THE PRICE OF IMPERFECT PROCESS REWARDS: BIAS FLOORS AND FINITE-BUDGET REVERSALS

Anonymous authors
Paper under double-blind review

ABSTRACT

A learned process reward model (PRM) provides feedback at each step, whereas its reward estimates may be inaccurate, which can lead to suboptimal policies. We compare policy learning from a fixed PRM with learning from unbiased outcome feedback under the same downstream rollout budget. On task families where the PRM preserves most preferences, we establish a reversal through matching bounds: process feedback has lower minimax regret at small positive budgets, while outcome feedback has lower minimax regret at large budgets. With an unknown transition model, the minimax regret of the process learner decreases toward an exact bias floor. Our lower bound accounts for the transition estimation error, establishing the reversal even when actions have an arbitrarily small effect on transition probabilities. In the known-transition family, for trajectories of H steps, the crossover budget scales as $\Theta(H^2/b^2)$ for the largest process regret b over a range of per-stage reward gap. For finite tabular MDPs with compact convex occupancy sets, we give exact bounds on the loss in cumulative true reward under policy comparisons with a specified tolerance in PRM value. These bounds account for potential shaping and reward scale, and policy constraints can make these class-specific certificates arbitrarily smaller than the corresponding global STARC guarantees. For a weighted combination of process and outcome rewards, the bias term in the regret bound is proportional to the PRM weight. Our tabular experiments vary PRM accuracy at a fixed true reward, examining how horizon length and transition estimation affect policy regret.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) uses a terminal verifier to evaluate completed responses. This signal provides little direct guidance about which intermediate decisions contributed to success (DeepSeek-AI, 2025; Shao et al., 2024). Learned process reward models (PRMs) provide feedback at each step and can guide policy optimization in long reasoning tasks (Lightman et al., 2024; Uesato et al., 2022; Wang et al., 2024). Existing theory compares the statistical efficiency of per-step and terminal feedback (Chen et al., 2025; Jia et al., 2025). Evaluations on ProcessBench reveal limitations in PRMs' ability to identify erroneous steps in mathematical reasoning (Zheng et al., 2025). The value of these intermediate scores depends on how they guide policy learning. Systematic overvaluation of particular steps by a PRM can steer policy optimization toward those steps, thereby altering the distribution of subsequent trajectories. Errors that appear small in isolated evaluations can hence have a larger effect on true performance (Pan et al., 2022; Gao et al., 2023; Laidlaw et al., 2025).

We compare two downstream training protocols on the same task family under a common rollout budget. The process learner uses a fixed PRM, while the outcome learner receives an unbiased terminal observation for each trajectory. The budget excludes the data used to train the PRM. We ask which protocol achieves lower policy regret as the training budget grows. When the transition model is unknown, both learners estimate how actions affect state transitions by collecting trajectories. For the process learner, more rollouts cannot help distinguish between different preferences of the true reward and the PRM. Outcome feedback can resolve this ambiguity, whereas each observation provides less guidance about intermediate decisions. The lower policy regret of the process learner can therefore persist until the outcome learner estimates reward preferences more accurately than the PRM.

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.

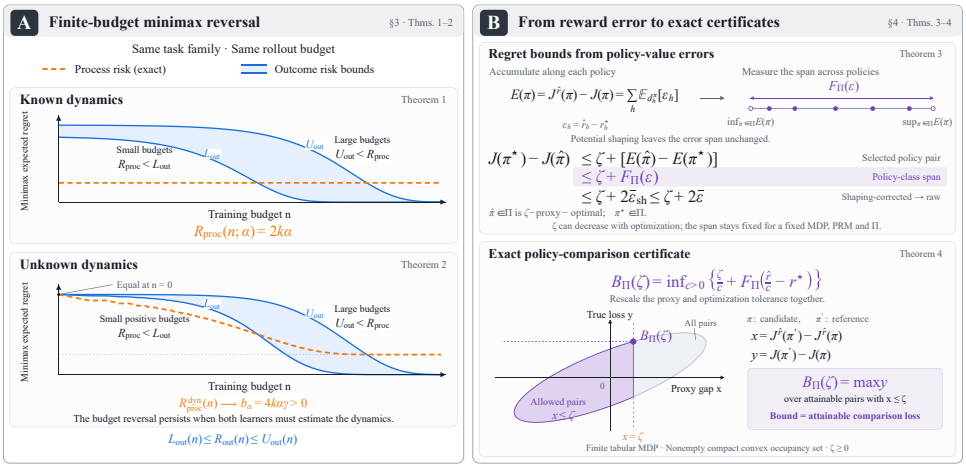

Figure 1: **Theoretical overview.** (A) On the constructed task families, the process–outcome comparison reverses with rollout budget. (B) The fixed-PRM bound and the exact scaled loss.

We construct task families in which the PRM preserves most local preferences with the ture reward and reverses some at unknown stages. With an unknown transition model, both learners estimate how actions affect state probabilities. The policy regret of the process learner decreases toward a nonzero bias floor as these estimates improve, while outcome learning eventually achieves lower regret. In section 3, we prove strict minimax comparisons at small and large budgets on the same task family. The bounds include a common term for errors in estimating the transition model. The reversal holds even when actions have an arbitrarily small effect on transition probabilities.

We relate the bias floor to reward error through the span of accumulated error over the policy class, denoted by F_Π (Section 4). A common shift in all policy values preserves their ordering, whereas unequal errors can make a policy with lower true return optimal under the PRM. The resulting bounds separate policy optimization error under the PRM from the loss caused by incorrect reward preferences. Potential shaping and positive scaling preserve policy ordering (Ng et al., 1999; Gleave et al., 2021; Skalse et al., 2023), hence we account for both when comparing rewards. For finite tabular MDPs with compact convex occupancy sets, we optimize the bound over positive PRM scales to obtain the maximum true-return loss allowed by an approximate PRM comparison. The hybrid objective combines process and outcome feedback, and the bias term in its regret bound is proportional to the PRM weight (Section 5). The experiments examine how PRM accuracy, horizon, and transition estimation affect the learning curves (Section 6). Figure 1 summarizes the finite-budget reversal, exact policy-comparison certificates, and the effect of policy constraints. Due to space constraints, we defer the discussion of related work to Appendix A.

Our contributions are summarized as follows.

- We prove that process feedback has lower minimax regret at small budgets, while outcome feedback has lower regret at large budgets. On the same task families with informative, imperfect PRMs, we give matching crossover bounds and characterize the exact bias floor under unknown transition model.
- We derive exact policy-comparison certificates for a specified policy class and PRM tolerance. They account for reward scale and potential shaping. Under policy constraints, they can be arbitrarily smaller than the corresponding global STARC guarantee even with bounded coverage ratios, while under the full class and zero tolerance they remain within a constant factor of the return-span normalization.
- We give regret guarantees for hybrid training with process and outcome feedback. The bias term is proportional to the PRM weight, and our budget-dependent hybrid learner achieves lower mean regret than both single-signal estimators on the fixed task used in our experiments.

2 PRELIMINARIES

In this section, we introduce the MDP framework and the two feedback protocols considered in this paper. We then define the policy regret and explain how the learning problem changes depending on whether the transition model is known.

2.1 MDP SETUP AND FEEDBACK PROTOCOLS

A Markov decision process (MDP) is specified by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \{P_h\}, \{r_h^\star\}, \rho)$ (Puterman, 1994). Here, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, and H is the horizon. For each stage $h \in \{1, \dots, H\}$, $P_h(\cdot \mid s, a)$ denotes the transition model, and $r_h^\star : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the expected true reward for that state–action pair. The distribution ρ specifies how the initial state is sampled. A trajectory $\tau = (s_1, a_1, \dots, s_H, a_H, s_{H+1})$ records the sequence of states and actions in a rollout, including the final state s_{H+1}, which is the deterministic terminal state. We define the cumulative reward along the trajectory as $R(\tau) = \sum_{h=1}^{H} r_h(s_h, a_h)$ under some reward model $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$, which can be the true reward or any other learned reward model. A Markov policy $\pi = \{\pi_h\}_{h=1}^{H}$ is a sequence of maps $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ denotes the space of probability distributions over actions. Starting from $s_1 \sim \rho$, a trajectory is generated by sampling $a_h \sim \pi_h(\cdot \mid s_h)$ and then $s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$ at each stage. Given a policy π, we denote the expected cumulative reward by:

$$J^r(\pi) = \mathbb{E}_{\tau \sim \pi}[R(\tau)] = \sum_{h=1}^{H} \mathbb{E}_{(s,a) \sim d_h^\pi}[r_h(s, a)], \tag{1}$$

where d_h^π is the state–action occupancy distribution at stage h, with $d_h^\pi(s, a) := \mathbb{P}^\pi(s_h = s, a_h = a)$. We abbreviate $J^{r^\star}(\pi)$ as $J(\pi)$ and assume that an optimal policy $\pi^\star \in \arg\max_\pi J(\pi)$ exists. We also assume that $R(\tau) \in [0, 1]$ for any trajectory.

In this paper, we focus on two feedback protocols in reinforcement learning: outcome feedback and process feedback. Outcome reward provides feedback for a completed trajectory (Jia et al., 2025). The learner observes the pair (τ, Y), where $Y \in [0, 1]$ is a single value of the outcome reward received after the final step. We allow this observation to be noisy and assume $\mathbb{E}[Y \mid \tau] = R(\tau)$, with $Y = R(\tau)$ in the noiseless case. Thus, the expected outcome reward is equal to the policy's expected cumulative true reward, i.e., $\mathbb{E}_{\pi,Y}[Y] = J(\pi)$. Under the process feedback protocol, the learner has access to a fixed learned PRM $\hat{r} = \{\hat{r}_h\}$. The PRM provides a reward at each step. For complex multi-step reasoning tasks, this finer-grained feedback can help identify suboptimal decisions and guide corrections to intermediate errors, even when the reasoning leads to a correct final answer (Lightman et al., 2024; Uesato et al., 2022; Wang et al., 2024).

In practice, a learned PRM generally includes estimation errors. Summation-based credit assignment can exacerbate reward hacking in PRM-guided policy optimization (Cheng et al., 2025). The relevant error is the deviation of the additive reward used for policy optimization from the true reward:

$$\hat{r}_h(s, a) = r_h^\star(s, a) + \varepsilon_h(s, a), \qquad \varepsilon_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}. \tag{2}$$

We measure the downstream training budget n by the number of complete trajectories collected for policy learning, excluding the data used to train the PRM. The PRM will remain fixed throughout policy training.

2.2 POLICY REGRET AND TRAINING BUDGET

We measure performance by the expected policy regret $\mathbb{E}[J(\pi^\star) - J(\hat{\pi})]$, where the expectation is taken over the training data and any randomization. At the same budget n, we compare the two feedback protocols through their minimax risk on the same task family. For a fixed learning algorithm, the worst-case risk is the largest expected regret over the tasks in that family. We then minimize this worst-case expected regret over all learning algorithms allowed by the protocol and budget. The learning problem also depends on whether the transition model $\{P_h\}$ is known. With known transition dynamics, the trained PRM is enough to select a PRM-optimal policy before collecting a subsequent policy-training trajectory for the process learner. With unknown transition dynamics, the transition model is not available to the learners. A trained PRM may assign a high reward to a state

while leaving the learner uncertain about which action is most likely to reach it. The outcome learner receives outcome feedback from which it can learn the true reward preferences.

3 Process and Outcome Learning

The fixed PRM supplies the process learner with step-level reward information before downstream training begins. We construct task families to determine how the value of this information changes with the training budget, first with known transitions and then with unknown transitions.

3.1 Learning with Known Transition Model

To isolate the role of reward information, consider an action-independent chain with $H \geq 4$ binary decisions $x_h \in \{-1, +1\}$. Every rollout visits the same sequence of states, so the reward preferences determine the optimal actions. Fix $1 \leq k \leq H/4$ and $0 < \alpha \leq 1/(4H)$. Each task has true action preferences $\omega \in \{-1, +1\}^H$ and PRM preferences z that differ from ω at exactly k unknown stages. We define the rewards by

$$r_h^{\omega}(x_h) = \tfrac{1}{2H} + \alpha\omega_h x_h, \qquad \hat{r}_h^{z}(x_h) = \tfrac{1}{2H} + \alpha z_h x_h. \tag{3}$$

Here k counts the incorrect PRM preferences, while α sets the action-reward gap 2α. The action opposite to ω_h incurs a true-return loss of exactly 2α at stage h. The outcome feedback follows $Y \mid x \sim \mathrm{Bernoulli}(1/2 + \alpha\langle\omega, x\rangle)$, so α also controls the outcome signal. Both cumulative returns lie in $[1/4, 3/4]$. For each protocol $q \in \{\mathrm{proc}, \mathrm{out}\}$, let $\mathcal{R}_q(n; \alpha)$ be the smallest worst-case expected regret over possibly randomized adaptive algorithms that use at most n trajectories and return a Markov policy. The two risks use the same task family in (3), with pairs (ω, z) that differ at exactly k stages. The theorem compares the exact process risk with outcome-feedback bounds and identifies the budgets at which each protocol has lower minimax regret.

Theorem 1 (Finite-budget feedback separation)**.** *For the family in* (3) *and every* $n \geq 0$*,*

$$\mathcal{R}_{\mathrm{proc}}(n; \alpha) = 2k\alpha,$$
$$\tfrac{27}{32} H\alpha e^{-64n\alpha^2/3} \leq \mathcal{R}_{\mathrm{out}}(n; \alpha) \leq H\alpha e^{-2n\alpha^2}. \tag{4}$$

There are universal constants $0 < c < C$ *such that process feedback has strictly smaller minimax regret for* $n \leq c\alpha^{-2}\log(H/k)$*, and outcome feedback has strictly smaller minimax regret for* $n \geq C\alpha^{-2}\log(H/k)$*.*

The process risk is constant because trajectories with different error locations have the same process observations. The learner cannot identify the k incorrect stages, so following the PRM is minimax optimal and the risk stays at the bias floor $2k\alpha$. On the other hand, outcome feedback reveals the true preferences. If one preference is reversed, the Bernoulli mean changes by 2α, hence identifying that preference requires $\Theta(\alpha^{-2})$ observations. Fewer PRM errors lower the process floor and therefore delay the reversal. We show the proof and the expected regret of the estimator in Appendices B.2–B.4.

3.2 Learning with Unknown Transition Model

Under unknown transition model, the learner must also learn which actions reach high-reward states. We therefore compare the two protocols on a family in which both use training trajectories. Divide the horizon into m decision–reward blocks with $H = 2m$. In block j, the learner chooses $x_j \in \{-1, +1\}$ and then observes reward state u_{j,v_j} with label $v_j \in \{-1, +1\}$. The transition signs $\theta \in \{-1, +1\}^m$ determine how actions affect these states, while the reward preferences $\omega \in \{-1, +1\}^m$ determine which state is better. The transition probabilities and true rewards are

$$\mathbb{P}(v_j = v \mid x_j) = \tfrac{1}{2} + \gamma\theta_j x_j v, \qquad r_{2j}^{\omega}(u_{j,v}, a) = \tfrac{1}{2m} + \alpha\omega_j v. \tag{5}$$

The parameter γ controls the effect of an action on state probabilities, while the two reward states differ by 2α. Decision states have zero reward, and actions at reward states have identical effects, so only the decision actions affect policy value.

The true-optimal action is $x_j = \theta_j\omega_j$; the opposite action incurs regret $4\alpha\gamma$. Each rollout provides one noisy observation $x_j v_j$ of θ_j, and we can estimate its sign by majority voting. The process

learner combines this estimate with z_j, while the outcome learner also estimates ω_j from state labels and outcome feedback. Let q_n be the probability that majority voting gives the wrong transition sign after n trajectories, with uniform tie-breaking. Even with known reward preferences, transition uncertainty results in regret $D_n = 4m\alpha\gamma q_n$. The theorem retains this shared term when comparing the two protocols.

Theorem 2 (Feedback separation with unknown dynamics)**.** *Let $m \geq 4$, $1 \leq k \leq m/4$, $0 < \gamma \leq 1/4$, and $0 < \alpha \leq 1/(4m)$. For every $n \geq 0$, define $q_n = \mathbb{P}[\mathrm{Bin}(n, 1/2+\gamma) < n/2] + \frac{1}{2}\mathbb{P}[\mathrm{Bin}(n, 1/2+\gamma) = n/2]$, $s_n = 1 - 2q_n$, and $D_n = 4m\alpha\gamma q_n$. Then*

$$\begin{aligned} \mathcal{R}^{\mathrm{dyn}}_{\mathrm{proc}}(n;\alpha,\gamma) &= 4\alpha\gamma\,[k + (m-2k)q_n], \\ \mathcal{R}^{\mathrm{dyn}}_{\mathrm{out}}(n;\alpha,\gamma) &\geq D_n + \tfrac{27}{16} s_n m\alpha\gamma e^{-64n\alpha^2/3}, \\ \mathcal{R}^{\mathrm{dyn}}_{\mathrm{out}}(n;\alpha,\gamma) &\leq D_n + 2s_n m\alpha\gamma e^{-2n\alpha^2}. \end{aligned} \tag{6}$$

For universal $0 < c < C$, process has strictly smaller minimax regret for $1 \leq n \leq c\alpha^{-2}\log(m/k)$, and outcome has strictly smaller minimax regret for $n \geq C\alpha^{-2}\log(m/k)$.

