy.

At $\varepsilon = 0$, more data from the same logger cannot distinguish the two reward worlds. We next examine how logging memory changes the information in the log while one-step coverage remains fixed.

What the logger's memory costs. We compare a history-dependent logger μ with its occupancy-equivalent Markovization $\bar{\mu}$, which preserves every one-step occupancy. This isolates the effect of logging memory on trajectory information. Define

$$\kappa_{\mathrm{mem}}(\mu) = \inf\{\kappa : \Sigma_{\bar{\mu}} \preceq \kappa \Sigma_{\mu}\}.$$

Only cross-time covariance blocks can differ between the two designs.

Theorem 3.3 (Price of logger memory; main form)**.** *For every Markov target policy π,*

$$\mathcal{C}_{\mathrm{out}}(\pi; \mu) \leq \kappa_{\mathrm{mem}}(\mu) \left[1 + \sum_{h=1}^{H} \chi^2(d_h^{\pi} \| d_h^{\mu})\right] \leq \kappa_{\mathrm{mem}}(\mu)[1 + H(C_{\mathrm{std}}(\pi; \mu) - 1)]. \tag{2}$$

For Markov logging, $\kappa_{\mathrm{mem}} = 1$. In the two-step XOR family, $\kappa_{\mathrm{mem}} = \max\{1, (2\varepsilon)^{-1}, [2(1-\varepsilon)]^{-1}\}$ and an $\Omega(\sigma\sqrt{\kappa_{\mathrm{mem}}/n})$ lower bound holds while $C_{\mathrm{std}} = 2$. Hence the price of logger memory in a uniform marginal-to-trajectory reduction is order-necessary. See Theorem C.6 for the exact likelihood-ratio comparison and endpoint cases.

Boundary with policy-generated logging. If the logger is Markov, then $\bar{\mu} = \mu$ and $\kappa_{\mathrm{mem}} = 1$, so (2) reduces to

$$\mathcal{C}_{\mathrm{out}}(\pi; \mu) \le 1 + \sum_{h=1}^{H} \chi^2(d_h^\pi \| d_h^\mu) \le 1 + H(C_{\mathrm{std}}(\pi; \mu) - 1).$$

Thus ordinary marginal coverage does control outcome geometry in the policy-generated regime studied by recent work. The separation in this paper begins exactly when the logger can carry cross-time memory that is invisible to those marginals.

Reduction through Markovization. The Markovized logger has the same within-time covariance blocks $\mathrm{diag}(d_h^\mu)$ as the original logger, so C_{std} and every other marginal occupancy statistic are unchanged. The Loewner comparison $\Sigma_{\bar{\mu}} \preceq \kappa_{\mathrm{mem}} \Sigma_\mu$ transfers the variational ratio (1) from μ to $\bar{\mu}$ at cost κ_{mem}. We use the change-of-trajectory-measure argument of Jia et al. (2025), with the sharper marginal bound of Yuan et al. (2025). This second-moment comparison quantifies the distortion between the two trajectory laws. XOR is sharp because its covariance matrices are simultaneously diagonal and the generalized eigenvalue ratio equals the worst trajectory likelihood ratio.

The global factor κ_{mem} also accounts for directions irrelevant to a particular target. Proposition C.7 gives a target-specific bound using κ_{rep}, which can remain finite even when κ_{mem} is infinite; Section 6 illustrates this case.

Representation boundary. A finite-memory logger can be made Markov by appending its memory variable to the state whenever that variable is observed. State augmentation changes the state dimension, trajectory features, and occupancy objects, thereby defining a different coverage problem. The separation and κ_{mem} are therefore relative to the declared state–feature representation.

4 Projected Pessimism without a Global Eigenvalue Condition

Section 3 shows that logged outcomes may leave multiple reward vectors consistent with the available information; finite-sample noise adds further uncertainty. We address both through pessimism: each policy is evaluated against the least favorable reward in a confidence set. The ridge guarantee (Algorithm 1, Definition D.1) requires global design concentration to obtain a population bound. We instead retain the reward box and project the residual onto the realized observation space.

Let $\Phi_n \in \mathbb{R}^{n \times d}$ have rows $\phi_i^\top$, let P_n be the projector onto $\mathrm{range}(\Phi_n)$, and $k_n = \mathrm{rank}(\Phi_n)$. With

$$\gamma_n(\delta) = 2\sigma\sqrt{2(k_n \log 5 + \log(1/\delta))},$$

define

$$\mathcal{R}_n(\delta) = \left\{ u \in [0,1]^d : \|P_n(Y - \Phi_n u)\|_2 \le \gamma_n(\delta) \right\}. \tag{3}$$

Projected residual confidence set. Only the component of $Y - \Phi_n u$ in $\mathrm{range}(\Phi_n)$ can change with u. The orthogonal component is noise that no reward vector can explain, so accounting for it would require a larger confidence radius. The projector P_n removes this component without restricting the reward vector: the box $[0,1]^d$ remains intact. When $\sigma = 0$, the set contains exactly the bounded rewards consistent with the observed trajectory outcomes.

Projected box-constrained pessimism (Algorithm 2). Given $\{(\phi_i, Y_i)\}_{i=1}^n$ and δ:

1. Form Φ_n, its column-space projector P_n, and $\gamma_n(\delta) = 2\sigma\sqrt{2(k_n \log 5 + \log(1/\delta))}$.
2. Build the reward set $\mathcal{R}_n(\delta)$ in (3). If it is empty, return an arbitrary feasible Markov policy and stop; otherwise define $L_n(d) = \min_{u \in \mathcal{R}_n(\delta)} \langle d, u \rangle$.
3. Return any feasible occupancy $\hat{d} \in \arg\max_{d \in \mathcal{K}} L_n(d)$ and a Markov policy inducing $\hat{d}$.

For fixed d, the inner problem is an SOCP. Moreover L_n is concave, a minimizer u_d is a supergradient, and the known-dynamics occupancy polytope $\mathcal{K}$ has a polynomial-size linear description. Thus standard conic/ellipsoid machinery yields polynomial-time optimization; the full computational statement is Theorem D.11. The reference implementation instead uses a master linear

program with nonlinear inner solves; its floating-point certificates and CPU timings (Appendix J) describe that implementation, not the oracle model of the theorem.

Algorithm inputs. Algorithm 2 uses the realized pairs (ϕ_i, Y_i), an upper bound on the noise scale σ, and the known dynamics only through the feasible occupancy polytope $\mathcal{K}$. It does *not* require Σ_μ, $\mathcal{C}_{\mathrm{out}}$, C_{std}, κ_{mem}, or $\lambda^+_{\min}(\Sigma_\mu)$ as inputs. Those quantities describe population difficulty in the analysis rather than tuning parameters of the procedure. For a fixed occupancy d, minimizing and maximizing $\langle d, u\rangle$ over $\mathcal{R}_n$ also gives the finite-sample analogue of the population identified interval; policy optimization simply maximizes its lower endpoint.

On the confidence event, Algorithm 2's regret is controlled by how much rewards in $\mathcal{R}_n(\delta)$ can disagree on the optimal policy's value. To bound this disagreement for a target occupancy d, define the realized identification width

$$\mathcal{W}_n(d;\delta) = \sup\left\{\langle d, h\rangle :\ h \in [-1,1]^d,\ \|\Phi_n h\|_2 \le 2\gamma_n(\delta)\right\}.$$

Theorem 4.1 (Finite-sample data-dependent bound; main form)**.** *With probability at least $1-\delta$, Algorithm 2 satisfies* $\mathrm{SubOpt}(\hat\pi) \le \mathcal{W}_n(d^{\pi^\star};\delta)$. *The bound is analytical and is evaluated at the unknown optimal occupancy. This statement contains no empirical-covariance inverse, population eigenvalue, or design burn-in condition (Theorem D.7).*

Its premise $r \in \mathcal{R}_n(\delta)$ holds conditionally on every realized design with probability at least $1-\delta$ (Lemma D.6). A population regret bound then requires finite $\mathcal{C}^\star_{\mathrm{out}}$ for an optimal policy, rather than a threshold on the global spectrum.

Theorem 4.2 (Burn-in-free expected learning; main form)**.** *Assume $n \ge 2$. Choose $\pi^\star$ among the optimal policies to minimize $\mathcal{C}_{\mathrm{out}}(\pi;\mu)$, and write $\mathcal{C}^\star_{\mathrm{out}} := \mathcal{C}_{\mathrm{out}}(\pi^\star;\mu) < \infty$ (cf. Remark D.7). If Algorithm 2 uses $\delta = 1/n$, then*

$$\mathbb{E}[\mathrm{SubOpt}(\hat\pi)] \le \left[4\sigma\sqrt{2(d\log 5 + \log n)} + \sqrt{dH}\right]\sqrt{\frac{\mathcal{C}^\star_{\mathrm{out}}}{n}} + \frac{H}{n}. \tag{4}$$

In particular, for fixed d, H and finite $\mathcal{C}_{\mathrm{out}}$, the expected suboptimality goes to zero without requiring a lower bound on $\lambda^+_{\min}(\Sigma_\mu)$. The exact data-dependent width bound appears in Theorem D.8.

High probability by block splitting. Theorem 4.2 is a one-shot guarantee in expectation. For high probability, let $B = \lceil \log_2(2/\delta)\rceil$ and $\alpha = \delta/(2B)$. Split the data into B blocks, run Algorithm 2 at confidence α on each, and select the largest pessimistic certificate among blocks with nonempty reward sets (score $-\infty$ for empty sets; if all are empty, return an arbitrary feasible Markov policy). Theorem D.9 shows that, for $n \ge 2B$, with probability at least $1-\delta$,

$$\mathrm{SubOpt}(\hat\pi) \le 2\sqrt{2B}\left(4\sigma\sqrt{2[d\log 5 + \log(1/\alpha)]} + \sqrt{dH}\right)\sqrt{\frac{\mathcal{C}^\star_{\mathrm{out}}}{n}},$$

again without a global eigenvalue condition or knowledge of $\mathcal{C}^\star_{\mathrm{out}}$.

Corollary 4.3 (Memory-aware burn-in-free learning)**.** *Let $A_n = 4\sigma\sqrt{2(d\log 5 + \log n)} + \sqrt{dH}$. Combining Theorems 3.3 and 4.2 gives*

$$\mathbb{E}[\mathrm{SubOpt}(\hat\pi)] \le A_n\sqrt{\frac{\kappa_{\mathrm{mem}}(\mu)[1 + H(C_{\mathrm{std}}(\pi^\star;\mu) - 1)]}{n}} + \frac{H}{n}.$$

For a Markov logger, $\kappa_{\mathrm{mem}} = 1$; with history-dependent logging, the same one-step coverage can therefore lead to a larger learning difficulty. Algorithm 2 itself uses neither κ_{mem} nor C_{std}. If $C_{\mathrm{std}}(\pi^\star;\mu) < \infty$, Proposition C.7 also yields the target-localized variant of the display with $\kappa_{\mathrm{rep}}(\pi^\star;\mu)$ in place of $\kappa_{\mathrm{mem}}(\mu)$; this can remain informative even when global memory distortion is infinite.

Proof idea. Write $d^\star = d^{\pi^\star}$. On the event $r \in \mathcal{R}_n$, pessimistic optimality reduces policy regret to the width of $\mathcal{R}_n$ in the single direction $d^\star$. Any difference $h = r - u$ between two feasible rewards lies in $[-1,1]^d$ and obeys $\|\Phi_n h\|_2 \le 2\gamma_n$. The score $g_i = \phi_i^\top \Sigma_\mu^\dagger d^\star$ of Proposition 2.1 satisfies $\mathbb{E}[g_i\phi_i] = d^\star$ and $\mathbb{E}[g_i^2] = \mathcal{C}^\star_{\mathrm{out}}$, and gives

$$\langle d^\star, h\rangle = \frac{1}{n}\sum_{i=1}^n g_i\phi_i^\top h + \left\langle d^\star - \frac{1}{n}\sum_{i=1}^n g_i\phi_i, h\right\rangle.$$

Table 2: Two complementary pessimistic guarantees. Ridge pessimism is high-probability once the design has concentrated; projected pessimism removes the global spectral threshold.

method	uncertainty set	population statement	extra requirement
Ridge (Alg. 1)	empirical-Gram ellipsoid	high probability, target leverage	design concentration / $\lambda^{+}_{\min}(\Sigma_\mu)$ burn-in
Projected box (Alg. 2)	$[0,1]^d$ intersected with realized observation-space residual	expected; boosted high probability via $\mathcal{C}^\star_{\mathrm{out}}$	no population eigenvalue threshold

Cauchy–Schwarz controls the first term by $2\gamma_n\|g\|_2/n$, whose expectation is $O(\gamma_n\sqrt{\mathcal{C}^\star_{\mathrm{out}}/n})$. The second is controlled in ℓ_1; independence across episodes and $\|\phi\|_2^2 = H$ give expected size at most $\sqrt{dH\mathcal{C}^\star_{\mathrm{out}}/n}$. The argument controls only the optimal-policy direction through second moments, avoiding global empirical-Gram approximation. The one-shot theorem uses this second-moment control in expectation; the boosted corollary turns the resulting constant-probability width control into high probability across independent blocks, without assuming a tail bound on g.

The ridge alternative, finite-policy refinements, and computational guarantees are collected in Appendix D.

5 Lower Bounds and Minimax Rates

Theorem 4.2 gives Algorithm 2 an upper bound with a noise-driven term of order $\sigma\sqrt{d\mathcal{C}_{\mathrm{out}}/n}$ and a low-noise term of order $\sqrt{dH\mathcal{C}_{\mathrm{out}}/n}$. We now ask whether either term can be improved uniformly over the tabular class. We establish lower bounds using two mechanisms: Gaussian testing captures the difficulty caused by outcome noise, while missing exposure captures the difficulty of learning when informative trajectories are absent from a finite log, even at zero noise. We first examine how horizon and logging memory interact, then combine both mechanisms in a full tabular construction.

On the K-block temporal XOR family, $H = 2K$ and, for $\varepsilon = 1/(2\kappa)$, the occupancy-equivalent Markov logger has the same one-step marginals and exact distortion $\kappa_{\mathrm{mem}} = \kappa$, while the deterministic optimal target satisfies

$$\mathcal{C}^\star_{\mathrm{out}} - 1 = H\kappa_{\mathrm{mem}}, \qquad C_{\mathrm{std}}(\pi^\star;\mu) = 2.$$

Gaussian testing on this family (Corollary E.5) yields a lower bound of order $\sigma H\sqrt{\kappa_{\mathrm{mem}}/n}$. A complementary two-state temporal-gating family keeps $\mathcal{C}^\star_{\mathrm{out}} = C$ fixed while H grows and yields, for every even H,

$$\Omega\left(\min\left\{H, C-1, \max\{1,\sigma\}\sqrt{\frac{H(C-1)}{n}}\right\}\right),$$

with $C_{\mathrm{std}} \le 2$. The two families belong to the same minimax class but expose different sources of difficulty: outcome noise and missing trajectory information.

For the full tabular class, let $H = 2K$, let $M \ge 1$, and set $d = 2H(M+1)$. Let $\mathfrak{M}_H(d, C, \sigma)$ be the class of instances in our model with this horizon and dimension, Gaussian outcome noise of variance σ^2, general valid history-dependent logging, and best-optimal-policy coefficient $\mathcal{C}^\star_{\mathrm{out}} \le C$. Write $\mathfrak{R}_{n,H}(d, C, \sigma)$ for its worst-case expected policy suboptimality.

Theorem 5.1 (General-even-horizon near-minimax rate; main form)**.** *Let $C \ge 2$, $c = C - 1$, $H = 2K$, $M \ge 1$, and $d = 2H(M+1)$. Suppose*

$$\sqrt{\frac{Mc}{n}} \le \min\{1, c\}, \qquad \sigma\sqrt{\frac{KMc}{n}} \le \min\{K, c\}.$$

Then

$$\mathfrak{R}_{n,H}(d, C, \sigma) \ge \frac{1}{16\sqrt{2}}\max\{\sigma, \sqrt{H}\}\sqrt{\frac{d(C-1)}{n}}. \tag{5}$$

Algorithm 2 gives, for every instance in the class,

$$\mathbb{E}[\mathrm{SubOpt}(\hat\pi)] \le \left[4\sigma\sqrt{2(d\log 5 + \log n)} + \sqrt{dH}\right]\sqrt{\frac{C}{n}} + \frac{H}{n}.$$

Consequently, when $d \geq \log n$, the two sides match up to universal constants:

$$\mathfrak{R}_{n,H}(d, C, \sigma) = \Theta\left(\max\{\sigma, \sqrt{H}\}\sqrt{\frac{dC}{n}}\right)$$

throughout the stated non-clipped regime. The exact construction and constants are Theorem E.10.

The noisy branch uses M persistent hard states, each carrying K independent XOR credit bits. Assouad testing then yields the $\sigma\sqrt{dC/n}$ term. The low-noise branch instead gives each hard state one horizon-coherent sign. If neither informative whole-horizon template is observed, the logged data cannot distinguish its sign, even at zero outcome noise. Replicating this missing-exposure mechanism across hard states yields the $\sqrt{dHC/n}$ term. Thus both terms of the upper bound are necessary over the stated minimax class. The $H = 2$ result in Theorem E.8 remains a sharp special case with an explicit globally clipped formula.

Identification and minimax regimes. When a target lies outside the logged span, observationally equivalent reward worlds can make nonvanishing policy error unavoidable (Theorem 3.2). Finite $\mathcal{C}^{\star}_{\mathrm{out}}$ suffices for burn-in-free expected consistency (Theorem 4.2). Over the class of Theorem 5.1, the upper and lower bounds characterize the worst-case rate; additional structure can reduce the dependence on ambient dimension.

Theorem 5.1 is a worst-case statement. On a structured neutral-padded XOR family with s active credit directions in arbitrary ambient dimension $d = 4K$ (Appendix E.4), a learner given the active support attains, in the stated Gaussian regime, the matching local rate $\Theta(\sigma\sqrt{s(\mathcal{C}^{\mathrm{rep}}_{\mathrm{out}} - 1)/n})$, where $\mathcal{C}^{\mathrm{rep}}_{\mathrm{out}} = 1 + s/\varepsilon$ is the coefficient of a designated optimal representative that plays the behavior marginal on neutral blocks (Theorem E.12). Separately, Proposition E.13 shows that on this exactly neutral family the active support need not be recovered for policy regret: its bounds are uniform in the ambient number of neutral blocks.

Scope beyond the core model. Appendix G develops the linear-feature geometry, Appendix M gives an estimated-dynamics ridge guarantee, and Appendix I treats general trajectory features; approximation, transition, or planning errors then enter alongside the core outcome-geometry term.

6 Experiments on Coverage and Learning

Figure 1 illustrates identification. The remaining controlled diagnostics examine how logging memory affects coverage and policy regret, when burn-in-free guarantees apply, and how regret scales with dimension.

Logging memory at fixed marginals. Figure 2 compares controlled logging laws on compiler-derived additive geometry with synthetic Gaussian learning noise. A random flag-budget logger and its occupancy-equivalent Markovization have identical one-step marginals to machine precision, yet their $\mathcal{C}_{\mathrm{out}}$ values are 18.296 and 11.618, respectively, a ratio of 1.5748. Across the prespecified sample sizes, the mean RMSE ratio is 1.2966 (bootstrap 95% CI $[1.2225, 1.3845]$), compared with the population prediction $\sqrt{18.296/11.618} = 1.2549$. Four prespecified budget laws with the same expected flag count also preserve these marginals while changing $\mathcal{C}_{\mathrm{out}}$ and target-error ratios (Figure 2, right).

Memory can instead favor a fixed target: in a target-aligned construction with uniform marginals, κ_{mem} diverges as the target's $\mathcal{C}_{\mathrm{out}}$ falls from 9 to 2.0088. At the endpoint, $\kappa_{\mathrm{mem}} = \infty$ but the target remains point-identified with $\mathcal{C}_{\mathrm{out}} = 2$ (Figure 17). This illustrates why the target-specific κ_{rep} refinement can remain informative when the global factor diverges.

Policy consequence at the Markov boundary. In the two-step XOR family, every cell has $C_{\mathrm{std}} = 2$ and Markov marginal bound 3; moving from the independent Markov cell $\varepsilon = 0.5$ to $\varepsilon = 0.02$ changes only cross-time dependence and raises $\mathcal{C}_{\mathrm{out}} = 1 + 1/\varepsilon$ from 3 to 51. With the ridge redistribution estimator ($\lambda = 0.05$), policy class, noise, sample size, and marginals fixed, mean regret at $n = 500$ rises from 0.00025 to 0.05375 for greedy planning and from 0.00025 to 0.12475

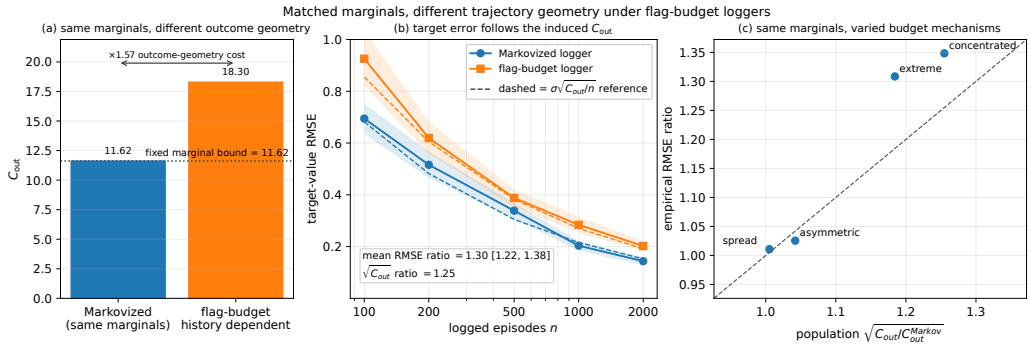

Figure 2: **Logging memory with matched marginals.** Left: a flag-budget logger and its Markovization have identical one-step marginals but different $\mathcal{C}_{\mathrm{out}}$. Middle: target-value RMSE with bootstrap uncertainty follows the corresponding $\sigma\sqrt{\mathcal{C}_{\mathrm{out}}/n}$ scales. Right: four budget laws preserve one-step marginals while changing trajectory geometry and target error. Geometry comes from the frozen $H = 8$ gcc table; learning noise is synthetic Gaussian.

for a finite-class ridge lower-confidence planner. Both planners are therefore sensitive to cross-time dependence; Appendix J reports paired intervals and the full sweep.

Learning beyond the ridge spectral threshold. At $n = 1000$, making an irrelevant XOR direction rare raises the sufficient global ridge burn-in from 556.2 to $99{,}831.8$ episodes while the optimal-target $\mathcal{C}_{\mathrm{out}}$ stays between 2 and 3. Algorithm 2 has mean suboptimality 0 in three cells and 8.50×10^{-5} in the fourth, without requiring a minimum-eigenvalue threshold (Figure 13).

Fixed-horizon minimax scaling. On the $H = 2$, $\mathcal{C}_{\mathrm{out}} = 5$ parallel-state family, absolute plug-in regret has log–log slope approximately 0.49 as d grows from 4 to 128, consistent with the $\sqrt{d}$ characterization; Algorithm 2 is certified for $d \leq 16$ (Figure 14).

7 Discussion and Limitations

With history-dependent logging, matching one-step occupancies is not enough: two loggers can contain different trajectory-level information and therefore have different identification properties and $\mathcal{C}_{\mathrm{out}}$. The logged span characterizes interior-reward value identification, finite optimal-policy $\mathcal{C}_{\mathrm{out}}$ supports expected learning without a global spectral threshold, and the stated non-clipped tabular lower bounds require both outcome noise and missing exposure (Theorem 5.1). For Markov loggers one-step coverage controls outcome geometry, whereas history-dependent cross-time coupling can dominate. Conversely, a fixed target can remain easy even when global memory distortion is infinite, motivating the target-localized κ_{rep} refinement. At the unidentifiable XOR endpoint, additional data from the same logger cannot recover the missing contrast (Theorem 3.2). Algorithm 2 requires neither $\mathcal{C}_{\mathrm{out}}$ nor κ_{mem} as an input.

Limitations. We assume independent episodes from one fixed logger, with history dependence only within episodes. Algorithm 2's burn-in-free guarantee assumes known dynamics; Appendix M gives an estimated-dynamics guarantee using ridge pessimism and design concentration; a burn-in-free unknown-dynamics treatment remains open. The memory comparison is relative to the declared state–feature representation, and $\kappa_{\mathrm{mem}}, \kappa_{\mathrm{rep}}$ need not be estimable from a finite log. Future work includes one-block high-probability control and joint treatments of unknown dynamics, approximation, nonadditivity, and model selection (Appendix K).

REPRODUCIBILITY STATEMENT

The appendices provide the complete assumptions, proofs, experimental definitions, sample sizes, noise laws, confidence-interval constructions, and all theorem-facing numerical values needed to interpret the results. The synthetic diagnostics use the fixed seeds reported with the artifact. For Algorithm 2, we report floating-point primal/dual max–min certificates. The exhaustive $H = 8$ compiler study uses a fixed table of 256 configurations with repeated gcc measurements. The optional $H = 16$ exercise reports machine-specific reference-implementation timing. The mathematical statements and experimental definitions are self-contained in the manuscript.

AI USE STATEMENT

We used generative AI tools to support manuscript preparation, literature discovery, research development, mathematical reasoning, and code implementation. Their specific roles are described below.

Writing and Presentation. AI assisted with drafting and revising manuscript text, organizing theoretical and empirical material, improving wording and readability, and checking LaTeX formatting and presentation consistency.

Retrieval and Discovery. AI supported literature discovery, related-work comparison, and preliminary novelty assessment. References and bibliographic information used in the manuscript were checked against the corresponding source publications.

Research Ideation and Execution. AI assisted with exploring research directions, refining methodological ideas, implementing and debugging research code, designing experiments, and interpreting empirical results. Experimental outputs were checked by rerunning the relevant scripts and comparing the reported quantities with the saved experiment outputs.

Mathematical Claims and Proofs. AI assisted with formulating and refining mathematical claims, developing proof strategies, checking assumptions, and identifying potential gaps. The resulting claims and proofs were subsequently checked against their stated assumptions and reexamined by the authors.

Synthetic Data. Generative AI was not used to generate synthetic datasets. All experimental data were obtained through the research procedures described in the paper.

AI-assisted mathematical arguments were checked against their assumptions; experimental code and figure-generation scripts were rerun against the saved outputs; and cited literature was checked against the corresponding source publications. The authors take responsibility for the final manuscript, results, proofs, and code.

REFERENCES

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24:2312–2320, 2011.

P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning*, 2004.

K. Amin, N. Jiang, and S. Singh. Repeated inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

P. Amortila, A. Huang, A. Krishnamurthy, and N. Jiang. A unifying view of coverage in linear off-policy evaluation. In *International Conference on Learning Representations*, 2026.

J. A. Arjona-Medina, M. Gillhofer, M. Widrich, T. Unterthiner, J. Brandstetter, and S. Hochreiter. RUDDER: Return decomposition for delayed rewards. *Advances in Neural Information Processing Systems*, 32, 2019.

H. Cao, S. N. Cohen, and L. Szpruch. Identifiability in inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12362–12373, 2021.

N. S. Chatterji, A. Pacchiano, P. L. Bartlett, and M. I. Jordan. On the theory of reinforcement learning with once-per-episode feedback. *Advances in Neural Information Processing Systems*, 34:3401–3412, 2021.

J. Chen and N. Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051, 2019.

X. Chen, H. Zhong, Z. Yang, Z. Wang, and L. Wang. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *International Conference on Machine Learning*, pages 3773–3793, 2022.

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

A. Cassel, H. Luo, A. Rosenberg, and D. Sotnikov. Near-optimal regret in linear MDPs with aggregate bandit feedback. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 5757–5791, 2024.

Y. Duan, Z. Jia, and M. Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709, 2020.

G. Elfving. Optimum allocation in linear regression theory. *The Annals of Mathematical Statistics*, 23(2):255–262, 1952.

Y. Efroni, N. Merlis, and S. Mannor. Reinforcement learning with trajectory feedback. In *AAAI Conference on Artificial Intelligence*, volume 35(8), pages 7288–7295, 2021.

S. Filippi, O. Cappé, A. Garivier, and C. Szepesvári. Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*, 23:586–594, 2010.

N. Felicioni, M. Ferrari Dacrema, M. Restelli, and P. Cremonesi. Off-policy evaluation with deficient support using side information. *Advances in Neural Information Processing Systems*, 35:30250–30264, 2022.

T. Fiez, L. Jain, K. G. Jamieson, and L. Ratliff. Sequential experimental design for transductive linear bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

D. J. Foster, A. Krishnamurthy, D. Simchi-Levi, and Y. Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. In *Proceedings of the 35th Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, page 3489, 2022.

T. Gangwani, Y. Zhou, and J. Peng. Learning guidance rewards with trajectory-space smoothing. *Advances in Neural Information Processing Systems*, 33:822–832, 2020.

B. Han, Z. Ren, Z. Wu, Y. Zhou, and J. Peng. Off-policy reinforcement learning with delayed rewards. In *International Conference on Machine Learning*, pages 8280–8303, 2022.

F. Chen, Z. Jia, A. Rakhlin, and T. Xie. Outcome-based online reinforcement learning: Algorithms and fundamental limits. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

R. Laroche and R. Tachet des Combes. On the occupancy measure of non-Markovian policies in continuous MDPs. In *International Conference on Machine Learning*, PMLR 202:18548–18562, 2023.

Z. Jia, A. Rakhlin, and T. Xie. Do we need to verify step by step? Rethinking process supervision from a theoretical perspective. In *International Conference on Machine Learning*, PMLR 267:27373–27398, 2025.

Y. Yuan, F. Chen, Z. Jia, A. Rakhlin, and T. Xie. Trajectory Bellman residual minimization: A simple value-based method for LLM reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

C. Jin, Z. Yang, Z. Wang, and M. I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143, 2020.

Y. Jin, Z. Yang, and Z. Wang. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pages 5084–5096, 2021.

N. Kallus and M. Uehara. Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020.

N. Kallus and A. Zhou. Confounding-robust policy improvement. *Advances in Neural Information Processing Systems*, 31, 2018.

N. Kallus and A. Zhou. Minimax-optimal policy learning under unobserved confounding. *Management Science*, 67(5):2870–2890, 2021.

C. Kausik, Y. Lu, K. Tan, M. Makar, Y. Wang, and A. Tewari. Offline policy evaluation and optimization under confounding. In *International Conference on Artificial Intelligence and Statistics*, pages 1459–1467, 2024.

K. Kim, S. Garg, K. Shiragur, and S. Ermon. Reward identification in inverse reinforcement learning. In *International Conference on Machine Learning*, volume 139, pages 5496–5505, 2021.

A. Kumar, A. Zhou, G. Tucker, and S. Levine. Conservative Q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

S. Levine, A. Kumar, G. Tucker, and J. Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

X. Liu, X. Dai, J. Zuo, S. Wang, C. Joe-Wong, J. C. S. Lui, and W. Chen. Offline learning for combinatorial multi-armed bandits. In *International Conference on Machine Learning*, PMLR 267:38251–38289, 2025.

Y. Liu, Y. Luo, Y. Zhong, X. Chen, Q. Liu, and J. Peng. Sequence modeling of temporal credit assignment for episodic reinforcement learning. *arXiv preprint arXiv:1905.13420*, 2019.

Y. Liu, A. Swaminathan, A. Agarwal, and E. Brunskill. Provably good batch reinforcement learning without great exploration. *Advances in Neural Information Processing Systems*, 33:1264–1274, 2020.

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

C. F. Manski. *Partial Identification of Probability Distributions*. Springer Series in Statistics. Springer-Verlag, New York, 2003.

R. Munos and C. Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9:815–857, 2008.

H. Namkoong, R. Keramati, S. Yadlowsky, and E. Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. *Advances in Neural Information Processing Systems*, 33:18819–18831, 2020.

A. Y. Ng and S. Russell. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, 2000.

A. Saha, A. Pacchiano, and J. Lee. Dueling RL: Reinforcement learning with trajectory preferences. In *International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 6263–6289, 2023.

R. Pathak, M. J. Wainwright, and L. Xiao. Noisy recovery from random linear observations: Sharp minimax rates under elliptical constraints. *The Annals of Statistics*, 52(6):2816–2850, 2024.

J. C. Perdomo, A. Krishnamurthy, P. L. Bartlett, and S. M. Kakade. A complete characterization of linear estimators for offline policy evaluation. *Journal of Machine Learning Research*, 24(284):1–50, 2023.

F. Pukelsheim. *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics, 2006.

P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, and S. Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

P. Rashidinejad, H. Zhu, K. Yang, S. Russell, and J. Jiao. Optimal conservative offline RL with general function approximation via augmented Lagrangian. In *International Conference on Learning Representations*, 2023.

X. Ren and T. Xie. When does trajectory-level supervision permit efficient offline reinforcement learning? *arXiv preprint arXiv:2606.18531*, 2026.

Z. Ren, R. Guo, Y. Zhou, and J. Peng. Learning long-term reward redistribution via randomized return decomposition. In *International Conference on Learning Representations*, 2022.

L. Shi, G. Li, Y. Wei, Y. Chen, and Y. Chi. Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International Conference on Machine Learning*, pages 19967–20025, 2022.

J. Skalse and A. Abate. Misspecification in inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2023.

J. Skalse, M. Farrugia-Roberts, S. Russell, A. Abate, and A. Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *International Conference on Machine Learning*, 2023.

A. Swaminathan, A. Krishnamurthy, A. Agarwal, M. Dudík, J. Langford, D. Jose, and I. Zitouni. Off-policy evaluation for slate recommendation. *Advances in Neural Information Processing Systems*, 30, 2017.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

R. Wang, D. P. Foster, and S. M. Kakade. What are the statistical limits of offline RL with linear function approximation? In *International Conference on Learning Representations*, 2021.

T. Xie and N. Jiang. $Q^\star$ approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, pages 550–559, 2020.

T. Xie, C.-A. Cheng, N. Jiang, P. Mineiro, and A. Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34:6683–6694, 2021.

W. Xiong, H. Zhong, C. Shi, C. Shen, L. Wang, and T. Zhang. Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent MDP and Markov game. In *International Conference on Learning Representations*, 2023.

Y. Xu, R. Wang, L. Yang, A. Singh, and A. Dubrawski. Preference-based reinforcement learning with finite-time guarantees. *Advances in Neural Information Processing Systems*, 33:18784–18794, 2020.

T. Xu, Y. Wang, S. Zou, and Y. Liang. Provably efficient offline reinforcement learning with trajectory-wise reward. *IEEE Transactions on Information Theory*, 70(9):6481–6518, 2024. doi:10.1109/TIT.2024.3427141.

B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

A. Zanette. Exponential lower bounds for batch reinforcement learning: Batch RL can be exponentially harder than online RL. In *International Conference on Machine Learning*, 2021.

W. Zhan, M. Uehara, N. Kallus, J. D. Lee, and W. Sun. Provable offline preference-based reinforcement learning. In *International Conference on Learning Representations*, 2024.

A. Zhou. Reward-relevance-filtered linear offline reinforcement learning. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3025–3033, 2024.

B. Zhu, M. I. Jordan, and J. Jiao. Principled reinforcement learning with human feedback from pairwise or K-wise comparisons. In *International Conference on Machine Learning*, volume 202, pages 43037–43067,

2023.

B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence*, pages 1433–1438, 2008.

Z. Zhang, Y. Chen, J. D. Lee, S. S. Du, and R. Wang. Minimax optimal regret bound for reinforcement learning with trajectory feedback. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 74692–74713, 2025.

A Related work

Trajectory-level supervision and the Markov boundary. The feedback model is established. Online work studies trajectory feedback, once-per-episode labels, and aggregate end-of-episode feedback (Efroni et al., 2021; Chatterji et al., 2021; Cassel et al., 2024; Zhang et al., 2025; Chen et al., 2025), while PARTED studies offline trajectory-wise reward with least-squares redistribution and pessimistic value iteration (Xu et al., 2024). For behavior- or Markov-policy data, Jia et al. (2025) introduce the change-of-trajectory-measure principle linking trajectory returns to step-level distribution shift. Yuan et al. (2025) sharpen this comparison to the marginal-χ^2 form used below. Ren and Xie (2026) develop pessimistic offline policy learning with matching lower bounds under Markov behavior-policy logging.

Assumption 2 instead permits any trajectory law generated by the environment and a valid history-dependent logger. Occupancy-equivalent Markovization is classical (Laroche and Tachet des Combes, 2023), and equal occupancy need not imply equal trajectory distributions. With one scalar outcome per trajectory, that difference enters the span and second moment of the logged trajectory features. The XOR and matched-marginal constructions isolate this boundary: they hold every declared one-step occupancy fixed while changing only cross-time coupling; at the independent Markov point the two regimes coincide.

Data-collection mechanisms with memory. History-dependent logging arises whenever the collection rule carries within-episode state beyond the environment state. Later decisions may depend on earlier logged choices, a remaining action budget or quota, or an episode-level randomized allocation. Such mechanisms can preserve every one-step visitation frequency while changing the joint trajectory law. The flag-budget and matched-marginal constructions in Appendix J instantiate this mechanism directly. Under outcome-only feedback, each scalar label is attached to the complete trajectory, so these cross-time couplings change the linear reward measurements available to the learner and therefore the trajectory span and second moment.

Linear estimability, design, and pseudoinverse OPE. The algebra behind target-functional coverage predates RL. In singular linear models, a contrast $c^\top\theta$ is estimable when $c \in \mathrm{range}(M)$, and $c^\top M^- c$ is the classical c-optimal variance criterion (Elfving, 1952; Pukelsheim, 2006). Transductive linear bandits likewise use design-dependent Mahalanobis geometry (Fiez et al., 2019). In contextual/slate OPE, pseudoinverse feature estimators exploit linear action structure to evaluate policies under support conditions weaker than full action support (Swaminathan et al., 2017; Felicioni et al., 2022). More recently, Amortila et al. (2026) develop target-dependent feature-dynamics coverage for linear OPE. Perdomo et al. (2023) give a complementary necessary-and-sufficient characterization of broad classes of linear estimators for offline policy evaluation under linear function approximation.

The span/pseudoinverse geometry itself is classical. Here the observation vectors are feasible MDP trajectories, and history-dependent logging can alter their design while leaving all one-step occupancies fixed. We trace this difference through identification, policy learning, and lower bounds. Offline combinatorial bandits provide another neighboring structured-action perspective (Liu et al., 2025).

Relative to these lines of work, the contribution is the outcome-only RL chain under a valid history-dependent logging law—sharp identification, the memory-distortion reduction, burn-in-free learning, and the even-horizon characterization—together with what changes when cross-time coupling varies at fixed one-step marginals.

Work	Logging/data class	Object and information condition	Necessary/sufficient identification?	Burn-in statement	Lower-bound / regime boundary
Xu et al. (2024) / PARTED	offline behavior-policy trajectories; trajectory-wise return	reward redistribution under trajectory-feature coverage	–	–	pessimistic upper bounds
Jia et al. (2025)	behavior-policy data; outcome vs process supervision	standard behavior-policy coverage	–	–	change-of-trajectory-measure comparison up to horizon factors
Yuan et al. (2025)	Markov-policy trajectory feedback	marginal χ^2 coverage and change of trajectory measure	–	–	sharpened Markov-regime trajectory comparison
Ren and Xie (2026)	offline Markov behavior-policy data	state–action coverage	–	–	pessimistic learning and matching lower bounds
Classical linear contrast/design literature (Elfving, 1952; Pukelsheim, 2006)	chosen linear measurements	contrast c; design matrix M; $c^\top M^\dagger c$	c estimable iff $c \in \mathrm{range}(M)$	–	classical contrast estimability and design variance
This paper	valid history-dependent trajectory logging	target value/sharp interval; trajectory span and second moment	necessary and sufficient at interior rewards; sharp partial identification	projected learning without global spectral burn-in	fixed-marginal memory separation and the stated general-even-H minimax regime

Table 3: Theorem-level comparison to the nearest statistical results. A dash denotes an item outside the comparison scope of the cited result.

Reward redistribution and credit assignment in practice. A large empirical literature converts an episodic return into per-step proxy rewards. RUDDER (Arjona-Medina et al., 2019) redistributes return via contribution analysis of a sequence model; IRCR (Gangwani et al., 2020) assigns each state–action pair the normalized return of the episodes containing it; randomized return decomposition (Ren et al., 2022) scales least-squares redistribution to long horizons by Monte-Carlo subsampling; related work treats delayed rewards in the off-policy setting (Han et al., 2022) and via sequence models (Liu et al., 2019). These methods target practical credit assignment rather than identification. Theorem C.1 characterizes target-value identification from logged outcomes, independently of the redistribution method used, and Corollary C.8 exhibits a regime in which no method can certify improvement over the logged behavior.

External-baseline availability. We considered PARTED, OPAC, and RUDDER for setting-matched trajectory-supervision comparisons (Xu et al., 2024; Ren and Xie, 2026; Arjona-Medina et al., 2019). At artifact preparation time, the available public implementations did not provide a setting-matched fixed-log baseline: PARTED and OPAC lacked a verified official implementation for this evaluation, while the available RUDDER implementation targets online interaction. The experiments therefore use the theorem-aligned baselines described in Appendix J.

Offline RL with per-step rewards. The statistical treatment of offline RL rests on pessimism: penalizing value estimates by data-dependent uncertainty yields bounds scaling with the concentrability of the *optimal* policy only (Jin et al., 2021; Rashidinejad et al., 2021; Xie et al., 2021; Liu et al., 2020; Levine et al., 2020), refining the uniform concentrability of the earlier fitted-value-iteration analyses (Munos and Szepesvári, 2008; Chen and Jiang, 2019; Xie and Jiang, 2020). Sharper tabular and linear results (Shi et al., 2022; Xiong et al., 2023; Rashidinejad et al., 2023) and practical algorithms (Kumar et al., 2020) follow the same template. The term *burn-in cost* is also used in model-based tabular offline RL for the sample threshold needed to enter the sample-optimal regime (Li et al., 2024). Here, *spectral burn-in* refers specifically to the sample threshold required for empirical trajectory-design domination in Lemma L.5. Our Algorithm 1 is the trajectory-feature analogue, but the relevant coverage notion changes from per-step density ratios to the spectrum of a trajectory-level second-moment matrix; Proposition C.4 shows the two notions are not comparable, and Appendix M shows that with unknown dynamics the transition error is governed by the *standard* coefficient while the reward error is governed by $\mathcal{C}_{\mathrm{out}}$.