On the $m-k$ blocks where the PRM is correct, the process learner chooses the suboptimal action when its transition estimate is wrong, with probability q_n. On the k incorrect blocks, it chooses the suboptimal action when the transition estimate is correct, with probability $1 - q_n$. Since most PRM preferences are correct, the process risk decreases with n and approaches the bias floor $b_\alpha = 4k\alpha\gamma$. We show the lower bound by conditioning on the transition observations in Appendix C.3.

4 Reward Error and Policy Regret

We fix the MDP, PRM, and policy class, and bound the resulting loss through differences in policy-value error, following the view that reward error should be evaluated through the policies and occupancies it affects (Fluri et al., 2025; Laidlaw et al., 2025). We consider the policy ordering based on expected cumulative rewards. The bounds then account for reward transformations that leave policy ordering unchanged and for constraints that restrict the available policy comparisons.

4.1 Policy-Relevant Error Span

Consider two policies with true values 0.6 and 0.5. A common shift of 0.1 preserves their ordering, whereas errors of -0.1 and $+0.1$, respectively, reverse it. Both PRMs have the same maximum absolute value error, but the second favors the policy with lower true value. The difference between the accumulated errors is therefore crucial for policy selection. The sum of the expected per-step errors under a policy gives the error in its value under the PRM. By linearity of expectation, for a policy π, we have

$$J^{\hat{r}}(\pi) - J(\pi) = \mathcal{E}(\pi), \qquad \mathcal{E}(\pi) := \sum\nolimits_{h=1}^{H} \mathbb{E}_{d_h^\pi}[\varepsilon_h]. \tag{7}$$

Thus, the PRM changes the value difference between π and π' by $\mathcal{E}(\pi) - \mathcal{E}(\pi')$. To bound this change for any learned policy, we take the supremum over all pairs in the policy class.

Definition 1 (Policy-relevant error span)**.** *For a nonempty policy class Π, define*

$$\mathrm{F}_\Pi(\varepsilon) := \sup_{\pi,\pi' \in \Pi} \big[\mathcal{E}(\pi) - \mathcal{E}(\pi')\big] = \sup_{\pi,\pi' \in \Pi} \sum\nolimits_{h=1}^{H} \big(\mathbb{E}_{d_h^\pi} - \mathbb{E}_{d_h^{\pi'}}\big)[\varepsilon_h]. \tag{8}$$

Equivalently, $\mathrm{F}_\Pi(\varepsilon) = \sup_{\pi\in\Pi} \mathcal{E}(\pi) - \inf_{\pi\in\Pi} \mathcal{E}(\pi)$. This is the range of policy-value errors over the policies available to the optimizer.

The span is zero whenever all policies in the class have the same accumulated error. Errors on state–action pairs that no policy in the class can reach do not affect the span. Constraints on action probabilities can limit the span even when every action remains reachable. For example, if the policy class requires equal probabilities for two actions at a state, equal and opposite reward errors cancel in every policy value. Both actions can remain reachable while their errors contribute nothing to the span.

For comparison, we write $\bar{\varepsilon} = \sum_h \|\varepsilon_h\|_\infty \le H\varepsilon_{\mathrm{prm}}$, where $\varepsilon_{\mathrm{prm}} = \max_h \|\varepsilon_h\|_\infty$. This quantity uniformly bounds the absolute PRM-induced value error of every policy, yielding $\mathrm{F}_\Pi(\varepsilon) \le 2\bar{\varepsilon}$. This guarantee is conservative because it depends only on error magnitudes and does not account for whether the errors change the relative policy ordering.

4.2 Shaping-Invariant Regret Bounds

Potential-based reward shaping is a classical way to redistribute reward across stages while preserving policy ordering (Ng et al., 1999; Wiewiora, 2003; Gleave et al., 2021; Skalse et al., 2023; Gupta et al., 2022). Ng et al. established the policy-invariance result for shaping terms generated by state potentials (Ng et al., 1999). In our setting, this invariance separates policy-independent shifts in value from reward errors that alter relative policy values.

Let $\{\Phi_h\}$ be bounded state potentials satisfying $\Phi_{H+1} = 0$. The corresponding shaping reward is $g_h^\Phi(s,a) = \mathbb{E}_{s'\sim P_h(\cdot|s,a)}[\Phi_{h+1}(s')] - \Phi_h(s)$. Under any policy, its expected cumulative contribution is $-\mathbb{E}_\rho \Phi_1(s_1)$, which is independent of the policy. The span hence remains unchanged after we subtract g^Φ from the error. We denote the shaping subspace by $\mathcal{G}$ and give the definition of this subspace and the cancellation identity in Appendix D.2. We define the shaping-corrected residual by

$$\bar{\varepsilon}^{\mathrm{sh}} \;=\; \inf_{\{\Phi_h\}:\,\Phi_{H+1}\equiv 0} \sum_{h=1}^{H} \left\|\varepsilon_h - g_h^\Phi\right\|_\infty \;\le\; \bar{\varepsilon}. \tag{9}$$

The residual measures how closely a potential shaping term approximates the reward error. The choice $\Phi = 0$ recovers the uniform error bound. For each choice of potential, the sum of the residual norms bounds the absolute accumulated residual error for each policy. The bound gives the factor of two for a pair of policies.

To include optimization error, we suppose $J^{\hat{r}}(\hat{\pi}) \ge \sup_{\pi\in\Pi} J^{\hat{r}}(\pi) - \zeta$. We call such a policy ζ-PRM-optimal (Appendix D.1). The tolerance ζ measures the optimization error. When $\zeta = 0$, the bound isolates the regret induced by the PRM preferences, which the error span bounds.

Theorem 3 (Regret bound for a fixed PRM)**.** *Let $\hat{\pi}$ be any ζ-PRM-optimal policy for the learned process reward $\hat{r}$, and suppose that $\pi^\star \in \Pi$. Then*

$$J(\pi^\star) - J(\hat{\pi}) \;\le\; \zeta \;+\; \big[\mathcal{E}(\hat{\pi}) - \mathcal{E}(\pi^\star)\big] \;\le\; \zeta + \mathrm{F}_\Pi(\varepsilon) \;\le\; \zeta \;+\; 2\,\bar{\varepsilon}^{\mathrm{sh}} \;\le\; \zeta + 2\bar{\varepsilon}. \tag{10}$$

The first inequality decomposes the regret into the PRM optimization tolerance and the value-error difference between $\hat{\pi}$ and $\pi^\star$. The supremum over policy pairs gives the span bound. The last two inequalities follow from uniform per-step norms and do not use policy occupancies. Proof details and tabular computation are given in Appendices D.4 and D.3. Additional training trajectories can reduce ζ, whereas the error bounds remain fixed for a given PRM and policy class. The bound therefore separates the two effects discussed in Section 3.2: optimization of the PRM reward by the learner and regret induced by PRM preferences.

Connection to the bias floors. In the task family with unknown transition model, the raw error bound is $\bar{\varepsilon} = 2k\alpha$, while the shaping-corrected residual is $\bar{\varepsilon}^{\mathrm{sh}} = 4k\alpha\gamma$. This residual equals the limiting process regret and reflects the effect of actions on the probability of reaching states that the PRM misranks. Appendix F derives these quantities and shows that no positive PRM scale gives a smaller residual. The certificate therefore recovers the bias scale used in the finite-budget comparison.

Horizon dependence and tightness. Aligned errors accumulate across stages and can produce a worst-case floor of order $H\varepsilon_{\mathrm{prm}}$. This floor matches the horizon dependence of the upper certificate, even after optimization over potential shaping and positive PRM scales (Appendix G.1). Small, repeated PRM errors can therefore cause persistent regret even when additional training removes the optimization error.

4.3 Exact Certificates and Policy Constraints

The optimization tolerance ζ is measured on the same scale as the cumulative PRM return. A positive rescaling of all PRM rewards multiplies every PRM value gap by the same factor. Thus, a policy that

is ζ-optimal for $\hat{r}$ is (ζ/c)-optimal for $\hat{r}/c$ for any $c > 0$. Theorem 3 therefore yields the following bound at scale c:

$$J(\pi^\star) - J(\hat{\pi}) \le \inf_{c>0} \left\{ \frac{\zeta}{c} + \mathrm{F}_\Pi\left(\frac{\hat{r}}{c} - r^\star\right) \right\} \le \zeta + \mathrm{F}_\Pi(\varepsilon), \tag{11}$$

where the last inequality uses $c = 1$. We denote this infimum by $\mathcal{B}_\Pi(\zeta)$. This certificate is unchanged by a common positive rescaling of the PRM reward and optimization tolerance, or by potential shaping of either reward. At zero tolerance, $\mathcal{B}_\Pi(0) = \mathrm{F}_\Pi^{\mathrm{ord}} \le 2\bar{\varepsilon}^{\mathrm{ord}}$, where $\mathrm{F}_\Pi^{\mathrm{ord}}$ and $\bar{\varepsilon}^{\mathrm{ord}}$ are the scale-optimized span and residual defined in Appendix D.3. Both zero-tolerance certificates vanish when $\hat{r} = c_0 r^\star + g^\psi$ with $c_0 > 0$. Appendix D.4 gives the proof.

The bound in (11) applies whenever a candidate policy has a PRM value at most ζ below that of a reference policy. To assess its tightness, we fix the rewards and maximize true-return loss over all policy pairs that satisfy this comparison. When the occupancy set is compact and convex, this maximum equals $\mathcal{B}_\Pi(\zeta)$. This condition holds for the full tabular Markov class and for mixtures of finitely many fixed policies.

Theorem 4 (Exact policy-comparison certificate)**.** *In a finite tabular MDP, suppose that the stacked occupancy set $K_\Pi = \{d^\pi : \pi \in \Pi\}$ is nonempty, compact, and convex. For every $\zeta \ge 0$,*

$$\mathcal{B}_\Pi(\zeta) = \max_{\substack{\pi,\pi' \in \Pi: \\ J^{\hat{r}}(\pi) \ge J^{\hat{r}}(\pi') - \zeta}} \big[J(\pi') - J(\pi) \big]. \tag{12}$$

The exactness concerns the largest loss over all policy pairs that satisfy the PRM comparison constraint. A particular approximately PRM-optimal output is a special case and can have smaller loss because it must also be close to the best PRM value in the class.

When $\pi^\star \in \Pi$, we use it as the reference policy in (12) to obtain the regret guarantee for approximate PRM optimization. In the proof, we write the occupancy difference as $v = d^{\pi'} - d^\pi$ and maximize $\langle r^\star, v \rangle$ subject to $\langle \hat{r}, v \rangle \le \zeta$. The inverse PRM scale $1/c$ serves as the Lagrange multiplier for this constraint and converts the PRM tolerance into true-return units. Appendix E.1 gives the detailed proof and explains the role of convexity.

Policy constraints and the STARC separation. At zero tolerance, STARC bounds the true-return loss in these comparisons over the full policy class (Skalse et al., 2024). We use the return span as both the normalization norm and the distance norm. After rescaling by the true-return span, the resulting distance lies within a factor of two of $\mathcal{B}_{\mathrm{all}}(0)$ (Appendix E.2). A restricted policy class Π excludes some comparisons and can yield a certificate arbitrarily smaller than the global guarantee. This unbounded separation compares a restricted class with the global full-class guarantee; under the same full policy class and zero tolerance, the two quantities remain within the constant-factor relation above. We also show a four-action construction with fixed rewards, true-return range one, and coverability at most four in Appendix E.3. The uniform policy provides a reference distribution with occupancy ratios of at most four.

5 Hybrid Policy Learning

On the task families in Section 3, the protocol with lower minimax regret changes with the training budget. Prior work combines process and outcome feedback in policy optimization (Setlur et al., 2025; Ding et al., 2026). We use Theorem 3 to quantify the contribution of PRM error when a learner combines both signals. For fixed reward scales and a PRM weight $\lambda \in [0, 1]$, we define the hybrid training objective as

$$J_\lambda(\pi) := (1 - \lambda) J(\pi) + \lambda J^{\hat{r}}(\pi) = J(\pi) + \lambda \mathcal{E}(\pi). \tag{13}$$

The hybrid reward has error $\lambda\varepsilon$, so Theorem 3 gives the following guarantee.

Corollary 5 (Hybrid reward bound)**.** *Let $\lambda \in [0, 1]$, $\zeta \ge 0$, and suppose that $\pi^\star \in \Pi$. If $\hat{\pi}_\lambda \in \Pi$ satisfies $J_\lambda(\hat{\pi}_\lambda) \ge \sup_{\pi \in \Pi} J_\lambda(\pi) - \zeta$ for $J_\lambda = J + \lambda\mathcal{E}$, then*

$$J(\pi^\star) - J(\hat{\pi}_\lambda) \le \zeta + \lambda \mathrm{F}_\Pi(\varepsilon) \le \zeta + 2\lambda\bar{\varepsilon}^{\mathrm{sh}}. \tag{14}$$

The bias term is proportional to λ: it equals the process-only term at $\lambda = 1$ and vanishes at the outcome-only endpoint $\lambda = 0$.

We prove the corollary in Appendix H.1. Table 1 compares this learner with the process majority and outcome moment estimators on a fixed task from the family in Section 3.2. All three methods use the same training budget, and the hybrid learner achieves lower mean regret than both baselines at all five evaluated budgets. The experimental details are given in Appendix I.1.4.

Table 1: Mean regret ($\times 10^{-3}$) on the fixed task with $m = 8$, $k = 1$, $\alpha = 1/256$, and $\gamma = 1/4$. Entries report means over 2,048 independent runs.

Budget n	Process	Outcome	Hybrid (Algorithm 1)
4,096	3.906	9.697	3.832
8,192	3.906	7.627	3.462
16,384	3.906	4.982	2.640
32,768	3.906	2.413	1.404
65,536	3.906	0.685	0.441

6 Experiments

The main experiments test the learning-curve predictions in Section 3. We vary PRM accuracy at a fixed true reward, compare sample requirements across horizons, and evaluate both learners under unknown dynamics. We measure performance by regret, $J(\pi^\star) - J(\hat{\pi})$, and report its mean and sample standard deviation across independent runs. Additional results on potential shaping are reported in Appendix I.2. We also evaluate the certificates from Section 4 on the same random MDPs, including the exact primal–dual certificate from Theorem 4.

For the 512 random MDPs, Table 2 compares the exact certificate with the best of the raw-error, shaping-corrected, fixed-scale span, and return-span-normalized bounds. The exact certificate agrees with the direct primal problem to numerical precision and is consistently tighter than these bounds. The improvement is modest but increases with the comparison tolerance.

6.1 PRM Accuracy at a Fixed True Reward

We set $H = 32$ and $\alpha = 1/128$ and fix one true sign vector in (3). The PRMs have $k \in \{1, 2, 4, 8\}$ incorrect rankings, and each larger error set contains the smaller ones. All comparisons use the same true task and the same 512 independently generated training datasets. The process learner has regret $2k\alpha$, and the outcome learner uses the moment estimator from Appendix B.3. Fewer PRM errors lower process regret while leaving the outcome learning curve unchanged (Figure 2(a)). The outcome learner needs an estimated 1,840 training trajectories at $k = 8$ and 14,372 at $k = 1$ to achieve lower mean regret than the process learner. Panel (b) gives the theoretical sample sizes, 1,863 and 14,211. We show all four comparisons and the bootstrap procedure in Appendix I.1.1.

6.2 Effect of Horizon Length

For $H \in \{8, 16, 32, 64\}$, we set $k = H/8$ and vary α over the fixed range $0 < \alpha \le \alpha_0 = 1/(4H)$. The maximum process regret is $b = 1/16$ at every horizon. At each sample size, we numerically maximize the expected regret of the moment estimator over this range. We generate 256 independent datasets at the maximizing value of α and evaluate the estimator on these datasets. We also maximize the unified minimax lower bound in (19) over the same range and plot the result. The regret curves nearly coincide when we divide regret by b and scale the sample size to nb^2/H^2 (Figure 2(c)). The outcome estimator requires 473, 1,893, 7,573, and 30,293 trajectories for its maximum expected regret to fall below b. When the horizon doubles, the estimator needs approximately four times as many trajectories, consistent with the H^2/b^2 dependence in Appendix B.5.

6.3 Learning with Unknown Dynamics

We use the family with unknown transitions from Section 3.2, with $m = 8$, $k = 1$, $\gamma = 1/4$, and $\alpha = 1/256$. We keep the true reward, transition signs, and PRM fixed across 512 independent runs.

Table 2: Mean exact certificate and best standard bound on 512 random MDPs

Tolerance	Exact $\mathcal{B}_\Pi(\zeta)$	Best standard bound	Reduction	Global max $\lvert W_\Pi - \mathcal{B}_\Pi \rvert$
0	0.01930	0.02050	5.9%	1.1×10^{-9}
0.05	0.02095	0.02248	6.8%	1.1×10^{-9}
0.10	0.02256	0.02446	7.8%	1.1×10^{-9}
0.20	0.02564	0.02833	9.5%	1.1×10^{-9}

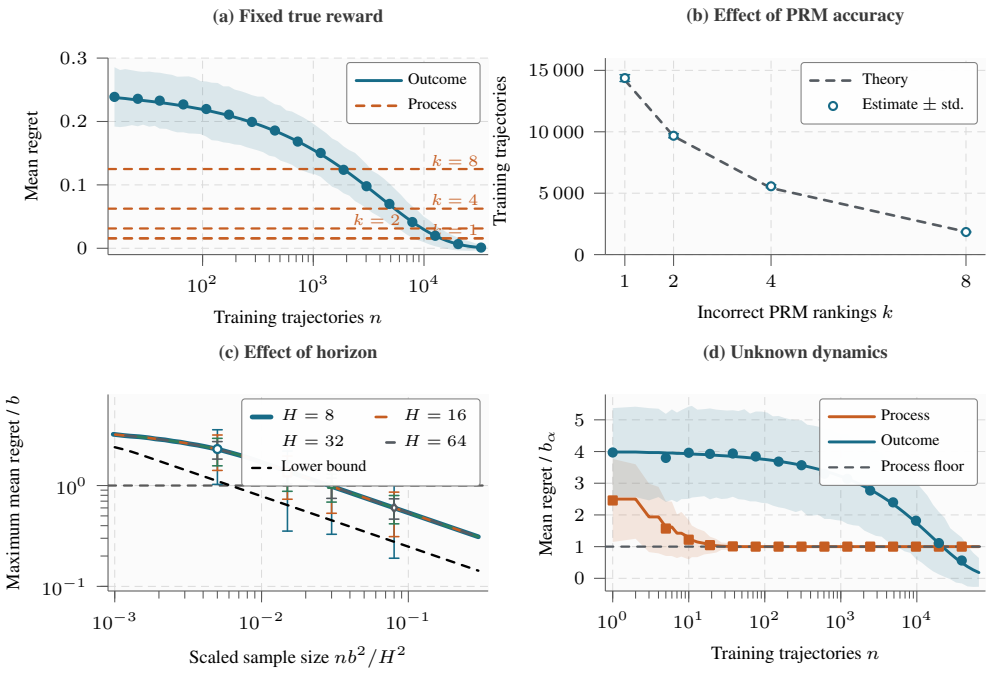

Figure 2: Process and outcome learning. (a) Mean regret at a fixed true reward. (b) Number of training trajectories required for outcome regret to fall below process regret. (c) Maximum expected regret of the moment estimator over reward gaps, with a minimax lower bound. (d) Mean regret under unknown transition model. Shading and error bars show one standard deviation across runs; panel (b) uses the standard deviation of bootstrap estimates.

Both learners estimate the transition signs, and the outcome learner also estimates the true reward preferences. Process regret decreases toward $b_\alpha = 1/256$, while outcome regret continues to fall (Figure 2(d)). We estimate that the outcome learner achieves lower mean regret than the process learner after 22,530 training trajectories. The expected regret formulas give 21,681 trajectories. We also vary the transition strength in Appendix I.1.3 to examine its effect on process learning.

7 Conclusion

We compare policy learning from an informative but imperfect PRM with learning from unbiased outcome feedback under the same rollout budget. We establish matching crossover scales and exact process bias floors for known and unknown transition dynamics. With unknown dynamics, the process learner improves its transition estimates but maintains a nonzero regret floor. Our certificates relate persistent regret to accumulated reward error within the policy class, accounting for potential shaping and reward scale. For compact convex occupancy sets in finite tabular MDPs, we characterize the maximum loss in expected true return allowed by an additive tolerance in PRM value comparisons. Policy constraints can therefore yield much smaller bounds than global STARC guarantees. Our tabular experiments support the predicted learning curves and show lower mean regret for the hybrid learner than for both baselines.

AI USE STATEMENT

We used generative AI tools primarily to polish the writing, check the theoretical proofs, and assist with the design and implementation of the experimental code. These tools also supported proof improvement and the analysis of the synthetic experiments. The authors take responsibility for the final manuscript, mathematical claims, and experimental code.

REPRODUCIBILITY STATEMENT

The appendices give the task constructions, information available to each learner, and proofs of the stated results. The experimental scripts provide fixed random seeds, parameter records, per-dataset regrets, standard deviations, and generated tables and figures. We also describe the controls and numerical checks in Appendix I.1.

REFERENCES

Niladri S. Chatterji, Aldo Pacchiano, Peter L. Bartlett, and Michael I. Jordan. On the theory of reinforcement learning with once-per-episode feedback. In *Advances in Neural Information Processing Systems*, volume 34, pp. 3401–3412, 2021.

Fan Chen, Zeyu Jia, Alexander Rakhlin, and Tengyang Xie. Outcome-based online reinforcement learning: Algorithms and fundamental limits. In *Advances in Neural Information Processing Systems*, volume 38, 2025. arXiv:2505.20268.

Jie Cheng, Gang Xiong, Ruixi Qiao, Lijun Li, Chao Guo, Junle Wang, Yisheng Lv, and Fei-Yue Wang. Stop summation: Min-form credit assignment is all process reward model needs for reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4382. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/be91eb86eb74efc055cff83e953f86ce-Abstract-Conference.html. arXiv:2504.15275.

Alon Cohen, Haim Kaplan, Tomer Koren, and Yishay Mansour. Online Markov decision processes with aggregate bandit feedback. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 1301–1329. PMLR, 2021. URL https://proceedings.mlr.press/v134/cohen21a.html.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948v1.

Ruiyi Ding, Yongxuan Lv, Xianhui Meng, Jiahe Song, Chao Wang, Chen Jiang, and Yuan Cheng. PRPO: Aligning process reward with outcome reward in policy optimization. *arXiv preprint arXiv:2601.07182*, 2026.

Yonathan Efroni, Nadav Merlis, and Shie Mannor. Reinforcement learning with trajectory feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7288–7295, 2021. doi: 10.1609/aaai.v35i8.16895.

Lukas Fluri, Leon Lang, Alessandro Abate, Patrick Forré, David Krueger, and Joar Max Viktor Skalse. The perils of optimizing learned reward functions: Low training error does not guarantee low regret. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 17306–17377. PMLR, 2025. URL https://proceedings.mlr.press/v267/fluri25a.html.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23h.html. arXiv:2210.10760.

Adam Gleave, Michael Dennis, Shane Legg, Stuart Russell, and Jan Leike. Quantifying differences in reward functions. In *International Conference on Learning Representations*, 2021. arXiv:2006.13900.

Abhishek Gupta, Aldo Pacchiano, Yuexiang Zhai, Sham M. Kakade, and Sergey Levine. Unpacking reward shaping: Understanding the benefits of reward engineering on sample complexity. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15281–15295, 2022. doi: 10.52202/068431-1112. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6255f22349da5f2126dfc0b007075450-Abstract-Conference.html.

Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason D. Lee, Wen Sun, Akshay Krishnamurthy, and Dylan J. Foster. Correcting the mythos of KL-regularization: Direct alignment without overoptimization via chi-squared preference optimization. In *International Conference on Learning Representations*, 2025a.

Jiawei Huang, Bingcong Li, Christoph Dann, and Niao He. Can RLHF be more efficient with imperfect reward models? A policy coverage perspective. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25438–25473. PMLR, 2025b. URL https://proceedings.mlr.press/v267/huang25o.html. arXiv:2502.19255.

Zeyu Jia, Alexander Rakhlin, and Tengyang Xie. Do we need to verify step by step? Rethinking process supervision from a theoretical perspective. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 27373–27398. PMLR, 2025. URL https://proceedings.mlr.press/v267/jia25f.html. arXiv:2502.10581.

Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Correlated proxies: A new definition and improved mitigation for reward hacking. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/dbe2cfe4767f3255160b73a36ae3162e-Abstract-Conference.html. arXiv:2403.03185.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. arXiv:2305.20050.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 278–287. Morgan Kaufmann, 1999.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*, 2022. arXiv:2201.03544.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 1994. doi: 10.1002/9780470316887.

Mohammad Rezaei, Jens Lehmann, and Sahar Vahdati. LLM reasoning with process rewards for outcome-guided steps. *arXiv preprint arXiv:2604.02341*, 2026.

Dhruv Rohatgi, Abhishek Shetty, Donya Saless, Yuchen Li, Ankur Moitra, Andrej Risteski, and Dylan J. Foster. Taming imperfect process verifiers: A sampling perspective on backtracking. In *International Conference on Learning Representations*, 2026. arXiv:2510.03149.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *International Conference on Learning Representations*, 2025. arXiv:2410.08146.

Shuning Shang, Hubert Strauss, Stanley Wei, Sanjeev Arora, and Noam Razin. When errors can be beneficial: A categorization of imperfect rewards for policy gradient. *arXiv preprint arXiv:2604.25872*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Joar Skalse, Lucy Farnik, Sumeet Ramesh Motwani, Erik Jenner, Adam Gleave, and Alessandro Abate. STARC: A general framework for quantifying differences between reward functions. In *International Conference on Learning Representations*, 2024. arXiv:2309.15257.

Joar Max Viktor Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32033–32058. PMLR, 2023. URL https://proceedings.mlr.press/v202/skalse23a.html.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Eric Wiewiora. Potential-based shaping and Q-value initialization are equivalent. *Journal of Artificial Intelligence Research*, 19:205–208, 2003. doi: 10.1613/JAIR.1190.

Blake Wulfe, Ashwin Balakrishna, Logan Ellis, Jean Mercat, Rowan McAllister, and Adrien Gaidon. Dynamics-aware comparison of learned reward functions. In *International Conference on Learning Representations*, 2022. arXiv:2201.10081.

Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D. Lee, and Wen Sun. Provable offline preference-based reinforcement learning. In *International Conference on Learning Representations*, 2024. arXiv:2305.14816.

Chenchen Zhang. From reasoning to agentic: Credit assignment in reinforcement learning for large language models. *arXiv preprint arXiv:2604.09459*, 2026.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying process errors in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1009–1024. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.50. URL https://aclanthology.org/2025.acl-long.50/. arXiv:2412.06559.

Banghua Zhu, Michael I. Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or K-wise comparisons. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 43037–43067. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhu23f.html. arXiv:2301.11270.

APPENDIX

A RELATED WORK

Process reward models learn to score individual reasoning steps using step-level supervision (Lightman et al., 2024; Uesato et al., 2022; Wang et al., 2024). ProcessBench evaluates whether models can identify the first erroneous step in a solution or determine that all steps are correct (Zheng et al., 2025). For policy optimization, Setlur et al. (2025) assign step rewards based on changes in the probability of a correct final answer under a separate prover policy. PURE uses min-form credit assignment to reduce reward hacking during PRM-guided training (Cheng et al., 2025), while PRPO and PROGRS combine process and outcome signals (Ding et al., 2026; Rezaei et al., 2026). Rohatgi et al. (2026) analyze test-time sampling with imperfect process verifiers. For a broader review of credit assignment in reinforcement learning for LLMs, see Zhang (2026).

Prior work studies RL with cumulative rewards or losses for each trajectory (Efroni et al., 2021; Cohen et al., 2021) and binary episode labels (Chatterji et al., 2021). In offline RL, Jia et al. (2025) compare total-reward feedback with true per-step rewards under reward realizability and state-action coverage. They show that outcome supervision is statistically no harder than process supervision, up to polynomial factors in the horizon. Chen et al. (2025) give online learning guarantees for outcome feedback and establish statistical separations from per-step reward feedback. Huang et al. (2025b) use imperfect reward models to improve sample efficiency in KL-regularized contextual-bandit RLHF. For data collection, their transfer procedure selects among source policies and a policy fitted to the preference data collected so far. Our minimax comparison identifies the budget ranges in which process or outcome learning has lower regret, including when both learners must estimate the transition dynamics.

Policy optimization can increase the expected proxy return while reducing the expected true return (Pan et al., 2022; Gao et al., 2023). Fluri et al. (2025) show that distribution shift can produce high policy regret despite low reward error under the data distribution. In preference-based learning, Zhu et al. (2023) and Zhan et al. (2024) derive policy guarantees under coverage assumptions, while Huang et al. (2025a) use regularization to control overoptimization. Laidlaw et al. (2025) analyze occupancy-measure regularization when proxy and true rewards are correlated under a reference policy. Shang et al. (2026) categorize reward errors by how they affect improvements in true return during policy-gradient learning. Our certificates measure how accumulated reward error affects comparisons within a policy class.

Potential-based reward shaping redistributes rewards across stages while preserving optimal policies (Ng et al., 1999). EPIC compares rewards up to shaping and positive scaling (Gleave et al., 2021), and DARD incorporates transition dynamics into the comparison (Wulfe et al., 2022). STARC gives a general family of normalized reward distances with upper and lower bounds on worst-case regret (Skalse et al., 2024). For compact convex occupancy sets in finite tabular MDPs, our certificate gives the maximum loss in expected true return allowed by an additive tolerance in PRM value comparisons. Policy constraints can then yield sharper guarantees than comparisons over the full policy class.

B FINITE-BUDGET FEEDBACK SEPARATION

We prove Theorem 1 and Corollary 7 by bounding the uncertainty left by each feedback protocol. For process learning, the unknowns are the locations of the PRM errors; for outcome learning, they are the true action preferences.

B.1 TASK FAMILY AND MINIMAX RISK

Fix $H \geq 4$, $1 \leq k \leq H/4$, and $0 < \alpha \leq 1/(4H)$. Let

$$\Omega_{H,k} := \left\{(\omega, z) \in \{-1,+1\}^H \times \{-1,+1\}^H : \left|\{h \in [H] : \omega_h \neq z_h\}\right| = k\right\}.$$

The state at stage h is s_h, and either action leads deterministically to s_{h+1}. The rewards on reachable pairs are those in (3); set both rewards to zero on unreachable pairs. Conditional on the training history and the current action trajectory, each outcome is a fresh Bernoulli observation with the

parameter specified in Section 3.1. Thus, $\mathbb{E}[Y \mid x] = R_\omega(x)$ and $R_\omega(x), \widehat{R}_z(x) \in [1/4, 3/4]$. The policy class is the full Markov class.

Both learners know H, k, α and the common dynamics. A process learner receives $\hat{r}^z$ and trajectories without true rewards or outcomes. An outcome learner receives trajectories and terminal outcomes, without $\hat{r}^z$. For a learner A_q using at most n trajectories under protocol q, define

$$\mathcal{R}_q(n;\alpha) := \inf_{\mathsf{A}_q} \sup_{(\omega,z)\in\Omega_{H,k}} \mathbb{E}_{\mathsf{A}_q,\omega,z}\big[J_\omega(\pi^\star_\omega) - J_\omega(\hat{\pi}_q)\big], \tag{15}$$

where the expectation is over learner randomization and observation noise, and J_ω averages over deployment of the output policy. Training policies may depend on previous observations. Each instance has a fixed PRM, and each PRM vector is compatible with every choice of k error locations.

B.2 Process Minimax Risk

Exact process minimax risk. The true-optimal policy follows ω_h and has value $1/2 + H\alpha$. Following the PRM signs z_h gives value $1/2 + (H-2k)\alpha$ and regret $2k\alpha$. This requires no process trajectories, so

$$\mathcal{R}_{\mathrm{proc}}(n;\alpha) \le 2k\alpha.$$

For a matching lower bound, fix z and draw $I = \{h : \omega_h \neq z_h\}$ uniformly among the k-subsets of $[H]$. Process transcripts have the same distribution for every I. Let p_h be the probability that the output follows z_h at stage h, averaging learner and deployment randomization. Then p_h is independent of I, and

$$\begin{aligned}\mathbb{E}_I\big[J_\omega(\pi^\star_\omega) - J_\omega(\hat{\pi}_{\mathrm{proc}})\big] &= 2\alpha \sum_{h=1}^{H} \left[\left(1 - \frac{k}{H}\right)(1-p_h) + \frac{k}{H} p_h\right] \\ &= 2k\alpha + 2\alpha\left(1 - \frac{2k}{H}\right)\sum_{h=1}^{H}(1-p_h).\end{aligned} \tag{16}$$

The final term is nonnegative because $k/H < 1/2$. This Bayes risk lower-bounds the minimax risk, so

$$\mathcal{R}_{\mathrm{proc}}(n;\alpha) \ge 2k\alpha.$$

Combining the two bounds gives the exact identity

$$\mathcal{R}_{\mathrm{proc}}(n;\alpha) = 2k\alpha \qquad \text{for every } n \ge 0.$$

□

The uniform policy, for comparison, has value $1/2$ and regret $H\alpha$.

B.3 Outcome Risk Bounds

Adaptive lower bound. Every true sign vector ω appears in $\Omega_{H,k}$, and outcome observations depend on ω alone. We can therefore average over a uniform ω to lower-bound the outcome minimax risk. We use binary testing and the KL chain rule for adaptive observations.