Lower bounds and barriers in offline RL. Information-theoretic limits for offline RL with per-step rewards include the linear-realizability barriers of Wang et al. (2021) and Foster et al. (2022), exponential separations between batch and online RL (Zanette, 2021), and minimax off-policy evaluation rates (Duan et al., 2020; Kallus and Uehara, 2020). Our Theorem E.1 isolates non-invertibility of the credit-assignment map as an obstruction that persists at every sample size. The Assouad construction of Section E.1 follows the standard hypercube recipe (Yu, 1997; Wainwright, 2019). Missing-exposure and random-design hardness are also established statistical mechanisms: closely related missing-mass behavior appears in offline contextual-bandit lower bounds (Rashidinejad et al., 2021), while Pathak et al. (2024) characterize minimax recovery under random linear operators. We use these mechanisms as building blocks; the contribution is the feasible-MDP construction that combines them to close the stated general-even-H, $\mathcal{C}_{\mathrm{out}}$ general-logging rate.

Preference-based learning and RLHF theory. Learning reward models from trajectory comparisons (Christiano et al., 2017) identifies only *differences* $\langle \phi(\tau) - \phi(\tau'), r \rangle$; cardinal outcomes identify the richer subspace $\mathcal{U}_\mu$ itself, including the component along $\mathbf{1}$ that preferences can never pin down (Remark F.2). On the theory side, Zhu et al. (2023) show for linear rewards under Bradley–Terry–Luce and Plackett–Luce models that plain MLE fails for policy learning while pessimistic MLE succeeds under a coverage condition — structurally parallel to our Theorem D.4, but with difference features and a logistic link. Reward-confidence-set construction followed by robust policy optimization is likewise established in offline preference-based RL (Zhan et al., 2024); the projected confidence-set result in Section D.1 instantiates this template with an outcome-design geometry and a burn-in-free $\mathcal{C}_{\mathrm{out}}$ guarantee under cardinal trajectory outcomes. Preference-based RL with finite-time guarantees (Xu et al., 2020; Chen et al., 2022; Saha et al., 2023) develops the online counterpart. Our Proposition F.1 treats binary cardinal outcomes with a known monotone link, where the identification theory transfers verbatim.

Inverse RL and identifiability of rewards. Inverse RL (Ng and Russell, 2000; Abbeel and Ng, 2004; Ziebart et al., 2008) assumes the behavior is (near-)optimal and asks which rewards rationalize it; we make no optimality assumption and instead observe outcomes. That literature has developed a precise account of when the reward is recoverable at all (Cao et al., 2021; Kim et al., 2021; Amin et al., 2017), and Skalse et al. (2023) characterize the *partially identifiable* reward set induced by demonstrations and by comparisons. Our Theorem C.1 is the analogous characterization for cardinal outcome data, but it differs in two ways: the ambiguity is driven by the behavior's trajectory support rather than by optimality invariances, and we carry the characterization through to finite-sample rates and matching lower bounds rather than stopping at the population level. Imitation learning uses no outcome signal at all; our Corollary C.8 shows that with a deterministic behavior policy in a deterministic environment, outcome-only data certifies nothing beyond imitation — a sharp boundary between the two problems.

Partial identification. The interval of Proposition C.2 is a sharp identified set in the sense of the partial-identification literature in econometrics (Manski, 2003). The same logic underlies confounding-robust policy learning, where one optimizes a worst case over an uncertainty set that does not shrink with sample size (Kallus and Zhou, 2018; 2021; Namkoong et al., 2020; Kausik et al., 2024). Our Proposition C.2 instantiates this in the outcome-only MDP and, unlike the sensitivity-analysis formulations, the uncertainty set is not a modeling choice: it is the exact identified set implied by $\mathcal{U}_\mu$, and the maximin decision over it coincides with the pessimistic algorithm rather than being imposed alongside it.

Technical tools. Our confidence sets are built from scratch (Lemma D.2) rather than invoking self-normalized martingale bounds (Abbasi-Yadkori et al., 2011); standard references for the concentration and minimax machinery are Wainwright (2019); Vershynin (2018); Lattimore and Szepesvári (2020). Proposition F.1 connects to generalized linear bandits (Filippi et al., 2010), and the linear-MDP structure we mention as a future direction is that of Jin et al. (2020).

B Formal Model and Coverage Definitions

Definition B.1 (Outcome-only offline RL instance)**.** *A finite-horizon MDP is a tuple* $M = (\mathcal{S}, \mathcal{A}, H, P, r, \rho)$ *with finite state space* $\mathcal{S}$ *(*$S := |\mathcal{S}|$*), finite action space* $\mathcal{A}$ *(*$A := |\mathcal{A}|$*), horizon*

$H \in \mathbb{N}$, time-inhomogeneous transition kernels $P = (P_h)_{h=1}^{H-1}$ with $P_h(\cdot \mid s,a) \in \Delta(\mathcal{S})$, initial distribution $\rho \in \Delta(\mathcal{S})$, and reward vector $r \in [0,1]^d$, $d := HSA$, indexed by triples (h,s,a). A trajectory is $\tau = (s_1, a_1, \ldots, s_H, a_H)$; its feature vector $\phi(\tau) \in \{0,1\}^d$ is the indicator of the visited triples $\{(h, s_h, a_h)\}_{h=1}^{H}$, so $\|\phi(\tau)\|_2^2 = H$ and $\langle \phi(\tau), \mathbf{1} \rangle = H$. The return is $G(\tau) := \langle \phi(\tau), r \rangle$.

Because rewards are indexed by (h,s,a), each trajectory visits each index at most once and ϕ is binary; the results extend verbatim to time-homogeneous rewards with integer-valued count features, at the cost of replacing H by $\|\phi\|_2^2 \le H^2$ in the constants.

Assumption 1 (Outcome model)**.** Each episode reveals the trajectory τ and a noisy outcome $Y = \langle \phi(\tau), r \rangle + \xi$, where, conditionally on τ, the noise ξ is mean-zero and σ-subgaussian: $\mathbb{E}[e^{u\xi} \mid \tau] \le e^{u^2\sigma^2/2}$ for all $u \in \mathbb{R}$. Per-step rewards are never observed.

Assumption 2 (Data)**.** The dataset $\mathcal{D} = \{(\tau_i, Y_i)\}_{i=1}^{n}$ consists of n independent episodes generated by the same environment (P, ρ) and a fixed stochastic logging rule μ that may depend on the complete within-episode history. Thus, for $h_h = (s_1, a_1, \ldots, s_h)$,

$$q_\mu(\tau) = \rho(s_1) \prod_{h=1}^{H} \mu_h(a_h \mid h_h) \prod_{h=1}^{H-1} P_h(s_{h+1} \mid s_h, a_h).$$

We use q_μ for the induced trajectory distribution. Markov logging is the special case $\mu_h(a \mid h_h) = \mu_h(a \mid s_h)$; arbitrary history-dependent correlation across actions is otherwise allowed. The logger may change action probabilities but never the environment transition law.

Assumption 2 is where this paper departs from Markov-logging analyses: the logger may correlate actions across time within an episode, so two loggers with identical one-step occupancies can induce different trajectory laws.

Assumption 3 (Known dynamics)**.** The transition kernels P and initial distribution ρ are known.

Assumption 3 matches the motivating applications — game rules, logistics network structure, and compiler semantics are known; the scoring function is not — and isolates the statistically novel component of the problem. Appendix M removes it: the additional transition error is governed by *standard* concentrability, with an explicit interaction with the reward-confidence radius (Corollary M.5).

For a (possibly stochastic, time-dependent Markov) policy π, the occupancy measure $d^\pi \in [0,1]^d$ is $d^\pi(h,s,a) := \mathbb{P}^\pi(s_h = s,\, a_h = a)$, so that $d^\pi = \mathbb{E}^\pi[\phi(\tau)]$, $\langle d^\pi, \mathbf{1} \rangle = H$, and the value is the linear functional

$$V^\pi \;=\; \mathbb{E}^\pi[G(\tau)] \;=\; \langle d^\pi, r \rangle. \tag{6}$$

We write $\pi^\star \in \arg\max_\pi V^\pi$ and $\mathrm{SubOpt}(\pi) := V^{\pi^\star} - V^\pi$. Given a candidate policy class Π (all time-dependent Markov policies unless stated otherwise), the goal is to select $\hat{\pi} \in \Pi$ from $\mathcal{D}$ with small $V_\Pi^\star - V^{\hat{\pi}}$ where $V_\Pi^\star := \max_{\pi \in \Pi} V^\pi$. This restriction loses no optimal value for the additive finite-horizon reward model; sections that study non-additive trajectory objectives state explicitly when history-dependent target policies are allowed.

Key objects. Let

$$\mathcal{U}_\mu \;:=\; \mathrm{span}\{\phi(\tau) : q_\mu(\tau) > 0\} \subseteq \mathbb{R}^d, \qquad \Sigma_\mu \;:=\; \mathbb{E}_{\tau \sim q_\mu}\big[\phi(\tau)\phi(\tau)^\top\big]. \tag{7}$$

Throughout, $\mathcal{U}_\mu$ describes which reward combinations the log can reveal, and Σ_μ describes how strongly they are represented: $v^\top \Sigma_\mu v = \mathbb{E}_\mu[\langle \phi(\tau), v \rangle^2]$ is the second moment of the logged measurements in direction v.

Lemma B.2. $\mathrm{range}(\Sigma_\mu) = \mathcal{U}_\mu$*, and the smallest nonzero eigenvalue* $\omega_{\min} := \lambda_{\min}^{+}(\Sigma_\mu)$ *is positive whenever* $\mathcal{U}_\mu \neq \{0\}$*. Moreover* $\phi(\tau) \in \mathcal{U}_\mu$ *for* q_μ*-almost every* τ*.*

Proof. $\Sigma_\mu = \sum_\tau q_\mu(\tau)\, \phi(\tau)\phi(\tau)^\top$ is a finite nonnegative combination of rank-one terms with ranges in $\mathcal{U}_\mu$, so $\mathrm{range}(\Sigma_\mu) \subseteq \mathcal{U}_\mu$. Conversely, if $v \in \mathcal{U}_\mu$ satisfies $v^\top \Sigma_\mu v = 0$, then $\mathbb{E}_\mu \langle \phi, v \rangle^2 = 0$, hence $\langle \phi(\tau), v \rangle = 0$ for all supported τ, i.e., $v \perp \mathcal{U}_\mu$; combined with $v \in \mathcal{U}_\mu$ this forces $v = 0$. Thus Σ_μ is positive definite on $\mathcal{U}_\mu$, which gives both the reverse range inclusion and $\omega_{\min} > 0$. The final claim holds because every supported $\phi(\tau)$ is one of the spanning vectors. □

For $v \in \mathcal{U}_\mu$ we write $\|v\|^2_{\Sigma_\mu^\dagger} := v^\top \Sigma_\mu^\dagger v$ with $\Sigma_\mu^\dagger$ the Moore–Penrose pseudoinverse.

Definition B.3 (Outcome concentrability)**.** *For a target policy π,*

$$\mathcal{C}_{\mathrm{out}}(\pi;\mu) \;:=\; \begin{cases} \|d^\pi\|^2_{\Sigma_\mu^\dagger} & \text{if } d^\pi \in \mathcal{U}_\mu, \\ +\infty & \text{otherwise.} \end{cases}$$

For comparison, the standard (per-step) concentrability is $C_{\mathrm{std}}(\pi;\mu) := \max_{(h,s,a):\, d^\pi(h,s,a)>0} d^\pi(h,s,a)/d^\mu(h,s,a)$, where $d^\mu(h,s,a) := \mathbb{P}_{q_\mu}(s_h = s, a_h = a)$ (with the convention $C_{\mathrm{std}} = \infty$ if some visited triple has $d^\mu = 0$).

The coefficient above has an exact variational interpretation. In the Markov-logging regime this variational form coincides with the additive change-of-trajectory-measure quantity used by Yuan et al. (2025); Lemma B.4 states it for every valid logger.

Lemma B.4 (Optimal additive change-of-measure constant)**.** *For every target occupancy d^π,*

$$\mathcal{C}_{\mathrm{out}}(\pi;\mu) = \sup_{v \in \mathbb{R}^d} \frac{\big(\mathbb{E}_\pi[\langle \phi(\tau), v\rangle]\big)^2}{\mathbb{E}_\mu[(\langle \phi(\tau), v\rangle)^2]},$$

with the convention that the ratio is $+\infty$ when the denominator is zero and the numerator is nonzero. Equivalently, $\mathcal{C}_{\mathrm{out}}(\pi;\mu)$ is the smallest constant C such that

$$\big(\mathbb{E}_\pi F_v(\tau)\big)^2 \le C\, \mathbb{E}_\mu[F_v(\tau)^2] \qquad \text{for every additive functional } F_v(\tau) = \langle \phi(\tau), v\rangle.$$

Proof. If $d^\pi \notin \mathrm{range}(\Sigma_\mu)$, there is $v \in \ker(\Sigma_\mu)$ with $\langle d^\pi, v\rangle \neq 0$, making the displayed supremum infinite. Otherwise write $d^\pi = \Sigma_\mu^{1/2} u$ on $\mathrm{range}(\Sigma_\mu)$. Cauchy–Schwarz gives $(\langle d^\pi, v\rangle)^2 \le (d^{\pi\top}\Sigma_\mu^\dagger d^\pi)(v^\top \Sigma_\mu v)$, with equality at $v = \Sigma_\mu^\dagger d^\pi$. Since $\langle d^\pi, v\rangle = \mathbb{E}_\pi F_v$ and $v^\top \Sigma_\mu v = \mathbb{E}_\mu F_v^2$, the claim follows. □

C Policy-Value Identification and Logging Memory

This appendix proves the identification results, introduces the XOR constructions used to study identification and memory, and develops the comparison between a history-dependent logger and its Markovization.

Two instances (M, r) and (M, r') sharing dynamics and behavior are *observationally equivalent* if they induce the same distribution of (τ, Y) for some admissible noise laws. Under Assumption 1 the conditional mean $m(\tau) := \mathbb{E}[Y \mid \tau] = \langle \phi(\tau), r\rangle$ is identified for every supported τ (and nothing else is: the noise law is unrestricted beyond subgaussianity). A functional of r is *identified* if it takes a common value across all reward vectors in the model class $[0,1]^d$ consistent with $\{m(\tau)\}_{\tau \in \mathrm{supp}(q_\mu)}$.

Theorem C.1 (Value identifiability)**.** *Fix an instance satisfying Assumptions 1–3 and a target policy π.*

- *(a)* ***(Sufficiency.)*** *If $d^\pi \in \mathcal{U}_\mu$, then V^π is identified: writing $d^\pi = \sum_{j=1}^k c_j\, \phi(\tau_j)$ with supported τ_j, we have $V^\pi = \sum_{j=1}^k c_j\, m(\tau_j)$, a function of identified quantities only.*
- *(b)* ***(Necessity.)*** *If $d^\pi \notin \mathcal{U}_\mu$ and $r \in (0,1)^d$, then V^π is not identified: there exists $r' \in [0,1]^d$, observationally equivalent to r, with $\langle d^\pi, r'\rangle \neq \langle d^\pi, r\rangle$.*
- *(c)* ***(Trajectory coverage suffices.)*** *If $\mathrm{supp}(\mathbb{P}^\pi) \subseteq \mathrm{supp}(q_\mu)$ as distributions over trajectories, then $d^\pi \in \mathcal{U}_\mu$, hence V^π is identified. Per-step coverage ($C_{\mathrm{std}}(\pi;\mu) < \infty$) does* not *imply $d^\pi \in \mathcal{U}_\mu$ (Proposition C.4 with $\varepsilon = 0$).*

Proof. (a) By definition of $\mathcal{U}_\mu$ there exist supported trajectories $\tau_1, \dots, \tau_k$ and reals $c_1, \dots, c_k$ with $d^\pi = \sum_j c_j \phi(\tau_j)$. Linearity of (6) gives $V^\pi = \langle d^\pi, r\rangle = \sum_j c_j \langle \phi(\tau_j), r\rangle = \sum_j c_j m(\tau_j)$. Every reward vector consistent with the identified means $\{m(\tau)\}$ produces the same right-hand side, so V^π is constant over the consistent set.

(b) Let $\Pi_{\mathcal{U}^\perp}$ denote orthogonal projection onto $\mathcal{U}_\mu^\perp$ and set $z := \Pi_{\mathcal{U}^\perp} d^\pi \neq 0$ (nonzero precisely because $d^\pi \notin \mathcal{U}_\mu$). Then $z \perp \mathcal{U}_\mu$ and $\langle d^\pi, z\rangle = \|z\|_2^2 > 0$. Since $r \in (0,1)^d$, there is $\epsilon > 0$ with $r + \epsilon z \in [0,1]^d$ (take $\epsilon < \min_i \min(r_i, 1-r_i)/\|z\|_\infty$, noting $z \neq 0$ so $\|z\|_\infty > 0$). Put $r' := r + \epsilon z$. For every supported τ, $\phi(\tau) \in \mathcal{U}_\mu$ (Lemma B.2), hence $\langle \phi(\tau), r'\rangle = \langle \phi(\tau), r\rangle = m(\tau)$: with the same noise law, (M, r') induces exactly the same law of (τ, Y), so the two instances are observationally equivalent. Yet $\langle d^\pi, r'\rangle - \langle d^\pi, r\rangle = \epsilon\|z\|_2^2 \neq 0$.

(c) $d^\pi = \mathbb{E}^\pi[\phi(\tau)] = \sum_{\tau \in \mathrm{supp}(\mathbb{P}^\pi)} \mathbb{P}^\pi(\tau)\, \phi(\tau)$ is a convex combination of vectors $\phi(\tau)$ with $\tau \in \mathrm{supp}(q_\mu)$, each of which lies in $\mathcal{U}_\mu$; a subspace is closed under linear combinations. The non-implication is witnessed by the construction of Proposition C.4 at $\varepsilon = 0$, where every triple (h, a) has $d^\mu = 1/2$ yet $d^{\pi^\star} \notin \mathcal{U}_\mu$. □

Remark C.1 (Boundary rewards — a degenerate case). The interiority hypothesis in (b) cannot be dropped entirely: the perturbation $r \pm \epsilon z$ must stay in $[0,1]^d$. If r lies on the boundary, replace z by $-z$ if necessary; a genuine obstruction occurs only if *both* directions exit the box, i.e., there are coordinates i, j in the support of z with (say) $r_i = 1, z_i > 0$ blocking $+z$ and $r_j = 0, z_j > 0$ blocking $-z$. The opposite sign configurations give the symmetric obstructions. In that measure-zero configuration the *specific* instance can be pinned down by the box constraints even though $d^\pi \notin \mathcal{U}_\mu$; the worst case over the model class (which is what lower bounds require) is unaffected, since one may always choose the interior representative $r = \frac{1}{2}\mathbf{1} + \epsilon' z$ as in Section E.

Remark C.2 (What identifiability does not require). Nothing in Theorem C.1 requires identifying the reward vector r itself. Reward identifiability is the special case $\mathcal{U}_\mu = \mathbb{R}^d$, which needs at least d linearly independent supported trajectories and typically fails; policy evaluation only needs the single direction d^π to be spanned.

When V^π is not identified it is still *partially* identified.

Proposition C.2 (Sharp identified interval; pessimism is maximin)**.** *Let $\mathcal{R}_0 := \{r' \in [0,1]^d : \langle \phi(\tau), r'\rangle = m(\tau)\ \forall \tau \in \mathrm{supp}(q_\mu)\}$ be the set of reward vectors consistent with the identified conditional means. Then:*

(a) $\mathcal{R}_0$ is a nonempty compact convex polytope, and the sharp identified set of V^π is the interval $[\underline{V}(\pi), \overline{V}(\pi)]$ with

$$\underline{V}(\pi) = \min_{r' \in \mathcal{R}_0} \langle d^\pi, r'\rangle, \qquad \overline{V}(\pi) = \max_{r' \in \mathcal{R}_0} \langle d^\pi, r'\rangle,$$

each a linear program. V^π is identified iff $\underline{V}(\pi) = \overline{V}(\pi)$.

(b) The maximin policy $\arg\max_{\pi \in \Pi} \min_{r' \in \mathcal{R}_0} \langle d^\pi, r'\rangle$ coincides with the policy maximizing the pessimistic value $\underline{V}(\pi)$: worst-case decision-making over the identified set is *pessimism with respect to the sharp set.*

Proof. (a) $\mathcal{R}_0$ is cut out from the compact convex box $[0,1]^d$ by finitely many affine equalities (the trajectory space is finite), hence is a compact convex polytope; it contains the true r, hence is nonempty. The map $r' \mapsto \langle d^\pi, r'\rangle$ is linear and continuous, so its image over $\mathcal{R}_0$ is a compact interval, whose endpoints are by definition the two stated linear programs. Every value in the interval is attained by some consistent instance (convexity), and no value outside it is, so the interval is the sharp identified set; the identification criterion is immediate. (b) is the definition of $\underline{V}$ read backwards: $\min_{r' \in \mathcal{R}_0} \langle d^\pi, r'\rangle = \underline{V}(\pi)$ for each π, so the two outer maximizations coincide. □

Proposition C.3 (Asymptotic bias of the least-squares plug-in)**.** *Let $P_{\mathcal{U}}$ be the orthogonal projector onto $\mathcal{U}_\mu$. The set of solutions of the population normal equations $\Sigma_\mu r' = \mathbb{E}_\mu[\phi Y]\ (= \Sigma_\mu r)$ is the affine set $r + \mathcal{U}_\mu^\perp$, whose minimum-$\ell_2$-norm element is $P_{\mathcal{U}} r$. Consequently the population least-squares plug-in value of a policy π equals $\langle d^\pi, P_{\mathcal{U}} r\rangle = V^\pi - \langle \Pi_{\mathcal{U}^\perp} d^\pi, \Pi_{\mathcal{U}^\perp} r\rangle$: its asymptotic bias is the inner product of the projection defects of d^π and r onto $\mathcal{U}_\mu^\perp$, and vanishes for every r iff $d^\pi \in \mathcal{U}_\mu$.*

Proof. $\Sigma_\mu r' = \Sigma_\mu r$ iff $r' - r \in \ker \Sigma_\mu = \mathrm{range}(\Sigma_\mu)^\perp = \mathcal{U}_\mu^\perp$ (Lemma B.2; Σ_μ is symmetric). The minimum-norm element of the affine set $r + \mathcal{U}_\mu^\perp$ is the orthogonal projection of the origin onto it, namely $r - \Pi_{\mathcal{U}^\perp} r = P_{\mathcal{U}} r$. Then $\langle d^\pi, P_{\mathcal{U}} r\rangle = \langle d^\pi, r\rangle - \langle d^\pi, \Pi_{\mathcal{U}^\perp} r\rangle = V^\pi - \langle \Pi_{\mathcal{U}^\perp} d^\pi, \Pi_{\mathcal{U}^\perp} r\rangle$, using

that $\Pi_{\mathcal{U}^\perp}$ is an orthogonal projector ($\langle x, \Pi y\rangle = \langle \Pi x, \Pi y\rangle$). The bias $\langle \Pi_{\mathcal{U}^\perp} d^\pi, \Pi_{\mathcal{U}^\perp} r\rangle$ vanishes for all $r \in [0,1]^d$ iff $\Pi_{\mathcal{U}^\perp} d^\pi = 0$ (choose $r = \frac{1}{2}\mathbf{1} + \epsilon \Pi_{\mathcal{U}^\perp} d^\pi$ otherwise, admissible for small $\epsilon > 0$). □

C.1 The ε-XOR family and the separation from per-step coverage

The following two-step instance is the running example of the paper; it turns the credit-assignment obstruction into a one-parameter family with closed-form constants.

The ε-XOR instance. $H = 2, S = 1, A = 2$ with actions $\{a, b\}$; the single state makes transitions trivial. Coordinates of $\mathbb{R}^4$ are ordered $(1a, 1b, 2a, 2b)$. Fix a gap $\Delta \in (0, 2]$ and set

$$z := e_{1a} - e_{1b} - e_{2a} + e_{2b}, \qquad r := \tfrac{1}{2}\mathbf{1} + \tfrac{\Delta}{4} z \in [0,1]^4.$$

The behavior draws the trajectory templates (a, a) and (b, b) with probability $(1-\varepsilon)/2$ each and the crossing templates (a, b) and (b, a) with probability $\varepsilon/2$ each, $\varepsilon \in [0, 1]$. (This joint law is induced by the independent uniform Markov policy only at $\varepsilon = 1/2$; for $\varepsilon \neq 1/2$ the two actions are correlated and require a history-dependent logger, which Assumption 2 permits.) The four deterministic policies are $(a, a), (b, b), (a, b), (b, a)$ with returns $1, 1, 1 + \Delta/2, 1 - \Delta/2$; thus $\pi^\star = (a, b)$ and every other deterministic policy is at least $\Delta/2$-suboptimal.

Proposition C.4 (Closed forms and separation)**.** *In the ε-XOR instance, define the orthonormal vectors $u_1 = \frac{1}{2}(1,1,1,1)$, $u_2 = \frac{1}{2}(1,-1,1,-1)$, $u_3 = \frac{1}{2}z$, $u_4 = \frac{1}{2}(1,1,-1,-1)$. Then:*

(a) $\phi(a,a) = u_1 + u_2$, $\phi(b,b) = u_1 - u_2$, $\phi(a,b) = u_1 + u_3$, $\phi(b,a) = u_1 - u_3$, and

$$\Sigma_\mu = u_1 u_1^\top + (1-\varepsilon)\, u_2 u_2^\top + \varepsilon\, u_3 u_3^\top.$$

For $0 < \varepsilon < 1$ its range is $\mathrm{span}\{u_1, u_2, u_3\}$; at $\varepsilon = 0$ it is $\mathrm{span}\{u_1, u_2\}$ and at $\varepsilon = 1$ it is $\mathrm{span}\{u_1, u_3\}$.

(b) For every $\varepsilon \in (0, 1]$, $\mathcal{C}_{\mathrm{out}}(\pi^\star; \mu) = 1 + 1/\varepsilon$ exactly. For $\varepsilon = 0$, $d^{\pi^\star} \notin \mathcal{U}_\mu$ and $\mathcal{C}_{\mathrm{out}}(\pi^\star; \mu) = \infty$.

(c) For every $\varepsilon \in [0, 1]$ and every policy π, $C_{\mathrm{std}}(\pi; \mu) \leq 2$, with equality for every deterministic policy (in particular for $\pi^\star$). Hence $\mathcal{C}_{\mathrm{out}}(\pi^\star)/C_{\mathrm{std}}(\pi^\star) = (1 + 1/\varepsilon)/2$ is unbounded as $\varepsilon \downarrow 0$ while the deterministic target has constant per-step concentrability.

Proof. (a) The expansions of the four ϕ's in the orthonormal basis $u_1, \ldots, u_4$ are verified coordinatewise; e.g., $\langle \phi(a,b), u_1\rangle = \frac{1}{2}(1+0+0+1) = 1$, $\langle \phi(a,b), u_2\rangle = \frac{1}{2}(1-0+0-1) = 0$, $\langle \phi(a,b), u_3\rangle = \frac{1}{2}(1-0-0+1) = 1$, $\langle \phi(a,b), u_4\rangle = 0$, giving $\phi(a,b) = u_1 + u_3$; the other three are analogous. Then, writing outer products in this basis,

$$\begin{aligned}
\Sigma_\mu &= \tfrac{1-\varepsilon}{2}\Big[(u_1+u_2)(u_1+u_2)^\top + (u_1-u_2)(u_1-u_2)^\top\Big] \\
&\quad + \tfrac{\varepsilon}{2}\Big[(u_1+u_3)(u_1+u_3)^\top + (u_1-u_3)(u_1-u_3)^\top\Big] \\
&= (1-\varepsilon)\big(u_1u_1^\top + u_2u_2^\top\big) + \varepsilon\big(u_1u_1^\top + u_3u_3^\top\big) \\
&= u_1u_1^\top + (1-\varepsilon)u_2u_2^\top + \varepsilon u_3u_3^\top.
\end{aligned}$$

because the cross terms $\pm(u_i u_j^\top + u_j u_i^\top)$ cancel between the paired trajectories. The eigenstructure and $\mathcal{U}_\mu$ follow since the u_i are orthonormal. (b) $d^{\pi^\star} = \phi(a,b) = u_1 + u_3$ (the policy is deterministic and transitions are trivial). For $0 < \varepsilon < 1$, $\Sigma_\mu^\dagger = u_1u_1^\top + (1-\varepsilon)^{-1}u_2u_2^\top + \varepsilon^{-1}u_3u_3^\top$. At $\varepsilon = 1$ the u_2 term is absent from the pseudoinverse because that eigenvalue is zero. In either case $\mathcal{C}_{\mathrm{out}} = \|u_1 + u_3\|^2_{\Sigma_\mu^\dagger} = 1 + 1/\varepsilon$. For $\varepsilon = 0$, $u_3 \notin \mathcal{U}_\mu$ so $d^{\pi^\star} \notin \mathcal{U}_\mu$. (c) At each step $h \in \{1, 2\}$ the behavior marginal of the action is uniform: e.g., at $h = 1$, $\mathbb{P}_\mu(a) = \frac{1-\varepsilon}{2} + \frac{\varepsilon}{2} = \frac{1}{2}$. Every policy's occupancy satisfies $d^\pi(h, \cdot, a') \leq 1$, so all ratios are at most $1/(1/2) = 2$, with equality for deterministic policies. □

C.2 MARKOVIZATION AND THE PRICE OF LOGGER MEMORY

The XOR family keeps all one-step marginals fixed while changing only their cross-time coupling. Replacing a history-dependent policy by an occupancy-equivalent Markov policy is a classical construction (Laroche and Tachet des Combes, 2023); our use of it is to expose the second-moment information that occupancy alone discards. Given any valid history-dependent logger μ, define its *Markovization* $\bar{\mu}$ by

$$\bar{\mu}_h(a \mid s) := \frac{d^\mu(h, s, a)}{\sum_{a'} d^\mu(h, s, a')}$$

whenever the denominator is positive, and arbitrarily otherwise. Define the *additive memory distortion*

$$\kappa_{\mathrm{mem}}(\mu) := \inf \left\{ \kappa \geq 0 : \ \Sigma_{\bar{\mu}} \preceq \kappa \, \Sigma_\mu \right\} \in [1, \infty], \tag{8}$$

where the infimum is $+\infty$ if no finite Loewner domination exists. Equivalently,

$$\kappa_{\mathrm{mem}}(\mu) = \sup_{v:\, v^\top \Sigma_\mu v > 0} \frac{v^\top \Sigma_{\bar{\mu}} v}{v^\top \Sigma_\mu v},$$

with value $+\infty$ if some $v \in \ker \Sigma_\mu$ has $v^\top \Sigma_{\bar{\mu}} v > 0$. The lower endpoint 1 follows whenever the factor is finite because both covariance matrices have trace H. Thus κ_{mem} is the smallest factor comparing second moments of *additive* trajectory functionals under the original logger and the Markov policy having exactly the same one-step occupancies. It is generally more targeted than a full trajectory likelihood ratio: if $q_{\bar{\mu}} \ll q_\mu$, then

$$\kappa_{\mathrm{mem}}(\mu) \leq \left\| \frac{dq_{\bar{\mu}}}{dq_\mu} \right\|_\infty . \tag{9}$$

Lemma C.5 (Markov change of trajectory measure)**.** *Let ν and π be Markov policies in the same finite-horizon MDP, with the same initial distribution and transition kernels. For arbitrary time-indexed measurable functions $f_h : \mathcal{S}_h \times \mathcal{A} \to \mathbb{R}$, define*

$$F(\tau) = \sum_{h=1}^{H} f_h(s_h, a_h).$$

Then

$$\big(\mathbb{E}_\pi F(\tau)\big)^2 \leq \left[1 + \sum_{h=1}^{H} \chi^2(d_h^\pi \| d_h^\nu) \right] \mathbb{E}_\nu[F(\tau)^2],$$

with the usual $+\infty$ convention when absolute continuity fails. This is the time-indexed additive specialization of the change-of-trajectory-measure principle of Jia et al. (2025), using the sharpened marginal-χ^2 form of Yuan et al. (2025).

Proof. Set $\bar{f}_{H+1} \equiv 0$ and, for $h \leq H$, define

$$\bar{f}_h(s) = \mathbb{E}_\nu \left[\sum_{t=h}^{H} f_t(s_t, a_t) \,\middle|\, s_h = s \right].$$

Let

$$\Delta_h = f_h(s_h, a_h) + \bar{f}_{h+1}(s_{h+1}) - \bar{f}_h(s_h).$$

The Markov property gives $\mathbb{E}_\nu[\Delta_h \mid s_h] = 0$. The resulting martingale-difference decomposition satisfies

$$F(\tau) = \bar{f}_1(s_1) + \sum_{h=1}^{H} \Delta_h$$

and

$$\mathbb{E}_\nu[F(\tau)^2] = \mathbb{E}_\nu[\bar{f}_1(s_1)^2] + \sum_{h=1}^{H} \mathbb{E}_\nu[\Delta_h^2].$$

For each h, the joint laws of (s_h, a_h, s_{h+1}) under π and ν share the same transition kernel. Their likelihood ratio therefore depends only on (s_h, a_h), and their χ^2 divergence equals $\chi^2(d_h^\pi \| d_h^\nu)$. Since $\mathbb{E}_\nu[\Delta_h] = 0$, the elementary χ^2 change-of-measure inequality gives

$$\big(\mathbb{E}_\pi \Delta_h\big)^2 \le \chi^2(d_h^\pi \| d_h^\nu)\, \mathbb{E}_\nu[\Delta_h^2].$$

The initial distribution is common, so

$$\big(\mathbb{E}_\pi \bar{f}_1(s_1)\big)^2 = \big(\mathbb{E}_\nu \bar{f}_1(s_1)\big)^2 \le \mathbb{E}_\nu[\bar{f}_1(s_1)^2].$$

Combining the decomposition above with Cauchy–Schwarz yields

$$\big(\mathbb{E}_\pi F\big)^2 \le \left[1 + \sum_{h=1}^{H} \chi^2(d_h^\pi \| d_h^\nu)\right] \left[\mathbb{E}_\nu[\bar{f}_1(s_1)^2] + \sum_{h=1}^{H} \mathbb{E}_\nu[\Delta_h^2]\right],$$

and the second bracket equals $\mathbb{E}_\nu[F^2]$. □

Theorem C.6 (Memory-sensitive reduction to marginal coverage)**.** *For every valid history-dependent logger μ and every target policy π from the Markov policy class of Section B:*

(a) $\bar{\mu}$ has exactly the same state–action occupancies as μ: $d^{\bar{\mu}} = d^\mu$. Consequently every per-step coverage quantity that depends only on these occupancies, including C_{std}, is unchanged by Markovization. Moreover the within-time diagonal blocks of $\Sigma_{\bar{\mu}}$ and Σ_μ are both $\mathrm{diag}(d_h^\mu)$; all differences between the two covariance matrices therefore live in cross-time blocks.

(b) The outcome coefficient obeys

$$\begin{aligned} \mathcal{C}_{\mathrm{out}}(\pi;\mu) &\le \kappa_{\mathrm{mem}}(\mu)\, \mathcal{C}_{\mathrm{out}}(\pi;\bar{\mu}) \\ &\le \kappa_{\mathrm{mem}}(\mu) \left[1 + \sum_{h=1}^{H} \chi^2(d_h^\pi \| d_h^\mu)\right] \\ &\le \kappa_{\mathrm{mem}}(\mu)\big[1 + H(C_{\mathrm{std}}(\pi;\mu) - 1)\big]. \end{aligned} \tag{10}$$

Consequently finite marginal coverage implies span identification whenever $\kappa_{\mathrm{mem}}(\mu) < \infty$.

(c) If $q_{\bar{\mu}} \ll q_\mu$, then the memory factor also satisfies the trajectory-level upper bound (9)*. Hence the generalized-eigenvalue factor κ_{mem} never pays more than the worst-case likelihood-ratio cost of replacing the original trajectory law by its Markovization, while ignoring trajectory discrepancies invisible to additive squared functionals.*

(d) If μ is Markov then $\bar{\mu} = \mu$ on all reachable states, $\kappa_{\mathrm{mem}}(\mu) = 1$, and (10) *reduces to the improved change-of-trajectory-measure bound of Yuan et al. (2025). For the ε-XOR logger,*

$$\kappa_{\mathrm{mem}}(\mu_\varepsilon) = \max\left\{1, \frac{1}{2\varepsilon}, \frac{1}{2(1-\varepsilon)}\right\}, \qquad 0 < \varepsilon < 1,$$

and $\kappa_{\mathrm{mem}} = \infty$ at $\varepsilon \in \{0, 1\}$. For $0 < \varepsilon < 1$ this value also equals $\|dq_{\bar{\mu}}/dq_{\mu_\varepsilon}\|_\infty$. In particular, as $\varepsilon \downarrow 0$, $C_{\mathrm{std}}(\pi^\star) = 2$ while $\mathcal{C}_{\mathrm{out}}(\pi^\star) = 1 + 1/\varepsilon = \Theta(\kappa_{\mathrm{mem}})$, so the memory factor in a uniform reduction from marginal coverage cannot be removed even at horizon two.

Proof roadmap. Part (a) is an induction on time showing that Markovization reproduces every one-step occupancy, so the two covariance matrices differ only in cross-time blocks. Part (b) transfers the variational ratio of Lemma B.4 from μ to $\bar{\mu}$ at cost κ_{mem} via the Loewner comparison, then applies the Markov-policy change-of-measure inequality to $\bar{\mu}$, then bounds χ^2 by C_{std}. Part (c) compares second moments under a bounded likelihood ratio. Part (d) uses the simultaneous diagonalization of the XOR covariances.

Proof. For (a), let $d_h^\mu(s) = \sum_a d^\mu(h, s, a)$. At $h = 1$ the state law is ρ under both loggers. If the state law agrees at step h, then by the definition of $\bar{\mu}$ its state–action law is $d_h^\mu(s)\bar{\mu}_h(a \mid s) =$

$d^\mu(h,s,a)$. Applying the common transition kernel P_h therefore gives the same next-state law. Induction proves $d^{\bar\mu} = d^\mu$. For the time-indexed one-hot feature map, the (h,h) covariance block has no off-diagonal entries and diagonal d_h^μ, proving the final statement.

Suppose first that $\kappa_{\mathrm{mem}} < \infty$. By $\Sigma_{\bar\mu} \preceq \kappa_{\mathrm{mem}}\Sigma_\mu$, every v with $v^\top\Sigma_{\bar\mu}v > 0$ satisfies

$$\frac{(d^{\pi\top}v)^2}{v^\top\Sigma_\mu v} \le \kappa_{\mathrm{mem}}(\mu)\frac{(d^{\pi\top}v)^2}{v^\top\Sigma_{\bar\mu}v}.$$

If $v^\top\Sigma_{\bar\mu}v = 0$, then either the numerator is zero or the right-hand variational constant is already infinite. Lemma B.4 therefore gives the first inequality in (b); it is trivial when $\kappa_{\mathrm{mem}} = \infty$. For the second inequality in (b), apply Lemma C.5 with $\nu = \bar\mu$ and $f_h(s,a) = v(h,s,a)$. Both π and $\bar\mu$ are Markov policies in the same MDP, and part (a) gives $d_h^{\bar\mu} = d_h^\mu$. Taking the supremum over v and using Lemma B.4 therefore gives

$$\mathcal{C}_{\mathrm{out}}(\pi;\bar\mu) \le 1 + \sum_{h=1}^{H}\chi^2(d_h^\pi\|d_h^{\bar\mu}) = 1 + \sum_{h=1}^{H}\chi^2(d_h^\pi\|d_h^\mu).$$

Finally, when $C_{\mathrm{std}} < \infty$,

$$1 + \chi^2(d_h^\pi\|d_h^\mu) = \sum_{s,a}\frac{d_h^\pi(s,a)^2}{d_h^\mu(s,a)} \le C_{\mathrm{std}}(\pi;\mu)\sum_{s,a}d_h^\pi(s,a) = C_{\mathrm{std}}(\pi;\mu),$$

which proves the last inequality. A finite right-hand side makes $\mathcal{C}_{\mathrm{out}}$ finite and hence implies span identification by Definition B.3 and Theorem C.1.

For (c), write $w = dq_{\bar\mu}/dq_\mu$. Every additive functional $F_v(\tau) = \langle\phi(\tau), v\rangle$ obeys

$$v^\top\Sigma_{\bar\mu}v = \mathbb{E}_\mu[w(\tau)F_v(\tau)^2] \le \|w\|_\infty\, v^\top\Sigma_\mu v.$$

Taking the smallest admissible Loewner factor proves (9).

For (d), Markov μ equals its Markovization on every state of positive occupancy, so the two trajectory laws and covariance matrices coincide. For XOR, $\bar\mu$ is the independent uniform logger, i.e. the $\varepsilon = 1/2$ member of the family. Proposition C.4(a) gives simultaneously diagonal covariance matrices in the basis (u_1,u_2,u_3,u_4): $\Sigma_{\bar\mu}$ has eigenvalues $(1,1/2,1/2,0)$ while Σ_{μ_ε} has $(1,1-\varepsilon,\varepsilon,0)$. The smallest Loewner domination factor is therefore the maximum of the three nonzero eigenvalue ratios shown above; at either endpoint one direction present under $\bar\mu$ lies in the kernel of Σ_μ. For $0 < \varepsilon < 1$, $q_{\bar\mu}$ gives probability $1/4$ to each of the four trajectories while q_{μ_ε} gives $(1-\varepsilon)/2$ to each matched trajectory and $\varepsilon/2$ to each crossing trajectory, so its maximum likelihood ratio is the same displayed factor. The final comparison uses Proposition C.4(b)–(c). □

Proposition C.7 (Target-representer localization)**.** *Let μ be a valid history-dependent logger, let $\bar\mu$ be its occupancy-equivalent Markovization, and let π be a Markov target policy. Write $A = \Sigma_\mu$, $B = \Sigma_{\bar\mu}$, and $d = d^\pi$. For $d \in \mathrm{range}(A)$ define*

$$\kappa_{\mathrm{rep}}(\pi;\mu) := \inf_{v:\,Av=d}\frac{v^\top Bv}{v^\top Av},$$

and set $\kappa_{\mathrm{rep}}(\pi;\mu) = +\infty$ otherwise. If $\mathcal{C}_{\mathrm{out}}(\pi;\bar\mu) < \infty$, then

$$\frac{\mathcal{C}_{\mathrm{out}}(\pi;\mu)}{\mathcal{C}_{\mathrm{out}}(\pi;\bar\mu)} \le \kappa_{\mathrm{rep}}(\pi;\mu).$$

Moreover, whenever $\kappa_{\mathrm{mem}}(\mu) < \infty$, $\kappa_{\mathrm{rep}}(\pi;\mu) \le \kappa_{\mathrm{mem}}(\mu)$. Consequently, under finite marginal coverage,

$$\mathcal{C}_{\mathrm{out}}(\pi;\mu) \le \kappa_{\mathrm{rep}}(\pi;\mu)\left[1 + \sum_{h=1}^{H}\chi^2(d_h^\pi\|d_h^\mu)\right] \le \kappa_{\mathrm{rep}}(\pi;\mu)\big[1 + H(C_{\mathrm{std}}(\pi;\mu) - 1)\big].$$

Thus the global factor in Theorem C.6 can be replaced, for a fixed target, by a certificate that ignores additive directions not needed to represent that target.

Proof. Assume $d \in \mathrm{range}(A)$ and put $C_\mu = d^\top A^\dagger d = \mathcal{C}_{\mathrm{out}}(\pi;\mu)$. Every feasible v with $Av = d$ satisfies $v^\top Av = v^\top d = C_\mu$, because feasible representers differ only by $\ker(A) \perp \mathrm{range}(A)$. Let $C_{\bar\mu} = d^\top B^\dagger d = \mathcal{C}_{\mathrm{out}}(\pi;\bar\mu)$. Since $C_{\bar\mu} < \infty$, $d \in \mathrm{range}(B)$, and PSD Cauchy–Schwarz gives

$$C_\mu^2 = (d^\top v)^2 \le C_{\bar\mu}\, v^\top Bv.$$

Taking the infimum over $Av = d$ proves $C_\mu / C_{\bar\mu} \le \kappa_{\mathrm{rep}}(\pi;\mu)$. If $B \preceq \kappa_{\mathrm{mem}}(\mu)A$, then every feasible v satisfies $v^\top Bv \le \kappa_{\mathrm{mem}}(\mu) v^\top Av$, proving the second inequality. The final display follows by applying the same Markov-policy change-of-trajectory-measure inequality used in Theorem C.6 to π and $\bar\mu$. □

Remark C.3 (XOR localization and scope). For the crossing target in the ε-XOR family,

$$\kappa_{\mathrm{rep}}^{\mathrm{cross}}(\varepsilon) = \frac{\varepsilon^2 + 1/2}{\varepsilon(\varepsilon+1)}, \qquad 0 < \varepsilon \le 1.$$

Hence $\kappa_{\mathrm{rep}}^{\mathrm{cross}}(\varepsilon) \sim (2\varepsilon)^{-1}$ as $\varepsilon \downarrow 0$, equals one at $\varepsilon = 1/2$, and tends to $3/4$ as $\varepsilon \uparrow 1$, even though the global κ_{mem} diverges because the matched direction disappears. When both outcome coefficients are finite, their ratio $\mathcal{C}_{\mathrm{out}}(\pi;\mu)/\mathcal{C}_{\mathrm{out}}(\pi;\bar\mu)$ is the exact fixed-target comparison; κ_{rep} is a representer-level upper certificate for it, not the smallest possible scalar for a fixed pair of covariances.

Corollary C.8 (Deterministic behavior certifies nothing)**.** *Suppose the dynamics are deterministic and the behavior is a single deterministic trajectory τ_0 (so $q_\mu = \delta_{\tau_0}$). Then $\mathcal{U}_\mu = \mathrm{span}\{\phi(\tau_0)\}$ is one-dimensional, and for any policy π whose (deterministic) trajectory differs from τ_0 in at least one step, V^π is not identified for any interior r. The only policies with identified value are those reproducing τ_0 exactly: outcome-only data cannot certify any improvement over imitation of the behavior.*

Proof. With deterministic dynamics and policy, $d^\pi = \phi(\tau_\pi)$ for the induced trajectory τ_π. Two distinct binary vectors with exactly H ones and identical support pattern per step are never positive multiples of one another unless equal: if $\phi(\tau_\pi) = c\,\phi(\tau_0)$ then comparing any coordinate where $\phi(\tau_0)$ is 1 gives $c \in \{0,1\}$ and $c \ne 0$ (both have H ones), so $c = 1$ and $\tau_\pi = \tau_0$. Hence $\tau_\pi \ne \tau_0$ implies $d^\pi \notin \mathrm{span}\{\phi(\tau_0)\} = \mathcal{U}_\mu$, and Theorem C.1(b) applies. □

Finally we record a structural fact about Σ_μ that will be used to compare Theorem C.1 with the coverage conditions assumed elsewhere in the literature.

Proposition C.9 (Σ_μ is always rank deficient)**.** *Write $\mathbf{1}_h := \sum_{s,a} e_{(h,s,a)}$. For every tabular OO-RL instance with $H \ge 2$ and* every *behavior q_μ,*

$$\mathrm{span}\{\mathbf{1}_h - \mathbf{1}_{h'} : h, h' \in [H]\} \;\subseteq\; \ker \Sigma_\mu, \qquad \mathrm{rank}\, \Sigma_\mu \;\le\; d - (H-1),$$

so $\lambda_{\min}(\Sigma_\mu) = 0$: no behavior, however exploratory, makes Σ_μ positive definite. Consequently the reward vector is not uniformly *point identified over $[0,1]^d$: every interior reward admits a nonzero sufficiently small perturbation $c(\mathbf{1}_h - \mathbf{1}_{h'})$ with the same data law. Particular boundary rewards can be pinned by the box constraints (Remark C.1). Policy values are unaffected by these structural perturbations because $\langle d^\pi, \mathbf{1}_h - \mathbf{1}_{h'}\rangle = 1 - 1 = 0$.*

Proof. Each trajectory takes exactly one action at each step, so $\langle \phi(\tau), \mathbf{1}_h\rangle = 1$ for every τ and every h, whence $\langle \phi(\tau), \mathbf{1}_h - \mathbf{1}_{h'}\rangle = 0$ identically and $\Sigma_\mu(\mathbf{1}_h - \mathbf{1}_{h'}) = \mathbb{E}[\phi\langle\phi, \mathbf{1}_h - \mathbf{1}_{h'}\rangle] = 0$. The $H-1$ vectors $\mathbf{1}_h - \mathbf{1}_1$, $h = 2, \dots, H$, are linearly independent because the $\mathbf{1}_h$ have disjoint supports. The two consequences are immediate from $\langle \phi(\tau), \mathbf{1}_h - \mathbf{1}_{h'}\rangle = 0$ and $\sum_{s,a} d^\pi(h,s,a) = 1$ for every h. □

Remark C.4 (Comparison with well-explored-dataset assumptions). Offline analyses of outcome-only feedback have assumed $\lambda_{\min}\big(\mathbb{E}_\mu[\Phi(\tau)\Phi(\tau)^\top]\big) \ge c > 0$ (Xu et al., 2024). By Proposition C.9 this is vacuous in the tabular setting: it fails for every instance with $H \ge 2$. Its natural repair is to restrict the requirement to $\mathcal{U}_\mu$, i.e. $\omega_{\min} = \lambda_{\min}^+(\Sigma_\mu) \ge \omega$. Even then the resulting guarantee is *uniform over policies*: for every π with $d^\pi \in \mathcal{U}_\mu$,

$$\mathcal{C}_{\mathrm{out}}(\pi;\mu) = \|d^\pi\|_{\Sigma_\mu^\dagger}^2 \;\le\; \frac{\|d^\pi\|_2^2}{\omega} \;\le\; \frac{\|d^\pi\|_\infty \|d^\pi\|_1}{\omega} \;\le\; \frac{H}{\omega},$$

so Theorem D.4 recovers a bound $O(\sigma\sqrt{dH/(\omega n)})$ that does not depend on the target policy at all. The single-policy coefficient can be far smaller: on the sparse family of Section E.4, $\omega = \min\{\varepsilon, 1-\varepsilon\}$ and $H = 2K$, so $H/\omega = 2K/\varepsilon$ whereas $\mathcal{C}^\star_{\mathrm{out}} = 1 + s/\varepsilon$; the ratio grows like $2K/s = d/(2s)$, and the uniform bound is therefore looser by a factor of order $\sqrt{d/s}$ — precisely the dimension factor that Theorem E.12 shows to be unnecessary.

D Estimation and Burn-in-Free Policy Learning

This appendix develops two pessimistic learners. Algorithm 1 (ridge) gives high-probability guarantees after a design burn-in (Theorems D.3–D.4); Algorithm 2 (projected box, Section D.1) removes the global spectral threshold. Finite-class refinements and tractability results follow.

Definition D.1 (Ridge estimator and pessimistic rule)**.** *Given $\mathcal{D}$ and a regularizer $\lambda > 0$, let*

$$\Lambda := \lambda I_d + \sum_{i=1}^n \phi_i \phi_i^\top, \qquad \hat{r} := \Lambda^{-1} \sum_{i=1}^n \phi_i Y_i, \qquad \phi_i := \phi(\tau_i).$$

Given a confidence radius $\beta > 0$ and a candidate class Π, Algorithm 1 *(pessimistic outcome-only policy optimization) returns*

$$\hat{\pi} \in \arg\max_{\pi \in \Pi} \widehat{V}(\pi), \qquad \widehat{V}(\pi) := \langle d^\pi, \hat{r} \rangle - \beta \, \|d^\pi\|_{\Lambda^{-1}}.$$

Occupancies d^π are computable exactly under Assumption 3.

Remark D.1 (Computation)*.* d^π ranges over the occupancy polytope (a polytope in $\mathbb{R}^d$ defined by the flow constraints of the MDP), $\langle d^\pi, \hat{r} \rangle$ is linear, and $\|d^\pi\|_{\Lambda^{-1}}$ is convex in d^π; hence $\widehat{V}$ is concave over the polytope and Algorithm 1 over the class of *all* policies is a second-order cone program. Some experiments enumerate deterministic policies on small instances; others optimize over the full occupancy polytope (Appendix J).

Lemma D.2 (Confidence set; self-contained)**.** *Under Assumptions 1–2, for any $\delta \in (0,1)$ and $\lambda > 0$, with probability at least $1-\delta$,*

$$\|\hat{r} - r\|_\Lambda \;\le\; \beta(\delta) \;:=\; 2\sigma\sqrt{2\big(d\log 5 + \log(1/\delta)\big)} \;+\; \sqrt{\lambda d}\,.$$

Proof. Write $\xi_i := Y_i - \langle \phi_i, r \rangle$ and $X := \sum_i \phi_i \xi_i$. From the normal equations, $\hat{r} - r = \Lambda^{-1} X - \lambda \Lambda^{-1} r$, hence

$$\|\hat{r} - r\|_\Lambda = \|\Lambda^{-1/2}(X - \lambda r)\|_2 \le \|\Lambda^{-1/2} X\|_2 + \lambda \|\Lambda^{-1/2} r\|_2 \le \|\Lambda^{-1/2} X\|_2 + \sqrt{\lambda}\,\|r\|_2,$$

using $\Lambda \succeq \lambda I$ so $\|\Lambda^{-1/2} r\|_2 \le \lambda^{-1/2}\|r\|_2$, and $\|r\|_2 \le \sqrt{d}$ since $r \in [0,1]^d$. It remains to bound $\|\Lambda^{-1/2} X\|_2$ with probability $1-\delta$.

Condition on the design $(\phi_1, \dots, \phi_n)$ (the noise variables are mean-zero and σ-subgaussian conditionally on the trajectories, and independent across i by Assumption 2). Fix a unit vector $v \in \mathbb{R}^d$ and set $c_i := \langle \Lambda^{-1/2} v, \phi_i \rangle$, so that $\langle v, \Lambda^{-1/2} X \rangle = \sum_i c_i \xi_i$ and

$$\sum_i c_i^2 = v^\top \Lambda^{-1/2} \Big(\sum_i \phi_i \phi_i^\top \Big) \Lambda^{-1/2} v \le v^\top \Lambda^{-1/2} \Lambda \, \Lambda^{-1/2} v = 1.$$

By Lemma L.1 (subgaussian linear combinations), $\sum_i c_i \xi_i$ is σ-subgaussian, so by Lemma L.2, for every $t \ge 0$, $\mathbb{P}(\langle v, \Lambda^{-1/2} X \rangle > t) \le e^{-t^2/(2\sigma^2)}$.

Let N be a $\frac{1}{2}$-net of the unit sphere of $\mathbb{R}^d$ with $|N| \le 5^d$ (Lemma L.3). For the maximizing unit vector u with $\langle u, \Lambda^{-1/2} X \rangle = \|\Lambda^{-1/2} X\|_2$, pick $v \in N$ with $\|u - v\|_2 \le \frac{1}{2}$; then

$$\|\Lambda^{-1/2} X\|_2 = \langle v, \Lambda^{-1/2} X \rangle + \langle u - v, \Lambda^{-1/2} X \rangle \le \max_{v \in N} \langle v, \Lambda^{-1/2} X \rangle + \tfrac{1}{2} \|\Lambda^{-1/2} X\|_2,$$

so $\|\Lambda^{-1/2} X\|_2 \le 2 \max_{v \in N} \langle v, \Lambda^{-1/2} X \rangle$. A union bound over N with $t = \sigma\sqrt{2(d\log 5 + \log(1/\delta))}$ gives $\mathbb{P}\big(\max_{v \in N} \langle v, \Lambda^{-1/2} X \rangle > t \big) \le 5^d e^{-t^2/(2\sigma^2)} = \delta$, and the bound holds conditionally on every design, hence unconditionally. Combining the displays proves the lemma. □

Theorem D.3 (Data-dependent suboptimality of Algorithm 1)**.** *Let* Π *be any candidate class and* $\pi^\star_\Pi \in \arg\max_{\pi\in\Pi} V^\pi$*. Run Algorithm 1 with* $\beta = \beta(\delta)$ *from Lemma D.2. With probability at least* $1-\delta$*,*

$$V^{\pi^\star_\Pi} - V^{\hat\pi} \;\le\; 2\,\beta(\delta)\,\|d^{\pi^\star_\Pi}\|_{\Lambda^{-1}}.$$

Proof. Work on the event $\mathcal{E} := \{\|\hat r - r\|_\Lambda \le \beta\}$ of Lemma D.2. For every policy π, by Cauchy–Schwarz in the Λ-inner product, $|\langle d^\pi, \hat r - r\rangle| \le \|d^\pi\|_{\Lambda^{-1}}\|\hat r - r\|_\Lambda \le \beta\|d^\pi\|_{\Lambda^{-1}}$. Hence, on $\mathcal{E}$: (i) $\widehat V(\pi) = \langle d^\pi, \hat r\rangle - \beta\|d^\pi\|_{\Lambda^{-1}} \le \langle d^\pi, r\rangle = V^\pi$ for all $\pi \in \Pi$ (validity of pessimism); and (ii) $\widehat V(\pi^\star_\Pi) \ge V^{\pi^\star_\Pi} - 2\beta\|d^{\pi^\star_\Pi}\|_{\Lambda^{-1}}$. Chaining $V^{\hat\pi} \overset{(i)}{\ge} \widehat V(\hat\pi) \ge \widehat V(\pi^\star_\Pi) \overset{(ii)}{\ge} V^{\pi^\star_\Pi} - 2\beta\|d^{\pi^\star_\Pi}\|_{\Lambda^{-1}}$ (the middle inequality is the optimality of $\hat\pi$ for $\widehat V$) completes the proof. □

To turn Theorem D.3 into a population rate we relate $\|d^{\pi^\star}\|_{\Lambda^{-1}}$ to $\mathcal{C}_{\rm out}$ through a design-concentration lemma (Lemma L.5 in Appendix L, proved via whitened features, a one-directional Chernoff bound, and a net over the sphere of $\mathcal{U}_\mu$): writing $k := \dim\mathcal{U}_\mu$ and $\omega_{\min} := \lambda^+_{\min}(\Sigma_\mu)$, if

$$n \;\ge\; n_0(\delta) \;:=\; \frac{8H}{\omega_{\min}}\Big(k\,\log\frac{24H}{\omega_{\min}} + \log\frac{1}{\delta}\Big),$$

then with probability at least $1-\delta$ one has $\sum_i \phi_i\phi_i^\top \succeq \frac{n}{4}\Sigma_\mu$ on $\mathcal{U}_\mu$, hence $\|v\|^2_{\Lambda^{-1}} \le \frac{4}{n}\|v\|^2_{\Sigma_\mu^\dagger}$ for every $v \in \mathcal{U}_\mu$.

Theorem D.4 (Population suboptimality bound)**.** *Suppose Assumptions 1–3 hold,* $\mathcal{C}^\star_{\rm out} := \mathcal{C}_{\rm out}(\pi^\star;\mu) < \infty$ *(which is a sufficient identification condition, and is necessary at interior reward instances / uniformly over the reward box by Theorem C.1 and Lemma B.2), and* $n \ge n_0(\delta/2)$ *from Lemma L.5. Run Algorithm 1 over all policies with* $\beta = \beta(\delta/2)$ *from Lemma D.2. With probability at least* $1-\delta$*,*

$$\mathrm{SubOpt}(\hat\pi) \;\le\; 4\,\beta(\delta/2)\,\sqrt{\frac{\mathcal{C}^\star_{\rm out}}{n}} \;=\; O\!\left(\big(\sigma\sqrt{d+\log(1/\delta)} + \sqrt{\lambda d}\big)\sqrt{\frac{\mathcal{C}^\star_{\rm out}}{n}}\right).$$

Proof. Intersect the events of Lemma D.2 (probability $\ge 1-\delta/2$) and Lemma L.5 (probability $\ge 1-\delta/2$); the intersection has probability $\ge 1-\delta$. Since $\mathcal{C}^\star_{\rm out} < \infty$ we have $d^{\pi^\star} \in \mathcal{U}_\mu$ (Definition B.3), so Lemma L.5 gives $\|d^{\pi^\star}\|^2_{\Lambda^{-1}} \le \frac{4}{n}\|d^{\pi^\star}\|^2_{\Sigma_\mu^\dagger} = \frac{4\mathcal{C}^\star_{\rm out}}{n}$. (Here we used that $\Lambda \succeq \lambda I + \sum_i \phi_i\phi_i^\top \succeq \sum_i \phi_i\phi_i^\top$ and that Λ is block-diagonal with respect to $\mathcal{U}_\mu \oplus \mathcal{U}_\mu^\perp$ — every ϕ_i lies in $\mathcal{U}_\mu$ almost surely by Lemma B.2, so $\sum_i \phi_i\phi_i^\top$ maps $\mathcal{U}_\mu \to \mathcal{U}_\mu$ and vanishes on $\mathcal{U}_\mu^\perp$, while λI is diagonal; hence for $v \in \mathcal{U}_\mu$, $\Lambda^{-1}v = (\Lambda|_{\mathcal{U}_\mu})^{-1}v$ and the restricted comparison suffices.) Plugging into Theorem D.3 with Π = all policies gives the first inequality; the second substitutes the value of β. □

Corollary D.5 (Memory-aware population guarantee)**.** *Suppose Assumptions 1–3 hold,* $n \ge n_0(\delta/2)$*, the logger has* $\kappa_{\rm mem}(\mu) < \infty$*, and the optimal target has* $C^\star_{\rm std} := C_{\rm std}(\pi^\star;\mu) < \infty$*. Run Algorithm 1 with* $\beta = \beta(\delta/2)$*. Then with probability at least* $1-\delta$*,*

$$\mathrm{SubOpt}(\hat\pi) \le 4\,\beta(\delta/2)\sqrt{\frac{\kappa_{\rm mem}(\mu)\,[1 + H(C^\star_{\rm std} - 1)]}{n}}.$$

Thus finite marginal coverage is sufficient for learning once the logger's additive memory distortion is finite. The sample-size condition still contains the global design/burn-in dependence through $\lambda^+_{\min}(\Sigma_\mu)$*; this corollary does not remove that dependence.*

Proof. Theorem C.6(b) makes $\mathcal{C}_{\rm out}(\pi^\star;\mu)$ finite and gives $\mathcal{C}_{\rm out}(\pi^\star;\mu) \le \kappa_{\rm mem}(\mu)[1 + H(C^\star_{\rm std} - 1)]$. Therefore Theorem D.4 applies, and substitution gives the displayed inequality. □

D.1 Burn-in-free projected confidence-set pessimism

The population conversion above is high probability, but it pays a global spectral threshold through $\lambda^+_{\min}(\Sigma_\mu)$. That threshold is an artifact of asking the empirical Gram matrix to dominate the population covariance on the *entire* logged span. For expected policy regret it can be avoided by keeping the reward box and working instead in the realized observation space.

Let $\mathbf{\Phi}_n \in \mathbb{R}^{n\times d}$ have rows $\phi_i^\top$, let $Y = (Y_1, \dots, Y_n)^\top$, and let P_n be the orthogonal projector onto $\mathrm{range}(\mathbf{\Phi}_n) \subseteq \mathbb{R}^n$. Write $k_n := \mathrm{rank}(\mathbf{\Phi}_n)$ and, for $\delta \in (0,1)$,

$$\gamma_n(\delta) := 2\sigma\sqrt{2\big(k_n \log 5 + \log(1/\delta)\big)}.$$

Define the projected reward confidence set

$$\mathcal{R}_n(\delta) := \left\{u \in [0,1]^d : \left\|P_n\big(Y - \mathbf{\Phi}_n u\big)\right\|_2 \le \gamma_n(\delta)\right\}. \tag{11}$$

Because $\mathbf{\Phi}_n u \in \mathrm{range}(\mathbf{\Phi}_n)$, only the component of the outcome noise that can be confused with a reward vector enters this set. Define

$$L_n(d) := \min_{u\in\mathcal{R}_n(\delta)} \langle d, u\rangle, \qquad \hat{d} \in \arg\max_{d\in\mathcal{K}} L_n(d),$$

and recover a Markov policy from $\hat{d}$ by the occupancy normalization of Theorem D.13. We call this *Algorithm 2*. On the exceptional event that $\mathcal{R}_n(\delta) = \varnothing$, Algorithm 2 returns an arbitrary policy.

Lemma D.6 (Projected reward confidence set)**.** *Under Assumptions 1–2, conditionally on every realized trajectory design,*

$$\mathbb{P}\big(r \in \mathcal{R}_n(\delta) \mid \tau_1, \dots, \tau_n\big) \ge 1 - \delta.$$

In particular, the statement requires no lower bound on any empirical or population eigenvalue.

Proof. Write $\xi = (\xi_1, \dots, \xi_n)^\top$, so $Y = \mathbf{\Phi}_n r + \xi$. Conditional on the trajectories, the coordinates of ξ are independent, mean-zero, and σ-subgaussian. For every unit vector $a \in \mathrm{range}(\mathbf{\Phi}_n)$, the scalar $a^\top\xi$ is therefore σ-subgaussian. A $1/2$-net of the unit sphere in the k_n-dimensional space $\mathrm{range}(\mathbf{\Phi}_n)$ has at most 5^{k_n} elements. Repeating the net argument of Lemma D.2 inside this subspace gives

$$\|P_n\xi\|_2 \le 2\sigma\sqrt{2\big(k_n\log 5 + \log(1/\delta)\big)} = \gamma_n(\delta)$$

with conditional probability at least $1-\delta$. Since $P_n(Y - \mathbf{\Phi}_n r) = P_n\xi$, the true reward belongs to (11) on that event. □

For a target occupancy d, define its realized box-constrained identification width

$$\mathcal{W}_n(d;\delta) := \sup\left\{\langle d, h\rangle :\ h \in [-1,1]^d,\ \|\mathbf{\Phi}_n h\|_2 \le 2\gamma_n(\delta)\right\}. \tag{12}$$

This quantity is finite even when the empirical Gram matrix is singular.

Theorem D.7 (Burn-in-free data-dependent bound)**.** *Run Algorithm 2 over all policies. Under Assumptions 1–3, with probability at least* $1-\delta$*,*

$$\mathrm{SubOpt}(\hat{\pi}) \le \mathcal{W}_n(d^{\pi^\star};\delta).$$

The statement has no design-concentration or minimum-eigenvalue condition.

Proof. Work on the event $r \in \mathcal{R}_n(\delta)$ from Lemma D.6. Because r is feasible, $L_n(d) \le \langle d, r\rangle$ for every $d \in \mathcal{K}$. Hence, by optimality of $\hat{d}$,

$$V^{\pi^\star} - V^{\hat{\pi}} \le V^{\pi^\star} - L_n(d^{\pi^\star}) = \max_{u\in\mathcal{R}_n(\delta)} \langle d^{\pi^\star}, r - u\rangle.$$

For every such u, put $h = r - u$. The reward box gives $h \in [-1,1]^d$. Moreover,

$$\|\mathbf{\Phi}_n h\|_2 = \|P_n\mathbf{\Phi}_n(r-u)\|_2 \le \|P_n(Y - \mathbf{\Phi}_n u)\|_2 + \|P_n(Y - \mathbf{\Phi}_n r)\|_2 \le 2\gamma_n(\delta),$$

where we used $P_n\mathbf{\Phi}_n = \mathbf{\Phi}_n$. Thus every difference $r - u$ in the preceding maximum is feasible in (12), which proves the claim. □

The next theorem converts this sample width directly to the population coefficient $\mathcal{C}_{\mathrm{out}}$ without first lower-bounding the empirical covariance. The one-shot statement is in expectation; Theorem D.9 below then amplifies its design-width control across independent blocks to obtain high probability without a global spectral condition.

Theorem D.8 (Burn-in-free population expectation bound)**.** *Assume* $n \ge 2$ *and* $\mathcal{C}^\star_{\mathrm{out}} := \mathcal{C}_{\mathrm{out}}(\pi^\star;\mu) < \infty$*. Run Algorithm 2 with* $\delta = 1/n$*. Then*

$$\mathbb{E}\big[\mathrm{SubOpt}(\hat{\pi})\big] \le \left[4\sigma\sqrt{2\big(d\log 5 + \log n\big)} + \sqrt{dH}\right]\sqrt{\frac{\mathcal{C}^\star_{\mathrm{out}}}{n}} + \frac{H}{n}. \tag{13}$$

Consequently, for fixed finite $(d, H, \mathcal{C}^\star_{\mathrm{out}})$*, Algorithm 2 is consistent without any lower bound on* $\lambda^+_{\min}(\Sigma_\mu)$ *and without a population design burn-in.*

Proof roadmap. Write $d^\star = d^{\pi^\star}$. On the confidence event, Theorem D.7 bounds regret by the realized width in direction $d^\star$. The obstacle is that, for small n, the empirical Gram matrix need not dominate Σ_μ on the whole logged span. The proof instead represents $d^\star$ through the score $g(\tau) = \phi(\tau)^\top \Sigma_\mu^\dagger d^\star$, whose second moment is $\mathcal{C}^\star_{\mathrm{out}}$: the width splits into a residual term, bounded by Cauchy–Schwarz against $\|\mathbf{\Phi}_n h\|_2$, and a representation error, bounded in ℓ_1 by independence across episodes. The complement of the confidence event costs at most H/n.

Proof. Let $d^\star = d^{\pi^\star}$, $\Sigma = \Sigma_\mu$, and $v := \Sigma^\dagger d^\star$. Finiteness of $\mathcal{C}^\star_{\mathrm{out}}$ implies $d^\star \in \mathrm{range}(\Sigma)$, hence $\Sigma v = d^\star$. For episode i put

$$g_i := \phi_i^\top v, \qquad z_n := d^\star - \frac{1}{n}\sum_{i=1}^n g_i \phi_i.$$

Then $\mathbb{E}[g_i\phi_i] = \Sigma v = d^\star$ and $\mathbb{E}[g_i^2] = v^\top \Sigma v = \mathcal{C}^\star_{\mathrm{out}}$. For any h feasible in (12),

$$\langle d^\star, h\rangle = \frac{1}{n}\sum_{i=1}^n g_i\, \phi_i^\top h + \langle z_n, h\rangle \le \frac{\|g\|_2}{n}\|\mathbf{\Phi}_n h\|_2 + \|z_n\|_1 \le \frac{2\gamma_n(\delta)}{n}\|g\|_2 + \|z_n\|_1.$$

Since $k_n \le d$, at $\delta = 1/n$ we may deterministically upper-bound $\gamma_n(1/n)$ by

$$\bar\gamma_n := 2\sigma\sqrt{2(d\log 5 + \log n)}.$$

Also,

$$\mathbb{E}\|g\|_2 \le \sqrt{\mathbb{E}\sum_{i=1}^n g_i^2} = \sqrt{n\mathcal{C}^\star_{\mathrm{out}}}.$$

The vectors $g_i\phi_i$ are i.i.d. with mean $d^\star$, so independence across episodes and $\|\phi_i\|_2^2 = H$ give

$$\mathbb{E}\|z_n\|_2^2 = \frac{1}{n}\mathbb{E}\|g_1\phi_1 - d^\star\|_2^2 \le \frac{1}{n}\mathbb{E}[g_1^2\|\phi_1\|_2^2] = \frac{H\mathcal{C}^\star_{\mathrm{out}}}{n}.$$

Therefore Cauchy–Schwarz yields $\mathbb{E}\|z_n\|_1 \le \sqrt{dH\mathcal{C}^\star_{\mathrm{out}}/n}$. On the confidence event, Theorem D.7 and the preceding bounds imply

$$\mathrm{SubOpt}(\hat\pi) \le \frac{2\bar\gamma_n}{n}\|g\|_2 + \|z_n\|_1.$$

On its complement, whose probability is at most $1/n$, the trivial reward range gives $\mathrm{SubOpt} \le H$. Taking expectations proves (13). □

Theorem D.9 (High-probability $\mathcal{C}_{\mathrm{out}}$ amplification)**.** *Fix $\delta \in (0,1)$ and let*

$$B := \left\lceil \log_2 \frac{2}{\delta} \right\rceil, \qquad m := \left\lfloor \frac{n}{B} \right\rfloor, \qquad \alpha := \frac{\delta}{2B}.$$

Assume $n \ge B$ and $\mathcal{C}^\star_{\mathrm{out}} < \infty$. Partition the first Bm iid episodes into B disjoint blocks and run Algorithm 2 on each at confidence α. Assign score $-\infty$ to a block with empty reward set; otherwise its score is its candidate's exact pessimistic value. Return a candidate from a block of largest score, breaking ties arbitrarily; if all reward sets are empty, return an arbitrary feasible Markov policy. With

$$A(\alpha) := 4\sigma\sqrt{2[d\log 5 + \log(1/\alpha)]} + \sqrt{dH},$$

we have, with probability at least $1-\delta$,

$$\mathrm{SubOpt}(\hat\pi) \le 2A(\alpha)\sqrt{\frac{\mathcal{C}^\star_{\mathrm{out}}}{m}}. \tag{14}$$

If additionally $n \ge 2B$, then

$$\mathrm{SubOpt}(\hat\pi) \le 2\sqrt{2B}\, A(\alpha)\sqrt{\frac{\mathcal{C}^\star_{\mathrm{out}}}{n}}.$$

No lower bound on $\lambda^+_{\min}(\Sigma_\mu)$ is required, and the learner does not need to know $\mathcal{C}^\star_{\mathrm{out}}$.

Proof. For a single block, repeat the width calculation in the proof of Theorem D.8 with block size m and confidence α. Because the block rank is at most d,

$$\mathbb{E}\big[\mathcal{W}_b(d^\star;\alpha)\big] \le A(\alpha)\sqrt{\frac{\mathcal{C}^\star_{\mathrm{out}}}{m}}.$$

Markov's inequality therefore makes $\mathcal{W}_b(d^\star;\alpha) \le 2A(\alpha)\sqrt{\mathcal{C}^\star_{\mathrm{out}}/m}$ with probability at least $1/2$. These width events depend only on disjoint trajectory blocks and are independent, so the probability that no block is good is at most $2^{-B} \le \delta/2$.

For each block, Lemma D.6 gives $\mathbb{P}(r \in \mathcal{R}_b(\alpha) \mid \mathrm{design}_b) \ge 1-\alpha$; hence a union bound makes all B reward sets valid with probability at least $1 - B\alpha = 1 - \delta/2$. On the intersection, choose a good block b_0. Validity makes every block reward set nonempty, so the fallback is not used on this intersection. Validity and Theorem D.7 imply

$$\max_{d\in\mathcal{K}} L_{b_0}(d) \ge L_{b_0}(d^\star) \ge V^\star - 2A(\alpha)\sqrt{\frac{\mathcal{C}^\star_{\mathrm{out}}}{m}}.$$

The selected block has at least this pessimistic score, while validity makes its score no larger than its candidate's true value. This proves (14). If $n \ge 2B$, then $m = \lfloor n/B \rfloor \ge n/(2B)$, giving the final display. □

Remark D.2 (Numerical certificates for Algorithm 2). The statistical statements above concern exact convex optimization. The reference implementation also returns a floating-point dual lower certificate and a master upper bound for each max–min solve. If every block has certified optimization gap at most η and block selection uses the dual lower certificate, the bound in (14) gains only an additive $+\eta$. These are numerical primal/dual certificates, not formal interval-arithmetic proofs. Since $B = O(\log(1/\delta))$, the boosted exact procedure remains polynomial-time under the oracle model of Theorem D.11.

Corollary D.10 (Burn-in-free memory-aware guarantee)**.** *Under the conditions of Theorem D.8, suppose $C^\star_{\mathrm{std}} := C_{\mathrm{std}}(\pi^\star;\mu) < \infty$. If $\kappa_{\mathrm{mem}}(\mu) < \infty$, then Algorithm 2 satisfies*

$$\mathbb{E}[\mathrm{SubOpt}(\hat\pi)] \le \Big[4\sigma\sqrt{2(d\log 5 + \log n)} + \sqrt{dH}\Big]\sqrt{\frac{\kappa_{\mathrm{mem}}(\mu)[1 + H(C^\star_{\mathrm{std}} - 1)]}{n}} + \frac{H}{n}.$$

Thus the memory-aware population comparison of Theorem C.6 can be converted to a finite-sample consistency bound without the global spectral burn-in required by the ridge high-probability corollary above. Under finite marginal coverage, the same bound also holds with $\kappa_{\mathrm{rep}}(\pi^\star;\mu)$ replacing $\kappa_{\mathrm{mem}}(\mu)$ by Proposition C.7, without assuming $\kappa_{\mathrm{mem}}(\mu) < \infty$.

Proof. Substitute Theorem C.6(b) into Theorem D.8. For the localized variant, substitute the target-specific inequality of Proposition C.7 instead. □

Remark D.3 (Sharp noiseless interpretation). If $\sigma = 0$, then $\gamma_n = 0$ and, on the outcome model,

$$\mathcal{R}_n = \{u \in [0,1]^d : \mathbf{\Phi}_n u = Y\}.$$

Thus Algorithm 2 is exactly robust optimization over the reward vectors consistent with the realized cardinal outcomes. It is a finite-sample analogue of the sharp population identified set in Proposition C.2, rather than an ellipsoidal relaxation based on a regularized point estimator.

Theorem D.11 (Algorithm 2 is polynomial-time tractable)**.** *For a fixed confidence level, $\mathcal{R}_n(\delta)$ is a compact convex (possibly empty) second-order-cone representable set, and its feasibility can be checked by an SOCP. Whenever it is nonempty, $L_n(d) = \min_{u\in\mathcal{R}_n} d^\top u$ is concave on $\mathcal{K}$, and an η-optimal occupancy for Algorithm 2 can be computed in time polynomial in the finite MDP/data description and $\log(1/\eta)$, under the standard rational-encoding model, using conic/ellipsoid optimization. A value and supergradient oracle for L_n requires only the SOCP $\min_{u\in\mathcal{R}_n} d^\top u$; any minimizer is a supergradient. The recovered occupancy induces a Markov policy as in Theorem D.13.*

Proof. The box constraints are linear and the single residual constraint in (11) is second-order conic, so the inner minimization is an SOCP. An infimum of linear functions of d is concave, and by Danskin's theorem any minimizing u is a supergradient of L_n. The occupancy set $\mathcal{K}$ has an explicit polynomial-size linear description. Therefore a standard ellipsoid method with the SOCP value/supergradient oracle maximizes L_n over $\mathcal{K}$ to additive accuracy η in polynomial time. Finally, the occupancy-to-policy normalization is identical to that in Theorem D.13. □

Remark D.4 (Algorithm 2 in context). Confidence-set pessimism and robust planning are established in offline RL and offline preference-based RL (Zhu et al., 2023; Zhan et al., 2024), and the blockwise amplification in Theorem D.9 is classical. Algorithm 2 specializes this template to outcome-only data through the projected observation-space set (11); its exact reduction, at $\sigma = 0$, to the reward vectors consistent with the realized outcomes (Remark D.3); and a $\mathcal{C}_{\mathrm{out}}$ analysis under general valid history-dependent logging without a global spectral burn-in. Theorem D.8 is the one-shot, use-all-data expectation result; Theorem D.9 trades a logarithmic sample-splitting and solve overhead for high probability. Complementary policy-generated results give high-probability guarantees under standard state–action coverage (Ren and Xie, 2026).

Remark D.5 (Degenerate cases of the ridge upper bound). (i) *Noiseless outcomes* ($\sigma = 0$): Lemma D.2 gives $\beta = \sqrt{\lambda d}$, so $\mathrm{SubOpt} \le 4\sqrt{\lambda d\, \mathcal{C}^\star_{\mathrm{out}}/n} \to 0$ as $\lambda \downarrow 0$; identification on $\mathcal{U}_\mu$ is exact once the supported trajectories have been observed. (ii) *Horizon one*: $\phi(\tau) = e_{(s_1,a_1)}$, so Σ_μ is diagonal with entries $d^\mu(1,s,a)$ and $\mathcal{C}_{\mathrm{out}}(\pi) = \sum_{s,a} d^\pi(s,a)^2/d^\mu(s,a) \le C_{\mathrm{std}}(\pi)\sum_{s,a} d^\pi(s,a) = C_{\mathrm{std}}(\pi)$: outcome-only feedback is never harder than standard offline learning at $H = 1$ (the outcome *is* the per-step reward), and the aliasing phenomenon of Proposition C.4 genuinely requires $H \ge 2$. (iii) $\mathcal{C}^\star_{\mathrm{out}} = \infty$: the bound is vacuous, necessarily so by Theorem E.1.

The dimension factor $\sqrt{d}$ inside β comes from certifying *all* directions of $\mathbb{R}^d$ simultaneously, whereas Algorithm 1 only ever evaluates the $|\Pi|$ linear functionals $\langle d^\pi, \hat{r} - r\rangle$. Making this explicit both improves the bound and, as Section E.1 shows, identifies the exact quantity that the lower bound will have to match.

Theorem D.12 (Refined radius for a finite policy class)**.** *Let Π be a finite set of policies fixed before the data are observed, and put*

$$\beta_\Pi(\delta) \;:=\; \sigma\sqrt{2\log\big(2|\Pi|/\delta\big)} \;+\; \sqrt{\lambda d}\,.$$

Then with probability at least $1-\delta$, simultaneously for every $\pi \in \Pi$,

$$\big|\langle d^\pi, \hat{r}\rangle - V^\pi\big| \;\le\; \beta_\Pi(\delta)\,\|d^\pi\|_{\Lambda^{-1}}.$$

Consequently Algorithm 1 run over Π with $\beta = \beta_\Pi(\delta)$ satisfies $V^{\pi^\star_\Pi} - V^{\hat{\pi}} \le 2\beta_\Pi(\delta)\|d^{\pi^\star_\Pi}\|_{\Lambda^{-1}}$ with probability at least $1-\delta$; and if in addition $n \ge n_0(\delta/2)$ and $\beta = \beta_\Pi(\delta/2)$, then with probability at least $1-\delta$,

$$V^{\pi^\star_\Pi} - V^{\hat{\pi}} \;\le\; 4\,\beta_\Pi(\delta/2)\sqrt{\frac{\mathcal{C}_{\mathrm{out}}(\pi^\star_\Pi;\mu)}{n}}\,.$$

For global suboptimality, add the approximation gap $V^\star - V^{\pi^\star_\Pi}$ to this bound; in particular the same right-hand side bounds $\mathrm{SubOpt}(\hat{\pi})$ whenever Π contains a globally optimal policy and $\pi^\star_\Pi$ is chosen among its in-class optimizers accordingly.