Fix a learner and two sign vectors ω, ω' that differ only at coordinate h. For any action trajectory x, let $p = R_\omega(x)$, $q = R_{\omega'}(x)$. Then $|p - q| = 2\alpha$ and $p, q \in [1/4, 3/4]$. Taylor's theorem applied to $u \mapsto \mathrm{kl}(u, q)$ gives, for some ξ between p and q,

$$\mathrm{kl}(p,q) = \frac{(p-q)^2}{2\xi(1-\xi)} \le \frac{8}{3}(p-q)^2.$$

Hence

$$\mathrm{kl}(R_\omega(x), R_{\omega'}(x)) \le \frac{32}{3}\alpha^2. \tag{17}$$

Conditional on any training history, the learner uses the same action-selection rule under the two tasks, so the conditional action probabilities cancel in the KL chain rule. Let P_ω and $P_{\omega'}$ denote the resulting transcript distributions, including the learner's random seed. Then

$$\mathrm{KL}(P_\omega \| P_{\omega'}) \le \sum_{i=1}^{n} \mathbb{E}_\omega\, \mathrm{kl}(R_\omega(x_i), R_{\omega'}(x_i)) \le \frac{32}{3} n\alpha^2 =: K_n. \tag{18}$$

Under an equal prior on this pair, any decision for the sign at stage h, including a randomized final action, has average error at least $(1 - \mathrm{TV}(P_\omega, P_{\omega'}))/2$. Let $A(P,Q) = \int \sqrt{dP\, dQ}$ be the Hellinger affinity. The Jensen and Cauchy–Schwarz inequalities give $A(P,Q) \geq e^{-\mathrm{KL}(P\|Q)/2}$ and $\mathrm{TV}(P,Q) \leq \sqrt{1 - A(P,Q)^2}$, respectively. Thus,

$$\mathrm{TV}(P_\omega, P_{\omega'}) \leq \sqrt{1 - e^{-K_n}}.$$

Average over the other $H-1$ signs and sum over h. Each wrong action costs 2α in true return, so

$$\mathcal{R}_{\mathrm{out}}(n;\alpha) \geq H\alpha \left(1 - \sqrt{1 - e^{-32n\alpha^2/3}}\right).$$

For $x \geq 0$, let $v = \sqrt{1 - e^{-x}} \in [0,1)$. Then

$$\frac{1 - \sqrt{1 - e^{-x}}}{e^{-2x}} = \frac{1}{(1-v)(1+v)^2}.$$

The function $(1-v)(1+v)^2$ is maximized on $[0,1]$ at $v = 1/3$, where it equals $32/27$. Therefore,

$$\mathcal{R}_{\mathrm{out}}(n;\alpha) \geq \frac{27}{32} H\alpha e^{-64n\alpha^2/3}. \tag{19}$$

Constructive upper bound. For $n \geq 0$ and $0 < \beta \leq 1/4$, let $Z_{n,\beta} \sim \mathrm{Bin}(n, 1/2+\beta)$ and define

$$p_n(\beta) := \mathbb{P}(Z_{n,\beta} < n/2) + \tfrac{1}{2}\mathbb{P}(Z_{n,\beta} = n/2). \tag{20}$$

In particular, $p_0(\beta) = 1/2$.

Lemma 6 (Binary testing error)**.** *For $n \geq 1$ and $0 < \beta \leq 1/4$,*

$$2p_n(\beta) \leq (1 - 4\beta^2)^{n/2} \leq e^{-2n\beta^2}. \tag{21}$$

Proof. Let P_+ and P_- be the laws of n independent Bernoulli observations with parameters $1/2+\beta$ and $1/2-\beta$. Majority voting with fair tie breaking is Bayes optimal under equal priors, with error $p_n(\beta)$. Therefore,

$$2p_n(\beta) = \sum_y \min\{P_+(y), P_-(y)\} \leq \sum_y \sqrt{P_+(y)P_-(y)} = (1 - 4\beta^2)^{n/2}.$$

The last equality follows by factoring over the independent observations, and $1 - u \leq e^{-u}$ gives the exponential bound. In particular, $p_n(\beta) < 1/2$ for $n \geq 1$. □

Sample n trajectories with independent uniform actions. Set $\widehat{\omega}_h$ to the sign of $\sum_i x_{ih}(2Y_i - 1)$, breaking ties independently and uniformly, and deploy the corresponding deterministic actions. For each h, the variables $V_{ih} := \omega_h x_{ih}(2Y_i - 1)$ are independent across trajectories, take values in $\{-1, +1\}$, and satisfy

$$\mathbb{E}V_{ih} = 2\omega_h \mathbb{E}\left[x_{ih}\left(Y_i - \frac{1}{2}\right)\right] = 2\alpha.$$

The last equality follows by averaging over the independent uniform action coordinates in $R_\omega(x)$. Thus $\mathbb{P}(V_{ih} = +1) = 1/2 + \alpha$, and the sign-estimation error at each coordinate is exactly $p_n(\alpha)$. Since every incorrect action incurs regret 2α,

$$\mathbb{E}\big[J_\omega(\pi^\star_\omega) - J_\omega(\hat{\pi}_{\mathrm{mom}})\big] = 2H\alpha\, p_n(\alpha) \qquad \text{for every } (\omega, z) \in \Omega_{H,k}. \tag{22}$$

Lemma 6, together with $p_0(\alpha) = 1/2$, therefore gives, for every $n \geq 0$,

$$\mathcal{R}_{\mathrm{out}}(n;\alpha) \leq 2H\alpha\, p_n(\alpha) \leq H\alpha e^{-2n\alpha^2}.$$

Together with (19), this proves the outcome risk bounds in (4).

B.4 FINITE-BUDGET REVERSAL

Strict minimax comparison. Write $\varrho = k/H \in (0, 1/4]$. The lower bound in (19) exceeds the exact process risk $2k\alpha$ whenever

$$n < \frac{3}{64\alpha^2} \log \frac{27}{64\varrho}. \tag{23}$$

The outcome upper bound is strictly below $2k\alpha$ whenever

$$n > \frac{1}{2\alpha^2} \log \frac{1}{2\varrho}. \tag{24}$$

For $0 < \varrho \le 1/4$,

$$\log \frac{27}{64\varrho} \ge \frac{1}{4} \log \frac{1}{\varrho},$$

because $\varrho^{3/4} \le 4^{-3/4} < 27/64$. Consequently, $c = 3/512$ and $C = 1$ give the strict comparisons in Theorem 1. The outcome minimax risk is nonincreasing in n, since additional observations may be ignored, whereas the process risk is constant. These comparisons locate the first budget at which outcome has strictly smaller minimax risk at scale $\Theta(\alpha^{-2} \log(H/k))$.

Moment-estimator comparison. By (22), on every task,

$$\mathbb{E}\big[J_\omega(\pi^\star_\omega) - J_\omega(\hat{\pi}_{\mathrm{mom}})\big] - 2k\alpha = 2\alpha\big(Hp_n(\alpha) - k\big).$$

Thus this estimator has strictly smaller regret than the PRM-optimal policy exactly when $p_n(\alpha) < k/H$. As the Bayes error of the binary experiment, $p_n(\alpha)$ is nonincreasing in n: additional observations can be ignored. The first strict crossing for this estimator is therefore the smallest n satisfying $p_n(\alpha) < k/H$.

B.5 A RANGE OF REWARD GAPS

Smaller reward gaps make individual errors less costly but also make action preferences harder to infer from outcomes. Taking the worst case over a fixed range accounts for both effects and expresses the crossover budget in terms of the maximal process regret.

Corollary 7 (Bias-dependent minimax crossover)**.** *Fix $0 < \alpha_0 \le 1/(4H)$ and include all amplitudes $0 < \alpha \le \alpha_0$, with α disclosed to both learners. Let $\mathcal{R}^{\le\alpha_0}_q(n)$ denote the minimax risk over this family and set $b = 2k\alpha_0$. For $n \ge 1$,*

$$\mathcal{R}^{\le\alpha_0}_{\mathrm{proc}}(n) = b, \qquad \mathcal{R}^{\le\alpha_0}_{\mathrm{out}}(n) = \Theta\Big(H \min\{\alpha_0, n^{-1/2}\}\Big). \tag{25}$$

Process has strictly smaller minimax risk for $n \le H^2/(1024b^2)$, and outcome has strictly smaller minimax risk for $n \ge H^2/b^2$.

Proof of Corollary 7. Fix the family with $(\omega, z) \in \Omega_{H,k}$ and $0 < \alpha \le \alpha_0$. The instance amplitude is disclosed to both learners. The process lower bound at $\alpha = \alpha_0$ is $b = 2k\alpha_0$, and following z has loss at most b throughout the family. Hence, $\mathcal{R}^{\le\alpha_0}_{\mathrm{proc}}(n) = b$.

For the outcome lower bound at $n \ge 1$, choose the admissible amplitude

$$\alpha_n := \min\{\alpha_0, 1/(16\sqrt{n})\}.$$

Then $(64/3)n\alpha_n^2 \le 1/12$. Restricting to this amplitude and applying (19) gives

$$\mathcal{R}^{\le\alpha_0}_{\mathrm{out}}(n) \ge \frac{27}{32} H\alpha_n e^{-1/12} \ge \frac{99}{128} H\alpha_n > \frac{3H}{4} \min\left\{\alpha_0, \frac{1}{16\sqrt{n}}\right\}, \tag{26}$$

where $e^{-1/12} \ge 1 - 1/12 = 11/12$. For the upper bound, define

$$\alpha_n^{\mathrm{up}} := \min\{\alpha_0, 1/(2\sqrt{n})\}.$$

The function $\alpha e^{-2n\alpha^2}$ is maximized over $(0, \alpha_0]$ at α_n^{up}. Therefore, the same moment estimator gives

$$\mathcal{R}^{\le\alpha_0}_{\mathrm{out}}(n) \le H\alpha_n^{\mathrm{up}} e^{-2n(\alpha_n^{\mathrm{up}})^2} \le H \min\left\{\alpha_0, \frac{1}{2\sqrt{en}}\right\}. \tag{27}$$

Equations (26)–(27) prove (25), with universal constants. At $n = 0$, symmetry gives the exact outcome risk $H\alpha_0$.

Because $k \le H/4$, we have $\alpha_0 \ge 2b/H$. If $1 \le n \le H^2/(1024b^2)$, then $1/(16\sqrt{n}) \ge 2b/H$, so

$$\mathcal{R}_{\mathrm{out}}^{\le \alpha_0}(n) \ge \frac{3}{2} b > \mathcal{R}_{\mathrm{proc}}^{\le \alpha_0}(n).$$

At $n = 0$, the same strict comparison follows from $H\alpha_0 \ge 2b$. If $n \ge H^2/b^2$, Equation (27) instead gives

$$\mathcal{R}_{\mathrm{out}}^{\le \alpha_0}(n) \le b/(2\sqrt{e}) < \mathcal{R}_{\mathrm{proc}}^{\le \alpha_0}(n).$$

These inequalities prove the claimed budget intervals and the matching scale $\Theta(H^2/b^2)$. □

Over the fixed amplitude range, gaps of order $n^{-1/2}$ determine the worst-case outcome rate at large budgets. For an individual task with a fixed amplitude, Equation (22) gives an exponential decrease in expected regret. At fixed H and k, the crossover scales as α^{-2}; varying the number of PRM errors at a fixed amplitude gives the logarithmic dependence in Theorem 1.

C Feedback Separation with Unknown Dynamics

In this family, the PRM ranks intermediate states, and both learners must estimate which actions reach them. The process learner uses these transition estimates to follow the PRM preferences. The outcome learner also uses terminal observations to estimate the true reward preferences.

C.1 Task Family and Information

Let $m \ge 4$, $1 \le k \le m/4$, $0 < \gamma \le 1/4$, and $0 < \alpha \le 1/(4m)$. The horizon is $H = 2m$, and the policy class is the full Markov class. At stage $2j - 1$, the learner chooses $x_j \in \{-1, +1\}$ at a decision state s_j. An unknown vector $\theta \in \{-1, +1\}^m$ determines the transition to a reward state u_{j,v_j}:

$$\mathbb{P}(v_j = v \mid x_j) = \frac{1}{2} + \gamma\theta_j x_j v, \qquad v \in \{-1, +1\}. \tag{28}$$

At stage $2j$, both actions have the same reward and lead to s_{j+1}, or to termination when $j = m$. The true and PRM rewards are zero at decision states and, at reward states, are

$$r_{2j}^{\omega}(u_{j,v}, a) = \frac{1}{2m} + \alpha\omega_j v, \qquad \hat{r}_{2j}^{z}(u_{j,v}, a) = \frac{1}{2m} + \alpha z_j v. \tag{29}$$

Here $\omega, z \in \{-1, +1\}^m$ differ at exactly k unknown coordinates. Conditional on the training history and the current completed trajectory, each outcome is a fresh draw

$$Y \mid x, v \sim \mathrm{Bernoulli}\left(\frac{1}{2} + \alpha\langle\omega, v\rangle\right). \tag{30}$$

At stage $2j - 1$, the reachable state is s_j; at stage $2j$, the reachable states are $u_{j,-1}$ and $u_{j,+1}$. Complete the MDP with a zero-reward absorbing state. At every other state–time pair, both rewards are zero and every action leads to that absorbing state.

Every true and PRM trajectory return lies in $[1/4, 3/4]$. Both learners know m, k, α, γ and the form of the family. They observe all actions and state labels v_j; the transition signs θ are unknown. The process learner receives the complete fixed reward function $\hat{r}^z$ and no outcomes. The outcome learner receives terminal outcomes and no PRM. Both may adapt their training policies to the observed history. Their minimax risks, denoted by $\mathcal{R}_q^{\mathrm{dyn}}(n; \alpha, \gamma)$, take the infimum over learners using at most n trajectories and returning a Markov policy, and the supremum over all triples (θ, ω, z) in this family.

Conditioning on the action in (28) gives $\mathbb{E}[v_j \mid x_j] = 2\gamma\theta_j x_j$. Thus, every Markov policy satisfies $J(\pi) = 1/2 + 2\alpha\gamma \sum_j \theta_j\omega_j \mathbb{E}_\pi[x_j]$. Maximizing each summand gives the true-optimal decision $x_j = \theta_j\omega_j$; replacing ω_j by z_j gives the PRM-optimal decision $x_j = \theta_j z_j$. A wrong decision costs

$4\alpha\gamma$ in expected true return. The uniform policy has regret $2m\alpha\gamma$, whereas exact PRM optimization has regret

$$b_\alpha = 4k\alpha\gamma. \tag{31}$$

To implement the PRM-optimal policy, the process learner must estimate the transition signs. Under a uniform prior on θ, either action at a decision state is equally likely to be optimal before any trajectories are observed.

C.2 Exact Process Minimax Risk

For fixed α and γ, abbreviate the binomial errors defined in (20) by $p_n = p_n(\alpha)$ and $q_n = p_n(\gamma)$, and write

$$s_n := 1 - 2q_n, \qquad D_n := 4m\alpha\gamma q_n.$$

In particular, $p_0 = q_0 = 1/2$ and $s_0 = 0$.

Exact process minimax risk. For block j of rollout i, the observation $W_{ij} := x_{ij}v_{ij}$ satisfies

$$\mathbb{P}(W_{ij} = w \mid \theta, \text{past}, x_{ij}) = \frac{1}{2} + \gamma\theta_j w.$$

Thus, conditional on θ, the entries of W are independent, and their law does not depend on the training actions. Let

$$\widehat{\theta}_j = \operatorname{sign}\left(\sum_{i=1}^{n} W_{ij}\right),$$

with independent fair tie breaking, and deploy $\widehat{x}_j = z_j\widehat{\theta}_j$. The transition estimate is wrong with probability q_n. Consequently, the deployed action is wrong with probability q_n at each of the $m-k$ correctly ranked coordinates and $1-q_n$ at each of the k reversed coordinates. Its regret is therefore

$$4\alpha\gamma\big[(m-k)q_n + k(1-q_n)\big] = 4\alpha\gamma\big[k + (m-2k)q_n\big].$$

This gives an upper bound on the process minimax risk.

For the matching lower bound, fix z, draw θ uniformly, and draw $I = \{j : z_j \neq \omega_j\}$ uniformly among the k-subsets of $[m]$, independently of θ. Pad any learner using at most n trajectories to exactly n without changing its output. Conditional on its independent random seed, W reconstructs the process transcript; conversely, the transcript determines W. Thus W contains all the information about θ, while neither W nor the seed provides information about I. In particular,

$$\mathbb{E}[\omega_j\theta_j \mid W] = \left(1 - \frac{2k}{m}\right) z_j\mathbb{E}[\theta_j \mid W].$$

Since $k < m/2$, the Bayes action is the majority-vote rule above. Its transition-sign error is q_n, so

$$\mathbb{E}\big|\mathbb{E}[\theta_j \mid W]\big| = 1 - 2q_n = s_n.$$

Its Bayes regret is

$$2m\alpha\gamma\left[1 - \left(1 - \frac{2k}{m}\right)s_n\right] = 4\alpha\gamma\big[k + (m-2k)q_n\big].$$

This lower-bounds the minimax risk. Combining both directions gives

$$\mathcal{R}^{\text{dyn}}_{\text{proc}}(n;\alpha,\gamma) = 4\alpha\gamma\big[k + (m-2k)q_n\big] = D_n + s_n b_\alpha. \tag{32}$$

This proves the process identity in Theorem 2.

C.3 Outcome Risk Bounds

Transition-aware adaptive lower bound. Draw θ and ω independently and uniformly; every pair is compatible with a PRM having k errors. Fix any learner using at most n rollouts, pad it to n without changing its output, and condition on its independent random seed ξ.