Proof. Fix $\pi \in \Pi$ and condition on the design $(\phi_1, \dots, \phi_n)$, so that Λ and d^π are deterministic and the noises $\xi_i = Y_i - \langle \phi_i, r\rangle$ remain independent, mean-zero and σ-subgaussian (Assumptions 1 and 2). As in the proof of Lemma D.2, $\hat{r} - r = \Lambda^{-1}X - \lambda\Lambda^{-1}r$ with $X = \sum_i \phi_i\xi_i$, so

$$\langle d^\pi, \hat{r} - r\rangle = \sum_{i=1}^n c_i\xi_i - \lambda\langle \Lambda^{-1}d^\pi, r\rangle, \qquad c_i := \langle \Lambda^{-1}d^\pi, \phi_i\rangle.$$

For the deterministic part, Cauchy–Schwarz in the Λ-inner product and $\Lambda \succeq \lambda I$ give $\lambda|\langle \Lambda^{-1}d^\pi, r\rangle| \le \lambda\|d^\pi\|_{\Lambda^{-1}}\|r\|_{\Lambda^{-1}} \le \sqrt{\lambda}\,\|r\|_2\,\|d^\pi\|_{\Lambda^{-1}} \le \sqrt{\lambda d}\,\|d^\pi\|_{\Lambda^{-1}}$. For the stochastic part,

$$\sum_{i=1}^n c_i^2 = (\Lambda^{-1}d^\pi)^\top\Big(\sum_i \phi_i\phi_i^\top\Big)(\Lambda^{-1}d^\pi) \le (\Lambda^{-1}d^\pi)^\top\Lambda(\Lambda^{-1}d^\pi) = \|d^\pi\|^2_{\Lambda^{-1}},$$

so by Lemma L.1 (after rescaling the coefficients by $\|d^\pi\|_{\Lambda^{-1}}$) the sum $\sum_i c_i\xi_i$ is $\sigma\|d^\pi\|_{\Lambda^{-1}}$-subgaussian, and Lemma L.2 applied to $\pm\sum_i c_i\xi_i$ yields

$$\mathbb{P}\Big(\big|\textstyle\sum_i c_i\xi_i\big| > \sigma\|d^\pi\|_{\Lambda^{-1}}\sqrt{2\log(2|\Pi|/\delta)}\Big) \;\le\; \tfrac{\delta}{|\Pi|}.$$

A union bound over the $|\Pi|$ policies gives the first display conditionally on every design, hence unconditionally. The two remaining in-class claims follow by repeating the chaining argument of Theorem D.3 (which used only the uniform-over-Π sandwich just established) and then the design step of Theorem D.4. The global decomposition follows by adding $V^\star - V^{\pi^\star_\Pi}$ to both sides. □

Remark D.6 (Dimension versus log-cardinality). Since the value V^π is linear in d^π and the occupancy polytope has the deterministic Markov policies as its vertices, taking Π to be the set of all deterministic policies loses nothing: $\max_{\pi\in\Pi} V^\pi = \max_\pi V^\pi$. For that choice $|\Pi| = A^{HS}$, so $\log|\Pi| = HS\log A = (d/A)\log A$, and Theorem D.12 replaces the factor $\sqrt{d\log 5}$ of Lemma D.2 by $\sqrt{2(d/A)\log A}$ — never worse, and smaller by a factor of order $\sqrt{A/\log A}$ for large action spaces. This refinement replaces the *stochastic* $\sqrt{d}$ term by $\sqrt{\log|\Pi|}$; the ridge-bias term $\sqrt{\lambda d}$ remains unless λ is tuned correspondingly. The construction of Section E.1 shows that $\log|\Pi| \asymp d$ can occur, so a dimension-sized stochastic term cannot be removed uniformly in general.

Remark D.7 (The coefficient may be minimized over optimal policies). The pessimism argument bounds $V^{\pi^\star_\Pi} - V^{\hat\pi}$ by $2\beta\|d^\pi\|_{\Lambda^{-1}}$ for *any* maximizer $\pi \in \arg\max_{\pi'\in\Pi} V^{\pi'}$, since the only property used is that the pessimistic objective at $\hat\pi$ is at least its value at π. When the optimal policy is not unique we may therefore replace $\mathcal{C}_{\mathrm{out}}(\pi^\star;\mu)$ throughout by

$$\mathcal{C}^\star_{\mathrm{out}} \;:=\; \min\big\{\mathcal{C}_{\mathrm{out}}(\pi;\mu)\;:\;\pi\in\Pi,\; V^\pi = \max_{\pi'\in\Pi} V^{\pi'}\big\}.$$

This refinement is consequential: in the family of Section E.4, the optimal policy is free to act arbitrarily on the reward-neutral part of the horizon, and the best and worst such choices separate a bounded coefficient from one that grows linearly in d.

Algorithm 1 is stated as a maximization over a policy class whose deterministic subset has A^{HS} members. The occupancy representation nevertheless yields a polynomial-size convex program.

Theorem D.13 (Algorithm 1 is a tractable convex program)**.** *Let $\mathcal{K}\subset\mathbb{R}^d$ be the occupancy polytope*

$$\mathcal{K} = \Big\{d\ge 0\;:\;\textstyle\sum_a d(1,s,a) = \rho(s)\,\forall s,\quad \sum_a d(h{+}1,s',a) = \sum_{s,a} P_h(s'|s,a)\,d(h,s,a)\,\forall h,s'\Big\},$$

whose elements are exactly the occupancy measures of Markov policies and whose vertices are the deterministic ones. Then:

- *(a) $L(d) := \langle d,\hat r\rangle - \beta\|d\|_{\Lambda^{-1}}$ is concave on $\mathbb{R}^d$, so maximizing L over $\mathcal{K}$ — a set with A^{HS} vertices but only HS linear equalities and HSA bound constraints — is a convex program; in epigraph form it is the second-order cone program $\max_{d\in\mathcal{K},\,t\ge0}\langle d,\hat r\rangle - \beta t$ subject to $\|\Lambda^{-1/2}d\|_2 \le t$.*
- *(b) The linear oracle $\max_{d\in\mathcal{K}}\langle d,c\rangle$ is one backward induction, at cost $O(HS^2A)$. Hence Frank–Wolfe started from any $d_0\in\mathcal{K}$ returns d_t with $\max_{\mathcal{K}} L - L(d_t) = O(1/t)$ after t oracle calls, for a total cost $O(t\,HS^2A)$ — polynomial in the description of the MDP.*
- *(c) If $\hat d\in\mathcal{K}$ is any η-maximizer of L and $\hat\pi(a\mid h,s) := \hat d(h,s,a)/\sum_{a'}\hat d(h,s,a')$, then on the event of Lemma D.2, $\mathrm{SubOpt}(\hat\pi)\le 2\beta\|d^{\pi^\star}\|_{\Lambda^{-1}} + \eta$, so Theorems D.3 and D.4 hold up to the additive η.*

Proof. (a) L is a linear functional minus a positive multiple of a norm; norms are convex, so L is concave. That $\mathcal{K}$ is exactly the set of Markov-policy occupancies, with deterministic policies as vertices, is the standard dual linear program for finite-horizon MDPs; the recovery map $\hat\pi(a|h,s) = \hat d(h,s,a)/\sum_{a'}\hat d(h,s,a')$ inverts it (any normalization is valid at unreached states). The epigraph reformulation is immediate.

(b) Maximizing $\langle d,c\rangle$ over $\mathcal{K}$ is the dual of the MDP with reward c, solved by backward induction: with $Q_h = c_h + P_hV_{h+1}$ and $V_h = \max_a Q_h$, the greedy deterministic policy's occupancy is a maximizer. For the convergence rate, $\nabla L(d) = \hat r - \beta\Lambda^{-1}d/\|d\|_{\Lambda^{-1}}$ is Lipschitz on $\mathcal{K}$ because $\|d\|_{\Lambda^{-1}}$ is bounded away from zero there: every $d\in\mathcal{K}$ has $\|d\|_1 = H$, hence $\|d\|_2 \ge H/\sqrt{d}$ and $\|d\|_{\Lambda^{-1}} \ge H/\sqrt{d\,\lambda_{\max}(\Lambda)} > 0$. Frank–Wolfe on a compact convex set with a concave, smooth objective then gives the standard $O(1/t)$ rate.

(c) Let $\pi^\star$ maximize V over Markov policies, so $d^{\pi^\star} \in \mathcal{K}$. Decompose

$$V^{\pi^\star} - V^{\hat{\pi}} = \underbrace{\big(V^{\pi^\star} - L(d^{\pi^\star})\big)}_{\le 2\beta\|d^{\pi^\star}\|_{\Lambda^{-1}}} + \underbrace{\big(L(d^{\pi^\star}) - L(\hat{d})\big)}_{\le \eta} + \underbrace{\big(L(\hat{d}) - V^{\hat{\pi}}\big)}_{\le 0},$$

where the first and third bounds hold simultaneously for all $d \in \mathcal{K}$ on the event of Lemma D.2, which controls $\|\hat{r} - r\|_\Lambda$ and hence certifies every direction at once, and the middle bound is η-optimality. □

Remark D.8 (A statistical–computational trade-off). Theorem D.13(c) uses the radius of Lemma D.2, which certifies all of $\mathbb{R}^d$ simultaneously and therefore carries the factor $\sqrt{d}$. The refined radius β_Π of Theorem D.12 is valid only on a fixed finite Π, and maximizing L over such a Π has no evident convex structure. At present, then, the $\sqrt{d}$ can be exchanged for $\sqrt{\log|\Pi|}$ only by giving up tractability, or tractability retained at the price of $\sqrt{d}$. The sparse family of Section E.4 escapes the dilemma because there the effective class is the vertex set of an $(s{+}1)$-dimensional polytope, on which the net-based radius is already $O(\sigma\sqrt{s}) = O(\sigma\sqrt{\log|\Pi|})$. Whether the two can be reconciled in general is open.

Remark D.9 (The pessimistic policy is genuinely stochastic). Because L is strictly concave along $\|\cdot\|_{\Lambda^{-1}}$, its maximizer over $\mathcal{K}$ is typically an interior point rather than a vertex, so $\hat{\pi}$ is a stochastic policy and enumerating deterministic policies does not solve the program exactly. This stochasticity can be essential: Proposition E.11(b) exhibits instances in which a stochastic optimal policy has a strictly smaller coverage coefficient than every deterministic one.

E Impossibility and Minimax Lower Bounds

This appendix proceeds from impossibility to rates. Theorem E.1 constructs observationally equivalent reward worlds for a target outside the logged span, so nonvanishing policy error is unavoidable for every algorithm at every sample size on that pair; Theorem E.2 and Corollary E.3 give two-point rates in C_{out} and κ_{mem}; the K-block family (Section E.1) exposes the dimension factor and the horizon–memory composition; the parallel-state class (Section E.2) decouples dimension from horizon at $H = 2$; Section E.3 combines Gaussian testing with missing exposure for general even H; and Section E.4 gives a structured subclass whose rate is not driven by ambient dimension.

Theorem E.1 (Exact impossibility without span coverage)**.** *Fix $\Delta \in (0, 2]$ and consider the 0-XOR pair of instances with rewards $r_\pm := \frac{1}{2}\mathbf{1} \pm \frac{\Delta}{4} z$ (Section C.1 with $\varepsilon = 0$), identical dynamics and behavior, and any common conditional noise law satisfying Assumption 1. Then the two instances induce exactly the same distribution of the dataset $\mathcal{D}$ for every n, while $V_+^\star = V_-^\star = 1 + \Delta/2$ and, for* every *policy π,*

$$\mathrm{SubOpt}_+(\pi) + \mathrm{SubOpt}_-(\pi) \;=\; \Delta.$$

Consequently, for every (possibly randomized) algorithm $\mathcal{A} : \mathcal{D} \mapsto \hat{\pi}$ and every sample size n,

$$\max_{s \in \{+,-\}} \mathbb{E}_s\big[\mathrm{SubOpt}_s(\hat{\pi})\big] \;\ge\; \frac{\Delta}{2}.$$

In particular, with $\Delta = 2$, no algorithm attains worst-case expected suboptimality below 1 at any sample size, even though every policy satisfies $C_{\mathrm{std}}(\pi; \mu) \le 2$ and every deterministic policy has equality.

Proof. Same data law. Under $\varepsilon = 0$ the supported trajectories are (a, a) and (b, b), whose features satisfy $\langle \phi, z \rangle = 0$ (Proposition C.4(a): $\phi(a, a) = u_1 + u_2 \perp u_3$ and likewise $\phi(b, b)$). Hence $\langle \phi(\tau), r_+ \rangle = \langle \phi(\tau), r_- \rangle = 1$ for both supported τ: conditional outcome means agree, the common noise law does the rest, and the trajectory marginal q_μ is shared; independence across episodes extends the equality to the law of $\mathcal{D}$.

Value identity. $r_+ + r_- = \mathbf{1}$, so for every policy $V_+(\pi) + V_-(\pi) = \langle d^\pi, \mathbf{1} \rangle = H = 2$. Both instances have optimal value $1 + \Delta/2$ (attained by (a, b) under r_+ and by (b, a) under r_-: returns computed in Section C.1 and its mirror image). Therefore $\mathrm{SubOpt}_+(\pi) + \mathrm{SubOpt}_-(\pi) = (1 + \frac{\Delta}{2} - V_+(\pi)) + (1 + \frac{\Delta}{2} - V_-(\pi)) = 2 + \Delta - 2 = \Delta$, for every π — including stochastic policies, since the identity is linear in d^π.

Algorithmic consequence. Let $\hat{\pi} = \mathcal{A}(\mathcal{D})$ (randomized algorithms: augment $\mathcal{D}$ with independent seeds; the argument is unchanged). Because the law of $\mathcal{D}$ is the same under both instances, the law of $\hat{\pi}$ is the same; call the common expectation $\mathbb{E}_0$. Then

$$\max_s \mathbb{E}_s[\mathrm{SubOpt}_s(\hat{\pi})] \geq \tfrac{1}{2}\big(\mathbb{E}_0[\mathrm{SubOpt}_+(\hat{\pi})] + \mathbb{E}_0[\mathrm{SubOpt}_-(\hat{\pi})]\big) = \tfrac{1}{2}\,\mathbb{E}_0[\Delta] = \frac{\Delta}{2}. \qquad \square$$

Theorem E.2 (Two-point lower bound under bounded $\mathcal{C}_{\mathrm{out}}$)**.** *Let the noise be Gaussian, $\xi \mid \tau \sim \mathcal{N}(0, \sigma^2)$ (which satisfies Assumption 1). For every $\mathcal{C} \geq 2$ and every $n \geq \sigma^2(\mathcal{C}-1)/4$, there exist two OO-RL instances, each with $\mathcal{C}_{\mathrm{out}}(\pi_s^\star; \mu) \leq \mathcal{C}$, $C_{\mathrm{std}}(\pi_s^\star; \mu) = 2$, and $C_{\mathrm{std}}(\pi; \mu) \leq 2$ for every policy, such that every (possibly randomized) algorithm satisfies*

$$\max_{s \in \{+,-\}} \mathbb{E}_s\big[\mathrm{SubOpt}_s(\hat{\pi})\big] \;\geq\; \frac{\sigma}{8}\sqrt{\frac{\mathcal{C}-1}{n}}.$$

Proof. Set $\varepsilon := 1/(\mathcal{C}-1) \in (0, 1]$ and use the ε-XOR pair $r_\pm = \frac{1}{2}\mathbf{1} \pm \frac{\Delta}{4} z$ with $\Delta := \sigma/\sqrt{n\varepsilon} \leq 2$ (the inequality is the assumed lower bound on n); the rewards lie in $[0,1]^4$ because $\Delta \leq 2$. By Proposition C.4(b), $\mathcal{C}_{\mathrm{out}}(\pi_\pm^\star) = 1 + 1/\varepsilon = \mathcal{C}$ under both signs (the optimal policy is (a,b) or (b,a); both have $\|u_1 \pm u_3\|^2_{\Sigma_\mu^\dagger} = 1 + 1/\varepsilon$), and $C_{\mathrm{std}}(\pi^\star; \mu) = 2$ by part (c).

Information bound. The two instances share q_μ; their episode laws differ only in the conditional outcome mean, which coincides on $(a,a), (b,b)$ (as in Theorem E.1) and differs by $|\langle \phi, r_+ - r_- \rangle| = \frac{\Delta}{2}|\langle \phi, z \rangle| = \Delta$ on each crossing trajectory ($\langle \phi(a,b), z \rangle = 2$, $\langle \phi(b,a), z \rangle = -2$). For Gaussians with common variance, $\mathrm{KL}(\mathcal{N}(m_+, \sigma^2) \,\|\, \mathcal{N}(m_-, \sigma^2)) = (m_+ - m_-)^2/(2\sigma^2)$. By the chain rule for KL over the pair (τ, Y) and then over n independent episodes,

$$\mathrm{KL}\big(\mathbb{P}_+^{\otimes n} \,\big\|\, \mathbb{P}_-^{\otimes n}\big) = n\,\mathbb{E}_{\tau \sim q_\mu}\left[\frac{(m_+(\tau) - m_-(\tau))^2}{2\sigma^2}\right] = n \cdot \varepsilon \cdot \frac{\Delta^2}{2\sigma^2} = \frac{1}{2},$$

using $\mathbb{P}_\mu(\text{crossing}) = \varepsilon$ and the choice of Δ. Pinsker's inequality gives $\mathrm{TV}(\mathbb{P}_+^{\otimes n}, \mathbb{P}_-^{\otimes n}) \leq \sqrt{\mathrm{KL}/2} = \frac{1}{2}$.

From testing to suboptimality. The identity $\mathrm{SubOpt}_+(\pi) + \mathrm{SubOpt}_-(\pi) = \Delta$ of Theorem E.1 holds verbatim for every ε (it used only $r_+ + r_- = \mathbf{1}$ and the common optimal value $1 + \Delta/2$). Let $f(\mathcal{D}) := \mathrm{SubOpt}_+(\hat{\pi}(\mathcal{D})) \in [0, \Delta]$. Then

$$\begin{aligned}\max_s \mathbb{E}_s[\mathrm{SubOpt}_s(\hat{\pi})] &\geq \tfrac{1}{2}\Big(\mathbb{E}_+[f] + \mathbb{E}_-[\Delta - f]\Big) = \tfrac{\Delta}{2} + \tfrac{1}{2}\big(\mathbb{E}_+[f] - \mathbb{E}_-[f]\big) \\ &\geq \tfrac{\Delta}{2} - \tfrac{\Delta}{2}\,\mathrm{TV}\big(\mathbb{P}_+^{\otimes n}, \mathbb{P}_-^{\otimes n}\big) \;\geq\; \frac{\Delta}{4} \;=\; \frac{\sigma}{4\sqrt{n\varepsilon}} \;=\; \frac{\sigma}{4}\sqrt{\frac{\mathcal{C}-1}{n}},\end{aligned}$$

where the total-variation step uses $|\mathbb{E}_+ f - \mathbb{E}_- f| \leq \|f\|_\infty \mathrm{TV} \leq \Delta\,\mathrm{TV}$ for the bounded measurable f (randomized algorithms: average over the independent internal seed first), and the final equality substitutes $1/\varepsilon = \mathcal{C} - 1$. This is stronger than the stated bound with the factor $1/8$, which we keep to absorb the truncation $\Delta = \min\{2, \sigma/\sqrt{n\varepsilon}\}$ needed in the boundary case $n = \lceil \sigma^2(\mathcal{C}-1)/4 \rceil$. $\square$

Corollary E.3 (The memory factor is statistically necessary)**.** *For every $\kappa \geq 1$ and every $n \geq \sigma^2\kappa/2$, there is a two-step OO-RL instance from the XOR family with*

$$\kappa_{\mathrm{mem}}(\mu) = \kappa, \qquad C_{\mathrm{std}}(\pi; \mu) \leq 2 \text{ for every policy},$$

such that every possibly randomized algorithm satisfies

$$\max_{s \in \{+,-\}} \mathbb{E}_s[\mathrm{SubOpt}_s(\hat{\pi})] \geq \frac{\sigma}{4\sqrt{2}}\sqrt{\frac{\kappa}{n}}.$$

At the singular endpoint $\kappa_{\mathrm{mem}} = \infty$, Theorem E.1 gives a non-vanishing lower bound for every sample size. Thus the $\sqrt{\kappa_{\mathrm{mem}}/n}$ dependence in the burn-in-free upper bound of Corollary D.10 cannot in general be removed, even when $H = 2$ and standard marginal concentrability is fixed.

Proof. Choose the XOR crossing probability $\varepsilon = 1/(2\kappa) \in (0, 1/2]$. By Theorem C.6(d),

$$\kappa_{\mathrm{mem}}(\mu_\varepsilon) = \max\left\{1, \frac{1}{2\varepsilon}, \frac{1}{2(1-\varepsilon)}\right\} = \kappa,$$

and Proposition C.4(c) gives the stated marginal coverage. The optimal target has $\mathcal{C}_{\mathrm{out}} = 1 + 1/\varepsilon = 1 + 2\kappa$. Apply Theorem E.2 with $\mathcal{C} = 1 + 2\kappa$; its sample-size requirement is exactly $n \geq \sigma^2\kappa/2$, and its lower bound becomes

$$\frac{\sigma}{8}\sqrt{\frac{2\kappa}{n}} = \frac{\sigma}{4\sqrt{2}}\sqrt{\frac{\kappa}{n}}.$$

The infinite-distortion statement is Theorem E.1. □

Remark E.1 (What is still missing). Theorems D.4 and E.2 match in their joint dependence on σ, n, and $\mathcal{C}_{\mathrm{out}}$, but not in the ambient dimension: the generic stochastic confidence term scales as $\sqrt{d + \log(1/\delta)}$, while Theorem D.12 replaces that stochastic term by $\sqrt{\log|\Pi|}$ but leaves the ridge-bias term $\sqrt{\lambda d}$. The two-point argument of Theorem E.2 cannot reveal either a dimension or log-cardinality factor because it has only two hypotheses. The ε-XOR family has $d = 4$ fixed, so it is structurally unable to certify a dimension factor. Section E.1 resolves this by replacing the two hypotheses with a hypercube of 2^K of them.

E.1 The multi-block construction: the dimension factor is necessary

The K-block XOR instance. Fix $K \geq 1$, $\varepsilon \in (0, 1]$, a gap $\Delta \in (0, 2]$, and a sign vector $\theta \in \{-1, +1\}^K$. Take $S = 1$, $\mathcal{A} = \{a, b\}$ and $H = 2K$, so that $d = HSA = 4K$; transitions are trivial. Group the steps into K consecutive blocks, block k consisting of steps $2k-1$ and $2k$, and order the four reward coordinates of block k as $\big((2k{-}1)a, (2k{-}1)b, (2k)a, (2k)b\big)$. Put

$$z_k := e_{(2k-1)a} - e_{(2k-1)b} - e_{(2k)a} + e_{(2k)b}, \qquad r_\theta := \tfrac{1}{2}\mathbf{1} + \frac{\Delta}{4}\sum_{k=1}^{K}\theta_k z_k \;\in\; [0,1]^d,$$

the membership holding because the z_k have disjoint supports and $\Delta \leq 2$. The behavior draws the action pair of each block *independently across blocks*, from $\{(a,a),(b,b)\}$ with probability $(1-\varepsilon)/2$ each and from $\{(a,b),(b,a)\}$ with probability $\varepsilon/2$ each. Each block is thus an independent copy of the ε-XOR instance of Section C.1, and $K = d/4$.

Proposition E.4 (Closed forms for the K-block family)**.** *Let $u_1^{(k)}, u_2^{(k)}, u_3^{(k)}, u_4^{(k)}$ be the orthonormal basis of the coordinates of block k defined as in Proposition C.4 (so $u_1^{(k)} = \frac{1}{2}(1,1,1,1)$ and $u_3^{(k)} = \frac{1}{2}z_k$ in block-local coordinates), and set $w := \sum_k u_1^{(k)}$, $\hat{w} := w/\sqrt{K}$. Then:*

(a) $\Sigma_\mu = K\,\hat{w}\hat{w}^\top + (1-\varepsilon)\sum_{k=1}^{K} u_2^{(k)}u_2^{(k)\top} + \varepsilon\sum_{k=1}^{K} u_3^{(k)}u_3^{(k)\top}$. *For $0 < \varepsilon < 1$ the nonzero eigenvalues are K (once), $1-\varepsilon$ (K times) and ε (K times), so $\dim\mathcal{U}_\mu = 2K+1$ and $\omega_{\min} = \min\{\varepsilon, 1-\varepsilon\}$. At $\varepsilon = 1$ the $u_2^{(k)}$ directions drop from the range and $\dim\mathcal{U}_\mu = K+1$.*

(b) For every policy π, writing $t_k(\pi) := \langle d^\pi, z_k\rangle \in [-2, 2]$,

$$V_\theta(\pi) \;=\; K + \frac{\Delta}{4}\sum_{k=1}^{K}\theta_k\, t_k(\pi), \qquad V_\theta^\star \;=\; K + \frac{\Delta K}{2},$$

the optimum being attained by the deterministic policy $\pi_\theta^\star$ that plays (a,b) in block k if $\theta_k = +1$ and (b,a) if $\theta_k = -1$.

(c) $\mathcal{C}_{\mathrm{out}}(\pi_\theta^\star; \mu) = 1 + K/\varepsilon$ for every *θ, while $C_{\mathrm{std}}(\pi; \mu) \leq 2$ for every policy π and every ε, with equality for deterministic policies.*

(d) Define the decoder $\hat{\theta}_k(\pi) := \operatorname{sign} t_k(\pi)$, with the convention $\operatorname{sign} 0 := +1$. Then for every policy π (stochastic or history-dependent),

$$\mathrm{SubOpt}_\theta(\pi) \;\geq\; \frac{\Delta}{2}\,\rho_{\mathrm{H}}\big(\hat{\theta}(\pi), \theta\big), \qquad \textit{and} \qquad \mathrm{SubOpt}_\theta(\pi) \leq \Delta K,$$

where ρ_{H} is the Hamming distance on $\{-1,+1\}^K$.

Proof. (a) Write $\phi(\tau) = \sum_k \phi_k(\tau)$ with ϕ_k supported on block k's four coordinates. Since the blocks are drawn independently,

$$\Sigma_\mu = \mathbb{E}\Big[\Big(\sum_k \phi_k\Big)\Big(\sum_l \phi_l\Big)^\top\Big] = \sum_k \mathbb{E}[\phi_k\phi_k^\top] + \sum_{k\neq l} \mathbb{E}[\phi_k]\,\mathbb{E}[\phi_l]^\top.$$

At every step the behavior plays each action with probability $\frac{1-\varepsilon}{2} + \frac{\varepsilon}{2} = \frac{1}{2}$, so $\mathbb{E}[\phi_k] = \frac{1}{2}(1,1,1,1) = u_1^{(k)}$ in block-local coordinates. By Proposition C.4(a), $\mathbb{E}[\phi_k\phi_k^\top] = u_1^{(k)}u_1^{(k)\top} + (1-\varepsilon)u_2^{(k)}u_2^{(k)\top} + \varepsilon u_3^{(k)}u_3^{(k)\top}$. Hence the u_1-terms assemble into $\sum_k u_1^{(k)}u_1^{(k)\top} + \sum_{k\neq l} u_1^{(k)}u_1^{(l)\top} = \big(\sum_k u_1^{(k)}\big)\big(\sum_l u_1^{(l)}\big)^\top = ww^\top = K\hat{w}\hat{w}^\top$, using $\|w\|^2 = K$ (the $u_1^{(k)}$ are orthonormal). The eigenvalues follow since $\hat{w}$, the $u_2^{(k)}$ and the $u_3^{(k)}$ are $2K+1$ mutually orthonormal vectors.

(b) $\langle d^\pi, \frac{1}{2}\mathbf{1}\rangle = \frac{1}{2}H = K$ for every policy, and $\langle d^\pi, \frac{\Delta}{4}\sum_k \theta_k z_k\rangle = \frac{\Delta}{4}\sum_k \theta_k t_k(\pi)$ by linearity. Each $t_k(\pi) = \langle d^\pi, z_k\rangle \in [-2,2]$ because $\|z_k\|_1 = 4$ and the occupancy puts total mass 1 on each of the two steps of block k; the value $t_k = 2\theta_k$ is achieved by the deterministic block choice (a,b) when $\theta_k = +1$ and (b,a) when $\theta_k = -1$, since by Proposition C.4(a) $\phi_k(a,b) = u_1^{(k)} + u_3^{(k)}$ and $\langle u_3^{(k)}, z_k\rangle = \frac{1}{2}\|z_k\|_2^2 = 2$. Maximizing each term separately (the blocks are unconstrained relative to each other) gives $V_\theta^\star = K + \Delta K/2$.

(c) By the expansion just used, $d^{\pi_\theta^\star} = \sum_k \big(u_1^{(k)} + \theta_k u_3^{(k)}\big) = w + \sum_k \theta_k u_3^{(k)}$, which lies in $\mathcal{U}_\mu$. From (a), for $0 < \varepsilon < 1$, $\Sigma_\mu^\dagger = K^{-1}\hat{w}\hat{w}^\top + (1-\varepsilon)^{-1}\sum_k u_2^{(k)}u_2^{(k)\top} + \varepsilon^{-1}\sum_k u_3^{(k)}u_3^{(k)\top}$; at $\varepsilon = 1$ the $u_2^{(k)}$ term is omitted. Since $\langle w, \hat{w}\rangle = \sqrt{K}$ and $\theta_k^2 = 1$,

$$\mathcal{C}_{\mathrm{out}}(\pi_\theta^\star;\mu) = \frac{(\sqrt{K})^2}{K} + \frac{1}{\varepsilon}\sum_{k=1}^K \theta_k^2 = 1 + \frac{K}{\varepsilon}.$$

For C_{std}: every step marginal of the behavior is uniform over $\{a,b\}$, so $d^\mu(h,s,a) = \frac{1}{2}$ for all (h,a), while any occupancy is at most 1; the ratio is at most 2, with equality for deterministic policies.

(d) By (b), $\mathrm{SubOpt}_\theta(\pi) = \frac{\Delta}{4}\sum_k \big(2 - \theta_k t_k(\pi)\big)$, and each summand lies in $[0,4]$ because $|t_k| \le 2$; this gives the upper bound ΔK. If $\hat{\theta}_k(\pi) \neq \theta_k$ then either $t_k > 0$ and $\theta_k = -1$, or $t_k \le 0$ and $\theta_k = +1$; in both cases $\theta_k t_k \le 0$ and the summand is at least 2. Hence $\mathrm{SubOpt}_\theta(\pi) \ge \frac{\Delta}{4}\cdot 2 \cdot \#\{k : \hat{\theta}_k \neq \theta_k\} = \frac{\Delta}{2}\rho_{\mathrm{H}}(\hat{\theta}(\pi),\theta)$. □

Corollary E.5 (Exact horizon–memory composition)**.** *In the K-block family, let $H = 2K$ and choose $\varepsilon = 1/(2\kappa)$ with $\kappa \ge 1$. The occupancy-equivalent Markovization has the same one-step marginals and*

$$\kappa_{\mathrm{mem}} = \kappa, \qquad \mathcal{C}_{\mathrm{out}}(\pi_\theta^\star;\mu) - 1 = H\kappa_{\mathrm{mem}}, \qquad C_{\mathrm{std}}(\pi_\theta^\star;\mu) = 2.$$

Under Gaussian outcome noise, whenever the sample-size condition of Theorem E.6 holds, every learner has a family member with

$$\mathbb{E}[\mathrm{SubOpt}] \ge \frac{\sigma H}{8\sqrt{2}}\sqrt{\frac{\kappa_{\mathrm{mem}}}{n}}.$$

Thus the marginal-to-trajectory comparison is jointly order-tight in horizon and logger memory on this temporal direct-sum family.

Proof. For $\varepsilon \le 1/2$, the repeat and crossing contrast eigenvalues in each block change from $1-\varepsilon$ and ε to $1/2$ under Markovization. Hence $\kappa_{\mathrm{mem}} = \max\{1, [2\varepsilon]^{-1}, [2(1-\varepsilon)]^{-1}\} = \kappa$. Proposition E.4(c) gives $\mathcal{C}_{\mathrm{out}} - 1 = K/\varepsilon = 2K\kappa = H\kappa$ and $C_{\mathrm{std}} = 2$. Substituting $d = 4K = 2H$ and $\mathcal{C}_{\mathrm{out}} - 1 = H\kappa_{\mathrm{mem}}$ in the lower bound of Theorem E.6 yields the displayed expression. □

The construction is designed so that the K blocks are *statistically independent but jointly scored*: a single outcome Y is the sum of all K block contributions plus noise, so the information about any one block is diluted, yet the losses add. This is precisely the situation Assouad's lemma is built for.

Theorem E.6 (Assouad lower bound; the dimension factor is necessary)**.** *Let the noise be Gaussian, $\xi \mid \tau \sim \mathcal{N}(0, \sigma^2)$. Fix $K \geq 1$, so $d = 4K$, and let $\mathcal{C} \geq 1 + K$ and*

$$n \;\geq\; \frac{\sigma^2(\mathcal{C}-1)}{4K}\,.$$

Then there is a family $\{I_\theta\}_{\theta\in\{-1,+1\}^K}$ of 2^K OO-RL instances, all sharing dynamics and behavior, each satisfying $\mathcal{C}_{\mathrm{out}}(\pi^\star_\theta;\mu) = \mathcal{C}$, $C_{\mathrm{std}}(\pi^\star_\theta;\mu) = 2$, and $C_{\mathrm{std}}(\pi;\mu) \leq 2$ for every policy, such that every (possibly randomized) algorithm $\hat\pi$ obeys

$$\max_\theta\; \mathbb{E}_\theta\big[\mathrm{SubOpt}_\theta(\hat\pi)\big] \;\geq\; \frac{\sigma}{16}\sqrt{\frac{d\,(\mathcal{C}-1)}{n}}.$$

Moreover, for the worst θ, $\mathrm{SubOpt}_\theta(\hat\pi) \geq \frac{\sigma}{32}\sqrt{d(\mathcal{C}-1)/n}$ with probability at least $1/16$.

Proof. Take the K-block family with

$$\varepsilon := \frac{K}{\mathcal{C}-1} \in (0,1], \qquad \Delta := \frac{\sigma}{\sqrt{n\varepsilon}} \;\leq\; 2,$$

where $\varepsilon \leq 1$ uses $\mathcal{C} \geq 1 + K$ and $\Delta \leq 2$ is exactly the assumed lower bound on n. By Proposition E.4(c), $\mathcal{C}_{\mathrm{out}}(\pi^\star_\theta;\mu) = 1 + K/\varepsilon = \mathcal{C}$ for every θ, and $C_{\mathrm{std}}(\pi^\star_\theta;\mu) = 2$.

Step 1: reduction to a hypercube estimation problem. Let $\hat\pi$ be any algorithm and define $\hat\theta := \hat\theta(\hat\pi) \in \{-1,+1\}^K$ by the decoder of Proposition E.4(d), a measurable function of the data (and of the algorithm's internal randomness). That proposition gives, for every θ, $\mathbb{E}_\theta[\mathrm{SubOpt}_\theta(\hat\pi)] \geq \frac{\Delta}{2}\mathbb{E}_\theta[\rho_{\mathrm{H}}(\hat\theta,\theta)]$.

Step 2: information between neighbors. Fix k and θ, and let $\theta' := \theta^{\oplus k}$. Both instances induce the same trajectory law q_μ (the behavior does not depend on the reward), and conditionally on τ the outcome is Gaussian with common variance σ^2 and means differing by

$$\langle \phi(\tau), r_\theta - r_{\theta'}\rangle = \frac{\Delta}{4}(\theta_k - \theta'_k)\langle \phi_k(\tau), z_k\rangle = \pm\frac{\Delta}{2}\langle \phi_k(\tau), z_k\rangle.$$

By Proposition C.4(a), $\langle \phi_k(\tau), z_k\rangle = \pm 2$ when block k is crossing (probability ε) and 0 otherwise, so the squared mean difference equals Δ^2 with probability ε and 0 otherwise. Using $\mathrm{KL}(\mathcal{N}(m,\sigma^2)\|\mathcal{N}(m',\sigma^2)) = (m-m')^2/(2\sigma^2)$, the chain rule for the pair (τ, Y), and tensorization over the n independent episodes,

$$\mathrm{KL}\big(\mathbb{P}_\theta^{\otimes n} \,\big\|\, \mathbb{P}_{\theta'}^{\otimes n}\big) = n\,\mathbb{E}_{\tau\sim q_\mu}\left[\frac{\langle \phi(\tau), r_\theta - r_{\theta'}\rangle^2}{2\sigma^2}\right] = \frac{n\,\varepsilon\,\Delta^2}{2\sigma^2} = \frac{1}{2},$$

by the choice of Δ. Pinsker's inequality Wainwright (2019) gives $\mathrm{TV}(\mathbb{P}_\theta^{\otimes n}, \mathbb{P}_{\theta'}^{\otimes n}) \leq \sqrt{\mathrm{KL}/2} = \frac{1}{2}$, uniformly over k and θ.

Step 3: Assouad. By Lemma L.6 (applied to the data distributions, with the algorithm's independent seed appended to the sample space),

$$\max_\theta \mathbb{E}_\theta\big[\rho_{\mathrm{H}}(\hat\theta,\theta)\big] \;\geq\; \frac{K}{2}\Big(1 - \max_{k,\theta}\mathrm{TV}\big(\mathbb{P}_\theta^{\otimes n}, \mathbb{P}_{\theta^{\oplus k}}^{\otimes n}\big)\Big) \;\geq\; \frac{K}{2}\cdot\frac{1}{2} = \frac{K}{4}.$$

Combining with Step 1, $\max_\theta \mathbb{E}_\theta[\mathrm{SubOpt}_\theta(\hat\pi)] \geq \Delta K/8$.

Step 4: substitution. Since $\varepsilon = K/(\mathcal{C}-1)$ and $d = 4K$,

$$\frac{\Delta K}{8} = \frac{K}{8}\cdot\frac{\sigma}{\sqrt{n\varepsilon}} = \frac{\sigma}{8}\sqrt{\frac{K(\mathcal{C}-1)}{n}} = \frac{\sigma}{8}\sqrt{\frac{d(\mathcal{C}-1)}{4n}} = \frac{\sigma}{16}\sqrt{\frac{d(\mathcal{C}-1)}{n}}.$$

Step 5: from expectation to constant probability. Let θ achieve the maximum and write $X := \mathrm{SubOpt}_\theta(\hat\pi)$, so $0 \leq X \leq M := \Delta K$ by Proposition E.4(d) and $\mathbb{E}X \geq M/8$. With $p := \mathbb{P}(X \geq M/16)$ we get $M/8 \leq \mathbb{E}X \leq (M/16)(1-p) + Mp \leq M/16 + pM$, whence $p \geq 1/16$; and $M/16 = \Delta K/16 = \frac{\sigma}{32}\sqrt{d(\mathcal{C}-1)/n}$. □

Corollary E.7 (Matching rate on the dense K-block family)**.** *Consider the K-block family of Theorem E.6, with Π the set of its 4^K deterministic policies (which contains $\pi^\star_\theta$), $\lambda \le \sigma^2/d$, and $n \ge \max\{n_0(\delta/2),\, \sigma^2(\mathcal{C}-1)/(4K)\}$. Then:*

(a) (Upper.) *Algorithm 1 with $\beta = \beta_\Pi(\delta/2)$ satisfies, with probability at least $1-\delta$,*

$$\mathrm{SubOpt}(\hat\pi) \;\le\; 4\Big(\sigma\sqrt{2\big(\tfrac{d}{4}\log 4 + \log(2/\delta)\big)} + \sigma\Big)\sqrt{\frac{\mathcal{C}}{n}} \;\le\; 9\,\sigma\sqrt{\frac{\big(d+\log(2/\delta)\big)\,\mathcal{C}}{n}}.$$

(b) (Lower.) *No algorithm can guarantee* $\mathrm{SubOpt} \le \frac{\sigma}{32}\sqrt{d(\mathcal{C}-1)/n}$ *with probability greater than* $15/16$ *uniformly over the family.*

Hence, for any fixed $\delta \in (0, 1/16)$ and $\mathcal{C} \ge 1 + d/4$, the upper and lower rates match up to absolute constants on this K-block family at order $\sigma\sqrt{d\,\mathcal{C}/n}$. In particular the dimension factor in the generic upper bound cannot be removed uniformly over all instances.

Proof. (a) Apply Theorem D.12 with $|\Pi| = 4^K = 4^{d/4}$, so $\log(2|\Pi|/\delta) = \frac{d}{4}\log 4 + \log(2/\delta)$, and $\sqrt{\lambda d} \le \sigma$ by the choice of λ; the numerical simplification uses $\frac{1}{2}\log 4 = \log 2 \le 0.7$ and $4(\sqrt{0.7d + 2\log(2/\delta)} + 1) \le 9\sqrt{d + \log(2/\delta)}$ for $d \ge 1$. Part (b) is the last statement of Theorem E.6. Finally, $\mathcal{C} \ge 1 + d/4$ implies $\mathcal{C} - 1 \ge \mathcal{C}/2$ for $\mathcal{C} \ge 2$, so the lower and upper displays have the same scaling on this family. □

E.2 A FIXED-HORIZON WORST-CASE CLASS

The temporal K-block family above makes the dimension factor visible, but it also couples d and H and necessarily has $\mathcal{C}_{\mathrm{out}} \gtrsim d$. The next construction removes both artifacts. It places the XOR blocks in *parallel states* rather than in consecutive time blocks, keeping $H = 2$ while letting d grow.

For $d = 4K$, let $\mathfrak{M}(d, \mathcal{C}, \sigma)$ be the class of all tabular OO-RL instances satisfying Assumptions 1–3 with horizon $H = 2$, tabular reward dimension d, conditionally Gaussian noise $\xi \mid \tau \sim \mathcal{N}(0, \sigma^2)$, and $\mathcal{C}^\star_{\mathrm{out}} \le \mathcal{C}$. The logger may be history dependent. Define its minimax expected suboptimality by

$$\mathfrak{R}_n(d, \mathcal{C}, \sigma) := \inf_{\widehat\pi} \sup_{I \in \mathfrak{M}(d,\mathcal{C},\sigma)} \mathbb{E}_I[\mathrm{SubOpt}_I(\widehat\pi)],$$

where the infimum is over all measurable offline learners.

Theorem E.8 (Uniform fixed-horizon minimax sandwich)**.** *Let $d = 4K$ for some integer $K \ge 2$, $\mathcal{C} \ge 2$, $\sigma > 0$, and $n \ge 2$. Then*

$$\min\left\{\frac{1}{4},\; \frac{\max\{1,\sigma\}}{16}\sqrt{\frac{d(\mathcal{C}-1)}{n}}\right\} \le \mathfrak{R}_n(d, \mathcal{C}, \sigma),$$

$$\mathfrak{R}_n(d, \mathcal{C}, \sigma) \le \min\left\{2,\; \left[4\sigma\sqrt{2(d\log 5 + \log n)} + \sqrt{2d}\right]\sqrt{\frac{\mathcal{C}}{n}} + \frac{2}{n}\right\}. \tag{15}$$

The Gaussian lower-bound family has $\mathcal{C}^\star_{\mathrm{out}} = \mathcal{C}$ and $C_{\mathrm{std}}(\pi;\mu) \le 2$ for every policy, with equality for deterministic policies; the noise-independent lower-bound family below has the same best-optimal-policy coefficient $\mathcal{C}^\star_{\mathrm{out}} = \mathcal{C}$ and benign one-step coverage. Consequently, if additionally $d \ge \max\{8, \log n\}$, then for every $\sigma > 0$,

$$\mathfrak{R}_n(d, \mathcal{C}, \sigma) = \Theta\left(\min\left\{1, \max\{1,\sigma\}\sqrt{\frac{d\mathcal{C}}{n}}\right\}\right)$$

with universal implicit constants.

Proof roadmap. The upper bound is Theorem D.8 at $H = 2$. The lower bound has two branches over the same class. In the Gaussian branch, K parallel XOR states carry independent sign bits and Assouad testing gives the σ term. In the exposure branch, hard states are rarely crossed, so an unexposed bit keeps a uniform posterior even at zero noise, giving the noise-independent term. Both constructions keep $\mathcal{C}^\star_{\mathrm{out}} = \mathcal{C}$ and benign one-step coverage.

Proof. Upper bound. For every instance in $\mathfrak{M}(d, \mathcal{C}, \sigma)$, Theorem D.8 with $H = 2$ gives

$$\mathbb{E}[\mathrm{SubOpt}(\hat{\pi})] \le \left[4\sigma\sqrt{2(d\log 5 + \log n)} + \sqrt{2d}\right]\sqrt{\frac{\mathcal{C}}{n}} + \frac{2}{n}$$

for Algorithm 2. The trivial reward range gives $\mathrm{SubOpt} \le H = 2$, so clipping the display at 2 proves the right-hand inequality.

Lower-bound family. Take $S = K$, $A = 2$, $H = 2$, initial distribution uniform on $\{1, \dots, K\}$, and the identity transition $P_1(s' \mid s, a) = \mathbf{1}\{s' = s\}$. Write the four coordinates of state k as $((1,k,a),(1,k,b),(2,k,a),(2,k,b))$ and define

$$z_k := e_{1,k,a} - e_{1,k,b} - e_{2,k,a} + e_{2,k,b}.$$

For a sign vector $\theta \in \{-1,+1\}^K$, set

$$r_\theta := \frac{1}{2}\mathbf{1} + \frac{\Delta}{4}\sum_{k=1}^{K}\theta_k z_k, \qquad \Delta := \min\left\{2,\ \sigma\sqrt{\frac{K}{n\varepsilon}}\right\}, \qquad \varepsilon := \frac{1}{\mathcal{C}-1} \in (0,1].$$

The disjoint supports of the z_k and $\Delta \le 2$ imply $r_\theta \in [0,1]^d$.

Conditioned on the initial state k, the logger draws the action pair (a,a) or (b,b) with probability $(1-\varepsilon)/2$ each and (a,b) or (b,a) with probability $\varepsilon/2$ each. Equivalently, it chooses the first action uniformly and at the second step repeats it with probability $1-\varepsilon$ and flips it with probability ε, so this is a valid history-dependent logging rule under the fixed transition law.

Let $u_1^{(k)}, u_2^{(k)}, u_3^{(k)}, u_4^{(k)}$ be the XOR basis of Proposition C.4 on state k's four coordinates. Because exactly one state block is active in an episode,

$$\Sigma_\mu = \frac{1}{K}\sum_{k=1}^{K}\left[u_1^{(k)}u_1^{(k)\top} + (1-\varepsilon)u_2^{(k)}u_2^{(k)\top} + \varepsilon u_3^{(k)}u_3^{(k)\top}\right].$$

The optimal Markov policy plays (a,b) in state k when $\theta_k = +1$ and (b,a) when $\theta_k = -1$, and hence

$$d^{\pi_\theta^\star} = \frac{1}{K}\sum_{k=1}^{K}\left(u_1^{(k)} + \theta_k u_3^{(k)}\right).$$

Taking the blockwise pseudoinverse therefore gives

$$\mathcal{C}_{\mathrm{out}}(\pi_\theta^\star;\mu) = \sum_{k=1}^{K}\frac{1}{K}\left(1 + \frac{1}{\varepsilon}\right) = 1 + \frac{1}{\varepsilon} = \mathcal{C}.$$

At either step the behavior occupancy of every (s,a) pair is $1/(2K)$, whereas any target occupancy is at most $1/K$; thus $C_{\mathrm{std}}(\pi;\mu) \le 2$ for every policy and deterministic policies attain 2.

For an arbitrary target policy define $t_k(\pi) := K\langle d^\pi, z_k\rangle \in [-2,2]$ and decode $\widehat{\theta}_k(\pi) := \operatorname{sign} t_k(\pi)$, with $\operatorname{sign} 0 := +1$. Directly from the reward definition,

$$V_\theta(\pi) = 1 + \frac{\Delta}{4K}\sum_{k=1}^{K}\theta_k t_k(\pi), \qquad V_\theta^\star = 1 + \frac{\Delta}{2},$$

and therefore

$$\mathrm{SubOpt}_\theta(\pi) \ge \frac{\Delta}{2K}\rho_{\mathrm{H}}(\widehat{\theta}(\pi),\theta), \qquad 0 \le \mathrm{SubOpt}_\theta(\pi) \le \Delta. \tag{16}$$

Fix neighboring parameters θ and $\theta' = \theta^{\oplus k}$. Their trajectory laws are identical. The conditional outcome means differ only when the initial state is k and the logger chooses a crossing pair; this event has probability ε/K, and on it the absolute mean difference is Δ. Thus the n-sample KL divergence is

$$\mathrm{KL}(\mathbb{P}_\theta^{\otimes n} \| \mathbb{P}_{\theta'}^{\otimes n}) = \frac{n\varepsilon\Delta^2}{2K\sigma^2} \le \frac{1}{2}.$$

The inequality is equality when $\Delta = \sigma\sqrt{K/(n\varepsilon)}$; if the minimum defining Δ selects 2, then $n\varepsilon/(K\sigma^2) < 1/4$, which gives the same conclusion. Pinsker and Lemma L.6 imply $\max_\theta \mathbb{E}_\theta[\rho_{\mathrm{H}}(\widehat{\theta}, \theta)] \geq K/4$ for every learner. Combining with (16),

$$\max_\theta \mathbb{E}_\theta[\mathrm{SubOpt}_\theta] \geq \frac{\Delta}{8} = \min\left\{\frac{1}{4},\ \frac{\sigma}{16}\sqrt{\frac{d(\mathcal{C}-1)}{n}}\right\},$$

which proves the left-hand inequality because the constructed family is a subset of $\mathfrak{M}(d, \mathcal{C}, \sigma)$.

Noise-independent exposure lower bound. It remains to show that random exposure is hard even when Gaussian noise is arbitrarily small. Let $L := K - 1$. Use the same K persistent states, but make states $1, \ldots, L$ hard and state K reward-neutral. Give each hard state initial mass q and the neutral state mass $q_0 := 1 - Lq$. On hard state k, use reward $\frac{1}{2}\mathbf{1} + (\Delta/4)\theta_k z_k$ and the same history-dependent logger, now with crossing probability ε; on the neutral state use constant reward $1/2$ and the independent-uniform logger. An optimal Markov target crosses according to θ_k on hard states and may copy the uniform behavior marginal on the neutral state. The blockwise pseudoinverse gives

$$\mathcal{C}^\star_{\mathrm{out}} = q_0 + Lq\left(1 + \frac{1}{\varepsilon}\right) = 1 + \frac{Lq}{\varepsilon}. \tag{17}$$

This is the minimum $\mathcal{C}_{\mathrm{out}}$ over the optimal-policy tie on the neutral state: if its probabilities of action a at the two steps are x, y, that block contributes

$$q_0\left[1 + 2(x + y - 1)^2 + 2(x - y)^2\right] \geq q_0,$$

with equality at $x = y = 1/2$ (and the zero-mass case $q_0 = 0$ is vacuous).

Put the uniform product prior on $\theta \in \{-1, +1\}^L$. If the dataset contains no crossing trajectory from hard state k, then every observed conditional mean is invariant to θ_k, so its posterior remains uniform even after conditioning on the complete dataset. For an arbitrary stochastic Markov output, let $t_k \in [-2, 2]$ be twice the difference between its action-a probabilities at the two steps. Its local regret is $(q\Delta/4)(2 - \theta_k t_k)$, hence conditional Bayes regret on an unexposed bit is exactly $q\Delta/2$. By linearity of expectation,

$$\mathfrak{R}_n(d, \mathcal{C}, \sigma) \geq \frac{\Delta}{2}\, Lq(1 - q\varepsilon)^n. \tag{18}$$

Set $\Delta = 2$ and enforce $Lq/\varepsilon = \mathcal{C} - 1$. If $n \leq L(\mathcal{C} - 1)/2$, choose $q = 1/L$ and $\varepsilon = 1/(\mathcal{C} - 1)$; then Bernoulli's inequality gives $(1 - q\varepsilon)^n \geq 1/2$ and (18) is at least $1/2$. Otherwise choose

$$\varepsilon = \sqrt{\frac{L}{2n(\mathcal{C} - 1)}}, \qquad q = \frac{(\mathcal{C} - 1)\varepsilon}{L},$$

so $q\varepsilon = 1/(2n)$, $q_0 \geq 0$, and $(1 - 1/(2n))^n \geq 1/2$. Since $L = K - 1 \geq K/2 = d/8$ for $K \geq 2$,

$$\mathfrak{R}_n(d, \mathcal{C}, \sigma) \geq \min\left\{\frac{1}{2}, \frac{1}{8}\sqrt{\frac{d(\mathcal{C} - 1)}{n}}\right\}.$$

Combining this with the Gaussian bound already proved yields the lower side of (15).

Finally, $\mathcal{C} - 1 \geq \mathcal{C}/2$ for $\mathcal{C} \geq 2$. If $d \geq \max\{8, \log n\}$, the upper bracket in (15) is $O(\max\{1, \sigma\}\sqrt{d})$ and $2/n$ is absorbed by $O(\sqrt{d\mathcal{C}/n})$. Constant clipping on both sides yields the stated all-noise Θ rate. □

Remark E.2 (Relation to policy-generated trajectory supervision). Theorem E.8 is a minimax statement for the general-logging class defined in this paper and is parameterized by the exact outcome geometry $\mathcal{C}_{\mathrm{out}}$. It complements trajectory-supervision minimax theory in the policy-generated setting, where Ren and Xie (2026) prove matching offline upper and lower bounds under state–action coverage and function-class assumptions. The distinction here is that the logger itself may carry cross-time dependence invisible to one-step occupancies. The missing-exposure mechanism parallels established results: related $\sqrt{S(C-1)/N}$ missing-mass behavior appears in offline contextual-bandit lower bounds (Rashidinejad et al., 2021), and random-design minimax effects are classical (Pathak et al., 2024). Here the mechanism completes the all-noise characterization for the exact $\mathcal{C}_{\mathrm{out}}$ general-logging class studied in this paper.

Remark E.3 (Why the parallel-state construction is needed). The temporal construction of Theorem E.6 needs $\mathcal{C} \geq 1 + K = 1 + d/4$: every one of its K consecutive blocks must carry a target crossing direction. Theorem E.8 shows that this coupling is not a property of the full OO-RL class. By placing the same hard bits in parallel states it keeps $H = 2$ and permits any $\mathcal{C} \geq 2$, while retaining the $\sqrt{d}$ worst-case lower bound. Section E.4 then makes the complementary point: a structured subclass with only s active directions can be much easier than this worst case, so the resulting $\sqrt{d}$ dependence is therefore worst-case: Section E.4 gives structured subclasses with only s active directions whose rates depend on s rather than ambient d.

E.3 General-even-horizon lower bounds and near-minimax characterization

The fixed-$H = 2$ theorem isolates dimension cleanly. Two questions arise as the horizon grows: whether low-noise hardness itself grows with H at fixed $\mathcal{C}_{\mathrm{out}}$, and whether the $\sqrt{H}$ term in the generic projected upper bound is necessary in the unrestricted tabular class.

A two-state temporal-gating family. Let $H = 2K$, $S = 2$, and $A = 2$. At the start of each two-step block, known action-independent dynamics draw a hard state with probability q and a reward-neutral state otherwise, and keep that state for the block. On the hard state the logger uses an ε-XOR action pair; on the neutral state it uses independent uniform actions. Blocks are independent across time. Put one sign $\theta_k \in \{-1, +1\}$ on the reward contrast in hard block k, and let the designated optimal policy cross according to that sign while matching the uniform logger on neutral states. The same blockwise diagonalization used above gives

$$\mathcal{C}_{\mathrm{out}}(\pi^\star; \mu) = 1 + \frac{Kq}{\varepsilon}, \qquad C_{\mathrm{std}}(\pi; \mu) \leq 2. \tag{19}$$

Thus any desired $C > 1$ is obtained by enforcing $q/\varepsilon = (C-1)/K$.

Theorem E.9 (General-even-H all-noise temporal lower bound)**.** *Let $H = 2K \geq 2$, $C > 1$, $n \geq 1$, and let the outcome noise be Gaussian with any $\sigma \geq 0$. Over the temporal-gating subclass above with $\mathcal{C}_{\mathrm{out}}(\pi^\star; \mu) \leq C$, every offline learner satisfies*

$$\sup_I \mathbb{E}_I[\mathrm{SubOpt}_I(\widehat{\pi})] \geq \frac{1}{16} \min\left\{ H, C-1, \max\{1, \sigma\}\sqrt{\frac{H(C-1)}{n}} \right\}.$$

All members used in the construction have benign one-step coverage $C_{\mathrm{std}} \leq 2$.

Proof. Write $a = (C-1)/K$, so $q/\varepsilon = a$. For the noise-independent calibration, take reward gap $\Delta = 2$ and choose

$$(\varepsilon, q) = \begin{cases} (1/2, a/2), & a < 2/n, \\ ((2na)^{-1/2}, a(2na)^{-1/2}), & 2/n \leq a \leq 2n, \\ (1/a, 1), & a > 2n. \end{cases}$$

Equation (19) gives $\mathcal{C}_{\mathrm{out}} = C$. A fixed hard sign is exposed in one episode only when its block is hard and crossing, an event of probability $q\varepsilon$. Conditional on no such crossing in the n episodes, the complete data law is identical under the two values of that sign, even when $\sigma = 0$. Under the uniform sign prior, the conditional Bayes regret per unseen sign is $q\Delta/2$. Summing over the K signs gives

$$\sup_I \mathbb{E}_I[\mathrm{SubOpt}_I] \geq Kq(1 - q\varepsilon)^n \geq \frac{1}{4} \min\left\{ K, C-1, \sqrt{\frac{K(C-1)}{n}} \right\},$$

where the final inequality follows by the three displayed calibrations. Since $H = 2K$, this is at least $\frac{1}{8}\min\{H, C-1, \sqrt{H(C-1)/n}\}$.

For the Gaussian branch choose the largest hard-state mass consistent with $\varepsilon \leq 1/2$: if $a \leq 2$, set $\varepsilon = 1/2$ and $q = a/2$; if $a > 2$, set $q = 1$ and $\varepsilon = 1/a$. Put $p = q\varepsilon$ and $\Delta_G = \min\{2, \sigma/\sqrt{np}\}$. Neighboring sign vectors then have n-sample KL divergence $np\Delta_G^2/(2\sigma^2) \leq 1/2$. The same policy-to-sign reduction and Assouad argument as in Theorem E.6 gives

$$\sup_I \mathbb{E}_I[\mathrm{SubOpt}_I] \geq \frac{1}{8} \min\left\{ K, C-1, \sigma\sqrt{\frac{K(C-1)}{n}} \right\} \geq \frac{1}{16} \min\left\{ H, C-1, \sigma\sqrt{\frac{H(C-1)}{n}} \right\}.$$

The two calibrations are different members of the same minimax subclass, so the worst-case risk is at least their maximum. Combining the two displays proves the theorem. □

Parallel states times temporal contrasts. The temporal theorem fixes $d = 4H$ and therefore does not by itself saturate the full low-noise term $\sqrt{dHC/n}$. To expose the missing dimension factor, let $M \geq 1$ hard persistent states and one neutral persistent state be chosen at the beginning of an episode, take $A = 2$, and retain $H = 2K$. Then $d = 2H(M + 1)$.

For the noisy family, a hard state carries K independent ε-XOR blocks and independent signs $\theta_{j,k}$. If the total hard-state mass is q, blockwise pseudoinversion gives $\mathcal{C}_{\mathrm{out}}(\pi^\star; \mu) = 1 + qK/\varepsilon$. With $c = C-1$ set $a = c/K$, $q = \min\{1, a/2\}$, $\varepsilon = q/a$, and let $p = (q/M)\varepsilon$ be the exposure probability of one hard bit. The Gaussian choice $\Delta_G = \min\{2, \sigma/\sqrt{np}\}$ keeps neighboring KL at most $1/2$ and yields

$$\sup_I \mathbb{E}_I[\mathrm{SubOpt}_I] \geq \frac{1}{8} \min\left\{K, c, \sigma\sqrt{\frac{KMc}{n}}\right\}. \tag{20}$$

For the low-noise family, each hard state s has one sign $\theta_s \in \{-1, +1\}$ across the horizon. Put $c_h = +1$ for $h \leq H/2$ and $c_h = -1$ afterward. For $0 < \Delta \leq 1$, set

$$r(h, s, a) = \tfrac{1}{2} + \tfrac{\Delta\theta_s c_h}{2}, \qquad r(h, s, b) = \tfrac{1}{2} - \tfrac{\Delta\theta_s c_h}{2}.$$

Every reward coordinate lies in $[0, 1]$. Conditional on s, the four whole-horizon action templates are $x_0 = (a, \ldots, a)$, $x_1 = (b, \ldots, b)$, x_+ choosing a in the first $H/2$ stages and b thereafter, and its complement x_-. Their logging probabilities are $(1 - \varepsilon)/2, (1 - \varepsilon)/2, \varepsilon/2, \varepsilon/2$. The conditional returns are

$$G(x_0) = G(x_1) = H/2, \qquad G(x_\pm) = H/2 \pm H\Delta\theta_s/2.$$

Thus x_0, x_1 contain no information about the sign, whereas the balanced policies differ by $H\Delta$; the lower-bound calibration takes $\Delta = 1$. With

$$q = \min\left\{1, \frac{c}{2}, \sqrt{\frac{cM}{2n}}\right\}, \qquad \varepsilon = \frac{q}{c},$$

the target has $\mathcal{C}_{\mathrm{out}} = 1 + q/\varepsilon = C$, each hard state is informative with probability $p = q\varepsilon/M \leq 1/(2n)$, and $(1 - p)^n \geq 1/2$. If neither informative template from state s is observed, the posterior sign remains uniform. Any stochastic output policy has conditional Bayes regret $(q/M)H\Delta/2$ at that state. Summing over states gives $qH\Delta(1 - p)^n/2$ and hence

$$\sup_I \mathbb{E}_I[\mathrm{SubOpt}_I] \geq \frac{H}{8} \min\left\{1, c, \sqrt{\frac{Mc}{n}}\right\}. \tag{21}$$

Both families have exact best-optimal-policy coefficient C and $C_{\mathrm{std}} = 2$.

Theorem E.10 (General-even-horizon full-class near-minimax sandwich)**.** *Let $H = 2K$, $M \geq 1$, $d = 2H(M + 1)$, $C \geq 2$, $c = C - 1$, and $\sigma \geq 0$. Let $\mathfrak{R}_{n,H}(d, C, \sigma)$ denote the minimax expected suboptimality over all tabular instances satisfying the core assumptions with these H, d, Gaussian outcome noise of variance σ^2, general valid history-dependent logging, and $\mathcal{C}^\star_{\mathrm{out}} \leq C$. If*

$$\sqrt{\frac{Mc}{n}} \leq \min\{1, c\}, \qquad \sigma\sqrt{\frac{KMc}{n}} \leq \min\{K, c\},$$

then

$$\frac{1}{16\sqrt{2}} \max\{\sigma, \sqrt{H}\}\sqrt{\frac{d(C-1)}{n}} \leq \mathfrak{R}_{n,H}(d, C, \sigma) \leq \left[4\sigma\sqrt{2(d\log 5 + \log n)} + \sqrt{dH}\right]\sqrt{\frac{C}{n}} + \frac{H}{n}. \tag{22}$$

Consequently, if $d \geq \log n$, then throughout this non-clipped regime

$$\mathfrak{R}_{n,H}(d, C, \sigma) = \Theta\left(\max\{\sigma, \sqrt{H}\}\sqrt{\frac{dC}{n}}\right)$$

with universal implicit constants.

Proof. The upper bound is Theorem D.8, since every instance in the class has an optimal policy with $\mathcal{C}_{\mathrm{out}} \le C$.

For the lower bound, the two parallel constructions above are members of the same minimax class but need not use the same logger. Their hard-coordinate dimension is $d_{\mathrm{hard}} = 4KM = 2HM$, and $d_{\mathrm{hard}}/d = M/(M+1) \ge 1/2$. Under the first non-clipping condition, (21) reduces to

$$\mathfrak{R}_{n,H} \ge \frac{H}{8}\sqrt{\frac{Mc}{n}} = \frac{1}{8\sqrt{2}}\sqrt{\frac{Hd_{\mathrm{hard}}c}{n}}.$$

Under the second, (20) gives

$$\mathfrak{R}_{n,H} \ge \frac{\sigma}{8}\sqrt{\frac{KMc}{n}} = \frac{\sigma}{16}\sqrt{\frac{d_{\mathrm{hard}}c}{n}}.$$

Taking the larger of the two class lower bounds and replacing d_{hard} by at least $d/2$ gives the conservative common constant $1/(16\sqrt{2})$ in (22).

If $d \ge \log n$, then $d\log 5 + \log n = O(d)$, so the upper bound is $O((\sigma + \sqrt{H})\sqrt{dC/n})$; moreover H/n is absorbed by the low-noise core because $d = 2H(M+1)$ and $C \ge 2$. Finally, $\sigma + \sqrt{H} \le 2\max\{\sigma, \sqrt{H}\}$ and $C - 1 \ge C/2$, which proves the rate statement. The restriction $d \ge \log n$ is used only to absorb the additive $\log n$ term in the upper-bound confidence radius into d. When $d < \log n$, the displayed lower bound remains valid while the current upper-bound analysis leaves an additional logarithmic gap in the noise-dependent term. Closing this gap remains open. □

Remark E.4 (Construction). Theorem E.10 combines classical statistical primitives—Assouad testing, missing-mass and random-design arguments—in a feasible MDP product construction. Under outcome-only feedback and a valid history-dependent logger, this construction makes the noise mechanism and the horizon-coherent exposure mechanism account simultaneously for the two scales in the projected $\mathcal{C}_{\mathrm{out}}$ upper bound. Theorem E.8 is the clean $H = 2$ full-class specialization.

E.4 A SPARSE FAMILY WHERE AMBIENT DIMENSION IS NOT THE RATE DRIVER

The (K, s)-sparse XOR instance. Fix $1 \le s \le K$, $\varepsilon \in (0, 1]$, $\Delta \in (0, 2]$, an arbitrary support $S \subseteq [K]$ with $|S| = s$, and signs $\theta \in \{-1, +1\}^S$. Take the K-block geometry of Section E.1 verbatim — $H = 2K$, $S_{\mathrm{MDP}} = 1$, $A = 2$, $d = 4K$, with the same block-independent behavior — but let only the blocks in S carry a gap:

$$r_{S,\theta} := \tfrac{1}{2}\mathbf{1} + \frac{\Delta}{4}\sum_{k\in S}\theta_k z_k \in [0,1]^d.$$

Blocks $k \notin S$ are *exactly reward-neutral*: every action sequence there earns exactly 1. The point of the family is that K, and hence the ambient dimension $d = 4K$, may be increased without limit while the decision-relevant reward variation remains confined to s blocks. The support S is fixed but need not be revealed to the learner.

A quick reading suggests that padding with irrelevant coordinates must be free. Two distinctions matter. First, the full-tabular coverage coefficient depends on *which* optimal policy one measures: a maximizer acting deterministically on neutral blocks has $\mathcal{C}_{\mathrm{out}}$ growing with d, whereas an equally optimal representative that plays the behavior marginal there has bounded $\mathcal{C}_{\mathrm{out}}$. Second, the generic ridge burn-in depends on the ambient design rank. The fixed-support reduction below removes that burn-in by passing to an s-dimensional sufficient statistic, while Proposition E.13 shows that revealing the support is not actually necessary for *policy regret* on this exact neutral-padded family.

Proposition E.11 (Closed forms for the sparse family)**.** *With the notation of Proposition E.4:*

(a) $V_{S,\theta}(\pi) = K + \frac{\Delta}{4}\sum_{k\in S}\theta_k t_k(\pi)$ and $V^\star_{S,\theta} = K + \Delta s/2$; the maximum is attained by any policy that crosses with sign θ_k on each active block, whatever it does on the neutral blocks.

(b) Among those maximizers, the representative that plays the behavior marginal (uniform) on every neutral block attains

$$\mathcal{C}^{\mathrm{rep}}_{\mathrm{out}} = 1 + \frac{s}{\varepsilon},$$

independently of K and hence of d. For $0 < \varepsilon < 1$, a maximizer choosing a repeated action pair on every neutral block has $\mathcal{C}_{\mathrm{out}} = 1 + s/\varepsilon + (K - s)/(1 - \varepsilon)$. Other deterministic neutral-block choices need not have this value: a crossing pair contributes $1/\varepsilon$, not $1/(1 - \varepsilon)$. At $\varepsilon = 1$, any nonzero neutral u_2 component lies outside $\mathcal{U}_\mu$, so the corresponding $\mathcal{C}_{\mathrm{out}}$ is infinite. Every policy satisfies $C_{\mathrm{std}} \le 2$, with equality for fully deterministic policies.

(c) $\mathrm{SubOpt}_{S,\theta}(\pi) \ge \frac{\Delta}{2}\rho_{\mathrm{H}}(\hat{\theta}_S(\pi), \theta)$ *for the decoder of Proposition E.4(d) restricted to S, and* $\mathrm{SubOpt}_{S,\theta}(\pi) \le \Delta s$.

(d) (Centered sufficient statistic for a fixed support.) *For every block put $\zeta_k(\tau) = \frac{1}{2}\langle \phi_k(\tau), z_k \rangle \in \{0, \pm 1\}$ and let $\zeta_S = (\zeta_k)_{k \in S} \in \mathbb{R}^s$. The common offset $K = H/2$ is known, and*

$$Y - K = \langle \zeta_S(\tau), \vartheta \rangle + \xi, \qquad \vartheta_k = \tfrac{\Delta}{2}\theta_k, \quad k \in S.$$

For the policy π_g crossing with signs $g \in \{-1, +1\}^S$ on active blocks and playing uniformly elsewhere, $V_{S,\theta}(\pi_g) - K = \langle g, \vartheta \rangle$. Moreover $\mathbb{E}_\mu[\zeta_S \zeta_S^\top] = \varepsilon I_s$, $\|\zeta_S\|^2 \le s$, $\|g\|^2_{(\varepsilon I_s)^{-1}} = s/\varepsilon = \mathcal{C}_{\mathrm{out}}^{\mathrm{rep}} - 1$, and $\|\vartheta\| \le \sqrt{s}$ because $\Delta \le 2$.

Proof. Parts (a) and (c) are the proof of Proposition E.4(b),(d) with the reward sum restricted to $k \in S$.

For (b), the uniform-on-neutral maximizer has occupancy $d^\star = w + \sum_{k \in S} \theta_k u_3^{(k)}$. Applying the pseudoinverse eigen-decomposition of Proposition E.4 gives

$$\mathcal{C}_{\mathrm{out}}(d^\star) = \frac{\langle w, \hat{w} \rangle^2}{K} + \frac{1}{\varepsilon} \sum_{k \in S} \theta_k^2 = 1 + \frac{s}{\varepsilon}.$$

For $0 < \varepsilon < 1$, each neutral block with a repeated deterministic pair adds one $u_2^{(k)}$ component and hence $(1-\varepsilon)^{-1}$ to the quadratic form; a crossing deterministic pair instead adds one $u_3^{(k)}$ component and contributes ε^{-1}. When $\varepsilon = 1$, the u_2 direction lies in the nullspace rather than the range, so its $\mathcal{C}_{\mathrm{out}}$ is infinite. Every behavior action marginal is uniform, giving $C_{\mathrm{std}} \le 2$.

For (d), $\langle \phi(\tau), r_{S,\theta} \rangle = K + (\Delta/2) \sum_{k \in S} \theta_k \zeta_k(\tau)$. Under the block-independent behavior, each ζ_k equals $+1$ and -1 with probability $\varepsilon/2$ each and 0 otherwise, independently across blocks; the covariance, target leverage, and norm claims follow. □

Theorem E.12 (Fixed-support matching rate on the sparse family)**.** *Let the noise be Gaussian. Fix $s \ge 1$, $\mathcal{C} \ge 1 + s$, any $K \ge s$, and any support $S \subseteq [K]$ with $|S| = s$. Let $n \ge \sigma^2(\mathcal{C} - 1)/(4s)$, take $\varepsilon = s/(\mathcal{C} - 1)$, and set $\Delta = \sigma/\sqrt{n\varepsilon} \le 2$. The resulting 2^s-member fixed-support subfamily has $\mathcal{C}_{\mathrm{out}}^{\mathrm{rep}} = \mathcal{C}$ and $C_{\mathrm{std}} \le 2$. Then:*

(a) Every algorithm obeys

$$\max_\theta \mathbb{E}_\theta[\mathrm{SubOpt}_\theta(\hat{\pi})] \ge \frac{\sigma}{8}\sqrt{\frac{s(\mathcal{C} - 1)}{n}},$$

and for the worst θ the suboptimality is at least $\frac{\sigma}{16}\sqrt{s(\mathcal{C} - 1)/n}$ with probability at least $1/16$.

(b) If the learner is given S, apply pessimistic ridge regression to the centered observations $Y - K$ and the s-dimensional features ζ_S from Proposition E.11(d), over $\Pi_S = \{\pi_g : g \in \{-1, +1\}^S\}$. If $\lambda \le \sigma^2/s$, then with probability at least $1 - \delta$,

$$\mathrm{SubOpt}(\hat{\pi}) \le 9\sigma\sqrt{\frac{(s + 2\log(2/\delta))\mathcal{C}}{n}},$$

provided $n \ge n_0(\delta/2)$, where one valid burn-in is

$$n_0(\delta) = \frac{8s}{\varepsilon}\left(s \log \frac{24s}{\varepsilon} + \log \frac{1}{\delta}\right).$$

In particular, this burn-in has no dependence on K or d.

Thus, when the support is revealed, the upper and lower local rates match up to absolute constants for every $d \geq 4s$.

Proof. Part (a) is the Assouad proof of Theorem E.6 restricted to the s active blocks. Neighboring signs differ in conditional outcome mean by Δ on an event of probability ε, so their n-sample KL divergence is $n\varepsilon\Delta^2/(2\sigma^2) = 1/2$. Pinsker and Assouad give expected Hamming error at least $s/4$; part (c) of Proposition E.11 converts this to $\Delta s/8$, yielding the displayed lower bound. The constant-probability form follows as in Theorem E.6 from $0 \leq \mathrm{SubOpt} \leq \Delta s$.

For (b), Proposition E.11(d) gives an s-dimensional linear model with $\|\zeta_S\|^2 \leq s$, $\mathbb{E}[\zeta_S \zeta_S^\top] = \varepsilon I_s$, and $\|\vartheta\| \leq \sqrt{s}$. Repeating the finite-policy proof of Theorem D.12 in these coordinates yields

$$\beta_{\Pi_S}(\delta/2) \leq \sigma\sqrt{2(s\log 2 + \log(4/\delta))} + \sqrt{\lambda s} \leq \sigma\sqrt{2(s\log 2 + \log(4/\delta))} + \sigma.$$

The design argument uses feature-square bound s and smallest positive eigenvalue ε, giving the displayed burn-in. The centered target leverage is $s/\varepsilon = \mathcal{C} - 1 \leq \mathcal{C}$, so standard pessimism chaining gives $\mathrm{SubOpt} \leq 4\beta_{\Pi_S}(\delta/2)\sqrt{\mathcal{C}/n}$; the numerical simplification gives 9. □

Proposition E.13 (Unknown support is free for policy regret on the neutral-padded family)**.** *For every block define*

$$T_k := \sum_{i=1}^{n} \zeta_{ik}(Y_i - K), \qquad \widehat{g}_k := \mathrm{sign}_+(T_k),$$

where $\mathrm{sign}_+(x) = +1$ *for* $x \geq 0$ *and* -1 *otherwise, and return the deterministic policy that crosses every block with sign* $\widehat{g}_k$*. The decoder is not given the active support,* s*, or* Δ*. Define*

$$\tau^2 := \sigma^2 + \frac{(s-1)\Delta^2}{4}, \qquad q_n := \left(1 - \varepsilon + \varepsilon \exp\left\{-\frac{\Delta^2}{8\tau^2}\right\}\right)^n,$$

with the natural convention that the exponential term is 0 *when* $\tau = 0$*. Under Assumptions 1–2, uniformly over all* $K \geq s$*, all supports* $S \subseteq [K]$ *with* $|S| = s$*, and all sign vectors,*

$$\mathbb{P}(\widehat{g}_k \neq \theta_k) \leq q_n, \qquad k \in S,$$

and

$$\mathbb{E}[\mathrm{SubOpt}(\widehat{\pi})] \leq s\Delta q_n, \qquad \mathbb{P}(\mathrm{SubOpt}(\widehat{\pi}) > 0) \leq s q_n.$$

Moreover,

$$q_n \leq \exp\left\{-\frac{n\varepsilon\Delta^2}{8\sigma^2 + (2s-1)\Delta^2}\right\}.$$

Consequently, for any $\delta \in (0,1)$ *it is sufficient that*

$$n \geq \frac{8\sigma^2 + (2s-1)\Delta^2}{\varepsilon\Delta^2} \log\frac{s}{\delta}$$

to return an optimal-value policy with probability at least $1 - \delta$*, with no ambient-*K *term.*

If additionally $\xi \sim N(0, \sigma^2)$ *with* $\sigma > 0$*, then*

$$\inf_{\widehat{\pi}} \sup_{\substack{S \subseteq [K],\, |S| = s \\ \theta \in \{-1,+1\}^S}} \mathbb{E}_{S,\theta}[\mathrm{SubOpt}(\widehat{\pi})] \geq \frac{s\Delta}{8} \exp\left\{-\frac{n\varepsilon\Delta^2}{2\sigma^2}\right\}.$$

Hence whenever $s\Delta^2 = O(\sigma^2)$*, the fixed-gap upper and lower bounds match the dependence*

$$s\Delta \exp\{-\Theta(n\varepsilon\Delta^2/\sigma^2)\}$$

uniformly in the ambient K*.*

Proof. Fix an active block $k \in S$ and put $I_i = \mathbf{1}\{|\zeta_{ik}| = 1\}$ and $M_k = \sum_i I_i$. Under the logger, $M_k \sim \mathrm{Binomial}(n, \varepsilon)$. On an informative episode, multiply the centered outcome by the observed crossing sign:

$$X_i := \zeta_{ik}(Y_i - K) = \frac{\Delta}{2}\theta_k + \frac{\Delta}{2} \sum_{\ell \in S \setminus \{k\}} \theta_\ell \zeta_{ik}\zeta_{i\ell} + \zeta_{ik}\xi_i.$$

Conditional on the full trajectory, $\zeta_{ik}\xi_i$ remains mean-zero and σ-subgaussian. Conditional only on $I_i = 1$, the other block variables are independent of block k, symmetric and bounded by one. Hoeffding's lemma therefore makes the sum of the $s-1$ cross terms subgaussian with variance proxy $(s-1)\Delta^2/4$. Thus $X_i - \Delta\theta_k/2$ is τ-subgaussian.

Conditional on $M_k = m$, the one-sided subgaussian tail gives

$$\mathbb{P}(\widehat{g}_k \neq \theta_k \mid M_k = m) \leq \exp\left\{-\frac{m\Delta^2}{8\tau^2}\right\}.$$

Averaging over the binomial crossing count yields exactly the displayed q_n upper bound. For the support-agnostic crossing policy,

$$\mathrm{SubOpt}(\widehat{\pi}) = \Delta \sum_{k\in S} \mathbf{1}\{\widehat{g}_k \neq \theta_k\};$$

decisions on S^c have zero value cost. Linearity of expectation and a union bound over only the s active coordinates give the two regret statements. Finally, $1 - e^{-a} \geq a/(1+a)$ with $a = \Delta^2/(8\tau^2)$ gives the simpler exponential bound and the high-probability sample threshold.

For the Gaussian lower bound, restrict the enlarged class to any fixed support. Pair two sign vectors differing only at active coordinate k. The logger law is identical under the pair, while the conditional Gaussian outcome means differ by $\Delta\zeta_k$; hence their n-sample KL divergence is $n\varepsilon\Delta^2/(2\sigma^2)$. Bretagnolle–Huber gives pair-averaged bit error at least one quarter of the exponential of the negative KL. The usual Assouad pairing therefore gives expected Hamming loss at least $(s/4)\exp\{-n\varepsilon\Delta^2/(2\sigma^2)\}$. Proposition E.11(c) converts this to the stated policy-regret lower bound. Since the unknown-support class contains this fixed-support subfamily, the same lower bound applies. □

Corollary E.14 (Ambient neutral padding can be free for policy regret)**.** *Under the same reward-box and sample-size conditions as Theorem E.12, set $\varepsilon = s/(\mathcal{C}-1)$ and $\Delta = \sigma/\sqrt{n\varepsilon}$. For the enlarged class in which the support S is unknown,*

$$\frac{\sigma}{8}\sqrt{\frac{s(\mathcal{C}-1)}{n}} \leq \inf_{\widehat{\pi}} \sup_{\substack{S\subseteq[K],\,|S|=s \\ \theta\in\{-1,+1\}^S}} \mathbb{E}_{S,\theta}[\mathrm{SubOpt}(\widehat{\pi})] \leq \sigma\sqrt{\frac{s(\mathcal{C}-1)}{n}},$$

uniformly for every $K \geq s$. Thus unknown support introduces no $\log K$ factor in policy regret on this neutral-padded family. The local upper order alone is range-trivial at this calibration; the substantive additional statement is the fixed-gap exponential error/exact-optimality characterization of Proposition E.13.

Proof. The lower bound restricts the enlarged class to any one fixed support and applies Theorem E.12(a). For the upper bound, the support-agnostic policy of Proposition E.13 satisfies $\mathrm{SubOpt} \leq \Delta s$ deterministically, and $\Delta s = \sigma\sqrt{s(\mathcal{C}-1)/n}$. □

Remark E.5 (Decision sparsity need not equal support recovery). A finite-class strategy that explicitly searches over supports is valid but unnecessary on this particular neutral-padded family. A learner may assign a crossing sign to every block: mistakes on inactive blocks change the occupancy but not the value. Proposition E.13 therefore removes the apparent $\log K$ support-search price for policy regret. This conclusion relies on exact reward neutrality and the block-independent logger, and is a family-specific statement about trajectory-outcome policy regret; decision-theoretic sparsity in offline RL more broadly is studied by Zhou (2024). The returned all-crossing policy can also have full-tabular $\mathcal{C}_{\mathrm{out}}$ growing with K because of its deterministic choices on neutral blocks; the dimension-free statement concerns decision value on the structured reward family and the designated optimal representative $\mathcal{C}_{\mathrm{out}}^{\mathrm{rep}} = 1 + s/\varepsilon$, not arbitrary tabular perturbations.

F Binary Outcomes and Preference Feedback

This appendix records how the identification theory changes under binary outcomes and preference feedback.

Proposition F.1 (Binary outcomes with a known monotone link)**.** *Suppose instead of Assumption 1 the outcome is binary, $Y \in \{0,1\}$, with $\mathbb{P}(Y = 1 \mid \tau) = f\big(\langle \phi(\tau), r\rangle\big)$ for a known, strictly increasing, continuous link $f : [0,H] \to [0,1]$. Then the conditional means identify $\langle \phi(\tau), r\rangle = f^{-1}(\mathbb{E}[Y \mid \tau])$ for every supported τ, and the entire identification theory of Section C — Theorem C.1, Proposition C.2, Proposition C.3 — holds verbatim for the* expected-return *functional $V^\pi = \langle d^\pi, r\rangle$.*

Proof. Strict monotonicity and continuity make f invertible on its range with continuous inverse, so the map $m(\tau) \mapsto f^{-1}(m(\tau)) = \langle \phi(\tau), r\rangle$ is a bijection between the identified conditional means and the linear measurements used throughout Section C; every argument there used only those linear measurements and the model class $[0,1]^d$. □

Remark F.1 (Unknown link; success probability as objective). Two genuine degeneracies appear beyond Proposition F.1. (i) If the link is unknown (even known to be monotone), only the ordering of the quantities $\langle \phi(\tau), r\rangle$ across supported trajectories is identified, not their scale; value differences of target policies are then generally not identified even under span coverage. (ii) If the objective is the success probability $\mathbb{E}^\pi[f(\langle \phi(\tau), r\rangle)]$ rather than the expected return, it is no longer a linear functional of d^π: for nonlinear f it depends on the entire law of the scalar $\langle \phi(\tau), r\rangle$ under π, so two policies with equal occupancy can have different success probabilities. A theory of OO-RL for distributional objectives is an open direction.

Remark F.2 (Comparison with preference feedback). A preference oracle on trajectory pairs (e.g., Bradley–Terry on returns) identifies, at best, the differences $\langle \phi(\tau) - \phi(\tau'), r\rangle$ over supported pairs, hence linear functionals of r in $\operatorname{span}\{\phi(\tau) - \phi(\tau')\} \subseteq \mathcal{U}_\mu$. This subspace never contains the direction of $\mathbf{1}$-components: $\langle \phi(\tau) - \phi(\tau'), \mathbf{1}\rangle = H - H = 0$ for all pairs, so $\frac{1}{\sqrt{d}}\mathbf{1}$-aligned reward mass is invisible to preferences while cardinal outcomes identify it whenever $\mathbf{1}$-components of supported features span it (always: $\langle \phi(\tau), \mathbf{1}\rangle = H \neq 0$). Cardinal outcome feedback is thus strictly more informative at the identification level; the practical trade is that preferences are often cheaper to collect.

G Linear function approximation

The outcome-geometry results extend directly from tabular rewards to linear reward features. The tabular model is the one-hot special case of the following p-dimensional formulation.