Generate the entire transition array $W = (W_{ij})$ in advance, independently conditional on θ, with $\mathbb{P}(W_{ij} = w \mid \theta) = 1/2 + \gamma\theta_j w$, and set $v_{ij} = x_{ij}W_{ij}$ when the learner chooses x_{ij}. Then W is

independent of ω even under adaptive actions. Given (W, ξ) and the preceding outcomes, all actions and states in the next rollout are determined; conversely, the transcript recovers W. The likelihood therefore factors as

$$\begin{aligned}\mathbb{P}(W, Y_{1:n} \mid \theta, \omega, \xi) = &\prod_{i,j}(1/2 + \gamma\theta_j W_{ij}) \\ &\times \prod_{i=1}^{n} R_\omega(v_i)^{Y_i}[1 - R_\omega(v_i)]^{1-Y_i},\end{aligned} \tag{33}$$

where $R_\omega(v) = 1/2 + \alpha\langle\omega, v\rangle$ and $v_i = v_i(W, Y_{<i}, \xi)$ is the same function of the observed data under every hidden parameter. The two factors depend only on θ and ω, respectively, so the product prior gives

$$\mathbb{P}(\theta, \omega \mid W, Y, \xi) = \mathbb{P}(\theta \mid W)\mathbb{P}(\omega \mid W, Y, \xi). \tag{34}$$

Write

$$A_j(W) = \mathbb{E}[\theta_j \mid W], \qquad B_j(W, Y, \xi) = \mathbb{E}[\omega_j \mid W, Y, \xi].$$

The posterior mean of the optimal action $\theta_j\omega_j$ is A_jB_j. A deployed action with conditional mean $a_j \in [-1, 1]$ has conditional Bayes regret $2\alpha\gamma(1 - a_jA_jB_j)$. Minimizing over a_j gives $2\alpha\gamma(1 - |A_jB_j|)$. The transition posterior is the same as in the process analysis; its Bayes error is q_n, so

$$\mathbb{E}_W|A_j(W)| = 1 - 2q_n = s_n. \tag{35}$$

We bound the reward uncertainty uniformly for each fixed (W, ξ). Reveal the other reward signs and define

$$\widetilde{B}_j = \mathbb{E}[\omega_j \mid W, Y, \xi, \omega_{-j}].$$

The tower property and conditional Jensen inequality give

$$\mathbb{E}[|B_j| \mid W, \xi] \le \mathbb{E}[|\widetilde{B}_j| \mid W, \xi]. \tag{36}$$

Fix $W = w$, ξ, and $\omega_{-j} = u$. The conditional prior on ω_j remains uniform. Let Q_+ and Q_- be the corresponding laws of $Y_{1:n}$ when $\omega_j = +1$ and $\omega_j = -1$. At any common outcome history $y_{<i}$, the state vector $v_i(w, y_{<i}, \xi)$ is fixed under both laws, and their conditional success probabilities are

$$\mu_i^{\pm}(y_{<i}) = \frac{1}{2} + \alpha\left(\sum_{\ell \ne j} u_\ell v_{i\ell} \pm v_{ij}\right).$$

They lie in $[1/4, 3/4]$ and differ in magnitude by 2α. The one-rollout bound (17) and the KL chain rule therefore give

$$\mathrm{KL}(Q_+\|Q_-) = \sum_{i=1}^{n} \mathbb{E}_{Q_+}\,\mathrm{kl}\big(\mu_i^+(Y_{<i}), \mu_i^-(Y_{<i})\big) \le \frac{32}{3}n\alpha^2 =: K_n. \tag{37}$$

This bound holds uniformly in w, ξ, u. Under the equal-prior mixture of Q_+ and Q_-,

$$\mathbb{E}[|\widetilde{B}_j| \mid W = w, \xi, \omega_{-j} = u] = \frac{1}{2}\sum_y |Q_+(y) - Q_-(y)| = \mathrm{TV}(Q_+, Q_-) \le \sqrt{1 - e^{-K_n}},$$

where the last step uses the affinity argument in Appendix B. Averaging over u and using (36) yields

$$\mathbb{E}[1 - |B_j| \mid W, \xi] \ge \ell_n, \qquad \ell_n := 1 - \sqrt{1 - e^{-32n\alpha^2/3}}. \tag{38}$$

Since A_j is a function of W, the conditional bound preserves the transition uncertainty:

$$\begin{aligned}\mathbb{E}[1 - |A_jB_j|] &= \mathbb{E}[1 - |A_j|] + \mathbb{E}[|A_j|\mathbb{E}[1 - |B_j| \mid W, \xi]] \\ &\ge 2q_n + s_n\ell_n.\end{aligned}$$

Multiplying by $2\alpha\gamma$ and summing over blocks lower-bounds the Bayes risk of every learner. The scalar inequality used in (19) gives $\ell_n \ge (27/32)e^{-64n\alpha^2/3}$. Hence

$$\begin{aligned}\mathcal{R}^{\mathrm{dyn}}_{\mathrm{out}}(n; \alpha, \gamma) &\ge D_n + 2s_nm\alpha\gamma\ell_n \\ &\ge D_n + \frac{27}{16}s_nm\alpha\gamma e^{-64n\alpha^2/3}.\end{aligned} \tag{39}$$

Constructive upper bound. Sample trajectories with independent uniform actions. Estimate θ_j from $\sum_i x_{ij} v_{ij}$ and ω_j from $\sum_i v_{ij}(2Y_i - 1)$, break all ties independently and fairly, and deploy $\widehat{\theta}_j \widehat{\omega}_j$. For a fixed task, let $U_{ij} = \theta_j x_{ij} v_{ij}$. For one rollout, the joint law factors as

$$\mathbb{P}(U = u, v) = 2^{-m} \prod_{j=1}^{m} (1/2 + \gamma u_j). \tag{40}$$

Thus U is independent of v and of (v, Y), and v is uniform on $\{-1, +1\}^m$. Consequently,

$$\mathbb{E}[\omega_j v_j (2Y - 1)] = 2\alpha\omega_j \mathbb{E}[v_j \langle \omega, v \rangle] = 2\alpha.$$

The reward-sign statistic is a sum of independent signs with success probability $1/2 + \alpha$, giving error p_n. The transition estimate has error q_n and is independent of the reward estimate. With independent tie breaking, their product is wrong exactly when one estimate is wrong and the other is correct. Thus the expected regret on every task is

$$\begin{aligned} R^{\mathrm{dyn}}_{\mathrm{mom}}(n; \alpha, \gamma) &= 4m\alpha\gamma \left[p_n + q_n - 2p_n q_n \right] \\ &= D_n + 4 s_n m \alpha \gamma p_n. \end{aligned} \tag{41}$$

Lemma 6, together with $p_0 = 1/2$, gives

$$\mathcal{R}^{\mathrm{dyn}}_{\mathrm{out}}(n; \alpha, \gamma) \le D_n + 2 s_n m \alpha \gamma e^{-2n\alpha^2}, \qquad n \ge 0.$$

Together with (39), this proves the outcome bounds in (6). At $n = 0$, both bounds equal $2m\alpha\gamma$ because $s_0 = 0$.

C.4 Finite-Budget Reversal

Strict minimax comparison. Let $\varrho = k/m \in (0, 1/4]$ and recall $\mathcal{R}^{\mathrm{dyn}}_{\mathrm{proc}} = D_n + s_n b_\alpha$. By Lemma 6, $q_n < 1/2$ and hence $s_n > 0$ for every $n \ge 1$ and $\gamma > 0$. Subtracting the process risk from the lower bound gives

$$\mathcal{R}^{\mathrm{dyn}}_{\mathrm{out}}(n; \alpha, \gamma) - \mathcal{R}^{\mathrm{dyn}}_{\mathrm{proc}}(n; \alpha, \gamma) \ge s_n m \alpha \gamma \left[\frac{27}{16} e^{-64 n \alpha^2 / 3} - 4\varrho \right]. \tag{42}$$

Thus process has strictly smaller minimax risk whenever

$$n < \frac{3}{64\alpha^2} \log \frac{27}{64\varrho}.$$

Similarly, the outcome upper bound gives

$$\mathcal{R}^{\mathrm{dyn}}_{\mathrm{out}}(n; \alpha, \gamma) - \mathcal{R}^{\mathrm{dyn}}_{\mathrm{proc}}(n; \alpha, \gamma) \le s_n m \alpha \gamma \left[2e^{-2n\alpha^2} - 4\varrho \right],$$

so outcome has strictly smaller minimax risk whenever

$$n > \frac{1}{2\alpha^2} \log \frac{1}{2\varrho}.$$

These are the same scalar comparisons as (23)–(24). Therefore $c = 3/512$ and $C = 1$ give the strict comparisons in Theorem 2. At zero budget both risks equal $2m\alpha\gamma$. The common transition term D_n cancels, while s_n is a positive common factor. The resulting sufficient budget intervals are uniform over $\gamma > 0$ and have scale $\alpha^{-2} \log(m/k)$.

Moment-estimator comparison. Subtracting (32) from (41) gives

$$R^{\mathrm{dyn}}_{\mathrm{mom}}(n; \alpha, \gamma) - \mathcal{R}^{\mathrm{dyn}}_{\mathrm{proc}}(n; \alpha, \gamma) = 4\alpha\gamma s_n (m p_n - k). \tag{43}$$

For $n \ge 1$, this estimator has strictly smaller expected regret exactly when $p_n(\alpha) < k/m$. The binary Bayes error $p_n(\alpha)$ is nonincreasing in n, since additional observations may be ignored. Thus the first strict crossing for this estimator is the smallest n satisfying this inequality; it is independent of γ. This is an estimator-level identity, not an exact characterization of the outcome minimax risk.

D Properties and Proofs of Bias Certificates

This appendix proves the identities and geometric properties used in the bias bounds and gives algorithms for computing the certificates in a finite tabular MDP.

D.1 Reward and Optimization Errors

Reward error is the difference between PRM scores and true per-step rewards. A uniform bound controls its effect under every policy. The maximum per-step error is $\varepsilon_{\mathrm{prm}} = \max_{h\in[H]} \|\varepsilon_h\|_\infty$, where the supremum is over all state–action pairs. Summing the stagewise error norms gives the bound

$$\bar{\varepsilon} \;=\; \sum_{h=1}^{H} \|\varepsilon_h\|_\infty \;\le\; H\,\varepsilon_{\mathrm{prm}}. \tag{44}$$

Summing the per-step bounds controls the absolute error in the return of every trajectory. Averaging over trajectories then gives the same bound on the error in any policy value.

Optimization error measures the gap between the achieved value and the best value in the candidate class Π under the training objective. The class may consist of fixed candidates or policies subject to constraints on action probabilities.

Definition 2 (ζ-optimal policy)**.** *Given a nonempty policy class Π, a reward r, and a tolerance $\zeta \ge 0$, a policy $\hat{\pi} \in \Pi$ is ζ-optimal for r if $J^r(\hat{\pi}) \ge \sup_{\pi\in\Pi} J^r(\pi) - \zeta$. For $r = \hat{r}$, we call it ζ-PRM-optimal.*

A finite-sample RL guarantee bounds the optimization error by $\zeta_{\mathrm{stat}}(n)$ after n rollouts. Setting $\zeta = 0$ isolates the loss caused by optimizing the PRM preferences. We keep ζ explicit to account for incomplete optimization when translating a PRM value guarantee into true policy regret.

D.2 Potential Shaping

Definition 3 (Potential shaping)**.** *A collection of per-step functions $\{g_h : \mathcal{S}\times\mathcal{A} \to \mathbb{R}\}$ is a potential shaping term if there exist bounded, measurable functions $\{\Phi_h : \mathcal{S} \to \mathbb{R}\}_{h=1}^{H+1}$ with $\Phi_{H+1} \equiv 0$ such that $g_h(s,a) = \mathbb{E}_{s'\sim P_h(\cdot|s,a)}[\Phi_{h+1}(s')] - \Phi_h(s)$ for every (h,s,a). This is the expected-reward, finite-horizon analogue of classical potential shaping (Ng et al., 1999). We denote the resulting linear subspace of per-step reward functions by $\mathcal{G} := \{g^\Phi : \Phi_{H+1} \equiv 0\}$.*

Setting the terminal potential to zero removes the contribution of the final state. The intermediate potentials cancel after taking expectations and summing over steps. For every policy,

$$\mathcal{E}(\pi) = -\mathbb{E}_{s_1\sim\rho}[\Phi_1(s_1)] + \sum_{h=1}^{H} \mathbb{E}_{d_h^\pi}\big[\varepsilon_h - g_h^\Phi\big]. \tag{45}$$

The first term depends only on the initial distribution, so it cancels in policy comparisons. Subtracting a potential shaping term therefore preserves the error difference.

D.3 Certificate Properties

Lemma 8 (Reward-difference identity)**.** *For any two reward functions r, r' on the same dynamics and any policy π,*

$$J^r(\pi) - J^{r'}(\pi) \;=\; \sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim d_h^\pi}\big[r_h(s,a) - r'_h(s,a)\big]. \tag{46}$$

In particular, taking $r = \hat{r}$ and $r' = r^\star$ gives $J^{\hat{r}}(\pi) - J(\pi) = \mathcal{E}(\pi)$, where $\mathcal{E}(\pi) := \sum_h \mathbb{E}_{d_h^\pi}[\varepsilon_h]$ is the expected accumulated PRM error along π.

Proof of Lemma 8. Fix a policy π. The trajectory law is determined by $(\rho, \pi, \{P_h\})$ and is the same under evaluation with r and r'. Linearity of expectation gives

$$J^r(\pi) - J^{r'}(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{h=1}^{H} r_h(s_h, a_h) - \sum_{h=1}^{H} r'_h(s_h, a_h)\right] \tag{47}$$

$$= \sum_{h=1}^{H} \mathbb{E}_{\tau\sim\pi}[r_h(s_h, a_h) - r'_h(s_h, a_h)] \tag{48}$$

$$= \sum_{h=1}^{H} \int_{\mathcal{S}\times\mathcal{A}} [r_h(s, a) - r'_h(s, a)]\, d_h^\pi(\mathrm{d}s, \mathrm{d}a). \tag{49}$$

This proves (46). Substituting $r = \hat{r}$, $r' = r^\star$, and $\hat{r} - r^\star = \varepsilon$ gives the expression for $\mathcal{E}(\pi)$. □

Lemma 9 (Shaping invariance of error differences)**.** *For any bounded potential $\{\Phi_h\}$ with $\Phi_{H+1} \equiv 0$, every policy satisfies Equation* (45)*. Subtracting the identities for two policies gives $\mathcal{E}(\hat{\pi}) - \mathcal{E}(\pi^\star) = \sum_h \left(\mathbb{E}_{d_h^{\hat{\pi}}} - \mathbb{E}_{d_h^{\pi^\star}}\right)[\varepsilon_h - g_h^\Phi]$. This expression is the same for every Φ. The residual in Equation* (9) *is the summed ℓ_∞ distance from the error to the shaping subspace and satisfies $\bar{\varepsilon}^{\mathrm{sh}} \le \bar{\varepsilon}$, with the latter inequality obtained by choosing $\Phi \equiv 0$.*

Proof of Lemma 9. Let μ_h^π be the state marginal of d_h^π, so $\mu_1^\pi = \rho$. If $(s_h, a_h) \sim d_h^\pi$ and $s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$, the marginal law of s_{h+1} is μ_{h+1}^π. By the tower property and Definition 3,

$$\mathbb{E}_{d_h^\pi}[g_h^\Phi] = \mathbb{E}_{d_h^\pi}[\mathbb{E}[\Phi_{h+1}(s_{h+1}) \mid s_h, a_h] - \Phi_h(s_h)] \tag{50}$$

$$= \mathbb{E}_{\mu_{h+1}^\pi}[\Phi_{h+1}] - \mathbb{E}_{\mu_h^\pi}[\Phi_h]. \tag{51}$$

Telescoping over h gives

$$\sum_{h=1}^{H} \mathbb{E}_{d_h^\pi}[g_h^\Phi] = \mathbb{E}_{\mu_{H+1}^\pi}[\Phi_{H+1}] - \mathbb{E}_{\mu_1^\pi}[\Phi_1] = -\mathbb{E}_\rho[\Phi_1], \tag{52}$$

because $\Phi_{H+1} \equiv 0$. Substituting $\varepsilon_h = (\varepsilon_h - g_h^\Phi) + g_h^\Phi$ into $\mathcal{E}(\pi)$ and applying the telescoping identity proves (45). The term $-\mathbb{E}_\rho[\Phi_1]$ is independent of π and cancels when the identities for two policies are subtracted. □

Proposition 10 (Span geometry)**.** *For every nonempty Π, F_Π is a seminorm whose kernel is exactly the set of errors for which $\mathcal{E}(\pi)$ is constant over Π. Its kernel contains every potential shaping term, and*

$$\mathrm{F}_\Pi(\varepsilon) \le 2\sum_{h=1}^{H} \|\varepsilon_h - g_h^\Phi\|_\infty, \qquad \textit{hence} \qquad \mathrm{F}_\Pi(\varepsilon) \le 2\bar{\varepsilon}^{\mathrm{sh}} \le 2\bar{\varepsilon}. \tag{53}$$

The first inequality is tight in general.

Proof of Proposition 10. For fixed dynamics, define $L_\pi(\varepsilon) = \mathcal{E}(\pi)$. This functional is linear in the error table, and

$$\mathrm{F}_\Pi(\varepsilon) = \sup_{\pi,\pi'\in\Pi} \big(L_\pi - L_{\pi'}\big)(\varepsilon).$$

This quantity is nonnegative because pairs with $\pi = \pi'$ are allowed, and $\mathrm{F}_\Pi(0) = 0$. Positive homogeneity follows directly for every scalar $c \ge 0$. For $c < 0$,

$$\sup_{\pi,\pi'} c(L_\pi - L_{\pi'})(\varepsilon) = |c| \sup_{\pi,\pi'}(L_{\pi'} - L_\pi)(\varepsilon) = |c|\mathrm{F}_\Pi(\varepsilon),$$

because the indexing set is invariant under exchanging (π, π') and (π', π). For any second error table ε', linearity and the inequality $\sup_x(f_x + g_x) \le \sup_x f_x + \sup_x g_x$ give

$$\mathrm{F}_\Pi(\varepsilon + \varepsilon') \le \mathrm{F}_\Pi(\varepsilon) + \mathrm{F}_\Pi(\varepsilon'),$$

so F_Π is a seminorm. If all $L_\pi(\varepsilon)$ are equal, every pairwise difference is zero, and hence $\mathrm{F}_\Pi(\varepsilon) = 0$. Conversely, suppose that $\mathrm{F}_\Pi(\varepsilon) = 0$. For every pair, both $L_\pi(\varepsilon) - L_{\pi'}(\varepsilon) \le 0$ and, after exchanging

the policies, $L_{\pi'}(\varepsilon) - L_\pi(\varepsilon) \le 0$. The two values must therefore be equal, which proves the kernel statement. Lemma 9 gives $L_\pi(g^\Phi) = -\mathbb{E}_\rho[\Phi_1]$ for every π, so every shaping term belongs to the kernel. More directly, for every pair π, π', the two shaping contributions cancel:

$$\begin{aligned}(L_\pi - L_{\pi'})(\varepsilon + g^\Phi) &= (L_\pi - L_{\pi'})(\varepsilon) + (L_\pi - L_{\pi'})(g^\Phi) \\ &= (L_\pi - L_{\pi'})(\varepsilon).\end{aligned}$$

Taking the supremum over the same set of pairs proves $\mathrm{F}_\Pi(\varepsilon + g^\Phi) = \mathrm{F}_\Pi(\varepsilon)$.