Setting. Let $\psi : [H] \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}^p$ be a known feature map with $\|\psi(h,s,a)\|_2 \le 1$, and suppose the reward is realizable,

$$r(h,s,a) \;=\; \langle \psi(h,s,a), w\rangle, \qquad w \in \mathbb{R}^p, \quad \|w\|_2 \le W.$$

The trajectory feature is the sum $\Phi(\tau) := \sum_{h=1}^{H} \psi(h, s_h, a_h) \in \mathbb{R}^p$, so the outcome model of Assumption 1 becomes $Y = \langle \Phi(\tau), w\rangle + \xi$, a linear measurement of w. For a policy π define the *successor feature*

$$\mu^\pi \;:=\; \mathbb{E}_\pi[\Phi(\tau)] \;=\; \Psi^\top d^\pi \in \mathbb{R}^p, \qquad \text{so that} \qquad V^\pi = \langle \mu^\pi, w\rangle,$$

where $\Psi \in \mathbb{R}^{d\times p}$ stacks the features. Set $\Sigma_\mu := \mathbb{E}_{\tau\sim q_\mu}[\Phi\Phi^\top] \in \mathbb{R}^{p\times p}$, $\mathcal{U}_\mu := \operatorname{range}(\Sigma_\mu)$ and $\mathcal{C}_{\mathrm{out}}(\pi;\mu) := \|\mu^\pi\|^2_{\Sigma_\mu^\dagger}$. Taking ψ to be the one-hot encoding of (h,s,a) recovers Sections 2–7 verbatim, with $p = d$ and $\mu^\pi = d^\pi$.

Theorems G.1 and G.2 assume exact realizability; Theorem G.3 allows a uniformly bounded per-step residual; and Remark G.1 shows that the lower-bound constructions are already linear-feature instances.

Theorem G.1 (Identification under bounded linear function approximation)**.** *Let $U := \operatorname{range}(\Sigma_\mu)$ and write $w_U = P_U w$, $\mu^\pi_\perp = P_{U^\perp}\mu^\pi$. The population law identifies w_U. Under the parameter constraint $\|w\| \le W$, the sharp identified set of the target value is*

$$\mathcal{I}_\pi = \big[\langle \mu^\pi, w_U\rangle - R\|\mu^\pi_\perp\|,\; \langle \mu^\pi, w_U\rangle + R\|\mu^\pi_\perp\|\big], \qquad R := \sqrt{W^2 - \|w_U\|^2}.$$

Consequently:

(a) If $\mu^\pi \in U$, then V^π is identified for every feasible parameter. Conversely, if $\mu^\pi \notin U$ and the instance has norm slack $\|w_U\| < W$ (in particular if $\|w\| < W$), then V^π is not point identified.

(b) The parameter w is uniformly point identified over the ball $\{\|w\| \le W\}$ iff $\Sigma_\mu \succ 0$. For a particular boundary instance with $\|w_U\| = W$, the ball itself can collapse the unidentified directions and make w unique even when Σ_μ is singular.

(c) In the tabular one-hot parameterization structural null directions are unavoidable (Proposition C.9); for a general feature map, $\Sigma_\mu \succ 0$ is possible whenever the supported trajectory features span $\mathbb{R}^p$.

Proof. The population conditional means identify $\langle \Phi(\tau), w\rangle$ on the support of q_μ, hence exactly the projection w_U. Any observationally equivalent feasible parameter is $w_U + v$ with $v \in U^\perp$ and $\|v\|^2 \le W^2 - \|w_U\|^2 = R^2$. Its target value is

$$\langle \mu^\pi, w_U\rangle + \langle \mu^\pi_\perp, v\rangle.$$

Cauchy–Schwarz gives the displayed interval, and both endpoints are attained by taking $v = \pm R\mu^\pi_\perp/\|\mu^\pi_\perp\|$ when the perpendicular component is nonzero. This proves sharpness and part (a). Part (b) follows because a nontrivial nullspace permits a nonzero v for every instance with norm slack, while a full-rank design has no such direction; the boundary exception is the case $R = 0$. Part (c) follows from the same span characterization. □

Theorem G.2 (Rates under function approximation)**.** *Let $\hat{w}$ be the ridge estimator on the features $\Phi(\tau_i)$ and run Algorithm 1 over the occupancy polytope with the penalty $\|\mu^\pi\|_{\Lambda^{-1}}$. Put $\beta_p(\delta) := 2\sigma\sqrt{2\big(p\log 5 + \log(1/\delta)\big)} + \sqrt{\lambda}\,W$. Then with probability at least $1-\delta$, $\mathrm{SubOpt}(\hat{\pi}) \le 2\beta_p(\delta)\|\mu^{\pi^\star}\|_{\Lambda^{-1}}$, and if moreover $n \ge n_0(\delta/2)$ with*

$$n_0(\delta) \;=\; \frac{8H^2}{\omega_{\min}}\Big(p\log\frac{24H^2}{\omega_{\min}} + \log\frac{1}{\delta}\Big), \qquad \omega_{\min} := \lambda^+_{\min}(\Sigma_\mu),$$

then with probability at least $1-\delta$,

$$\mathrm{SubOpt}(\hat{\pi}) \;\le\; 4\,\beta_p(\delta/2)\sqrt{\frac{\mathcal{C}^\star_{\mathrm{out}}(\mu)}{n}} = O\left(\Big[\sigma\sqrt{p+\log(1/\delta)} + \sqrt{\lambda}W\Big]\sqrt{\frac{\mathcal{C}^\star_{\mathrm{out}}}{n}}\right).$$

In particular this is $O(\sigma\sqrt{p\mathcal{C}^\star_{\mathrm{out}}/n})$ when δ is fixed and $\lambda W^2 \lesssim \sigma^2 p$. The finite-class proof of Theorem D.12 applies with the stochastic term depending on $|\Pi|$ and the deterministic ridge term $\sqrt{\lambda}W$ in place of $\sqrt{\lambda d}$, and Theorem D.13 continues to hold: the objective remains concave on the occupancy polytope.

Proof. The proofs of Lemma D.2, Theorem D.3 and Theorem D.4 use only that Y_i is a linear measurement of an unknown vector against an observed feature, corrupted by independent σ-subgaussian noise, plus an almost-sure bound on the feature norm; they therefore apply verbatim in $\mathbb{R}^p$. The norm bound is $\|\Phi(\tau)\|_2 \le \sum_h \|\psi(h, s_h, a_h)\|_2 \le H$, so $B = H^2$ in Remark L.1, which gives the stated n_0; and $\mathrm{tr}\Sigma_\mu = \mathbb{E}\|\Phi\|^2 \le H^2$ likewise. The deterministic part of the confidence radius is $\sqrt{\lambda}\|w\|_2 \le \sqrt{\lambda}W$. For tractability, write the objective as a function of the occupancy $d \in \mathcal{K}$: $L(d) = \langle \Psi^\top d, \hat{w}\rangle - \beta\|\Psi^\top d\|_{\Lambda^{-1}} = \langle d, \Psi\hat{w}\rangle - \beta\|d\|_M$ with $M := \Psi\Lambda^{-1}\Psi^\top \succeq 0$; a linear functional minus a seminorm is concave, and $\mathcal{K}$ is unchanged, so Theorem D.13 applies with the same linear oracle. □

Theorem G.3 (Misspecification with a certified residual bound)**.** *Suppose instead $r(h,s,a) = \langle \psi(h,s,a), w\rangle + \eta(h,s,a)$ with $\|\eta\|_\infty \le \varepsilon_{\mathrm{mis}}$, and run Algorithm 1 with the inflated radius $\beta_p(\delta/2) + H\varepsilon_{\mathrm{mis}}\sqrt{n}$. Then, on the corresponding confidence event and for $n \ge n_0(\delta/2)$,*

$$\mathrm{SubOpt}(\hat{\pi}) \;\le\; 4\beta_p(\delta/2)\sqrt{\frac{\mathcal{C}^\star_{\mathrm{out}}}{n}} \;+\; 4H\varepsilon_{\mathrm{mis}}\sqrt{\mathcal{C}^\star_{\mathrm{out}}} \;+\; 2H\varepsilon_{\mathrm{mis}}.$$

Proof. Write $V^\pi = \langle \mu^\pi, w\rangle + \langle d^\pi, \eta\rangle$ and note $|\langle d^\pi, \eta\rangle| \le \|d^\pi\|_1 \|\eta\|_\infty = H\varepsilon_{\mathrm{mis}}$, since every occupancy has $\|d^\pi\|_1 = H$. Hence the linear surrogate $\tilde{V}^\pi := \langle \mu^\pi, w\rangle$ satisfies $|V^\pi - \tilde{V}^\pi| \le H\varepsilon_{\mathrm{mis}}$ uniformly, and comparing the true optimum with the surrogate optimum costs at most $2H\varepsilon_{\mathrm{mis}}$.

For the estimation error, the outcome is $Y_i = \langle \Phi_i, w\rangle + b_i + \xi_i$ with $b_i := \langle \phi(\tau_i), \eta\rangle$ deterministic given τ_i and $|b_i| \le H\varepsilon_{\mathrm{mis}}$. Repeating the decomposition in the proof of Theorem D.12,

$$\langle \mu^\pi, \hat{w} - w\rangle = \sum_i c_i \xi_i + \sum_i c_i b_i - \lambda \langle \Lambda^{-1}\mu^\pi, w\rangle, \qquad c_i = \langle \Lambda^{-1}\mu^\pi, \Phi_i\rangle,$$

and by Cauchy–Schwarz together with $\sum_i c_i^2 \le \|\mu^\pi\|^2_{\Lambda^{-1}}$,

$$\Big|\sum_i c_i b_i\Big| \le H\varepsilon_{\mathrm{mis}} \sum_i |c_i| \le H\varepsilon_{\mathrm{mis}}\sqrt{n}\,\Big(\sum_i c_i^2\Big)^{1/2} \le H\varepsilon_{\mathrm{mis}}\sqrt{n}\,\|\mu^\pi\|_{\Lambda^{-1}}.$$

So the confidence statement holds with the inflated radius $\beta_p + H\varepsilon_{\mathrm{mis}}\sqrt{n}$, and running Algorithm 1 with that radius gives, for the surrogate values, $\tilde{V}^{\pi^\star} - \tilde{V}^{\hat{\pi}} \le 2(\beta_p + H\varepsilon_{\mathrm{mis}}\sqrt{n})\|\mu^{\pi^\star}\|_{\Lambda^{-1}}$. Bounding $\|\mu^{\pi^\star}\|_{\Lambda^{-1}} \le 2\sqrt{\mathcal{C}^\star_{\mathrm{out}}/n}$ as in Theorem D.4 turns the second term into $4H\varepsilon_{\mathrm{mis}}\sqrt{\mathcal{C}^\star_{\mathrm{out}}}$, and adding the $2H\varepsilon_{\mathrm{mis}}$ of the surrogate comparison completes the proof. □

Remark G.1 (Lower bounds under linear function approximation). The constructions of Section E also apply beyond tabular instances that happen to be realizable: their sufficient statistics are *exactly* per-step feature maps. For the (K, s)-sparse family of Section E.4, set $p = s + 1$ and

$$\psi_0(h,a) := \tfrac{1}{2}\ \forall (h,a), \qquad \psi_k(2k{-}1,a) = \tfrac{1}{2},\ \psi_k(2k{-}1,b) = -\tfrac{1}{2},\ \psi_k(2k,a) = -\tfrac{1}{2},\ \psi_k(2k,b) = \tfrac{1}{2}$$

for $k = 1, \dots, s$, all other entries zero. Summing along a trajectory gives $\Phi(\tau) = (K, \zeta_1(\tau), \dots, \zeta_s(\tau))$ — exactly the sufficient statistic of Proposition E.11(d) — and $r_\theta = \langle \psi, w\rangle$ with $w = (1, \frac{\Delta}{2}\theta)$, while $\|\psi(h,a)\|_2 \le 1/\sqrt{2}$. The dense family of Section E.1 is the case $s = K$, with $p = K + 1$. Consequently Theorems E.6 and E.12 also give lower-bound constructions for linear function approximation. On these families, the lower and upper bounds have matching dependence, up to constants and the stated burn-ins, on the active feature dimension and outcome coverage. The sparse construction makes the relevant feature dimension explicit: $p = s + 1$, while the ambient tabular dimension $d = 4K$ can grow independently.

Remark G.2 (Tightness of misspecification amplification). Theorem G.3 contains a $\sqrt{\mathcal{C}^\star_{\mathrm{out}}}\,\varepsilon_{\mathrm{mis}}$ term arising from the Cauchy–Schwarz control of the misspecification contribution. Figure 10 supports the predicted linear dependence on $\varepsilon_{\mathrm{mis}}$ and shows a larger fitted slope under poorer coverage. Across the tested regimes, however, the observed slope ratio is smaller than the corresponding $\sqrt{\mathcal{C}^\star_{\mathrm{out}}}$ ratio. Determining the tight worst-case dependence of misspecification error on $\mathcal{C}^\star_{\mathrm{out}}$ remains open.

H Compiler Autotuning Case Study

Every preceding theorem-facing experiment uses synthetic data with known ground truth. This section complements those diagnostics with a frozen real-system measurement table. The gcc study instantiates the same outcome-only observation interface: each complete flag configuration is observed together with one scalar performance outcome. We use the fitted additive surrogate to study trajectory-span geometry, evaluate policy selection against empirical conditional-mean outcomes, and measure whole-configuration residuals to quantify nonadditive flag interactions. Theorem G.3 provides the corresponding bounded misspecification analysis for per-step residuals, while the residual diagnostic here records interaction strength directly at the configuration level.

The instance. Compiler flag selection provides a concrete outcome-only observation interface. A configuration is a sequence of H binary decisions, one per flag, on top of `-O1`; the analyst observes the complete configuration and one measured speedup for the resulting binary, with no per-flag reward attribution. In our notation $S = 1$, $A = 2$, and $d = 2H$; a deterministic policy is a flag subset. The additive surrogate

$$V(\pi) = \sum_h r(h, a_h)$$

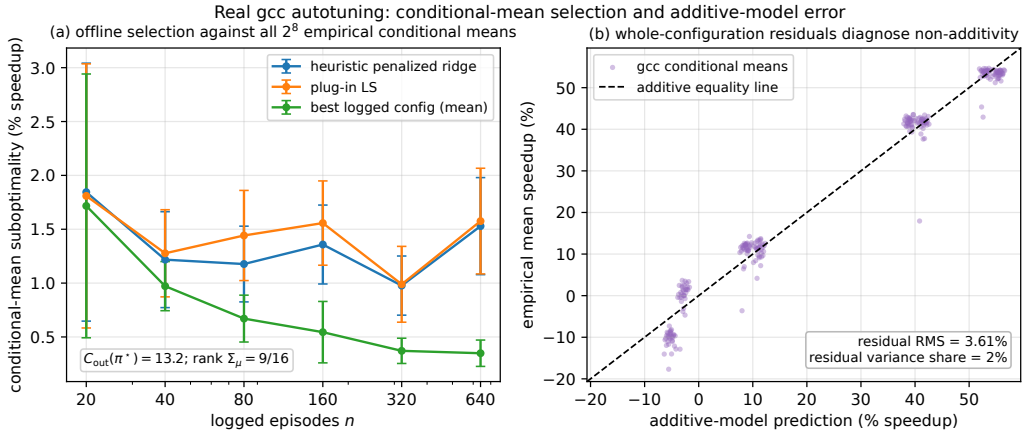

Figure 3: Compiler autotuning with frozen gcc measurements (20 independent seeds; Student-t 95% CIs). Values are empirical conditional-mean speedups over nine timing replicas for each of all 2^8 configurations. Left: heuristic penalized-ridge selection, plug-in least squares, and the best logged configuration. Right: empirical conditional-mean value against the least-squares additive prediction; departures from the diagonal quantify whole-configuration nonadditivity.

represents flag contributions in the trajectory-feature geometry, while configuration-level residuals record interactions among flags. We use gcc 13.3 on a loop-and-call benchmark kernel, with $H = 8$ flags and all $2^8 = 256$ configurations timed nine times.. For each configuration, the value used for learning and evaluation is the empirical conditional-mean speedup across its nine replicas. The logging policy enables each flag independently with probability $q = 0.35$. The mean within-configuration sample standard deviation is 2.16 percentage points and is reported as a descriptive variability measure. The baseline mean runtime is 41.46 ms and the best mean runtime is 18.80 ms (mask 175), a 54.66% speedup.

The structural predictions hold exactly. Proposition C.9 predicts that Σ_μ is rank deficient by at least $H - 1$ for every behavior. Measured on the real design matrix: $\operatorname{rank} \Sigma_\mu = 9$ out of $d = 16$, a deficit of exactly $7 = H - 1$. The reward vector is thus not point identified in this system — no logging policy could make it so — while the identifiability condition of Theorem C.1 does hold for the optimal configuration, with $\mathcal{C}_{\text{out}}(\pi^\star) = 13.22$.

Whole-configuration additive residual. The additive surrogate is accurate in aggregate but can have large local error. Its residual over the 256 empirical conditional means has RMS 3.61 percentage points and maximum magnitude 22.89 points; the residual accounts for 2.35% of the value variance. Policy selection depends on differences near the optimum rather than total explained variance, so this residual matters for policy selection.

Empirical plateau. From $n = 40$ through $n = 640$, the heuristic penalized-ridge and plug-in curves fluctuate around roughly 1–1.6% conditional-mean suboptimality, while the best-logged-configuration baseline reaches 0.35%. Both ridge-based methods optimize an additive surrogate, and the observed plateau coincides with configuration-level deviations from that surrogate.

Near-optimal geometry of the instance. On this benchmark, enabling all eight flags achieves 53.69% speedup against the optimum's 54.66%, so the all-flags configuration already lies within about one percentage point of the optimum. The resulting near-optimal gaps are therefore small relative to some of the observed whole-configuration residuals.

The frozen exhaustive $H = 8$ table supplies the real-system diagnostics reported in the paper. The optional `--live-h16` command generates a new machine-specific $H = 16$ timing dataset and is included as a reference-implementation timing exercise.

Empirical takeaways. The case study yields three observations. First, the rank obstruction of Proposition C.9 appears exactly in the real design. Second, an additive surrogate can explain most aggregate value variation while still making errors that are large relative to near-optimal gaps. Third, the best-logged-configuration baseline provides a useful comparison when additive credit assignment is approximate.

I Nonadditive Returns: Identification, Planning, and Model Selection

Section H shows that an additive surrogate can have small aggregate residual yet errors large relative to near-optimal gaps. We now extend the linear observation geometry to general trajectory features and separate the statistical role of the feature representation from the computational structure required for planning.

I.1 Setting

Let $\mathcal{T}$ denote the set of trajectories and let $\Psi : \mathcal{T} \to \mathbb{R}^p$ be a known feature map, subject only to $\|\Psi(\tau)\|_2 \le B_\Psi$. The return is

$$R(\tau) \;=\; \langle \Psi(\tau), w\rangle, \qquad w \in \mathbb{R}^p,\ \|w\|_2 \le W, \qquad Y = R(\tau) + \xi.$$

Nothing requires Ψ to decompose along the horizon. Writing $m^\pi := \mathbb{E}_\pi[\Psi(\tau)]$ for the expected trajectory feature we still have $V^\pi = \langle m^\pi, w\rangle$, and we set $\Sigma_\mu := \mathbb{E}_{\tau\sim q_\mu}[\Psi\Psi^\top]$, $\mathcal{U}_\mu := \mathrm{range}(\Sigma_\mu)$ and $\mathcal{C}_{\mathrm{out}}(\pi;\mu) := \|m^\pi\|^2_{\Sigma_\mu^\dagger}$ exactly as before. The additive model studied in the main text and preceding appendices is the special case $\Psi(\tau) = \sum_h \psi(h, s_h, a_h)$, in which $m^\pi = \Psi^\top d^\pi$ is a *linear image of the occupancy measure*. This linearity is what enables optimization directly over the occupancy polytope in the additive case.

I.2 Identification and estimation with general trajectory features

Theorem I.1 (Additivity-free restatement)**.** *For an arbitrary feature map Ψ as above, let $U := \mathrm{range}(\Sigma_\mu)$ and $w_U = P_U w$. Then:*

(a) The population law identifies w_U. Under the constraint $\|w\|_2 \le W$, the sharp identified interval for V^π is

$$\left[\langle m^\pi, w_U\rangle - R\|P_{U^\perp} m^\pi\|_2,\ \langle m^\pi, w_U\rangle + R\|P_{U^\perp} m^\pi\|_2\right], \qquad R = \sqrt{W^2 - \|w_U\|_2^2}.$$

Hence $m^\pi \in U$ is sufficient for point identification and is necessary for every instance with norm slack. It is also necessary for identification uniformly over the parameter ball. Suitable interior pairs outside the span give the same type of impossibility construction as Theorem E.1. At boundary instances, the parameter constraint can collapse the unidentified directions, as in Theorem G.1.

(b) The confidence radii of Lemma D.2 and Theorem D.12, the data-dependent pessimistic bound, and the population design argument extend with (d^π, d) replaced by (m^π, p) and feature bound B_Ψ. With the function-approximation radius $\beta_p = O(\sigma\sqrt{p + \log(1/\delta)} + \sqrt{\lambda}W)$, the population bound has order

$$O\Big(\Big[\sigma\sqrt{p + \log(1/\delta)} + \sqrt{\lambda}W\Big]\ \sqrt{\mathcal{C}^\star_{\mathrm{out}}/n}\Big)$$

after the corresponding design burn-in.

(c) The lower-bound constructions of Section E remain valid special cases of this feature model; their matching claims remain family-specific as stated there.

Proof. Part (a) is the argument of Theorem G.1 with the per-step origin of the feature vector removed: the supported values $\langle \Psi(\tau), w\rangle$ identify exactly $P_U w$, and optimization over the remaining

Euclidean ball in $U^\perp$ gives the interval. Part (b) follows because the concentration proofs depend on linear noisy measurements, a parameter-norm bound, and an almost-sure feature-norm bound. Part (c) follows because the additive constructions are contained in the arbitrary-trajectory-feature model. □

Remark I.1 (Where additivity enters). Additivity enters three parts of the development.

1. Proposition C.2 gives a linear-program representation of the identified interval in the tabular model. For general Ψ, the achievable feature expectations $\{m^\pi\}$ form the convex hull of trajectory features, but need not admit the same $O(HS)$ occupancy description.

2. Theorem D.13 uses the linear relation between m^π and the occupancy measure to obtain a concave objective on the occupancy polytope. Theorem I.3 shows that general pairwise trajectory interactions can instead make planning NP-hard.

3. Proposition C.9 is specific to the one-hot tabular parameterization. General feature maps can have $\Sigma_\mu \succ 0$ whenever their supported trajectory features span $\mathbb{R}^p$, as characterized by Theorem G.1(c).

Identification, observational impossibility, confidence sets, and the family-specific lower-bound constructions extend by replacing tabular occupancies with the corresponding trajectory-feature expectations. Their computational realization depends on the structure of the chosen feature map.

I.3 Model class, floor, and the bias–variance trade-off

Since the statistical guarantees hold for any Ψ, the analyst is free to choose it, and the choice is a trade-off. Fix a nested family $\Psi_1, \Psi_2, \dots$ of feature maps with dimensions $p_1 \le p_2 \le \cdots$ — for binary decisions, the natural choice is interaction order: Ψ_o contains all monomials of degree at most o in the action indicators, of dimension $p_o = \sum_{k \le o} \binom{H}{k}$.

Choosing a trajectory feature class trades approximation against statistical and computational complexity. Richer interaction classes can improve empirical projection accuracy while increasing feature dimension and planning cost. Theorem G.3 gives the certified misspecification guarantee for uniformly bounded per-step residuals. For the more general trajectory-feature classes below, we study approximation empirically and characterize complementary selection and lower-bound mechanisms in Sections I.7–I.9.

I.4 Planning becomes NP-hard

Before the general statement, the smallest possible example is worth recording: a single pairwise term already leaves the additive class, and by a quantifiable margin.

Lemma I.2 (The smallest non-additive interaction)**.** *For $A = 2$ and two steps $i < j$, the function $\tau \mapsto \mathbf{1}\{a_i \ne a_j\}$ is not of the form $f(a_i) + g(a_j)$; moreover its ℓ_∞ distance to the additive class is exactly $1/2$.*

Proof. Suppose $\mathbf{1}\{a_i \ne a_j\} = f(a_i) + g(a_j)$. Evaluating at the four action pairs gives $f(0) + g(0) = 0$, $f(0) + g(1) = 1$, $f(1) + g(0) = 1$ and $f(1) + g(1) = 0$. Adding the first and fourth gives $f(0) + f(1) + g(0) + g(1) = 0$; adding the second and third gives the same left-hand side equal to 2, a contradiction. For the distance, every additive h satisfies the parallelogram identity $h(0,0) - h(0,1) - h(1,0) + h(1,1) = 0$, whereas the same alternating sum of $\mathbf{1}\{a_i \ne a_j\}$ equals -2. Hence the error $e := \mathbf{1}\{a_i \ne a_j\} - h$ has $|e(0,0)| + |e(0,1)| + |e(1,0)| + |e(1,1)| \ge 2$, so $\|e\|_\infty \ge 1/2$; the constant $h \equiv 1/2$ attains it. □

Theorem I.3 (Computational phase transition at additivity)**.** *Consider planning with known dynamics and known reward parameter w: given (Ψ, w), output $\arg\max_\pi \langle m^\pi, w \rangle$.*

(a) If Ψ is additive, the problem is solved exactly by one backward induction, in time $O(HS^2A)$ (Theorem D.13).

(b) *If Ψ is allowed to contain degree-at-most-2 polynomial features in the action indicators, the problem is NP-hard, already for $S = 1$, $A = 2$, deterministic dynamics and $p \le \binom{H}{2}$ features each bounded by 1.*

(c) *Randomization does not help: the optimum over stochastic Markov policies equals the optimum over deterministic ones, so (b) is not an artifact of restricting to deterministic policies.*

(d) *The problem inherits the APX-hardness of* MAX-CUT*: there is a constant $\rho < 1$ such that approximating the optimal value within ρ is NP-hard.*

Proof. (b) Reduce from MAX-CUT. Given a graph $G = (V, E)$ with $|V| = H$ and nonnegative weights ω_{ij}, build the instance with $S = 1$, $\mathcal{A} = \{0, 1\}$, horizon H and trivial dynamics; a deterministic policy is a bit-string $b \in \{0, 1\}^H$, which we read as a bipartition. Let Ψ have one coordinate per edge,

$$\Psi_{ij}(\tau) \;=\; \mathbf{1}\{a_i \neq a_j\} \;=\; b_i + b_j - 2b_ib_j,$$

a degree-2 polynomial in the action indicators with $\Psi_{ij} \in \{0, 1\}$, and let $w_{ij} = \omega_{ij}$. Then $V(b) = \sum_{(i,j)\in E} \omega_{ij}\mathbf{1}\{b_i \neq b_j\}$ is exactly the cut value, so an optimal policy is a maximum cut. The reduction is polynomial and the returns are deterministic given the action sequence, so the optimal policy may be taken deterministic.

(c) For a Markov policy on $S = 1$ the actions at distinct steps are independent, so with $x_h := \mathbb{P}(a_h = 1)$,

$$\mathbb{E}_\pi \Psi_{ij} = x_i + x_j - 2x_ix_j,$$

and V is a *multilinear* function of $x \in [0, 1]^H$: affine in each coordinate with the others fixed. A multilinear function on a box attains its maximum at a vertex — fix all but one coordinate and optimize the resulting affine function at an endpoint, repeating H times — so $\max_{x\in[0,1]^H} V = \max_{b\in\{0,1\}^H} V$. Hence randomizing cannot improve the optimum and the hardness of (b) applies to the stochastic problem. (The stochastic-policy advantage in Remark D.9 therefore arises from the pessimistic penalty term; the value objective here remains multilinear.)

(d) Immediate from (b), since the reduction preserves the objective value exactly, and MAX-CUT is APX-hard. □

Remark I.2 (Worst-case hardness and empirical planning). Theorem I.3 establishes worst-case computational hardness. On the random instances of the same pairwise family used in Section J, ten-restart local search attains a lowest per-horizon mean ratio of approximately 99.7% of the optimum. By comparison, an additive model fitted to the same returns attains mean ratios of 38–65% across the tested horizons. These experiments separate the observed model-class approximation effect from the worst-case planning complexity characterized by Theorem I.3.

I.5 BOUNDED TEMPORAL RANGE RESTORES A DYNAMIC PROGRAM

The natural structural assumption for sequential problems is that interactions are local in time.

Definition I.4 (Interaction range)**.** *Ψ has* range *k if the return decomposes as $R(\tau) = \sum_h \varrho_h\big((s, a)_{h-k:h}\big)$, i.e. into terms each depending on at most $k + 1$ consecutive steps. Range 0 is additivity.*

Theorem I.5 (Range-k returns are additive on an augmented state)**.** *Let Ψ have range k. At step h, let the augmented state retain the current state together with the previous at most k state–action pairs,*

$$\tilde{s}_h := \big(s_h, (s_{h-k}, a_{h-k}), \ldots, (s_{h-1}, a_{h-1})\big),$$

with the obvious truncation near the beginning of the episode. Thus $\tilde{S} \le S^{k+1}A^k$. Then:

(a) *The return is an* additive *function of the augmented trajectory, so standard finite-horizon planning applies on the augmented MDP.*

(b) *One backward induction therefore solves planning in time $O(H\tilde{S}^2A) = O(HS^{2k+2}A^{2k+1})$ under this crude cardinality bound — exponential in the interaction range k, but linear in the horizon for fixed k, S, A.*

(c) The required policy may depend on this finite memory. Memoryless Markov policies on the original state are in general strictly suboptimal: for $k \geq 1$ there are instances on which every such policy loses a constant fraction of the optimum.

(d) The augmentation changes the planning representation while leaving the statistical feature dimension unchanged. If the original trajectory feature map has dimension p, the feature-space statistical results of Section G continue to depend on that original p, while planning operates on the augmented state space.

Proof. At step h, the pair $(\tilde{s}_h, a_h)$ contains every state and action appearing in $(s,a)_{h-k:h}$, so each ϱ_h can be written as an ordinary one-step reward $\tilde{r}_h(\tilde{s}_h, a_h)$. The augmented transition shifts the retained window, appends (s_h, a_h), and samples s_{h+1} from the original kernel. This proves (a); (b) is ordinary backward induction, and (d) follows because the statistical regression still uses the original feature vector $\Psi(\tau) \in \mathbb{R}^p$.

For (c), take $k = 1$, $H = 2$, $\mathcal{A} = \{0,1\}$, and two initial states $s_1 \in \{A, B\}$ with probability $\frac{1}{2}$ each, both transitioning deterministically to a single state C. Let the return be $\mathbf{1}\{a_1 = f(s_1)\} + \mathbf{1}\{a_2 = a_1\}$ with $f(A) = 0$, $f(B) = 1$. A history-dependent policy plays $a_1 = f(s_1)$ and then $a_2 = a_1$, achieving 2. A Markov policy chooses a_2 as a function of $s_2 = C$ alone, which carries no information about s_1 and hence none about the optimal a_1; whatever it plays, $\Pr(a_2 = a_1) \leq \frac{1}{2}$ unless it makes a_1 constant, in which case it forfeits $\frac{1}{2}$ on the first term. Either way its value is at most $\frac{3}{2}$. □

Remark I.3 (Other tractable classes). Bounded temporal range gives one explicit tractable class beyond additivity. More generally, bounded-treewidth interaction factor graphs admit standard junction-tree or variable-elimination dynamic programs. Theorem I.5 makes the temporal-memory case explicit and quantifies its dependence on the interaction range.

I.6 Why the model-free baseline sometimes wins

Section H found the best-logged-configuration baseline competitive with the model-based estimators. Its statistical mechanism is different: performance is governed by the logging mass on near-optimal trajectories and by replication of the configurations that appear in the log. The following result characterizes this mechanism directly.

Theorem I.6 (Best-logged configuration)**.** *Let $q_\alpha := \mathbb{P}_{\tau\sim q_\mu}\big(R(\tau) \geq R^\star - \alpha\big)$ be the logging policy's mass on α-optimal trajectories, and suppose the outcome noise is σ-subgaussian. Consider the estimator that averages the outcomes of each distinct logged configuration and returns the maximizer.*

(a) If $n \geq q_\alpha^{-1}\log(1/\delta)$ then with probability $\geq 1-\delta$ the log contains an α-optimal configuration; if in addition each distinct configuration is replicated $r \geq 2\sigma^2\alpha^{-2}\log(2N/\delta)$ times, where N is the number of distinct logged configurations, the returned configuration is 3α-optimal.

(b) Conversely, any algorithm whose output is restricted to configurations appearing in the log is, with probability at least $(1-q_\alpha)^n$, no better than α-suboptimal. Hence $n = \Omega(q_\alpha^{-1})$ is necessary.

Part (a) therefore requires both sufficient near-optimal exposure, through q_α, and sufficient replication of the distinct configurations used for empirical comparison.

Proof. (a) The number of α-optimal draws is binomial with success probability q_α, and $(1-q_\alpha)^n \leq e^{-nq_\alpha} \leq \delta$ under the stated n. Given r replications, a union bound over the N distinct configurations with the subgaussian tail gives that every configuration's empirical mean is within α of its value; the maximizer of the empirical means is then within 2α of the best *logged* value, which is itself within α of $R^\star$. (b) If no α-optimal configuration appears — an event of probability $(1-q_\alpha)^n$ — then every output is worse than $R^\star - \alpha$ by definition. □

Remark I.4 (Reading Section H through Theorem I.6). On the measured $H = 8$ autotuning instance, the logger assigns nontrivial mass to near-optimal configurations, which helps explain why

the best-logged baseline becomes competitive at the sample sizes in Figure 3. For larger horizons, Theorem I.6 identifies the actual near-optimal mass q_α as the governing quantity; its scaling depends on both the logger and the cardinality of the near-optimal set.

I.7 Choosing among candidate classes with held-out evaluation

Empirical fit, estimation width, and planning cost can vary across candidate feature classes. Held-out evaluation provides a finite-class comparison among the candidate policies produced by these classes. Sample splitting fixes the candidates for evaluation, so the stochastic confidence cost depends on $\sqrt{\log O}$ rather than $\sqrt{p_O}$. The result below compares the realized candidates using an independent well-specified evaluator.

Theorem I.7 (Held-out selection among interaction orders)**.** *Let $\Psi_1 \subseteq \cdots \subseteq \Psi_O$ be nested feature maps, and suppose the largest map is well specified, $R^\star = \langle \Psi_O, w_O \rangle$ pointwise. Assume a known bound $W_O \geq \|w_O\|_2$. Split the data into independent halves $\mathcal{D}_1, \mathcal{D}_2$ of size $n/2$. Train any candidate policy $\hat{\pi}_o$ from $\mathcal{D}_1$ for each o. On $\mathcal{D}_2$, form the ridge estimate $\hat{w}_O$ with Gram matrix Λ_O, set*

$$\beta_O := \sigma\sqrt{2\log(4O/\delta)} + \sqrt{\lambda}\, W_O,$$

and choose

$$\hat{o} \in \arg\max_{o \in [O]} \left\{ \langle m^{\hat{\pi}_o}, \hat{w}_O \rangle - \beta_O \|m^{\hat{\pi}_o}\|_{\Lambda_O^{-1}} \right\}.$$

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

$$\mathrm{SubOpt}(\hat{\pi}_{\hat{o}}) \leq \min_{o \in [O]} \left\{ \mathrm{SubOpt}(\hat{\pi}_o) + 2\beta_O \|m^{\hat{\pi}_o}\|_{\Lambda_O^{-1}} \right\}.$$

The bound requires no population design burn-in and compares the realized candidate policies*. Their population approximation quality remains a property of the corresponding candidate-training procedures.*

Proof. Condition on $\mathcal{D}_1$. The candidates are then a fixed set of at most O policies. Since Ψ_O is well specified, the finite-class ridge confidence bound on the independent half $\mathcal{D}_2$ gives, simultaneously for every candidate,

$$\left| \langle m^{\hat{\pi}_o}, \hat{w}_O \rangle - V^{\hat{\pi}_o} \right| \leq \beta_O \|m^{\hat{\pi}_o}\|_{\Lambda_O^{-1}} =: \tfrac{1}{2}\omega_o.$$

Writing $\mathrm{LCB}_o = \langle m^{\hat{\pi}_o}, \hat{w}_O \rangle - \frac{1}{2}\omega_o$, we have $\mathrm{LCB}_o \leq V^{\hat{\pi}_o} \leq \mathrm{LCB}_o + \omega_o$. The definition of $\hat{o}$ therefore implies, for every o,

$$V^{\hat{\pi}_{\hat{o}}} \geq \mathrm{LCB}_{\hat{o}} \geq \mathrm{LCB}_o \geq V^{\hat{\pi}_o} - \omega_o,$$

and subtracting from $V^\star$ proves the result. □

Remark I.5 (Scope of the selector). A saturated largest map makes the held-out evaluator well specified but can make its widths large. The theorem remains valid because it is fully data dependent; a small width is a property of the held-out design and the particular candidate, not something guaranteed merely by having few model classes. A moderate-dimensional well-specified largest map can of course yield substantially tighter widths.

Remark I.6 (What the finite-class refinement buys). The stochastic part of the held-out confidence radius scales with $\sqrt{\log O}$ because only O fixed candidate functionals are queried. Certification cost still reflects the deterministic ridge term $\sqrt{\lambda}W_O$ and the candidate-specific leverage $\|m^{\hat{\pi}_o}\|_{\Lambda_O^{-1}}$.

Remark I.7 (Selection under uneven evaluation coverage). Held-out widths are largest where the evaluation log carries the least information. In the autotuning case study, valuable configurations are relatively rare under the logger, so a candidate with narrower evaluation width can be selected even when another fixed-order method has higher population value. The realized widths on the right-hand side of Theorem I.7 quantify this trade-off directly.

I.8 THE FLOOR CANNOT BE ESCAPED FROM INSIDE THE CLASS

A separate lower-bound question remains: can a procedure restricted to a feature class escape the class's approximation obstruction as the sample size grows? The following result shows that it need not. The argument is the one behind Theorem E.1, transplanted from occupancy space to feature space.

Theorem I.8 (The approximation floor is unavoidable)**.** *Fix a feature map Ψ and a behavior q_μ. Let ξ be a trajectory-independent noise variable with a common symmetric law in the two instances, and let $w \in \mathbb{R}^p$. Suppose $\eta : \mathcal{T} \to \mathbb{R}$ satisfies*

$$\mathbb{E}_{q_\mu}\big[\eta(\tau) \,\big|\, \Psi(\tau)\big] = 0 \qquad \textit{and} \qquad \mathrm{Law}\big(\eta(\tau) \,\big|\, \Psi(\tau)\big) \textit{ is symmetric},$$

and consider the two instances $R_\pm := \langle \Psi, w\rangle \pm \eta$. Then:

(a) The two instances induce identical *laws for the observation $(\Psi(\tau), Y)$, at every n.*

(b) Consequently, for any algorithm $\hat\pi$ that is measurable with respect to $\{(\Psi(\tau_i), Y_i)\}_{i\le n}$ and its own randomness,

$$\max_{\varsigma\in\{+,-\}} \mathbb{E}_\varsigma\big[\mathrm{SubOpt}_\varsigma(\hat\pi)\big] \;\ge\; G(\Psi) \;:=\; \tfrac{1}{2}\big(V_+^\star + V_-^\star\big) - \max_\pi \langle m^\pi, w\rangle,$$

for every n, where $V_\pm^\star$ are the two optimal values.

(c) w is the $L^2(q_\mu)$ projection coefficient of $R_\pm$ onto the class. If $\mathrm{Floor}_\pm(\Psi) := V_\pm^\star - \max_\pi\langle m^\pi, w\rangle$, then

$$G(\Psi) = \tfrac{1}{2}\big(\mathrm{Floor}_+(\Psi) + \mathrm{Floor}_-(\Psi)\big).$$

Thus at least one of the two indistinguishable instances has a non-vanishing class approximation gap of at least $G(\Psi)$.

Proof. (a) The trajectory law q_μ is common to both instances, since the behavior does not depend on the reward. Conditionally on τ, $Y = \langle \Psi(\tau), w\rangle \pm \eta(\tau) + \xi$. Condition further on $\Psi(\tau)$: the first term is fixed, and by hypothesis the law of $\eta(\tau)$ given $\Psi(\tau)$ is symmetric. Since the independent noise ξ has the same symmetric law in both instances, $+\eta(\tau) + \xi$ and $-\eta(\tau) + \xi$ have the same conditional law. Therefore the joint law of $(\Psi(\tau), Y)$ agrees, and by independence across episodes so does the law of the whole sample.

(b) By (a) the output $\hat\pi$ has the same law under both instances; write $\mathbb{E}$ for the common expectation. For any fixed policy π,

$$V_+^\pi + V_-^\pi = \langle m^\pi, w\rangle + \mathbb{E}_\pi\eta + \langle m^\pi, w\rangle - \mathbb{E}_\pi\eta = 2\langle m^\pi, w\rangle \;\le\; 2\max_{\pi'}\langle m^{\pi'}, w\rangle.$$

Hence

$$\begin{aligned}\mathbb{E}_+[\mathrm{SubOpt}_+(\hat\pi)] + \mathbb{E}_-[\mathrm{SubOpt}_-(\hat\pi)] &= V_+^\star + V_-^\star - \mathbb{E}\big[V_+^{\hat\pi} + V_-^{\hat\pi}\big] \\ &\ge V_+^\star + V_-^\star - 2\max_\pi\langle m^\pi, w\rangle = 2G(\Psi).\end{aligned}$$

and the maximum of two numbers is at least their average.

(c) $\mathbb{E}[\eta \mid \Psi] = 0$ implies $\mathbb{E}[\eta\,\Psi_j] = 0$ for every coordinate j, i.e. $\eta \perp \mathrm{span}(\Psi)$ in $L^2(q_\mu)$; hence the projection of $R_\pm$ on the class is $\langle \Psi, w\rangle$ in both cases. Substituting the common projected-policy value into the definitions gives the displayed average-floor identity. □

Remark I.8 (Information structure of the approximation lower bound). Theorem I.8 applies to procedures measurable with respect to the class statistic $\Psi(\tau)$ and the observed outcomes. Full-trajectory replay uses additional information in τ and follows the near-optimal-exposure mechanism of Theorem I.6. Held-out selection in Theorem I.7 defines a third regime: it compares a finite set of realized candidate policies through an independent well-specified evaluator. These results characterize three distinct information structures for handling trajectory-feature approximation.