For any fixed Φ and any $\pi, \pi' \in \Pi$, shaping invariance and the fact that each d_h^π is a probability measure give

$$\mathcal{E}(\pi) - \mathcal{E}(\pi') = \sum_{h=1}^{H} (\mathbb{E}_{d_h^\pi} - \mathbb{E}_{d_h^{\pi'}})[\varepsilon_h - g_h^\Phi] \tag{54}$$

$$\le \sum_{h=1}^{H} \left(\left|\mathbb{E}_{d_h^\pi}[\varepsilon_h - g_h^\Phi]\right| + \left|\mathbb{E}_{d_h^{\pi'}}[\varepsilon_h - g_h^\Phi]\right| \right) \tag{55}$$

$$\le 2 \sum_{h=1}^{H} \|\varepsilon_h - g_h^\Phi\|_\infty. \tag{56}$$

Taking the supremum over pairs of policies and then the infimum over Φ proves (53). The aligned binary-chain error in Theorem 15 attains equality.

□

The following auxiliary quantities record the zero-tolerance certificate in Section 4.3 and the scale-optimized residuals used in Appendices F and G.1.

Definition 4 (Scale-aware policy and residual certificates)**.** *For each PRM scale $c > 0$, define*

$$\varepsilon^{(c)} := c^{-1}\hat{r} - r^\star, \qquad D_{\mathrm{sh}}(c) := \inf_{g^\Phi \in \mathcal{G}} \sum_{h=1}^{H} \|\varepsilon_h^{(c)} - g_h^\Phi\|_\infty. \tag{57}$$

We define $\mathrm{F}_\Pi^{\mathrm{ord}}(\hat{r}, r^\star) := \inf_{c>0} \mathrm{F}_\Pi(\varepsilon^{(c)})$ and $\bar{\varepsilon}^{\mathrm{ord}}(\hat{r}, r^\star) := \inf_{c>0} D_{\mathrm{sh}}(c)$. Thus, $\mathrm{F}_\Pi^{\mathrm{ord}}(\hat{r}, r^\star) = \mathcal{B}_\Pi(0)$, and $D_{\mathrm{sh}}(1) = \bar{\varepsilon}^{\mathrm{sh}}$. The first certificate measures policy-value differences. The second bounds these differences through per-step residuals. Both certificates account for the PRM scale and use true-return units.

Corollary 11 (Scale- and shaping-invariant certificates)**.** *The scale-aware certificates satisfy*

$$\mathrm{F}_\Pi^{\mathrm{ord}}(\hat{r}, r^\star) \le 2\bar{\varepsilon}^{\mathrm{ord}}(\hat{r}, r^\star), \tag{58}$$

and both certificates are unchanged by multiplying $\hat{r}$ by a positive constant or adding a potential shaping term to either reward. In particular, both vanish if $\hat{r} = c_0 r^\star + g^\psi$ for some $c_0 > 0$.

Proof of Corollary 11. For each $c > 0$, Proposition 10 gives

$$\mathrm{F}_\Pi(\varepsilon^{(c)}) \le 2D_{\mathrm{sh}}(c). \tag{59}$$

Taking the infimum over $c > 0$ proves $\mathrm{F}_\Pi^{\mathrm{ord}}(\hat{r}, r^\star) \le 2\bar{\varepsilon}^{\mathrm{ord}}(\hat{r}, r^\star)$.

Consider the transformed rewards

$$\widetilde{r} = a\hat{r} + g^\psi, \qquad \widetilde{r}^\star = r^\star + g^\chi, \qquad a > 0.$$

For every $c > 0$,

$$\frac{a\hat{r} + g^\psi}{c} - (r^\star + g^\chi) = \varepsilon^{(c/a)} + g^{\psi/c - \chi}. \tag{60}$$

The shaping term belongs to the kernel of F_Π by Proposition 10. Since $\mathcal{G}$ is a linear subspace, adding an element of $\mathcal{G}$ also leaves the distance to $\mathcal{G}$ unchanged. Finally, $c \mapsto c/a$ is a bijection of $(0, \infty)$. Taking infima in (60) therefore proves positive-PRM-rescaling invariance and shaping invariance of both certificates.

If $\hat{r} = c_0 r^\star + g^\psi$ with $c_0 > 0$, then at $c = c_0$,

$$\varepsilon^{(c_0)} = c_0^{-1}(c_0 r^\star + g^\psi) - r^\star = g^{\psi/c_0} \in \mathcal{G}. \tag{61}$$

Both $\mathrm{F}_\Pi(\varepsilon^{(c_0)})$ and $D_{\mathrm{sh}}(c_0)$ are therefore zero. Since the two certificates are infima of nonnegative quantities, both scale-aware certificates equal zero.

□

Corollary 12 (Finite-tabular computation)**.** *In a finite tabular MDP, $\bar{\varepsilon}^{\mathrm{sh}}$ is the value of a linear program. For the full Markov policy class, $\mathrm{F}_\Pi = \max_\pi J^\varepsilon(\pi) - \min_\pi J^\varepsilon(\pi)$ can be computed by solving the MDP twice, with rewards ε and $-\varepsilon$.*

Proof of Corollary 12. In a finite tabular MDP, the shaping-corrected residual is the optimal value of the linear program

$$\begin{aligned} \min_{\{\Phi_h\},\{t_h\}} \quad & \sum_{h=1}^{H} t_h \\ \text{subject to} \quad & -t_h \le \varepsilon_h(s,a) - \sum_{s'} P_h(s' \mid s,a)\Phi_{h+1}(s') + \Phi_h(s) \le t_h, \quad \forall (h,s,a), \\ & t_h \ge 0, \qquad \Phi_{H+1} \equiv 0. \end{aligned} \tag{62}$$

The middle expression in the constraint is $\varepsilon_h(s,a) - g_h^\Phi(s,a)$. For fixed Φ, the inequalities require t_h to bound its absolute value at every (s,a). The smallest feasible value is $\|\varepsilon_h - g_h^\Phi\|_\infty$. Minimizing $\sum_h t_h$ over Φ yields (9). Choosing $\Phi \equiv 0$ and $t_h = \|\varepsilon_h\|_\infty$ establishes feasibility. The objective is bounded below by zero, so the finite-dimensional linear program attains its optimum at a pair (Φ, t).

For the full Markov class,

$$\mathrm{F}_\Pi(\varepsilon) = \max_\pi J^\varepsilon(\pi) - \min_\pi J^\varepsilon(\pi) = \max_\pi J^\varepsilon(\pi) + \max_\pi J^{-\varepsilon}(\pi),$$

where $-\min_\pi J^\varepsilon(\pi) = \max_\pi J^{-\varepsilon}(\pi)$. To compute the two maxima, set $V_{H+1}^+ = V_{H+1}^- = 0$ and apply backward induction:

$$V_h^+(s) = \max_a \left\{ \varepsilon_h(s,a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)} V_{h+1}^+(s') \right\}, \tag{63}$$

$$V_h^-(s) = \max_a \left\{ -\varepsilon_h(s,a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)} V_{h+1}^-(s') \right\}. \tag{64}$$

A maximizing action exists at every state because the action set is finite. The resulting deterministic Markov policies attain both maxima. Consequently, $\max_\pi J^\varepsilon(\pi) = \mathbb{E}_\rho V_1^+$ and $\max_\pi J^{-\varepsilon}(\pi) = \mathbb{E}_\rho V_1^-$, which gives the required computation. □

D.4 Proofs of the Bias Bounds

Proof of Theorem 3. By Lemma 8, $J^{\hat{r}}(\pi) - J(\pi) = \mathcal{E}(\pi)$, where $\mathcal{E}(\pi) \coloneqq \sum_{h=1}^H \mathbb{E}_{d_h^\pi}[\varepsilon_h]$. We decompose the suboptimality as follows:

$$\begin{aligned} J(\pi^\star) - J(\hat{\pi}) &= \left[J(\pi^\star) - J^{\hat{r}}(\pi^\star)\right] + \left[J^{\hat{r}}(\pi^\star) - J^{\hat{r}}(\hat{\pi})\right] + \left[J^{\hat{r}}(\hat{\pi}) - J(\hat{\pi})\right] \\ &= -\mathcal{E}(\pi^\star) + \left[J^{\hat{r}}(\pi^\star) - J^{\hat{r}}(\hat{\pi})\right] + \mathcal{E}(\hat{\pi}). \end{aligned} \tag{65}$$

Because $\pi^\star \in \Pi$, Definition 2 gives $J^{\hat{r}}(\hat{\pi}) \ge \sup_{\pi \in \Pi} J^{\hat{r}}(\pi) - \zeta \ge J^{\hat{r}}(\pi^\star) - \zeta$, or, equivalently, $J^{\hat{r}}(\pi^\star) - J^{\hat{r}}(\hat{\pi}) \le \zeta$. Substituting this into (65) yields

$$J(\pi^\star) - J(\hat{\pi}) \le \zeta + \mathcal{E}(\hat{\pi}) - \mathcal{E}(\pi^\star). \tag{66}$$

By Definition 1,

$$\mathcal{E}(\hat{\pi}) - \mathcal{E}(\pi^\star) \le \mathrm{F}_\Pi(\varepsilon). \tag{67}$$

Proposition 10 gives $\mathrm{F}_\Pi(\varepsilon) \le 2\bar{\varepsilon}^{\mathrm{sh}} \le 2\bar{\varepsilon}$. The pairwise inequality holds for every potential, so taking the infimum in (9) establishes the result on a general state space.

Combining (66), (67), and (53) establishes every inequality in (10):

$$J(\pi^\star) - J(\hat{\pi}) \le \zeta + \mathrm{F}_\Pi(\varepsilon) \le \zeta + 2\bar{\varepsilon}^{\mathrm{sh}} \le \zeta + 2\bar{\varepsilon}.$$

Setting $\zeta = \zeta_{\text{stat}}(n)$ gives the finite-sample bound on the event where the PRM-optimization guarantee holds. For fixed dynamics and a fixed policy class, the error span and shaping-corrected residual are deterministic functions of the PRM error and remain constant as n varies. □

Proof of Equation (11) *and the invariance properties of* $\mathcal{B}_\Pi$. Fix $c > 0$. By linearity of policy value in the reward and positivity of c, a policy that is ζ-optimal for $\hat{r}$ is (ζ/c)-optimal for $\hat{r}/c$. Applying Theorem 3 with error $\varepsilon^{(c)} = \hat{r}/c - r^\star$ gives

$$J(\pi^\star) - J(\hat{\pi}) \le \frac{\zeta}{c} + \mathrm{F}_\Pi(\varepsilon^{(c)}). \tag{68}$$

Taking the infimum over $c > 0$, and then using $c = 1$ as a feasible choice, proves (11):

$$J(\pi^\star) - J(\hat{\pi}) \le \mathcal{B}_\Pi(\zeta) \le \zeta + \mathrm{F}_\Pi(\varepsilon).$$

For $a > 0$, simultaneously replace $(\hat{r}, \zeta)$ by $(a\hat{r}, a\zeta)$. With $u = c/a$,

$$\inf_{c>0} \left\{ \frac{a\zeta}{c} + \mathrm{F}_\Pi\left(\frac{a\hat{r}}{c} - r^\star \right) \right\} = \inf_{u>0} \left\{ \frac{\zeta}{u} + \mathrm{F}_\Pi\left(\frac{\hat{r}}{u} - r^\star \right) \right\}. \tag{69}$$

Thus positive rescaling preserves the certificate only when the PRM and its optimization tolerance are rescaled together.

If potential shaping terms are added to either reward, then for every $c > 0$ the transformed error differs from $\varepsilon^{(c)}$ by an element of $\mathcal{G}$, as in (60) with $a = 1$. Proposition 10 therefore leaves the span term unchanged at every scale, while ζ/c is unchanged. Hence $\mathcal{B}_\Pi(\zeta)$ is invariant to potential shaping of either reward for every $\zeta \ge 0$.

At zero tolerance, Definition 4 and Corollary 11 give

$$\mathcal{B}_\Pi(0) = \mathrm{F}_\Pi^{\text{ord}} \le 2\bar{\varepsilon}^{\text{ord}}.$$

The same corollary shows that both zero-tolerance certificates vanish when $\hat{r} = c_0 r^\star + g^\psi$ for some $c_0 > 0$. □

E Exact Certificates and the STARC Separation

For compact convex occupancy sets, policy-comparison loss is a convex optimization problem with a linear objective. The scale-aware certificate is its dual value. This representation also explains how constraints on occupancy differences produce the separation in Theorem 14.

E.1 Exact Comparison Loss

Stack the occupancies and rewards over (h, s, a), so $J^r(\pi) = \langle r, d^\pi \rangle$. Define the difference set

$$D_\Pi := K_\Pi - K_\Pi = \{x - y : x, y \in K_\Pi\}. \tag{70}$$

Under the assumptions of Theorem 4, this set is compact, convex, and symmetric about zero. Its support function is the policy span: for every reward table e,

$$\mathrm{F}_\Pi(e) = \max_{v \in D_\Pi} \langle e, v \rangle. \tag{71}$$

Proof of Theorem 4. Writing $v = d^{\pi'} - d^\pi$, the right-hand side of (12) is

$$W_\Pi(\zeta) := \max_{v \in D_\Pi:\, \langle \hat{r}, v \rangle \le \zeta} \langle r^\star, v \rangle. \tag{72}$$

The feasible set contains zero and is compact, so the maximum exists. Its Lagrangian, with multiplier $t \ge 0$, is

$$\mathcal{L}(v, t) = \langle r^\star, v \rangle + t\big(\zeta - \langle \hat{r}, v \rangle\big).$$

Maximizing over D_Π and using symmetry gives the dual objective

$$\max_{v \in D_\Pi} \mathcal{L}(v, t) = t\zeta + \mathrm{F}_\Pi(r^\star - t\hat{r}) = t\zeta + \mathrm{F}_\Pi(t\hat{r} - r^\star). \tag{73}$$

If $\zeta > 0$, zero is strictly feasible for the inequality in (72). It lies in the relative interior of D_Π, since D_Π is convex and symmetric. If $\zeta = 0$ and $\mathrm{F}_\Pi(\hat{r}) > 0$, symmetry gives a point $v_0 \in D_\Pi$ with $\langle \hat{r}, v_0 \rangle < 0$. A convex combination of v_0 and zero with positive weights gives a strictly feasible point in the relative interior. Strong duality therefore applies in both cases. If $\mathrm{F}_\Pi(\hat{r}) = 0$, then $\langle \hat{r}, v \rangle = 0$ for every $v \in D_\Pi$. The constraint then holds throughout D_Π, so the primal value is $\mathrm{F}_\Pi(r^\star)$ and the dual infimum is $\inf_{t \geq 0}\{t\zeta + \mathrm{F}_\Pi(r^\star)\} = \mathrm{F}_\Pi(r^\star)$. Hence, for every $\zeta \geq 0$,

$$W_\Pi(\zeta) = \inf_{t \geq 0}\{t\zeta + \mathrm{F}_\Pi(t\hat{r} - r^\star)\} = \inf_{c > 0}\{\zeta/c + \mathrm{F}_\Pi(\hat{r}/c - r^\star)\} = \mathcal{B}_\Pi(\zeta). \tag{74}$$

Compactness of D_Π makes its support function continuous, so including $t = 0$ does not change the infimum over $t > 0$. The substitution $t = 1/c$ then gives the second equality. A zero minimizing multiplier corresponds to the limit $c \to \infty$; the positive-scale infimum need not be attained. The primal maximum is attained because its feasible set is compact. Every feasible v is the difference of two occupancies in K_Π, so a maximizing pair of policies attains $W_\Pi(\zeta)$. This proves (12). □

Equation (12) characterizes loss over all admissible policy comparisons. A ζ-PRM-optimal output satisfies its comparison condition with every reference policy in Π, including a true-optimal policy when one is present. The maximum over approximately optimal outputs can be smaller because their PRM values must be close to the class maximum. Equation (79) computes both quantities on the separation family.

Convexity determines which policy comparisons the identity covers. Policy values are linear in occupancies, so replacing an arbitrary occupancy set by its closed convex hull leaves every span, and therefore $\mathcal{B}_\Pi$, unchanged. The exact characterization applies to this convex hull. For example, take three actions with $r^\star = (1, 1/2, 0)$ and $\hat{r} = (1, 9/10, 0)$, and let Π contain only the three deterministic policies. Their orderings agree, so the largest loss of a PRM-improving comparison within Π is zero. The certificate is $\mathcal{B}_\Pi(0) = 2/5$: over the convex hull, the mixture $(9/10, 0, 1/10)$ and the second action have the same PRM value $9/10$ and true values $9/10$ and $1/2$, respectively. This comparison attains the certificate. Allowing mixtures thus changes the set of comparisons described by the identity.

For a polytope $K_\Pi = \mathrm{conv}\{d_1, \ldots, d_K\}$, the exact certificate can be computed by a linear program. Let $q_i = \langle \hat{r}, d_i \rangle$ and $r_i = \langle r^\star, d_i \rangle$. Then

$$\mathcal{B}_\Pi(\zeta) = \min_{t \geq 0,\, u, \ell \in \mathbb{R}} \big\{ t\zeta + u - \ell : \ \ell \leq tq_i - r_i \leq u, \ i = 1, \ldots, K \big\}. \tag{75}$$

For fixed t, minimizing $u - \ell$ gives the span of the vertex values. Linear rewards attain their extrema at vertices, which proves the identity. Allowing $t = 0$ preserves the infimum by continuity, as in (74).

E.2 Return-Span Normalization

We compare scale optimization with return-span normalization on a fixed policy class. For rewards r and q, write $\mathcal{B}_\Pi(0; q, r) = \inf_{t > 0} \mathrm{F}_\Pi(tq - r)$ and $\Delta_\Pi(r) = \mathrm{F}_\Pi(r)$. Define $s_\Pi(r) = r/\Delta_\Pi(r)$ when $\Delta_\Pi(r) > 0$ and $s_\Pi(r) = 0$ otherwise, and let

$$\mathcal{D}_\Pi(r, q) := \Delta_\Pi(r)\, \mathrm{F}_\Pi\big(s_\Pi(q) - s_\Pi(r)\big). \tag{76}$$

Proposition 13 (Equivalence to span normalization)**.** *For every nonempty policy class with bounded returns,*

$$\tfrac{1}{2}\mathcal{D}_\Pi(r, q) \leq \mathcal{B}_\Pi(0; q, r) \leq \mathcal{D}_\Pi(r, q). \tag{77}$$

In the setting of STARC, with Π equal to the full policy class, using the return span both as the normalization norm and as the final norm on canonical reward differences gives the distance $\mathrm{F}_\Pi(s_\Pi(q) - s_\Pi(r))$.

Proof. Suppose first that $a = \mathrm{F}_\Pi(r) > 0$ and $b = \mathrm{F}_\Pi(q) > 0$. Then $\mathcal{D}_\Pi(r, q) = \mathrm{F}_\Pi((a/b)q - r)$, so the scale $t = a/b$ proves the upper bound. For any $t > 0$, the triangle and reverse triangle inequalities for the seminorm give

$$\begin{aligned} \mathcal{D}_\Pi(r, q) &\leq \mathrm{F}_\Pi(tq - r) + \mathrm{F}_\Pi((a/b - t)q) \\ &= \mathrm{F}_\Pi(tq - r) + |a - tb| \\ &\leq 2\mathrm{F}_\Pi(tq - r). \end{aligned}$$

Taking the infimum over t proves the lower bound. If $a = 0$, then $\mathrm{F}_\Pi(tq - r) = t\mathrm{F}_\Pi(q)$ and both quantities in (77) are zero. If $a > 0$ and $b = 0$, both are equal to a.

For the full policy class in the STARC setting, the return span is a norm on the image of a canonicalization (Skalse et al., 2024, Proposition 4). Canonicalization preserves policy-value differences. Thus, applying that norm to normalized canonical rewards gives $\mathrm{F}_\Pi(s_\Pi(q) - s_\Pi(r))$, including the specified convention for zero spans. Multiplication by $\Delta_\Pi(r)$ expresses the distance in true-return units. □

E.3 Four-Action Separation

For a finite policy class or a compact class of occupancy measures, define

$$C_{\mathrm{cov}}(\Pi) := \inf_{\mu=\{\mu_h\}} \max_{h\in[H]} \sup_{\pi\in\Pi} \max_{s,a} \frac{d_h^\pi(s,a)}{\mu_h(s,a)}, \tag{78}$$

where a positive numerator and zero denominator are interpreted as $+\infty$, while $0/0$ is interpreted as zero. In the one-step construction below, this definition reduces to the coverage coefficient of a class of action distributions.

Theorem 14 (Separation from global STARC guarantees)**.** *There is a one-step, four-action MDP with fixed rewards $r^\star, \hat{r} \in [0,1]^4$ and compact convex policy classes Π_η, $0 < \eta \le 1/8$. Each class contains a true-optimal policy, has true-return range one, and satisfies $C_{\mathrm{cov}} \le 4$, with C_{cov} defined in* (78)*. For $0 \le \zeta \le 1/8$,*

$$\mathcal{B}_{\Pi_\eta}(\zeta) = 2\eta + 2\zeta, \qquad \max_{\substack{\pi\in\Pi_\eta:\\ \pi \text{ is } \zeta\text{-PRM-optimal}}} \big[J(\pi^\star) - J(\pi)\big] = \eta/2 + 2\zeta. \tag{79}$$

Every STARC metric d on the absorbing-state extension satisfies $U_d \Delta_{\mathrm{all}}(r^\star) d(r^\star, \hat{r}) \ge 1/2$, where U_d is its global soundness constant and $\Delta_{\mathrm{all}}(r^\star)$ is the full-class true-return range.

Proof of Theorem 14. Consider one initial state and four actions, each followed by termination. The reward vectors are fixed throughout the construction:

$$r^\star = (1, 0, 1/2, 1/2), \qquad \hat{r} = (3/4, 1/4, 1, 0). \tag{80}$$

Identify a policy with its action distribution $p \in \Delta_4$ and, for $0 < \eta \le 1/8$, set

$$\Pi_\eta := \{p \in \Delta_4 : |p_3 - p_4| \le \eta\}. \tag{81}$$

This class is compact and convex. The uniform policy $u = (1/4, 1/4, 1/4, 1/4)$ lies in its interior relative to the simplex because all its coordinates are positive and $|u_3 - u_4| < \eta$. The deterministic policies selecting the first and second actions both belong to the class. They have true values one and zero, respectively, so the true-return range is one and the first action is globally true-optimal.

Using the uniform policy as the reference distribution gives $p_i/u_i \le 4$ for every action and every policy in Π_η. The exact coverability coefficient in (78) is

$$C_{\mathrm{cov}}(\Pi_\eta) = \sum_{i=1}^{4} \max_{p\in\Pi_\eta} p_i = 1 + 1 + (1+\eta)/2 + (1+\eta)/2 = 3 + \eta. \tag{82}$$

To obtain the first equality, write $m_i = \max_{p\in\Pi_\eta} p_i$. Any reference distribution μ with $m_i \le C\mu_i$ must have $\sum_i m_i \le C$, and $\mu_i = m_i / \sum_j m_j$ attains equality. The coordinate maxima follow from the simplex constraint and $|p_3 - p_4| \le \eta$.

Let $v(p) = p_3 - p_4$. Direct calculation gives

$$J^{\hat{r}}(p) = \tfrac{1}{2}J(p) + \tfrac{1}{4} + \tfrac{1}{2}v(p), \qquad J^{2\hat{r}-r^\star}(p) = \tfrac{1}{2} + v(p). \tag{83}$$

Since $v(p)$ ranges from $-\eta$ to η, the choice $c = 1/2$ gives

$$\mathcal{B}_{\Pi_\eta}(\zeta) \le 2\zeta + \mathrm{F}_{\Pi_\eta}(2\hat{r} - r^\star) = 2\zeta + 2\eta. \tag{84}$$

For the reverse inequality, consider

$$p' = (1-\eta, 0, 0, \eta), \qquad p = (1 - 3\eta - 2\zeta, 2\eta + 2\zeta, \eta, 0). \tag{85}$$

For the stated ranges of η and ζ, these are probability distributions in Π_η. They satisfy $J^{\hat{r}}(p) = J^{\hat{r}}(p') - \zeta$ and $J(p') - J(p) = 2\eta + 2\zeta$. Theorem 4, together with (84), proves the first equality in (79).

To compute the loss under approximate PRM optimization, observe that

$$J^{\hat{r}}(p) = \tfrac{3}{4} - \tfrac{1}{2}p_2 - \tfrac{1}{2}p_4 + \tfrac{1}{4}v(p) \le \tfrac{3}{4} + \tfrac{1}{4}\eta.$$

Equality holds only at $p_0 = (1-\eta, 0, \eta, 0)$, so the PRM optimum within the class is unique. If p is ζ-PRM-optimal, then (83) implies

$$\tfrac{1}{2}J(p) + \tfrac{1}{4} + \tfrac{1}{2}v(p) \ge \tfrac{3}{4} + \tfrac{1}{4}\eta - \zeta, \qquad \text{and hence} \qquad J(p) \ge 1 - \eta/2 - 2\zeta.$$

The policy $p_\zeta = (1-\eta-2\zeta, 2\zeta, \eta, 0)$ belongs to Π_η and has PRM value exactly ζ below the optimum. Its true loss is $\eta/2 + 2\zeta$, proving the second equality in (79). The PRM-return range over the class is $(1+\eta)/2$, since its minimum is $1/4 - \eta/4$ at $(0, 1-\eta, 0, \eta)$.

For comparison with the original discounted formulation of STARC, add an absorbing terminal state with zero reward after the first action and choose any fixed discount factor in $(0, 1)$. All policy returns remain those in (80), and both states are reachable. The full-class true-return range is $\Delta_{\text{all}}(r^\star) = 1$. Let e_i denote the deterministic policy selecting action i at the initial state. The PRM prefers e_3 to e_1, while their true values differ by $1/2$. Applying STARC soundness (Skalse et al., 2024, Definition 5 and Theorem 1) to this policy pair therefore gives

$$\tfrac{1}{2} = J(e_1) - J(e_3) \le U_d \Delta_{\text{all}}(r^\star) d(r^\star, \hat{r}). \tag{86}$$

The rewards, dynamics, and discount factor are fixed as η varies, so this global certificate is independent of η. At zero tolerance, its ratio to $\mathcal{B}_{\Pi_\eta}(0)$ is at least $1/(4\eta)$, which diverges as $\eta \to 0$. The actual optimization loss is $\eta/2$ and the class-specific certificate is 2η. □

The product $U_d \Delta_{\text{all}}(r^\star) d(r^\star, \hat{r})$ in (86) expresses the global guarantee in true-return units and accounts for rescaling of the metric. Computing return spans on the restricted occupancy set gives the corresponding class-specific comparison in Proposition 13.

F Reward Certificates for the Unknown-Transition Family

The following calculations use the family defined in Appendix C.

Reward certificates. For $I = \{j : z_j \neq \omega_j\}$, the nonzero reward errors are $\varepsilon_{2j}(u_{j,v}, a) = 2\alpha z_j v$ for $j \in I$. Thus $\bar{\varepsilon} = 2k\alpha$, and

$$\mathcal{E}(\pi) = 4\alpha\gamma \sum_{j \in I} z_j \theta_j \mathbb{E}_\pi[x_j].$$

The full Markov class can choose each decision independently, giving

$$\mathrm{F}_\Pi(\varepsilon) = 4k\alpha\gamma - (-4k\alpha\gamma) = 8k\alpha\gamma = 2b_\alpha.$$

For the shaping correction, set $\Phi_{2j}(u_{j,v}) = -\varepsilon_{2j}(u_{j,v}, a)$ and all other potential values to zero. Then, for $j \in I$,

$$(\varepsilon - g^\Phi)_{2j} = 0, \qquad (\varepsilon - g^\Phi)_{2j-1}(s_j, x_j) = 4\alpha\gamma z_j \theta_j x_j,$$

with zero residual elsewhere. Hence $\bar{\varepsilon}^{\text{sh}} \le b_\alpha$; the reverse inequality follows from $\mathrm{F}_\Pi(\varepsilon) \le 2\bar{\varepsilon}^{\text{sh}}$.

For inverse PRM scale $t > 0$,

$$\mathrm{F}_\Pi(t\hat{r} - r^\star) = 4\alpha\gamma\left[(m-k)|t-1| + k(t+1)\right].$$

Since $k < m/2$, this expression is minimized at $t = 1$, giving $\mathcal{B}_\Pi(0) = 2b_\alpha$. Applying the span–residual inequality at each scale gives $\mathcal{B}_\Pi(0)/2 \le \bar{\varepsilon}^{\text{ord}} \le \bar{\varepsilon}^{\text{sh}} = b_\alpha$. Consequently,

$$\bar{\varepsilon} = 2k\alpha, \qquad \bar{\varepsilon}^{\text{sh}} = \bar{\varepsilon}^{\text{ord}} = b_\alpha, \qquad \mathrm{F}_\Pi(\varepsilon) = \mathcal{B}_\Pi(0) = 2b_\alpha. \tag{87}$$

Thus the corrected residuals equal the limiting process regret, whereas (32) also includes finite-budget transition-estimation error.

G Horizon Dependence and Tightness of Reward-Error Bounds

G.1 Bias-Floor Lower Bound and Horizon Dependence

Theorem 15 (Bias-floor lower bound). *Fix $H \geq 1$ and $\varepsilon_{\mathrm{prm}} \in (0, \frac{1}{2H}]$, and let Π be the full Markov policy class. There is a family $\{\mathcal{M}_\omega\}_{\omega \in \{+,-\}^H}$ with $|\mathcal{A}| = 2$ and deterministic transitions, together with a single PRM $\hat{r}$ shared across the family. The construction satisfies the normalization $R \in [0,1]$ in Section 2. On every instance, $\bar{\varepsilon}^{\mathrm{ord}} = \bar{\varepsilon}^{\mathrm{sh}} = \bar{\varepsilon} = H\varepsilon_{\mathrm{prm}}$ and $\mathrm{F}_\Pi^{\mathrm{ord}} = \mathrm{F}_\Pi(\varepsilon^\omega) = 2H\varepsilon_{\mathrm{prm}}$. Under the process-only protocol, the observation consists of $\hat{r}$ and the common state transitions; $r^\star$ and the outcomes are latent. Writing J_ω for the true value on $\mathcal{M}_\omega$, any possibly randomized procedure A with output $\hat{\pi} \in \Pi$ satisfies*

$$\sup_\omega \mathbb{E}_{\mathsf{A}}[J_\omega(\pi^\star_\omega) - J_\omega(\hat{\pi})] \;\geq\; H\,\varepsilon_{\mathrm{prm}} \;=\; \bar{\varepsilon}^{\mathrm{sh}}. \tag{88}$$

The expectation is over the internal randomization of the procedure; J_ω already averages over the trajectory generated by its output policy.

Proof of Theorem 15. Let $H \geq 1$ and $\varepsilon_{\mathrm{prm}} \in (0, \frac{1}{2H}]$. Consider a deterministic chain with state space $\mathcal{S} = \{s_1, \ldots, s_{H+1}\}$, initial distribution $\rho = \delta_{s_1}$, and two actions $\mathcal{A} = \{a_+, a_-\}$. For every $h \in [H]$ and either action, set $P_h(s_{h+1} \mid s_h, a) = 1$; the state s_{H+1} is terminal. Define transitions from unreachable state–time pairs arbitrarily, but identically across the family. Thus, s_h is the unique reachable state at decision step h, and the transitions are action-independent on the reachable chain. The per-step rewards $r^\star_h$ and $\hat{r}_h$ are defined below on the reachable state–action pairs (s_h, a). On all unreachable pairs, both rewards are set to zero. Consequently, the ℓ_∞ norms $\|\varepsilon_h\|_\infty$ in the definition of $\bar{\varepsilon}$ are attained on reachable pairs. We index the family by $\omega = (\omega_1, \ldots, \omega_H) \in \{+,-\}^H$ and write J_ω for the true value on $\mathcal{M}_\omega$. In instance $\mathcal{M}_\omega$, the true per-step rewards are

$$r^\star_h(s_h, a_{\omega_h}) = 2\varepsilon_{\mathrm{prm}}, \qquad r^\star_h(s_h, a_{-\omega_h}) = 0, \qquad h = 1, \ldots, H, \tag{89}$$

where $a_{-\omega_h}$ denotes the action other than a_{ω_h}. The return satisfies the required bound because $R(\tau) \leq \sum_h 2\varepsilon_{\mathrm{prm}} = 2H\varepsilon_{\mathrm{prm}} \leq 1$ (Section 2), and each per-step reward satisfies $2\varepsilon_{\mathrm{prm}} \leq 1/H \leq 1$. In $\mathcal{M}_\omega$, the optimal policy selects a_{ω_h} at step h, and hence

$$J_\omega(\pi^\star_\omega) = \sum_{h=1}^{H} 2\varepsilon_{\mathrm{prm}} = 2H\varepsilon_{\mathrm{prm}} \qquad \text{for every } \omega. \tag{90}$$

We use a single constant process reward for all instances:

$$\hat{r}_h(s_h, a_+) = \hat{r}_h(s_h, a_-) = \varepsilon_{\mathrm{prm}}, \qquad h = 1, \ldots, H. \tag{91}$$

In every $\mathcal{M}_\omega$, the absolute per-step error is $|\varepsilon_{\mathrm{prm}} - 2\varepsilon_{\mathrm{prm}}| = \varepsilon_{\mathrm{prm}}$ on the good action and $|\varepsilon_{\mathrm{prm}} - 0| = \varepsilon_{\mathrm{prm}}$ on the bad action. Thus, $\max_h \|\hat{r}_h - r^\star_h\|_\infty = \varepsilon_{\mathrm{prm}}$ and $\bar{\varepsilon} = H\varepsilon_{\mathrm{prm}}$ uniformly over the family. Because $\hat{r}$ is independent of ω and constant across actions, it contains no information about which action is good.