I.9 What remains open

Theorem I.8 gives a population obstruction for class-measurable procedures, while Theorem I.7 gives finite-sample held-out comparison among realized candidates. A population oracle inequality across misspecified classes would connect these two levels by controlling how each candidate-training procedure approaches its class-specific population optimum. Other open questions include removing the conditional-symmetry construction from the floor lower bound, sharpening planning complexity for structured interactions beyond the augmented-state bound of Theorem I.5, and combining model selection with tractable planning when candidate classes have unbounded interaction range.

J Additional Experiments and Implementation Details

The synthetic experiments have known ground truth; the gcc case study uses a fixed table of repeated timing measurements. Unless stated otherwise, synthetic diagnostics aggregate 20 independent seeds and report two-sided Student-t 95% confidence intervals over seed means; where blocks or MDPs are the independent experimental unit, we use the stated block or hierarchical bootstrap. All diagnostics use fixed random seeds.

Figure 3 and the nonadditive gcc diagnostics use empirical conditional means of the repeated $H = 8$ measurements. Its penalized-ridge curve is treated as an engineering heuristic. The matched-marginal diagnostics use a separately declared synthetic Gaussian learning-noise model on a box-normalized additive surrogate extracted from the same $H = 8$ table. Figures 11–12 use the estimands and calibration procedures stated with each experiment.

Table 4 labels each experiment as an identity check, lower-bound calibration, stress test, or implementation diagnostic.

Table 4: Questions, diagnostics, and their role.

Question (results)	Diagnostic	Role
Identification (Thm. C.1, Props. C.2, C.3)	Fig. 1	population identities; finite-sample illustration
Rates in n and $\mathcal{C}_{\mathrm{out}}$ (Lem. D.2, Thm. D.4)	Fig. 4	scaling illustration; certificate coverage
Impossibility (Thm. E.1)	Fig. 5	exact floor check
Beyond XOR	Fig. 6	calibration/robustness diagnostic
Dimension factor; horizon–memory identities (Prop. E.4, Thm. E.6, Cor. E.5)	Fig. 7	closed-form checks; lower-bound calibration
Structured sparsity (Prop. E.11, Thm. E.12, Prop. E.13)	Fig. 8; unknown-support check	closed forms; ambient-dimension check
Tractability (Thm. D.13)	Fig. 9	optimizer check; timings machine-dependent
Linear features (Thms. G.1–G.3)	Fig. 10	stress test
Non-additivity; held-out selection (App. I)	Figs. 11, 12	stress tests
Burn-in-free learning (Thm. D.8)	Fig. 13	guarantee-applicability diagnostic
Fixed-H minimax (Thm. E.8)	Fig. 14	calibrated scaling diagnostic
Algorithm 2 implementation	Tab. 5	reference-implementation timing
Logger memory (Thm. C.6, Prop. C.7)	Figs. 2, 15, 16, 17; Tab. 6	matched-marginal diagnostics
Real system	Fig. 3 (App. H)	case study

Figure 1 (verifies Theorem C.1, Propositions C.2 and C.3). On the 0-XOR instance, the plug-in value error of the covered policy (a, a) decays at the $n^{-1/2}$ rate, while the error for $\pi^\star = (a, b)$ plateaus exactly at the asymptotic bias $|\langle \Pi_{\mathcal{U}^\perp} d^{\pi^\star}, \Pi_{\mathcal{U}^\perp} r\rangle| = \Delta/2 = 0.5$ predicted by Proposition C.3; the sharp identified interval computed by the linear programs of Proposition C.2 is $[0, 2]$ while the true value is 1.5.

Figure 4 (verifies Lemma D.2 and Theorem D.4). On the ε-XOR family, ridge least-squares plug-in value-estimation error $|\langle d^{\pi^\star}, \hat{r} - r\rangle|$ (with $\lambda = 1$, as specified for Experiment 2) tracks the predicted shape $\sigma\sqrt{2\mathcal{C}_{\mathrm{out}}/n}$ both as a function of n (at $\varepsilon = 0.1$, $\mathcal{C}_{\mathrm{out}} = 11$) and as a function of $\mathcal{C}_{\mathrm{out}} = 1 + 1/\varepsilon$ at fixed n (with $C_{\mathrm{std}} = 2$ throughout, so the variation is driven purely by outcome concentrability); the certificate $\beta\|d^{\pi^\star}\|_{\Lambda^{-1}}$ of Lemma D.2 upper-bounds the realized error in 100% of runs, above the 95% nominal level.

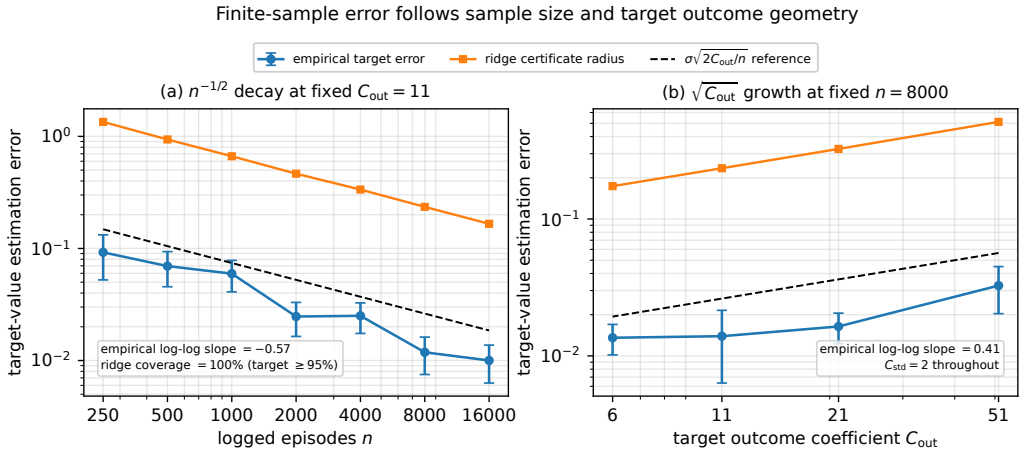

Figure 4: Target-policy value-estimation diagnostic (not policy regret): ridge least-squares plug-in error with $\lambda = 1$ scales with n and $\mathcal{C}_{\text{out}}$ (20 independent seeds per cell; Student-t 95% CIs).

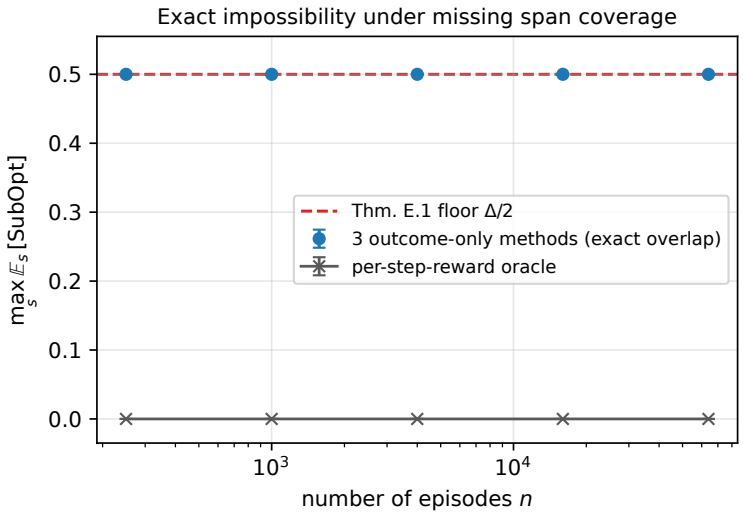

Figure 5: Exact impossibility without span coverage: the plotted worst-case estimand is $\max_s \mathbb{E}_s[\text{SubOpt}_s]$ (20 seeds, 95% CIs).

Figure 5 (verifies Theorem E.1). On the 0-XOR pair $\{r_+, r_-\}$ with $\Delta = 1$, every outcome-only method (pessimism, plug-in least squares, uniform credit assignment) sits exactly on the predicted floor $\Delta/2 = 0.5$ for the theorem estimand $\max_s \mathbb{E}_s[\text{SubOpt}_s]$ at every sample size from 250 to 64,000, while an oracle observing noisy *per-step* rewards drives suboptimality to zero. The maximum is taken after the seed mean within each hypothesis; it is not $\mathbb{E}[\max_s \text{SubOpt}_s]$.

Figure 6 (robustness beyond the XOR family). On 12 random tabular MDPs ($H = 3$, $S = 3$, $A = 2$) with a uniform behavior policy, Algorithm 1 is optimized over the full occupancy polytope using its theorem radius, while plug-in optimization is exact. With the MDP as the experimental unit and 20 seeds nested within each MDP, the IQM suboptimalities at $n = 500/2000/8000$ are $0.1964/0.0440/0.0105$ for theorem-calibrated pessimism and $0.0010/0.00017/ \approx 0$ for plug-in; uniform credit gives $0.0071/0.0052/0.0038$ and top-quantile cloning $0.0259/0.0021/0.00049$. Thus this experiment serves as a calibration and robustness diagnostic: on these instances, the theorem-calibrated radius is substantially more conservative than the plug-in fit.

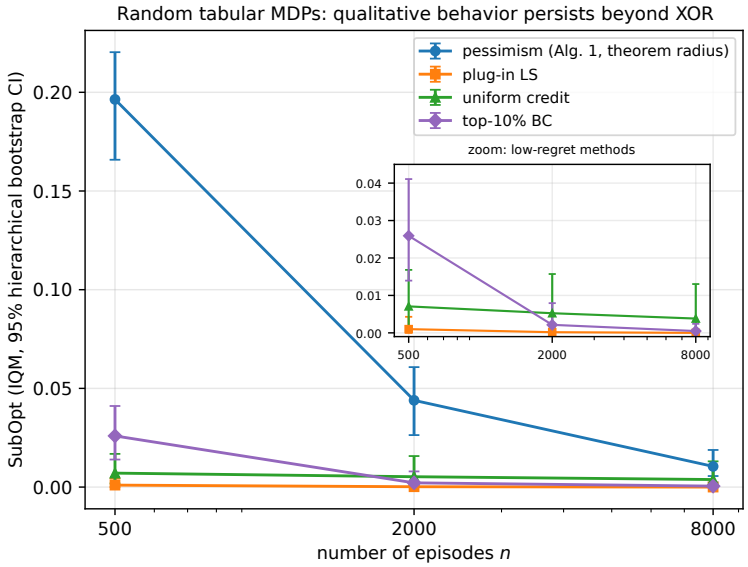

Figure 6: Random tabular MDPs: IQM with hierarchical MDP→seed bootstrap 95% CIs over 12 MDPs × 20 nested seeds. The Algorithm 1 curve uses the theorem radius and full occupancy-polytope optimization.

Figure 7 (verifies Proposition E.4 and Theorem E.6). On the K-block family with $\varepsilon = 0.25$, $\sigma = 0.5$, $n = 4000$ and $K \in \{1, \dots, 8\}$ (so $d = 4K$ up to 32 and $\mathcal{C}_{\text{out}} = 1 + K/\varepsilon$ up to 33), the closed forms of Proposition E.4 are confirmed numerically to machine precision: $\mathcal{C}_{\text{out}}$, $\dim \mathcal{U}_\mu = 2K + 1$ and $V^\star$ agree exactly with the formulas for every K tested. Policy optimization is exact (all 4^K deterministic policies are enumerated). At the Assouad calibration $\Delta = \sigma/\sqrt{n\varepsilon}$, the per-block decoding error rate is flat in K and stays above the floor $1/4$ of Lemma L.6 (panel b, solid), and consequently $\mathbb{E}[\text{SubOpt}]/\Delta$ grows *linearly* in K, tracking the theoretical slope $K/8$ from above — measured slope 0.293 for the plug-in estimator, versus the guaranteed 0.125 (panel a). Raising the gap to 4Δ collapses the per-block error to a few percent and flattens the growth (panel b, dashed), confirming that the construction is calibrated at the correct scale rather than at an accidentally hard one. Pessimism is uniformly more conservative than the plug-in here, as expected: at the detection threshold its penalty dominates the signal and it abstains, which costs a constant factor but preserves the linear growth in K. The same frozen experiment also records theorem-facing population checks used by Corollary E.5 and Theorems E.9–E.10: for $H \in \{2, 4, 6, 8, 12, 16\}$ the direct-sum identity $\mathcal{C}_{\text{out}} - 1 = H\kappa_{\text{mem}}$ holds numerically; the fixed-$C = 9$ temporal and parallel constructions preserve $\mathcal{C}_{\text{out}} = 9$ and $C_{\text{std}} = 2$ on their declared grids. These records provide numerical identity checks for the corresponding closed-form constructions.

Figure 8 (fixed-support diagnostic for Proposition E.11 and Theorem E.12). On the sparse family with $\varepsilon = 0.25$, $\sigma = 0.5$, $n = 4000$ and the gap calibrated at $\Delta = \sigma/\sqrt{n\varepsilon}$, we vary the ambient dimension over $d = 4K \in \{4, \dots, 128\}$ at fixed sparsity $s \in \{1, 2, 4\}$. The closed form $\mathcal{C}^\star_{\text{out}} = 1 + s/\varepsilon$ is confirmed to machine precision at every (K, s), and $\mathbb{E}[\text{SubOpt}]/\Delta$ is flat in d: regressing it on $\log_2 d$ gives slopes $+0.046$, 0.000 and -0.005 for $s = 1, 2, 4$ — small relative to the sampling variation across a 32-fold increase in d — while staying above the Assouad guarantees $s/8$. The dense family ($s = K$) plotted on the same axes grows from 0.35 to 2.45 over the same range. Read at fixed $d = 32$, the four families with $s = 1, 2, 4, 8$ give $0.40, 0.45, 1.50, 2.45$: the difficulty tracks s, not d.

Unknown-support diagnostic for Proposition E.13. With random active supports of size $s = 4$ at the local Theorem E.12 calibration, the regret normalized by $\sigma\sqrt{s(\mathcal{C}^{\text{rep}}_{\text{out}} - 1)/n}$ remains approximately flat as K grows from 16 to 256 (coefficient of variation 0.144; endpoint growth $0.936\times$). In a separate fixed-gap check at $n = 2000$, the observed mean regrets for $K = 16, 64, 256$ are

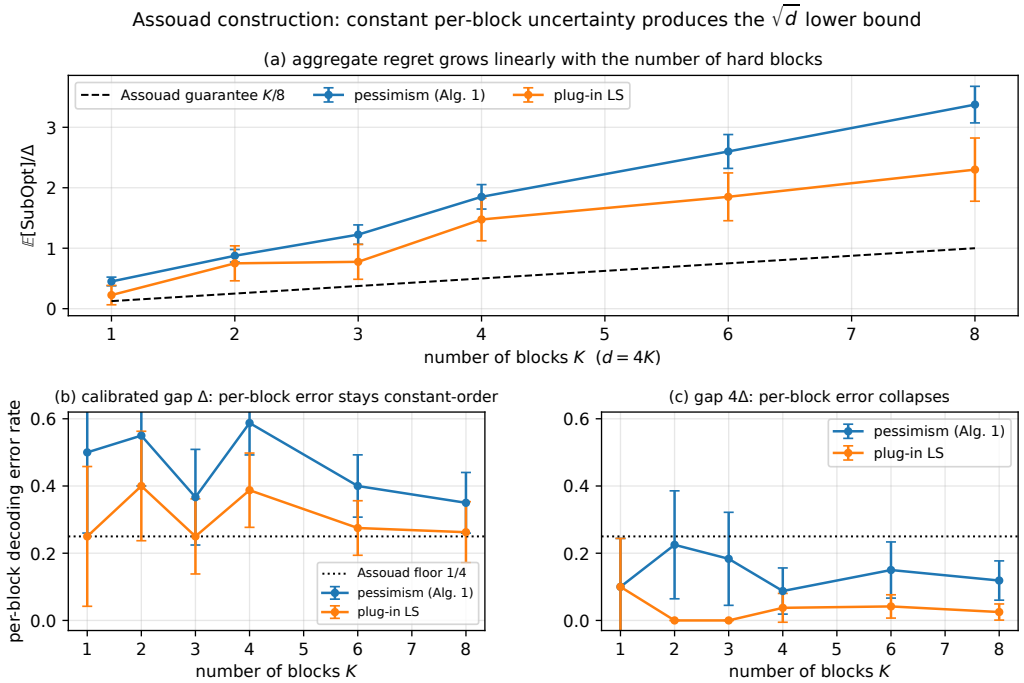

Figure 7: Assouad multi-block construction: per-block error is $\Theta(1)$ while total suboptimality grows linearly in $K = d/4$ (20 independent seeds per cell; Student-t 95% CIs; θ drawn uniformly from the hypercube, the same average the Assouad argument bounds).

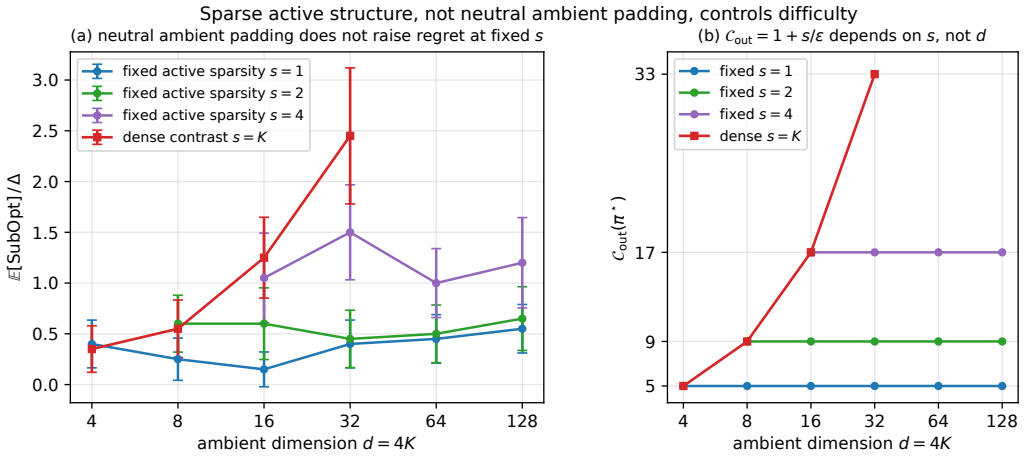

Figure 8: The regime $\mathcal{C}_{\mathrm{out}} = o(d)$. Left: expected suboptimality at the detection threshold is flat in the ambient dimension at fixed sparsity, while the dense family grows. Right: the concentrability coefficient is exactly $1 + s/\varepsilon$ regardless of d (20 independent seeds; Student-t 95% CIs).

$0.015, 0.030, 0.025$, while the common analytic upper bound from Proposition E.13 is 0.1152. The purpose of this diagnostic is only to test the predicted absence of ambient-K dependence; the proposition itself is finite-sample and nonasymptotic.

Figure 9 (verifies Theorem D.13). On the K-block geometry we compare Frank–Wolfe (300 dynamic-program calls) against exhaustive enumeration of all 2^H deterministic policies. The linear oracle is first checked against brute-force maximization on random MDPs with $S \in \{2, 3\}$, agreeing

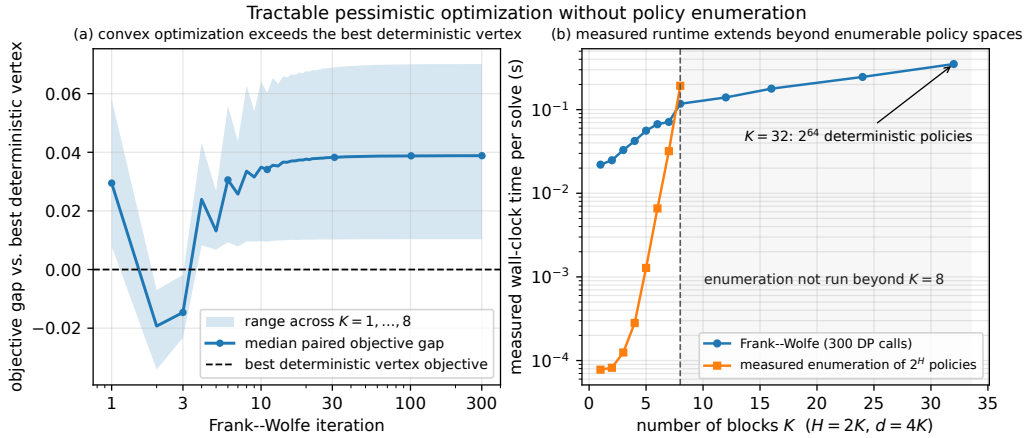

Figure 9: Pessimism is tractable. Left: the median paired Frank–Wolfe objective gap relative to the best deterministic-policy objective, with the range over $K = 1, \ldots, 8$; positive values show that the stochastic occupancy solution is better in the pessimistic objective. Right: measured wall-clock time per solve; enumeration is shown only where it was actually run (machine-dependent; 20 seeds per cell).

to 10^{-9}. The comparison is paired seed-by-seed. Frank–Wolfe attains a pessimistic objective at least as large as the best deterministic policy at every K, with mean paired excess increasing from $+0.0103$ at $K = 1$ to $+0.0701$ at $K = 8$, confirming Remark D.9: the maximizer is not a vertex. The paired true-suboptimality difference stays within 0.0066 in magnitude. Enumeration becomes rapidly impractical, whereas the $K = 32$ convex solve ($H = 64$, $d = 128$, 1.8×10^{19} vertices) remains subsecond in the reported CPU run. These wall-clock values characterize the reference implementation on the stated machine.

Figure 10 (stress-tests Theorems G.1, G.2 and G.3). The embedding of Remark G.1 is checked exactly: for (K, s) up to $(32, 4)$ the per-step feature map reproduces the tabular reward vector to machine precision ($\leq 10^{-16}$) and returns the same $V^\star$, with $\|\psi\|_2 \leq 0.708$ and $p = s + 1$ against ambient $d = 4K$ up to 128. A separate rank diagnostic with $H = 8$ and $A = 3$ uses a generic feature map and is full rank for every tested $p \leq 12$, showing that the tabular structural deficiency of Proposition C.9 need not persist under generic features.

For the finite-class confidence radius we use the realized valid norm bound $W = \|w\|_2$ for every instance. Because target leverage drifts mildly with p, panel (a) plots $\mathrm{SubOpt}/\sqrt{\mathcal{C}_{\mathrm{out}}}$ rather than raw suboptimality. Its log–log slope versus p is $+0.538$, against the $+0.5$ reference from Theorem G.2. Panel (b) sweeps misspecification under common random numbers. The fitted slope is $+0.396$ with 95% block-bootstrap CI $[+0.218, +0.589]$ at $\mathcal{C}_{\mathrm{out}} = 6.72$, and $+0.522$ $[+0.265, +0.822]$ at $\mathcal{C}_{\mathrm{out}} = 21.20$. Their ratio 1.32 is smaller than the $\sqrt{\mathcal{C}_{\mathrm{out}}}$ ratio 1.78, and the intervals overlap. The experiment supports the predicted linear dependence on misspecification and the ordering with coverage; the tight worst-case $\sqrt{\mathcal{C}_{\mathrm{out}}}$ amplification remains open (Remark G.2).

Figure 11 (stress-tests Section I). Lemma I.2 is confirmed numerically: the ℓ_∞ distance from $\mathbf{1}\{a_i \neq a_j\}$ to the additive class is 0.5000.

Panel (a) uses the real gcc measurement cache with interaction orders $1, 2, 3$ ($p = 9, 37, 93$), but unlike the robust engineering study of Figure 3 it defines each configuration's population value by the empirical conditional *mean* of its timing replicas. The resulting $L^2(q_\mu)$ projection floors are respectively 0.583%, 2.252%, and 0.634%, while the corresponding fit R^2 values increase from 0.977 to 0.986 to 0.994. Better predictive fit therefore need not produce a monotone policy floor because the maximizing configuration can move. The finite-log candidates in this panel use plain ridge plug-in fits; no full-cache noise proxy enters their selection rule. At $n = 2560$, order-$1/2/3$ mean suboptimalities are $0.939\%/1.970\%/0.742\%$, while the best-logged baseline reaches 0.422%.

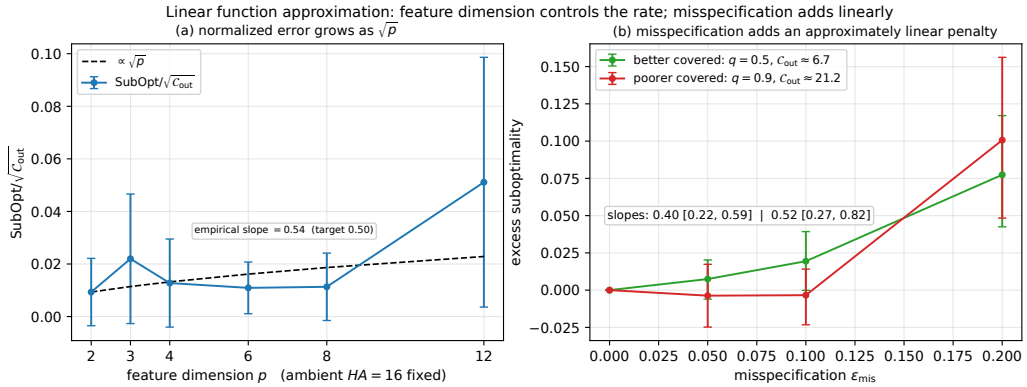

Figure 10: Linear function approximation. Panel (a): $\mathrm{SubOpt}/\sqrt{\mathcal{C}_{\mathrm{out}}}$ versus feature dimension p has log–log slope 0.538, against the predicted $1/2$; each finite-class radius uses the realized valid parameter-norm bound $W = \|w\|_2$ (20 independent seeds; Student-t 95% CIs). Panel (b): misspecification enters approximately linearly, with a larger fitted slope under poorer coverage (12 independent blocks $\times$ 20 nested seeds; paired block-bootstrap 95% CIs).

Panel (b) instantiates the reduction of Theorem I.3 on random graphs, averaging over 20 independent runs per horizon. An additive projection of the true returns attains mean ratios of only 38%–65% of the maximum cut for $H \in [8, 16]$, while ten-restart local search on the correct pairwise objective has a lowest per-horizon mean ratio of approximately 99.7%.

Panel (c) tests a state-independent special case of the finite-memory construction in Theorem I.5. The augmented dynamic program agrees with brute-force enumeration to 10^{-9} wherever the latter is run ($H \leq 18$), while avoiding enumeration beyond that range. The wall-clock values report reference-implementation timing on the stated machine. Under the mean-valued gcc target, the logger places mass $q_{1\%} = 0.052$ on 1%-optimal configurations, explaining the competitiveness of the best-logged baseline on this 256-configuration instance. Theorem I.6 identifies q_α as the quantity governing the corresponding exposure requirement at other horizons.

Figure 12 (stress-tests Theorem I.7). Candidates from interaction orders $1, 2, 3$ are trained on the first half of the log by plain ridge plug-in argmax, with no confidence-radius noise proxy used in candidate construction. The independent second half uses the saturated one-hot evaluator from Theorem I.7, with the common finite-support Hoeffding proxy $\sigma_{\mathrm{eval}} = 91.1598$ obtained from half the maximum observed conditional range.

For a deterministic configuration with held-out count n_π, the saturated ridge predictor is $\sum_{i:\pi_i=\pi} Y_i/(\lambda + n_\pi)$ and the confidence width is $\beta_O/\sqrt{\lambda + n_\pi}$. The finite table defines this controlled benchmark, so the exact $\|w_O\|_2$ is supplied to the evaluator as its norm bound. At $n = 2560$, selected/best-fixed/hindsight mean suboptimalities are $1.063\%/0.733\%/0.658\%$; order 1 is selected in 80% of seeds.

Figure 13 (direct burn-in diagnostic). At $n = 1000$, decreasing the probability ε of an *irrelevant* XOR direction from 0.5 to 0.005 makes the sufficient global ridge burn-in increase from 556.2 to $99{,}831.8$ episodes, while the target coefficient stays bounded ($\mathcal{C}_{\mathrm{out}} = 3.000, 2.111, 2.020, 2.005$). Across these cells, Algorithm 2 remains near-optimal, with mean suboptimality $0, 0, 0, 8.50 \times 10^{-5}$; its dual-certified maximum numerical gaps are respectively $2.35 \times 10^{-9}, 3.35 \times 10^{-8}, 4.08 \times 10^{-8}, 6.69 \times 10^{-6}$. The target $\mathcal{C}_{\mathrm{out}}$ stays small while the sufficient global ridge burn-in threshold grows sharply. The ridge and plug-in learners also have near-zero realized regret in these cells (panel a). The diagnostic therefore isolates guarantee applicability: the target-local premise of the burn-in-free result remains benign while the global design condition required by the ridge population bound becomes increasingly stringent.

Figure 14 (direct fixed-H minimax diagnostic). On the parallel-state family, $H = 2$ and $\mathcal{C}_{\mathrm{out}} = 5$ are fixed while $d = 4K$ grows from 4 to 128. Plug-in suboptimality has log–log

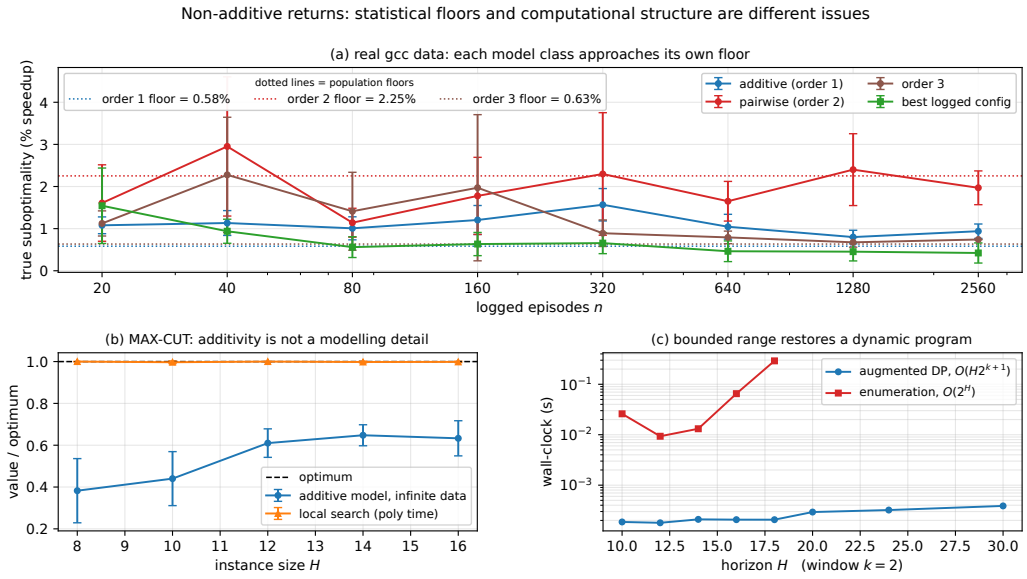

Figure 11: Beyond additive returns. Panels (a)–(b) use 20 independent runs with Student-t 95% CIs. Panel (a): interaction-order models on the gcc cache using empirical conditional-mean values; predictive fit and policy floor need not be monotone together. Panel (b): an additive projection on true MAX-CUT returns can lose substantial value even when local search on the correct pairwise objective is near-exact on these random instances. Panel (c): an augmented dynamic program is exact on a bounded-range special case wherever brute force is affordable.

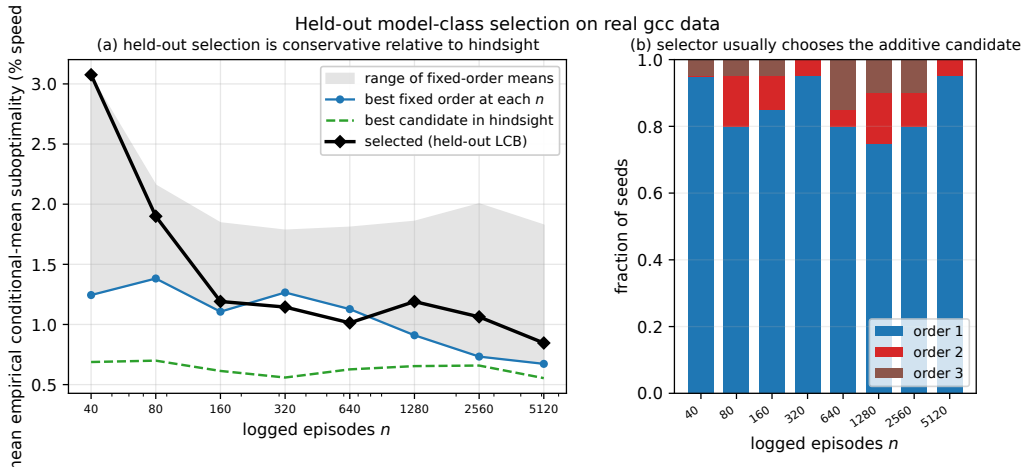

Figure 12: Held-out choice among interaction orders on the finite gcc measurement table (20 seeds). Candidates are plain ridge plug-in policies trained on one half of the log; the independent saturated evaluator uses the finite-class LCB of Theorem I.7 with uniform finite-support Hoeffding proxy $\sigma_{\text{eval}} = 91.16$. Left: mean selected suboptimality, with the fixed-order candidate envelope and hindsight reference. Right: selection frequencies.

slope 0.487256 against the predicted $1/2$, and the coefficient of variation of $\mathbb{E}[\text{SubOpt}]/\Delta$ is 0.085580. Because the gap Δ is calibrated at the lower-bound scale, the normalized error is the informative quantity. Its near-constancy over the tested dimensions is consistent with constant-order per-bit decoding error. On the prespecified Algorithm 2 subset $d = 4, 8, 16$, mean suboptimalities are $0.011030, 0.015075, 0.021707$ and the maximum dual-certified numerical gaps are $1.978, 1.999, 1.977 \times 10^{-4}$. The $K = 1$ point is a boundary diagnostic; Theorem E.8 is stated for $K \geq 2$.

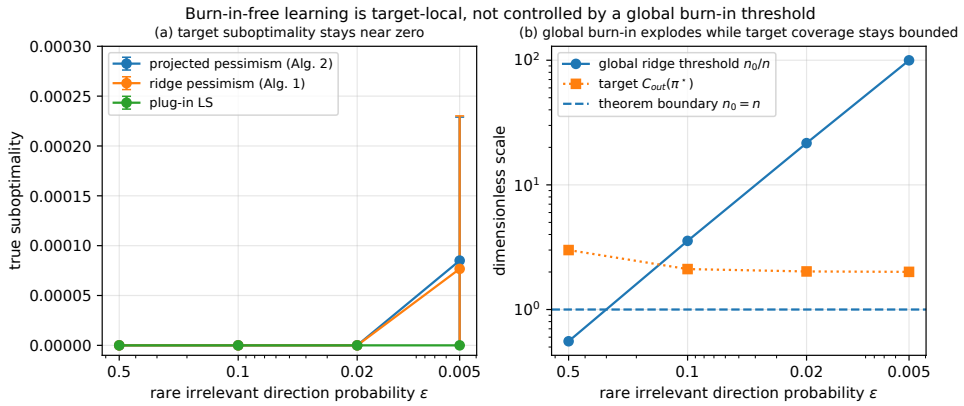

Figure 13: Target-local versus global design geometry. The target remains easy while a weak irrelevant direction drives the sufficient global ridge burn-in upward. Algorithm 2 gaps use the floating-point primal/dual numerical certificate (20 seeds, 95% bootstrap CIs).

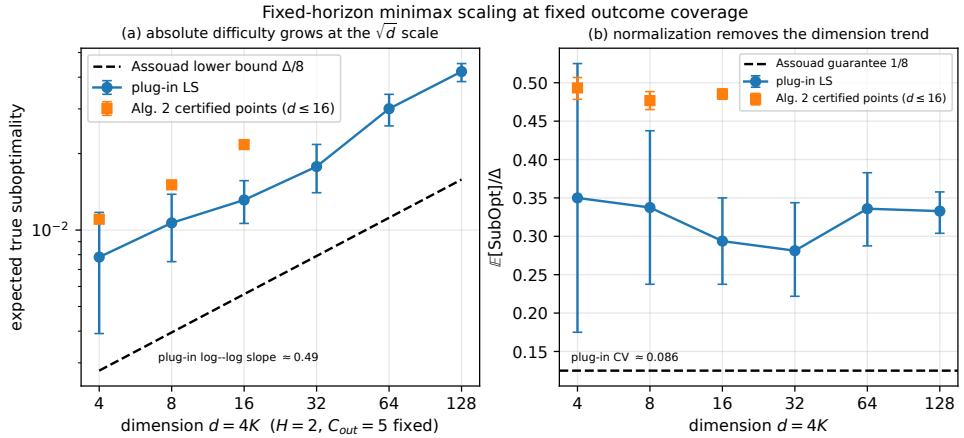

Figure 14: Fixed-horizon minimax diagnostic with $H = 2$ and $\mathcal{C}_{\text{out}} = 5$ fixed. Absolute suboptimality follows the $\sqrt{d}$ calibration and normalized error is approximately constant; Algorithm 2 is shown only on the prespecified certified subset.

Figure 2 (mechanism-based flag-budget logger). The main diagnostic uses a logging mechanism rather than a correlation interpolation. Each episode draws $B \in \{2, 3\}$ with probabilities $(1/2, 1/2)$ and then chooses a B-subset of the eight compiler flags uniformly without replacement. Sequentially, the probability of enabling a flag depends on how many earlier flags have already been selected, so the logger is history dependent but outcome independent. Its occupancy-equivalent Markovization enables each flag independently with $q = \mathbb{E}[B]/8 = 0.3125$; the maximum marginal discrepancy is 5.55×10^{-17}.

The target is the additive-surrogate optimum (mask 43, four enabled flags), which has zero trajectory probability under the budget logger but remains point identified. The Markovized law has $\mathcal{C}_{\text{out}} = 11.618$ (equal to its one-step marginal bound) while the budget law has $\mathcal{C}_{\text{out}} = 18.296$, a factor 1.575 larger; $C_{\text{std}} = 3.2$ is unchanged, $\kappa_{\text{mem}} = 6.875$, and $\kappa_{\text{rep}} = 3.841$. With synthetic Gaussian outcome noise $\sigma = 2$, 200 seeds, and $n \in \{100, 200, 500, 1000, 2000\}$, the two RMSE curves have log–log slopes -0.503 and -0.536. Their mean RMSE ratio is 1.2966 with a 4000-replicate bootstrap 95% interval $[1.2225, 1.3845]$, compared with the population prediction $\sqrt{18.296/11.618} = 1.2549$. A prespecified same-marginal mechanism ablation keeps $\mathbb{E}[B] = 2.5$ fixed while changing the budget law: the four $\sqrt{\mathcal{C}_{\text{out}}/\mathcal{C}_{\text{out}}^{\text{Markov}}}$ ratios are approximately $1.255, 1.005, 1.184, 1.042$, with empirical RMSE ratios $1.348, 1.011, 1.309, 1.025$, respec-

tively. The additive surrogate is fitted once from the frozen gcc conditional-mean table; this experiment takes no new compiler measurements.

Numerical scalability of Algorithm 2. Theorem D.11 concerns an SOCP value/supergradient oracle with conic/ellipsoid optimization. The numerical reference uses a HiGHS master linear program, SLSQP inner solves, linear program special cases, and a scalar-dual L-BFGS-B fallback. These timings therefore characterize the reference implementation, not the theorem's oracle implementation.

For the $H = 2$, $\mathcal{C}_{\mathrm{out}} = 5$ ParallelStateXOR family at $n = 2000$, a five-dimension pilot and ten-seed-per-dimension sweep used a 300-second subprocess ceiling and a 10,800-second total clock including pilot, sweep, launch overhead, and aggregation. The complete pilot and sweep finished in 505.332 seconds. Table 5 gives the Algorithm 2 sweep wall times and logical master/inner calls. All 50 sweep runs met the 2×10^{-4} numerical gap tolerance, with no final box, cone, or certificate-order violations. Ten SLSQP failures at $d = 4$ or $d = 8$ invoked the fallback and completed successfully. All five pilot instances also met their certificates. A separate deliberately failing pilot sentinel produced an exception and 54 explicit certification-stop records; these are excluded from the 50 sweep runs. The reference Algorithm 1 was run for a fixed 300 Frank–Wolfe iterations on the same instances; those timings have no equivalent optimality certificate and are not used for a solver-quality comparison.

Table 5: Algorithm 2 CPU sweep: ten seeds per dimension; wall time in seconds as median [IQR]. Call counts are medians.

d	Wall time, median [IQR]	Master calls	Inner calls
4	0.073 $[0.068, 0.086]$	10	11
8	0.123 $[0.113, 0.197]$	26.5	27.5
16	0.304 $[0.296, 0.404]$	73	74
24	13.046 $[12.437, 14.287]$	141.5	142.5
32	22.246 $[21.907, 23.643]$	220.5	221.5

The CPU environment was Python 3.12.13, NumPy 2.4.4, Linux 7.0.11 x86_64, Intel Xeon Gold 6128 at 3.40 GHz. The reported grid covers $d \leq 32$ and characterizes the reference implementation on this hardware.

Single-stage conic implementation cross-check. As an implementation cross-check, we also tested a single-stage conic formulation obtained from the minimax/Bellman-dual reduction and recovered occupancies from Bellman multipliers. On nine paired instances at $d \in \{4, 8, 16\}$ (three seeds per dimension), all 18 reference/conic method runs met the 2×10^{-4} numerical target, with no failures or timeouts, and all nine paired certificate intervals overlapped. The maximum absolute objective difference was 1.15325×10^{-4} and the largest conic feasibility/recovery residual was 1.2862×10^{-6}. Median reference/conic wall times were $0.084119/0.004818$, $0.222483/0.006294$, and $0.380871/0.007547$ seconds for $d = 4, 8, 16$, respectively. These small paired pilots characterize numerical agreement and reference-implementation timing for $d \in \{4, 8, 16\}$. The $d = 24, 32$ reference grid was not rerun with the conic formulation.

Figure 15 (Markov-boundary bridge). The two-step ε-XOR family isolates the boundary with the nearest policy-generated trajectory-supervision regime. For every $\varepsilon \in \{0.5, 0.25, 0.10, 0.05, 0.02\}$, one-step action marginals are uniform, so the deterministic target has $C_{\mathrm{std}} = 2$ and the Markov marginal χ^2 bound is fixed at 3. The cell $\varepsilon = 0.5$ is exactly the independent Markov logger; for smaller ε, only cross-time dependence changes and $\mathcal{C}_{\mathrm{out}} = 1 + 1/\varepsilon$ takes values $3, 5, 11, 21, 51$.

Each of the 500 seeds fits ridge-regularized least-squares trajectory-return reward redistribution once with $\lambda = 0.05$. The greedy planner and a finite-class ridge lower-confidence planner use the same Φ, outcomes, fitted reward, Gram matrix, and four deterministic policies. The outcome-noise law, n, and one-step marginals are common to the two arms. At $n = 500$, the greedy mean regrets for $\varepsilon = (0.5, 0.25, 0.10, 0.05, 0.02)$ are $0.00025, 0.00200, 0.01725, 0.02550, 0.05375$; their bootstrap 95% intervals are, respectively, $[0, 0.00075]$, $[0.00075, 0.00350]$, $[0.01250, 0.02275]$,

$[0.01975, 0.03175]$, and $[0.04550, 0.06275]$. The paired lower-confidence-planner mean regrets are $0.00025, 0.02500, 0.10150, 0.12025, 0.12475$. Table 6 reports their paired differences. At the Markov cell ($\varepsilon = 0.5$), the two planners have identical regret on all 500 paired seeds, yielding a paired mean difference of 0 and a paired-bootstrap 95% interval of $[0, 0]$. The lower-confidence planner has higher mean regret in every history-dependent cell, with each paired 95% interval excluding zero. Both tested planners exhibit sensitivity to cross-time dependence, despite using different downstream policy-selection rules. The comparison is limited to these two planners on a common fitted reward; neither reproduces a complete published trajectory-supervision method.

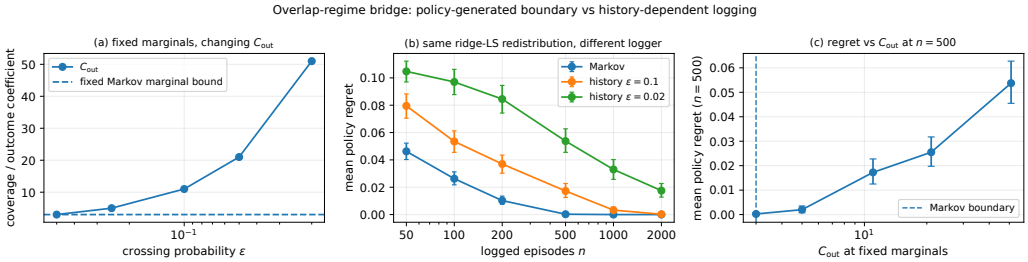

Figure 15: Markov-boundary bridge under matched one-step marginals. Left: $\mathcal{C}_{\text{out}}$ departs from the fixed Markov marginal bound as cross-time dependence increases. Middle: the same ridge-regularized LS trajectory-return redistribution estimator ($\lambda = 0.05$) is evaluated across sample sizes. Right: at $n = 500$, policy regret rises with $\mathcal{C}_{\text{out}}$ although C_{std} and the Markov marginal bound remain fixed.

Table 6: Markov-boundary paired downstream-planning control at $n = 500$, 500 seeds, $\lambda = 0.05$. Each arm uses the same fitted ridge reward and four deterministic policies. Difference intervals are paired-bootstrap 95% intervals across all cells; the difference is lower-confidence minus greedy regret.

ε	Greedy	Lower confidence	Difference	Difference 95% CI
0.50	0.00025	0.00025	0	$[0, 0]$
0.25	0.00200	0.02500	0.02300	$[0.01900, 0.02725]$
0.10	0.01725	0.10150	0.08425	$[0.07800, 0.09025]$
0.05	0.02550	0.12025	0.09475	$[0.08850, 0.10100]$
0.02	0.05375	0.12475	0.07100	$[0.06275, 0.07925]$

Figure 16 (adverse matched-marginal logger memory). We next hold all one-step flag marginals fixed and vary only cross-time coupling on the $H = 8$ compiler-derived additive geometry. The reference Markov logger switches each flag on independently with probability $q = 0.35$. The history-dependent endpoint samples one latent Bernoulli(q) decision for each adjacent flag pair and repeats it within the pair; mixtures indexed by ρ preserve every per-flag marginal exactly. The real conditional-mean table is first least-squares projected onto the additive class and then positively affine-normalized so the per-step surrogate reward lies in $[0, 1]^{16}$ without changing policy ordering. Its raw projection residual has RMS 3.6068 and sup norm 22.8937 in percentage-speedup units; learning uses declared synthetic Gaussian noise with $\sigma = 2$.

The fixed Markov marginal bound is 10.582418 and $C_{\text{std}} = 2.857$ throughout. As ρ increases from 0 to .99, $\mathcal{C}_{\text{out}}$ rises from 10.582 to 443.167, κ_{mem} from 1 to 100, and the target-local factor κ_{rep} from 1 to 99.191. Target-value MAE has correlation 0.9932 with $\sqrt{\mathcal{C}_{\text{out}}}$ and log–log slope 0.547 versus $\mathcal{C}_{\text{out}}$. A separate noiseless $n = 200$ diagnostic shows the sharp finite-sample target-value identified width growing to 2.5676 at $\rho = .99$; at $\rho = 1$ the target leaves the logged span and the population sharp width is 3.9853.

Figure 17 (target-local memory can help). The complementary construction starts from the independent uniform Markov logger and mixes toward a history-dependent logger supported only on the additive-optimal configuration and its bitwise complement. Every flag therefore remains marginally

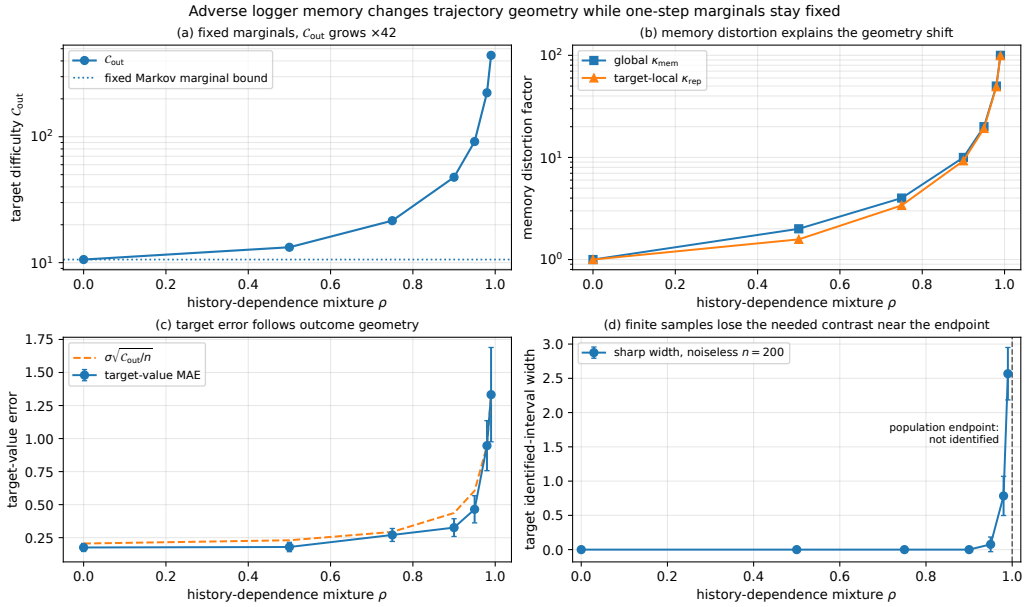

Figure 16: Adverse matched-marginal logger memory on compiler-derived additive geometry. Panel (a): $\mathcal{C}_{\text{out}}$ grows while the Markov marginal bound stays fixed. Panel (b): κ_{mem} and κ_{rep} expose the memory distortion. Panel (c): target-value error follows $\sigma\sqrt{\mathcal{C}_{\text{out}}/n}$. Panel (d): the sharp noiseless finite-sample identified width grows as the needed trajectory contrast becomes rare; the $\rho = 1$ endpoint is population-unidentified.

Bernoulli$(1/2)$ for every ρ, but the cross-time coupling becomes extreme. In closed form,

$$\mathcal{C}_{\text{out}}(\rho) = 1 + \frac{H}{1 + (H-1)\rho}, \qquad \kappa_{\text{mem}}(\rho) = \frac{1}{1-\rho} \quad (\rho < 1).$$

The numerical closed forms agree to 1.05×10^{-11}. Thus increasing memory makes the *global* distortion arbitrarily bad while the fixed target becomes easier: $\mathcal{C}_{\text{out}}$ decreases from 9 to 2.0088 by $\rho = .99$. At $\rho = 1$, $\kappa_{\text{mem}} = \infty$ but the target remains point identified with $\mathcal{C}_{\text{out}} = 2$, $\kappa_{\text{rep}} = 0.5625$, and zero sharp target-value width. This is precisely the regime for which the target-representer refinement is more informative than the global factor.

K Scope and Open Questions

The core outcome-geometry results extend beyond the tabular parameterization. Section G develops the p-dimensional linear-feature formulation under exact realizability and uniformly bounded misspecification. The lower-bound constructions also admit linear-feature representations, including sparse families in which the active feature dimension can remain small while the ambient tabular dimension grows. Extending these guarantees to learned feature maps, unbounded feature misspecification, and nonlinear function classes remains open.

Section I extends the linear observation geometry to general trajectory features and identifies a computational boundary. Pairwise trajectory interactions can make planning NP-hard (Theorem I.3), while bounded temporal interaction range admits an augmented-state dynamic program (Theorem I.5). The gcc case study quantifies whole-configuration additive error through empirical conditional means, and Figure 11 shows that improved L^2 predictive fit can coexist with a worse policy floor. Theorem I.8 characterizes an approximation obstruction for class-measurable procedures, whereas Theorem I.7 gives held-out comparison among realized candidate policies. A population oracle inequality across misspecified classes remains open.

The memory-sensitive reduction of Theorem C.6 quantifies the second-moment change induced by replacing a history-dependent trajectory law with its occupancy-equivalent Markovization. For

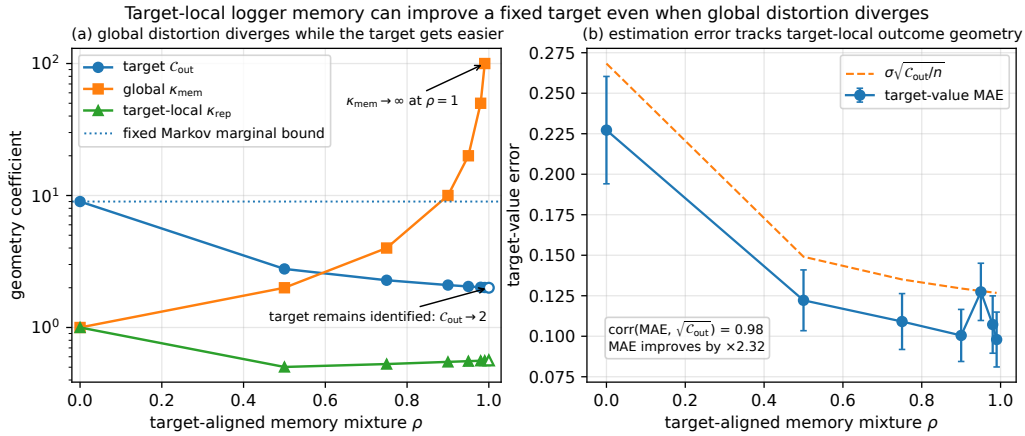

Figure 17: Target-aligned matched-marginal memory on the same compiler-derived additive surrogate. Left: global memory distortion diverges while target $\mathcal{C}_{\text{out}}$ falls; the target-local representer certificate stays finite. Right: target-value MAE is lower near the target-aligned endpoint, with non-monotonic variation across intermediate mixtures. The $\rho = 1$ logger is globally singular but still point-identifies the fixed target.

Markov loggers, the inequality of Yuan et al. (2025) controls $\mathcal{C}_{\text{out}}$ through one-step χ^2 coverage. Proposition C.7 further localizes the comparison to additive representers relevant to a fixed target. When both outcome coefficients are finite, their ratio $\mathcal{C}_{\text{out}}(\pi;\mu)/\mathcal{C}_{\text{out}}(\pi;\bar{\mu})$ gives the exact fixed-target comparison, while κ_{rep} supplies a representer-level upper certificate. Population-valid inference for κ_{mem} and κ_{rep} from a single finite log requires control of population support beyond the realized sample and remains an open statistical problem.

The projected confidence-set method removes the global spectral threshold from the expected learning guarantee: Theorem D.8 gives one-shot expected control from finite $\mathcal{C}_{\text{out}}$, and Theorem D.9 converts the corresponding constant-probability width control into high probability through blockwise confidence amplification. On the temporal direct-sum family, Corollary E.5 gives the exact relation between horizon and logger-memory distortion. Theorem E.9 provides complementary fixed-$\mathcal{C}_{\text{out}}$ hard instances for general even horizons, and Theorem E.10 combines Gaussian testing with horizon-coherent missing exposure to obtain the near-minimax tabular rate under its stated non-clipping conditions. The two mechanisms contribute the scales $\sigma\sqrt{d\mathcal{C}_{\text{out}}/n}$ and $\sqrt{dH\mathcal{C}_{\text{out}}/n}$, respectively. Theorem E.8 gives the sharp $H = 2$ specialization.

The worst-case tabular rate coexists with substantially easier structured instances. On the neutral-padded sparse family, only s reward-relevant directions determine policy regret, and Proposition E.13 shows that exact support recovery is unnecessary under exact neutrality. Approximately neutral directions may reintroduce an ambient search cost; characterizing this adaptive transition remains open. The generic finite-policy refinement replaces the stochastic $\sqrt{d}$ dependence by $\sqrt{\log|\Pi|}$ while retaining the deterministic ridge contribution.

Statistical and computational refinements also expose a remaining trade-off (Remark D.8). The occupancy-polytope optimizer is polynomial-time for the generic ridge radius, while the sharp finite-class radius is formulated for a fixed candidate class. Algorithm 2 supplies a second polynomial-time construction through an SOCP value/supergradient oracle. Establishing a genuine statistical–computational separation would require corresponding computational lower bounds.

Appendix M gives a tabular extension with estimated dynamics. Its transition contribution depends on standard coverage and interacts explicitly with the reward-confidence scale through $\beta/\sqrt{\lambda}$; sharpening the horizon and state dependence is an open direction. The numerical Algorithm 2 results characterize the reference implementation on one CPU and the reported finite-dimensional grid, while Theorem D.11 gives the separate polynomial-time oracle statement. Binary outcomes admit the identification result of Proposition F.1; finite-sample generalized-linear-model rates and the distributional objectives of Remark F.1 remain open.

A further extension concerns multi-agent settings in which other agents' actions are unobserved and only aggregate outcomes are logged. Such data combine trajectory-level credit assignment with partial observability of the joint action process and provide a natural setting for extending the identification analysis.

L Concentration and Testing Lemmas

Lemma L.1 (Subgaussian linear combinations)**.** *Let $\xi_1, \dots, \xi_n$ be independent, each mean-zero and σ-subgaussian, and let $c \in \mathbb{R}^n$ with $\sum_i c_i^2 \le 1$. Then $\sum_i c_i \xi_i$ is σ-subgaussian.*

Proof. For $u \in \mathbb{R}$, by independence, $\mathbb{E}\exp(u\sum_i c_i\xi_i) = \prod_i \mathbb{E}\exp(uc_i\xi_i) \le \prod_i \exp(u^2 c_i^2 \sigma^2/2) = \exp\left(\frac{u^2\sigma^2}{2}\sum_i c_i^2\right) \le \exp(u^2\sigma^2/2)$. □

Lemma L.2 (Subgaussian tail)**.** *If Z is mean-zero σ-subgaussian, then $\mathbb{P}(Z > t) \le e^{-t^2/(2\sigma^2)}$ for all $t \ge 0$.*

Proof. Chernoff: for $u > 0$, $\mathbb{P}(Z > t) \le e^{-ut}\mathbb{E}e^{uZ} \le e^{-ut+u^2\sigma^2/2}$; optimize $u = t/\sigma^2$. □

Lemma L.3 (Nets of the unit sphere)**.** *For every $\epsilon \in (0,1)$, the unit sphere of a k-dimensional Euclidean space admits an ϵ-net N (in Euclidean distance) with $|N| \le (1 + 2/\epsilon)^k$. In particular a $\frac{1}{2}$-net of size at most 5^k exists.*

Proof. Let N be a maximal ϵ-separated subset of the sphere; maximality makes it an ϵ-net (any sphere point within distance ϵ of some net point, else it could be added). The open balls of radius $\epsilon/2$ centered at the points of N are disjoint and contained in the ball of radius $1 + \epsilon/2$ centered at the origin. Comparing volumes in $\mathbb{R}^k$: $|N|(\epsilon/2)^k \le (1 + \epsilon/2)^k$, i.e., $|N| \le (1 + 2/\epsilon)^k$. □

Lemma L.4 (Lower-tail Chernoff for bounded sums)**.** *Let $W_1, \dots, W_n$ be i.i.d. with $W_i \in [0, B]$ and $m := \mathbb{E}W_1 > 0$. Then $\mathbb{P}\big(\sum_i W_i \le \frac{nm}{2}\big) \le e^{-nm/(8B)}$.*

Proof. Set $V_i := W_i/B \in [0,1]$, $\bar{m} := m/B$. For $u > 0$, using $e^{-uv} \le 1 - (1 - e^{-u})v$ for $v \in [0,1]$ (convexity of $v \mapsto e^{-uv}$, equality at $v \in \{0,1\}$) and $1 + x \le e^x$:

$$\mathbb{E}e^{-uV_i} \le 1 - (1 - e^{-u})\bar{m} \le \exp\big(-(1 - e^{-u})\bar{m}\big).$$

Chernoff with threshold $n\bar{m}/2$: $\mathbb{P}(\sum V_i \le n\bar{m}/2) \le \exp\big(un\bar{m}/2 - n(1-e^{-u})\bar{m}\big)$. Take $u = \ln 2$: the exponent is $n\bar{m}(\frac{\ln 2}{2} - \frac{1}{2}) = -n\bar{m}\cdot\frac{1-\ln 2}{2} \le -n\bar{m}/8$ (since $(1 - \ln 2)/2 \approx 0.153 \ge 1/8$). Substitute $\bar{m} = m/B$. □

Lemma L.5 (Design concentration on $\mathcal{U}_\mu$)**.** *Let $k := \dim \mathcal{U}_\mu$ and $\omega_{\min} := \lambda^+_{\min}(\Sigma_\mu)$. If $n \ge n_0(\delta) := \frac{8H}{\omega_{\min}}(k \log \frac{24H}{\omega_{\min}} + \log\frac{1}{\delta})$, then with probability at least $1 - \delta$, the sample second-moment matrix satisfies $\sum_{i=1}^n \phi_i\phi_i^\top \succeq \frac{n}{4}\Sigma_\mu$ as quadratic forms on $\mathcal{U}_\mu$; consequently, for every $v \in \mathcal{U}_\mu$, $\|v\|^2_{\Lambda^{-1}} \le \frac{4}{n}\|v\|^2_{\Sigma_\mu^\dagger}$.*

Proof of Lemma L.5. Write $M := \sum_{i=1}^n \phi_i\phi_i^\top$ and let $\Psi := (\Sigma_\mu^\dagger)^{1/2}$, restricted to $\mathcal{U}_\mu$ where it is the inverse square root of the positive definite operator $\Sigma_\mu|_{\mathcal{U}_\mu}$. Define the whitened features $\psi_i := \Psi\phi_i \in \mathcal{U}_\mu$ (recall $\phi_i \in \mathcal{U}_\mu$ almost surely, Lemma B.2); then $\mathbb{E}[\psi\psi^\top] = \Psi\Sigma_\mu\Psi = I_{\mathcal{U}_\mu}$ (the identity on $\mathcal{U}_\mu$), and

$$\|\psi_i\|_2^2 = \phi_i^\top\Sigma_\mu^\dagger\phi_i \le \frac{\|\phi_i\|_2^2}{\omega_{\min}} = \frac{H}{\omega_{\min}} =: B.$$

The claim $M \succeq \frac{n}{4}\Sigma_\mu$ on $\mathcal{U}_\mu$ is equivalent to $Q := \sum_i \psi_i\psi_i^\top \succeq \frac{n}{4}I$ on $\mathcal{U}_\mu$, i.e., to $\lambda^{\mathcal{U}}_{\min}(Q) := \min_{v\in\mathcal{U}_\mu, \|v\|=1} v^\top Q v \ge n/4$.

Pointwise bound. Fix a unit $v \in \mathcal{U}_\mu$ and set $W_i := \langle\psi_i, v\rangle^2 \in [0, B]$ with $\mathbb{E}W_i = v^\top\mathbb{E}[\psi\psi^\top]v = 1$. By Lemma L.4, $\mathbb{P}\big(v^\top Q v \le n/2\big) \le e^{-n/(8B)} = e^{-n\omega_{\min}/(8H)}$.

Net argument. Let N be an ϵ-net of the unit sphere of $\mathcal{U}_\mu$ with $\epsilon := \omega_{\min}/(8H) \in (0,1)$ and $|N| \le (1+2/\epsilon)^k$ (Lemma L.3). Deterministically, $\|Q\|_{\mathrm{op}} \le \sum_i \|\psi_i\|^2 \le nB$. For any unit $u \in \mathcal{U}_\mu$ and its net approximant v ($\|u-v\| \le \epsilon$),

$$u^\top Q u \ge v^\top Q v - |u^\top Q u - v^\top Q v| \ge v^\top Q v - 2\epsilon\|Q\|_{\mathrm{op}} \ge v^\top Q v - 2\epsilon nB = v^\top Q v - \frac{n}{4},$$

where the middle step uses $|u^\top Q u - v^\top Q v| = |(u-v)^\top Q u + v^\top Q(u-v)| \le 2\|u-v\|\|Q\|_{\mathrm{op}}$ and the last step substitutes $\epsilon B = \omega_{\min}/(8H)\cdot H/\omega_{\min} = 1/8$. On the event $\{\min_{v\in N} v^\top Q v \ge n/2\}$ we conclude $\lambda^{\mathcal{U}}_{\min}(Q) \ge n/2 - n/4 = n/4$.

Union bound. Using $1 + 2/\epsilon = 1 + 16H/\omega_{\min} \le 24H/\omega_{\min}$ (valid because $\omega_{\min} \le \mathrm{tr}\Sigma_\mu = \mathbb{E}\|\phi\|^2 = H$, so $16H/\omega_{\min} \ge 16$ and $1 + x \le \frac{17}{16}x \le \frac{3}{2}x$ for $x \ge 16$),

$$\mathbb{P}\big(\lambda^{\mathcal{U}}_{\min}(Q) < n/4\big) \le |N|\, e^{-n\omega_{\min}/(8H)} \le \exp\Big(k\log\tfrac{24H}{\omega_{\min}} - \tfrac{n\omega_{\min}}{8H}\Big) \le \delta$$

whenever $n \ge n_0(\delta)$ as stated. Finally, on this event, for $v \in \mathcal{U}_\mu$: Λ is block-diagonal with respect to $\mathcal{U}_\mu \oplus \mathcal{U}_\mu^\perp$ (as argued in the proof of Theorem D.4) and $\Lambda|_{\mathcal{U}_\mu} \succeq M|_{\mathcal{U}_\mu} \succeq \frac{n}{4}\Sigma_\mu|_{\mathcal{U}_\mu}$, whence, by the standard order-reversal of inverses of positive definite operators on $\mathcal{U}_\mu$ ($0 \prec X \preceq Y \Rightarrow Y^{-1} \preceq X^{-1}$, a consequence of $Y^{-1/2}XY^{-1/2} \preceq I$), $\|v\|^2_{\Lambda^{-1}} = v^\top(\Lambda|_{\mathcal{U}_\mu})^{-1}v \le \frac{4}{n}\, v^\top(\Sigma_\mu|_{\mathcal{U}_\mu})^{-1}v = \frac{4}{n}\|v\|^2_{\Sigma_\mu^\dagger}$. □

Remark L.1 (The lemma is not specific to trajectory indicators). The proof above uses the horizon H only through the two facts $\|\phi\|^2 \le H$ almost surely and $\mathrm{tr}\Sigma_\mu = \mathbb{E}\|\phi\|^2 = H$. Consequently Lemma L.5 holds verbatim for *any* feature map with $\|\phi\|^2 \le B$ almost surely, with H replaced by B throughout (including in n_0). The same is true of Theorem D.12, whose proof uses only that the observations are linear measurements of an unknown vector corrupted by independent subgaussian noise. We use this in Section E.4 to pass to a reduced coordinate system in which the burn-in, like the rate, is free of the ambient dimension.

The last lemma is the hypercube ("Assouad") reduction used in Theorem E.6. We include a proof since the constant matters for the matching statement of Corollary E.7.

Lemma L.6 (Assouad)**.** *Let $\Theta = \{-1,+1\}^K$ and let $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ be probability measures on a common measurable space. For $\theta \in \Theta$ and $k \in [K]$ write $\theta^{\oplus k}$ for θ with its k-th coordinate flipped. Then for every measurable $\hat\theta$ taking values in $\{-1,+1\}^K$,*

$$\max_{\theta\in\Theta} \mathbb{E}_\theta\big[\rho_{\mathrm{H}}(\hat\theta,\theta)\big] \;\ge\; \frac{K}{2}\Big(1 - \max_{k\in[K],\,\theta\in\Theta} \mathrm{TV}\big(\mathbb{P}_\theta, \mathbb{P}_{\theta^{\oplus k}}\big)\Big).$$

The bound also holds for randomized $\hat\theta$, i.e. measurable functions of (ω, U) with U an independent random seed.

Proof. Bounding the maximum by the average over the 2^K hypercube vertices and exchanging the two sums,

$$\max_\theta \mathbb{E}_\theta[\rho_{\mathrm{H}}(\hat\theta,\theta)] \;\ge\; 2^{-K}\sum_{\theta\in\Theta}\sum_{k=1}^K \mathbb{P}_\theta(\hat\theta_k \ne \theta_k) = \sum_{k=1}^K 2^{-K}\sum_{\theta\in\Theta}\mathbb{P}_\theta(\hat\theta_k \ne \theta_k).$$

Fix k and partition Θ into the 2^{K-1} pairs $\{\theta, \theta^{\oplus k}\}$; in each pair label by θ the element with $\theta_k = +1$. With $A := \{\hat\theta_k = -1\}$,

$$\mathbb{P}_\theta(\hat\theta_k \ne \theta_k) + \mathbb{P}_{\theta^{\oplus k}}(\hat\theta_k \ne \theta^{\oplus k}_k) = \mathbb{P}_\theta(A) + \mathbb{P}_{\theta^{\oplus k}}(A^c) = 1 - \big(\mathbb{P}_{\theta^{\oplus k}}(A) - \mathbb{P}_\theta(A)\big) \ge 1 - \mathrm{TV}\big(\mathbb{P}_\theta, \mathbb{P}_{\theta^{\oplus k}}\big),$$

by the definition of total variation as the supremum of measure differences over events. Summing over the 2^{K-1} pairs,

$$2^{-K}\sum_{\theta\in\Theta}\mathbb{P}_\theta(\hat\theta_k \ne \theta_k) \;\ge\; 2^{-K}\cdot 2^{K-1}\cdot \min_\theta\big(1 - \mathrm{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta^{\oplus k}})\big) \;=\; \tfrac{1}{2}\min_\theta\big(1 - \mathrm{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta^{\oplus k}})\big),$$

and summing over $k \in [K]$ gives the claim. For randomized $\hat\theta$, replace $\mathbb{P}_\theta$ by $\mathbb{P}_\theta \otimes \mathcal{L}(U)$; since $\mathrm{TV}(P\otimes\nu, Q\otimes\nu) = \mathrm{TV}(P,Q)$ for a common factor ν, the right-hand side is unchanged. □

M Policy Learning with Estimated Dynamics

This appendix drops Assumption 3 and estimates P from the same n episodes. Two sources of error then appear: reward uncertainty, still governed by outcome coverage $\mathcal{C}_{\mathrm{out}}$, and transition uncertainty, controlled through observed transitions and visit counts and hence by standard coverage C_{std}. Because transitions are observed, we assume in this appendix that trajectories follow the Markov dynamics (P, ρ), while actions may still be chosen by an arbitrary history-dependent logging rule; transition noise is Markov given (h, s_h, a_h).

Let $N_h(s,a)$ be the visit count of (h,s,a) and $\hat{P}_h(\cdot \mid s,a)$ the empirical kernel (arbitrary, say uniform, on unvisited pairs). For a policy π, let $\hat{d}^\pi$ be its occupancy under $(\hat{P}, \rho)$. Fix $\delta \in (0,1)$ and define

$$L := S\log 2 + \log\big(2d\,n/\delta\big), \qquad c_h(s,a) := \min\Big\{2,\ \sqrt{2L/\max\{N_h(s,a),1\}}\Big\}.$$

Lemma M.1 (ℓ_1 concentration of empirical kernels)**.** *With probability at least $1-\delta/2$, simultaneously for all (h,s,a) with $h \le H-1$: $\|\hat{P}_h(\cdot \mid s,a) - P_h(\cdot \mid s,a)\|_1 \le c_h(s,a)$.*

Proof. For a fixed (h,s,a) and a fixed count value $N = N_h(s,a) = \nu \ge 1$: conditionally on the ν visit times, the observed next states are i.i.d. draws from $P_h(\cdot \mid s,a)$ (Markov transition noise; this is the standard conditioning argument for tabular counts). For any subset $B \subseteq \mathcal{S}$, Hoeffding's inequality for Bernoulli averages gives $\mathbb{P}(\hat{P}(B) - P(B) > t) \le e^{-2\nu t^2}$. Since $\|\hat{P} - P\|_1 = 2\max_{B\subseteq\mathcal{S}}(\hat{P}(B) - P(B))$ (choose $B = \{s' : \hat{P}(s') > P(s')\}$), a union bound over the 2^S subsets yields

$$\mathbb{P}\big(\|\hat{P} - P\|_1 > 2t \,\big|\, \nu\big) \le 2^S e^{-2\nu t^2}.$$

Set $2t = \sqrt{2L'/\nu}$ with $L' := S\log 2 + \log(1/\delta')$: the right side is $\le \delta'$. Union over the at most d triples and the at most n possible values of $\nu \in \{1,\dots,n\}$ with $\delta' := \delta/(2dn)$ gives the claim with $L = S\log 2 + \log(2dn/\delta)$; for $\nu = 0$ the trivial bound $\|\hat{P} - P\|_1 \le 2$ is used, matching the cap in c_h. □

Lemma M.2 (Simultaneous visit-count event)**.** *With L as above, let $\mathcal{E}_{\mathrm{count}}$ be the event that, simultaneously for all (h,s,a) with $h \le H-1$ and $nd_h^\mu(s,a) \ge 8L$,*

$$N_h(s,a) \ge \tfrac{1}{2} n d_h^\mu(s,a).$$

Then $\mathbb{P}(\mathcal{E}_{\mathrm{count}}) \ge 1-\delta/2$. On this event, all such triples satisfy $c_h(s,a) \le \sqrt{4L/(nd_h^\mu(s,a))}$.

Proof. For each fixed triple, the visit indicators across the n episodes are i.i.d. Bernoulli with mean $d_h^\mu(s,a)$, although actions may depend on history within an episode. Lemma L.4, with $B = 1$, bounds the probability of a count below half its mean by $\exp[-nd_h^\mu(s,a)/8] \le e^{-L}$ on the specified triples. A union bound over at most d triples gives probability at most $de^{-L} \le \delta/2$. Substituting $N_h(s,a) \ge nd_h^\mu(s,a)/2$ into the definition of c_h proves the final bound. □

Lemma M.3 (Simulation lemma, symmetric form)**.** *For every policy π and every h, writing $d_h^\pi, \hat{d}_h^\pi \in \Delta(\mathcal{S}\times\mathcal{A})$ for the step-h occupancies under P and $\hat{P}$,*

$$\|\hat{d}_h^\pi - d_h^\pi\|_1 \ \le\ \sum_{g=1}^{h-1} \mathbb{E}_{(s,a)\sim\hat{d}_g^\pi}\big[\|\hat{P}_g(\cdot \mid s,a) - P_g(\cdot \mid s,a)\|_1\big],$$

and the same bound holds with the expectation taken under d_g^π instead.

Proof. Let $\mu_h, \hat{\mu}_h \in \Delta(\mathcal{S})$ be the step-h state distributions; occupancies are $d_h(s,a) = \mu_h(s)\pi_h(a \mid s)$, so $\|\hat{d}_h - d_h\|_1 \le \|\hat{\mu}_h - \mu_h\|_1$ (multiplying by the common $\pi_h(\cdot \mid s)$ and summing over a cannot increase the ℓ_1 distance: $\sum_{s,a}|\hat{\mu}_h(s) - \mu_h(s)|\pi_h(a \mid s) = \sum_s |\hat{\mu}_h(s) - \mu_h(s)|$). We prove $\|\hat{\mu}_h - \mu_h\|_1 \le \sum_{g<h}\mathbb{E}_{\hat{d}_g}\|\hat{P}_g - P_g\|_1$ by induction on h; the base case $h = 1$ is $0 = 0$ ($\hat{\mu}_1 = \mu_1 = \rho$). For the step, decompose

$$\hat{\mu}_{h+1}(s') - \mu_{h+1}(s') = \sum_{s,a}\big(\hat{d}_h(s,a) - d_h(s,a)\big)P_h(s' \mid s,a) + \sum_{s,a}\hat{d}_h(s,a)\big(\hat{P}_h(s' \mid s,a) - P_h(s' \mid s,a)\big).$$

Taking absolute values and summing over s', the first sum is at most $\|\hat{d}_h - d_h\|_1$ because each row of P_h sums to one. The second is

$$\mathbb{E}_{(s,a)\sim\hat{d}_h}\|\hat{P}_h(\cdot \mid s,a) - P_h(\cdot \mid s,a)\|_1.$$

Apply the induction hypothesis and $\|\hat{d}_h - d_h\|_1 \le \|\hat{\mu}_h - \mu_h\|_1$. Swapping which kernel is factored out in the decomposition (write $\hat{d}_h\hat{P}_h - d_hP_h = (\hat{d}_h - d_h)\hat{P}_h + d_h(\hat{P}_h - P_h)$) yields the version with expectations under d_g. □

Theorem M.4 (Data-dependent bound with estimated dynamics)**.** *Define the transition bonus* $b(\pi) := \sum_{h=2}^{H}\sum_{g=1}^{h-1}\mathbb{E}_{(s,a)\sim\hat{d}_g^{\pi}}[c_g(s,a)]$ $\big(\le H\sum_{g=1}^{H-1}\mathbb{E}_{\hat{d}_g^{\pi}}[c_g]\big)$*, computable from data, and run the modified pessimistic rule*

$$\hat{\pi} \in \arg\max_{\pi} \langle \hat{d}^{\pi}, \hat{r}\rangle - \beta(\delta/2)\,\|\hat{d}^{\pi}\|_{\Lambda^{-1}} - b(\pi),$$

with $\hat{r}, \Lambda$ *as in Definition D.1 (these use only observed features and outcomes, hence are unaffected by* $\hat{P}$*). With probability at least* $1-\delta$*,*

$$\mathrm{SubOpt}(\hat{\pi}) \;\le\; 2\beta(\delta/2)\,\|\hat{d}^{\pi^\star}\|_{\Lambda^{-1}} \;+\; 2\,b(\pi^\star).$$

Proof. Intersect the event $\mathcal{E}_1$ of Lemma D.2 at level $\delta/2$ with the event $\mathcal{E}_2$ of Lemma M.1. On $\mathcal{E}_1 \cap \mathcal{E}_2$, for *every* policy π:

$$\begin{aligned}\big|\langle \hat{d}^{\pi}, \hat{r}\rangle - V^{\pi}\big| &\le \big|\langle \hat{d}^{\pi}, \hat{r} - r\rangle\big| + \big|\langle \hat{d}^{\pi} - d^{\pi}, r\rangle\big| \\ &\le \beta(\delta/2)\,\|\hat{d}^{\pi}\|_{\Lambda^{-1}} + \sum_{h=1}^{H}\|\hat{d}_h^{\pi} - d_h^{\pi}\|_1 \cdot \|r\|_\infty \;\le\; \beta(\delta/2)\,\|\hat{d}^{\pi}\|_{\Lambda^{-1}} + b(\pi), \quad (23)\end{aligned}$$

where the first term uses Cauchy–Schwarz in the Λ-norm exactly as in Theorem D.3 (valid for the arbitrary fixed vector $\hat{d}^{\pi}$ once we condition on the data), and the second combines Lemma M.3 (expectations under $\hat{d}$, the computable side), Lemma M.1 to replace each $\|\hat{P}_g - P_g\|_1$ by $c_g(s,a)$, $\|r\|_\infty \le 1$, and the definition of b. Display (23) says the objective of the modified rule is a uniform lower bound on V^{π} (validity), and is at least $V^{\pi} - 2\beta\|\hat{d}^{\pi}\|_{\Lambda^{-1}} - 2b(\pi)$ (accuracy). The chaining argument of Theorem D.3, verbatim with the extra bonus, gives the claim. □

Corollary M.5 (Population rate with the occupancy-estimation interaction)**.** *Assume additionally* $C_{\mathrm{std}}^{\star} := C_{\mathrm{std}}(\pi^\star;\mu) < \infty$*, let* L *be as above, and suppose the burn-in condition of Lemma L.5 holds for* $d^{\pi^\star}$*. Define*

$$T_\star := H\sum_{g=1}^{H-1}\mathbb{E}_{(s,a)\sim d_g^{\pi^\star}}[c_g(s,a)].$$

On the intersection of the confidence, transition-concentration, count, and design events (of probability at least $1-3\delta$ *by the stated confidence allocations and a union bound),*

$$\|\hat{d}^{\pi^\star} - d^{\pi^\star}\|_{1,\mathrm{tot}} \le T_\star, \qquad b(\pi^\star) \le \sum_{j=1}^{H-1}\frac{(H-j)(H-j+1)}{2}\mathbb{E}_{d_j^{\pi^\star}}[c_j] \le \frac{H-1}{2}T_\star,$$

and

$$T_\star \le H^2 C_{\mathrm{std}}^{\star}\left(\sqrt{\frac{4LSA}{n}} + \frac{16SAL}{n}\right).$$

Consequently,

$$\mathrm{SubOpt}(\hat{\pi}) \le 4\beta(\delta/2)\sqrt{\frac{\mathcal{C}_{\mathrm{out}}^{\star}}{n}} + \left(H - 1 + \frac{2\beta(\delta/2)}{\sqrt{\lambda}}\right)T_\star.$$

The transition term depends on the horizon and on standard coverage; its interaction with the reward-confidence radius is explicit. For fixed d *and confidence level with nonzero stochastic noise, when* $T_\star = O(n^{-1/2})$*, the term* $\beta T_\star/\sqrt{\lambda}$ *vanishes if* $n\lambda_n \to \infty$*. Fixed positive* λ *is one sufficient choice under this bound; more generally, the condition requires* λ_n *to decay more slowly than* $1/n$*.*

Proof. By the true-occupancy version of Lemma M.3 and the event of Lemma M.1,

$$\sum_{h=1}^{H} \|\hat{d}_h^{\pi^\star} - d_h^{\pi^\star}\|_1 \le \sum_{h=2}^{H} \sum_{g<h} \mathbb{E}_{d_g^{\pi^\star}}[c_g] \le T_\star.$$

Put $a_g = \mathbb{E}_{d_g^{\pi^\star}}[c_g]$. Since $0 \le c_g \le 2$, for probability distributions p, q, $|\mathbb{E}_p c_g - \mathbb{E}_q c_g| \le \|p - q\|_1$. The true-occupancy simulation bound therefore gives

$$\begin{aligned} b(\pi^\star) &= \sum_{g=1}^{H-1} (H-g) \mathbb{E}_{\hat{d}_g^{\pi^\star}}[c_g] \\ &\le \sum_{g=1}^{H-1} (H-g) \left(a_g + \sum_{j<g} a_j \right) \\ &= \sum_{j=1}^{H-1} \frac{(H-j)(H-j+1)}{2} a_j \le \frac{H-1}{2} T_\star. \end{aligned}$$

For the population bound on $T_\star$, use the same count split as in the standard tabular analysis. On triples satisfying $n d_g^\mu(s,a) \ge 8L$, the simultaneous count event of Lemma M.2 gives $c_g(s,a) \le \sqrt{4L/(n d_g^\mu(s,a))}$, so

$$\sum_{s,a} d_g^{\pi^\star}(s,a) c_g(s,a) \le C_{\mathrm{std}}^\star \sqrt{\frac{4LSA}{n}}.$$

On the remaining triples, $c_g \le 2$ and $d_g^{\pi^\star} \le C_{\mathrm{std}}^\star d_g^\mu < 8 C_{\mathrm{std}}^\star L/n$, so their total contribution is at most $16 C_{\mathrm{std}}^\star SAL/n$. Summing over g and using the outer factor H proves the displayed bound on $T_\star$.

Finally Theorem M.4 and the triangle inequality give

$$\mathrm{SubOpt}(\hat{\pi}) \le 2\beta \|d^{\pi^\star}\|_{\Lambda^{-1}} + \frac{2\beta}{\sqrt{\lambda}} \|\hat{d}^{\pi^\star} - d^{\pi^\star}\|_{1,\mathrm{tot}} + 2b(\pi^\star).$$

Lemma L.5 bounds the first norm by $2\sqrt{\mathcal{C}_{\mathrm{out}}^\star/n}$, and the first two displays above bound the remaining terms, yielding the claim. □

The decomposition is therefore slightly less separable than in the known-dynamics case: $\mathcal{C}_{\mathrm{out}}$ controls the reward/credit-assignment term, while C_{std} controls occupancy estimation. Evaluating the reward confidence radius at an estimated occupancy introduces the explicit $\beta/\sqrt{\lambda}$ interaction in Corollary M.5.