For any policy π, $J^{\hat{r}}(\pi) = H\varepsilon_{\mathrm{prm}}$, so every policy is PRM-optimal. Every instance supplies the same observed input $(\hat{r}, P)$ to A, and the output-policy distribution is therefore identical across ω. Let $p_h = \mathbb{E}_{\mathsf{A}}[\hat{\pi}_h(a_+ \mid s_h)]$, where the expectation averages the internal randomization of the procedure. Because s_h is the unique reachable state at step h, this marginal determines the expected reward of the deployed policy at that step.

To average over the family, let $\omega \sim \mathrm{Unif}(\{+,-\}^H)$, so the coordinates are independent and $\mathbb{P}(\omega_h = +) = \frac{1}{2}$. At step h, the expected true reward collected by $\hat{\pi}$ is

$$\begin{aligned}\mathbb{E}_\omega \mathbb{E}_{\mathsf{A}} \mathbb{E}_{a_h \sim \hat{\pi}_h(\cdot \mid s_h)}[r^\star_h(s_h, a_h)] &= 2\varepsilon_{\mathrm{prm}}\Big(\mathbb{P}(\omega_h = +)p_h + \mathbb{P}(\omega_h = -)(1 - p_h)\Big) \\ &= 2\varepsilon_{\mathrm{prm}}\Big(\tfrac{1}{2}p_h + \tfrac{1}{2}(1 - p_h)\Big) = \varepsilon_{\mathrm{prm}},\end{aligned} \tag{92}$$

which is independent of p_h. Summing over h gives $\mathbb{E}_\omega \mathbb{E}_{\mathsf{A}}[J_\omega(\hat{\pi})] = H\varepsilon_{\mathrm{prm}}$. Because $J_\omega(\pi^\star_\omega) = 2H\varepsilon_{\mathrm{prm}}$ for all ω,

$$\sup_\omega \mathbb{E}_{\mathsf{A}}[J_\omega(\pi^\star_\omega) - J_\omega(\hat{\pi})] \;\geq\; \mathbb{E}_\omega \mathbb{E}_{\mathsf{A}}[J_\omega(\pi^\star_\omega) - J_\omega(\hat{\pi})] = 2H\varepsilon_{\mathrm{prm}} - H\varepsilon_{\mathrm{prm}} = H\varepsilon_{\mathrm{prm}}. \tag{93}$$

This establishes the worst-case persistent floor $H\varepsilon_{\mathrm{prm}} = \bar{\varepsilon}$ for every process-only procedure, including exact PRM optimization. The residual calculation below gives $\bar{\varepsilon}^{\mathrm{sh}} = \bar{\varepsilon} = H\varepsilon_{\mathrm{prm}}$. Hence, the floor is $\Theta(\bar{\varepsilon}^{\mathrm{sh}})$ and matches the sharp upper bound $2\bar{\varepsilon}^{\mathrm{sh}}$ of Theorem 3 up to a factor of 2.

For a fixed instance ω, the all-bad policy accumulates error $+H\varepsilon_{\mathrm{prm}}$, whereas the all-good policy accumulates error $-H\varepsilon_{\mathrm{prm}}$. Therefore, $\mathrm{F}_\Pi(\varepsilon) \geq 2H\varepsilon_{\mathrm{prm}}$. Conversely, the accumulated error of every policy lies in $[-H\varepsilon_{\mathrm{prm}}, H\varepsilon_{\mathrm{prm}}]$, so no pair can differ by more than $2H\varepsilon_{\mathrm{prm}}$. It follows that $\mathrm{F}_\Pi(\varepsilon) = 2H\varepsilon_{\mathrm{prm}} = 2\bar{\varepsilon}^{\mathrm{sh}}$, which attains (53) with equality.

The scale-aware certificates have the same values. The construction gives $J^{\hat{r}}(\pi) = H\varepsilon_{\mathrm{prm}}$ for every policy, while the full Markov class contains the all-good and all-bad policies and has attainable true-return range $[0, 2H\varepsilon_{\mathrm{prm}}]$. Therefore, for every $c > 0$,

$$\mathrm{F}_\Pi(\varepsilon^{(c)}) = \mathrm{span}_{\pi\in\Pi}\left\{\frac{H\varepsilon_{\mathrm{prm}}}{c} - J_\omega(\pi)\right\} = 2H\varepsilon_{\mathrm{prm}}. \tag{94}$$

Taking the infimum gives $\mathrm{F}_\Pi^{\mathrm{ord}} = 2H\varepsilon_{\mathrm{prm}}$. Corollary 11, feasibility of $c = 1$, and feasibility of the zero potential yield

$$H\varepsilon_{\mathrm{prm}} = \frac{1}{2}\mathrm{F}_\Pi^{\mathrm{ord}} \leq \bar{\varepsilon}^{\mathrm{ord}} \leq \bar{\varepsilon}^{\mathrm{sh}} \leq \bar{\varepsilon} = H\varepsilon_{\mathrm{prm}}. \tag{95}$$

Thus all inequalities are equalities, so $\bar{\varepsilon}^{\mathrm{ord}} = \bar{\varepsilon}^{\mathrm{sh}} = \bar{\varepsilon} = H\varepsilon_{\mathrm{prm}}$ and positive scale optimization does not improve the certificate in this construction.

□

H Hybrid Policy Learning

H.1 Hybrid Reward Objectives

Proof of Corollary 5. The assumed ζ-optimality and $\pi^\star \in \Pi$ imply

$$J_\lambda(\hat{\pi}_\lambda) \geq \sup_{\pi\in\Pi} J_\lambda(\pi) - \zeta \geq J_\lambda(\pi^\star) - \zeta, \tag{96}$$

and hence

$$J_\lambda(\pi^\star) - J_\lambda(\hat{\pi}_\lambda) \leq \zeta. \tag{97}$$

Substituting $J_\lambda(\pi) = J(\pi) + \lambda\mathcal{E}(\pi)$ for both policies and rearranging gives

$$J(\pi^\star) - J(\hat{\pi}_\lambda) = J_\lambda(\pi^\star) - J_\lambda(\hat{\pi}_\lambda) + \lambda\big[\mathcal{E}(\hat{\pi}_\lambda) - \mathcal{E}(\pi^\star)\big]. \tag{98}$$

Both policies belong to Π, so Definition 1 gives $\mathcal{E}(\hat{\pi}_\lambda) - \mathcal{E}(\pi^\star) \leq \mathrm{F}_\Pi(\varepsilon)$. Combining this inequality with (97) and (98), and using $\lambda \geq 0$, yields

$$J(\pi^\star) - J(\hat{\pi}_\lambda) \leq \zeta + \lambda\mathrm{F}_\Pi(\varepsilon). \tag{99}$$

Proposition 10 gives $\mathrm{F}_\Pi(\varepsilon) \leq 2\bar{\varepsilon}^{\mathrm{sh}}$. Multiplication by $\lambda \geq 0$ preserves this inequality and proves the second bound in (14). □

H.2 Finite-Budget Hybrid Learning

We now analyze an empirical hybrid objective on the unknown-dynamics family in Section 3.2. The learner receives the fixed PRM preferences z and terminal outcomes from the same $n \geq 1$ training trajectories. It knows the public parameters m, k, α, γ and samples actions independently and uniformly at every decision block. It computes

$$S_j = \sum_{i=1}^{n} v_{ij}(2Y_i - 1), \qquad \widehat{\theta}_j = \mathrm{sign}\left(\sum_{i=1}^{n} x_{ij}v_{ij}\right), \tag{100}$$

and selects

$$\widehat{x}_j^{\mathrm{hyb}} = \widehat{\theta}_j\,\mathrm{sign}(S_j + t_{\mathrm{prm}}z_j), \qquad t_{\mathrm{prm}} = \frac{\log((m-k)/k)}{\log((1/2+\alpha)/(1/2-\alpha))}. \tag{101}$$

All sign ties use independent uniform randomization. The rule keeps the PRM preference at block j when $z_j S_j > -t_{\mathrm{prm}}$ and reverses it when $z_j S_j < -t_{\mathrm{prm}}$. The threshold follows from a likelihood-ratio comparison of the outcome distributions for correct and incorrect PRM preferences, weighted by $1 - k/m$ and k/m, respectively; see the proof of Proposition 16.

The statistic $S_j/(2n)$ is an unbiased estimate of the true reward coefficient $\alpha\omega_j$. The PRM supplies the coefficient αz_j. Set

$$\lambda_n = \frac{t_{\mathrm{prm}}}{2n\alpha + t_{\mathrm{prm}}}. \tag{102}$$

Their weighted estimate satisfies

$$(1-\lambda_n)\frac{S_j}{2n} + \lambda_n \alpha z_j = \frac{\alpha}{2n\alpha + t_{\mathrm{prm}}}(S_j + t_{\mathrm{prm}} z_j). \tag{103}$$

Under the estimated transition model $\widehat{P}(v_j = v \mid x_j = x) = 1/2 + \gamma\widehat{\theta}_j xv$, each action in (101) maximizes the expected reward with this coefficient. Thus the rule optimizes an empirical version of (13). The weight λ_n decreases with the training budget, so outcome evidence gradually replaces the initial PRM preference.

Algorithm 1 PRM-guided hybrid learning

Require: PRM preferences z; public parameters m, k, α; budget $n \geq 1$.
1: Collect n trajectories and terminal outcomes with independent uniform actions at each block.
2: Set $t_{\mathrm{prm}} = \log((m-k)/k)/\log((1/2+\alpha)/(1/2-\alpha))$.
3: Use independent uniform randomization for all sign ties.
4: **for** each block $j = 1, \ldots, m$ **do**
5: Compute $S_j = \sum_{i=1}^n v_{ij}(2Y_i - 1)$ and $\widehat{\theta}_j = \mathrm{sign}(\sum_{i=1}^n x_{ij} v_{ij})$.
6: Set $\widehat{x}_j^{\mathrm{hyb}} = \widehat{\theta}_j \,\mathrm{sign}(S_j + t_{\mathrm{prm}} z_j)$.
7: **end for**
8: **return** The policy with decision actions $\widehat{x}_j^{\mathrm{hyb}}$.

To state the exact risk, write $\varrho = k/m$ and let

$$f_n^{\pm}(b) = \mathbb{P}[\mathrm{Bin}(n, 1/2 \pm \alpha) = b], \qquad e_n^{\mathrm{hyb}} = \sum_{b=0}^{n} \min\{(1-\varrho) f_n^{+}(b), \varrho f_n^{-}(b)\}. \tag{104}$$

The quantity e_n^{hyb} is the mean probability of an incorrect reward preference across the m blocks. We retain $p_n = p_n(\alpha)$, $q_n = p_n(\gamma)$, $s_n = 1 - 2q_n$, and $D_n = 4m\alpha\gamma q_n$ from Appendix C.2.

Proposition 16 (Finite-budget hybrid improvement)**.** *Under the parameter conditions of Theorem 2, fix any task (θ, ω, z) and an integer $n \geq 1$. Let $R_{\mathrm{hyb}}^{\mathrm{dyn}}(n; \alpha, \gamma)$ be the expected regret of* (101) *under the sampling procedure above. Then*

$$\begin{aligned} e_n^{\mathrm{hyb}} &\leq \min\{k/m, p_n\}, \\ R_{\mathrm{hyb}}^{\mathrm{dyn}}(n; \alpha, \gamma) &= D_n + 4 s_n m \alpha \gamma e_n^{\mathrm{hyb}} \\ &\leq \min\{\mathcal{R}_{\mathrm{proc}}^{\mathrm{dyn}}(n; \alpha, \gamma), R_{\mathrm{mom}}^{\mathrm{dyn}}(n; \alpha, \gamma)\}. \end{aligned} \tag{105}$$

The last inequality is strict for $n > t_{\mathrm{prm}}$. The outcome comparator is the moment estimator in (41)*; the process comparator is the majority estimator that attains* (32)*.*

Proof. Fix a block j. Equation (40) makes v_i uniform on $\{-1, +1\}^m$ and gives

$$\mathbb{E}[v_{ij}(2Y_i - 1)] = 2\alpha\mathbb{E}[v_{ij}\langle\omega, v_i\rangle] = 2\alpha\omega_j.$$

Since $v_{ij}(2Y_i - 1)$ is a sign, $z_j v_{ij}(2Y_i - 1) = 1$ has probability $1/2 + \alpha$ when $z_j = \omega_j$ and $1/2 - \alpha$ when $z_j = -\omega_j$. The count $B_j = (n + z_j S_j)/2$ therefore has mass function f_n^{+} or f_n^{-}, respectively.

Every fixed task has $m - k$ correct PRM preferences and k incorrect ones. After averaging the preference error over blocks, a rule that keeps z_j at count b contributes $\varrho f_n^{-}(b)$, while a rule that

reverses it contributes $(1-\varrho)f_n^+(b)$. These weights follow from the fixed counts of the two types of blocks. For $\kappa_\alpha = (1/2+\alpha)/(1/2-\alpha) > 1$, the likelihood ratio is

$$\frac{f_n^+(b)}{f_n^-(b)} = \kappa_\alpha^{2b-n}.$$

The smaller contribution comes from keeping z_j exactly when $2b - n \geq -t_{\mathrm{prm}}$, with either decision optimal at equality. Hence (101) has mean preference error e_n^{hyb}. The rule that always keeps z_j has mean error ϱ, and the rule that uses $\mathrm{sign}(S_j)$ has mean error p_n. The pointwise minimum in (104) cannot exceed the error of either rule, which proves the first inequality in (105).

Equation (40) also makes the transition statistics independent of all (v_i, Y_i). With independent tie randomization, the transition error at each block is independent of the reward-preference error. If the latter has probability a_j, the action error has probability $q_n(1-a_j) + (1-q_n)a_j = q_n + s_n a_j$. Each wrong action costs $4\alpha\gamma$, and $\sum_j a_j = m e_n^{\mathrm{hyb}}$. The expected regret is therefore $D_n + 4 s_n m \alpha\gamma e_n^{\mathrm{hyb}}$. Substituting $e_n^{\mathrm{hyb}} \leq \varrho$ and $e_n^{\mathrm{hyb}} \leq p_n$ gives the two comparisons.

For strictness relative to process learning, suppose $n > t_{\mathrm{prm}}$. At $b = 0$, the weighted error contribution of reversing the PRM preference is strictly smaller than that of retaining it. Both binomial laws assign positive mass to this event, so $e_n^{\mathrm{hyb}} < \varrho$.

For the outcome comparison, the parameter conditions imply

$$\frac{1-\varrho}{\varrho} \geq 3, \qquad 1 < \kappa_\alpha \leq \frac{9}{7}, \qquad t_{\mathrm{prm}} > 1.$$

If n is even, $b = n/2$ has positive mass under both laws. At this count, the hybrid rule keeps z_j and has smaller weighted error than the uniform tie rule of the outcome estimator. If n is odd, the count $b = (n-1)/2$ gives $z_j S_j = -1$. Since $t_{\mathrm{prm}} > 1$, the hybrid rule again keeps z_j, while the outcome estimator reverses it and incurs strictly larger weighted error. Thus $e_n^{\mathrm{hyb}} < p_n$ for every $n \geq 1$. Finally, $\gamma > 0$ implies $q_n < 1/2$ and hence $s_n > 0$. Both regret comparisons are strict when $n > t_{\mathrm{prm}}$. □

I Experimental Details and Additional Results

I.1 Experimental Details

I.1.1 Effect of PRM Accuracy

We use the family in (3) with $H = 32$, $\alpha = 1/128$, and one true sign vector ω drawn using seed 20260911. A random permutation of the stages determines the PRM errors: the model with k errors flips the first k coordinates in this permutation. The true rewards, transitions, and outcome distribution are identical across $k \in \{1, 2, 4, 8\}$, and each PRM remains fixed during training.

For each of 512 independent runs, we sample actions uniformly and generate terminal outcomes from the Bernoulli model. All four PRM comparisons use the same training data. We evaluate 65 logarithmically spaced sample sizes from 16 to 32,768, rounding to integers and removing duplicates. At sample size n, the moment estimator uses the first n observations and breaks ties independently and uniformly at random. We compute the regret of the learned policy from the known task parameters. Variation across runs comes from the sampled training data and random tie breaking.

At each sample size, we report the mean and sample standard deviation of regret across the 512 runs. We estimate the number of training trajectories needed for mean outcome regret to fall below process regret by locating the first such sample size and interpolating linearly in log sample size between adjacent measurements. We compute the standard deviation of this estimate from 2,000 bootstrap resamples with seed 20260924. Each resample retains the measurements of each selected run at all sample sizes and for all PRMs. Table 3 reports this bootstrap standard deviation and the first integer sample size at which expected outcome regret is strictly below process regret.

I.1.2 Effect of Horizon Length

For each $H \in \{8, 16, 32, 64\}$, set $k = H/8$ and $\alpha_0 = 1/(4H)$. At every sample size, we consider the same range $0 < \alpha \leq \alpha_0$, with process minimax regret $b = 2k\alpha_0 = 1/16$. The maximum expected

k	Process regret	Theory	Estimate	Std.
1	0.015625	14,211	14,372	276
2	0.031250	9,639	9,676	165
4	0.062500	5,421	5,564	118
8	0.125000	1,863	1,840	51

Table 3: Effect of PRM accuracy on a fixed true task. The theory and estimate columns give the number of training trajectories needed for outcome regret to fall below process regret. The last column reports the standard deviation of the bootstrap estimates.

regret of the moment estimator over this range and the outcome minimax lower bound are

$$U_H(n) = \max_{0<\alpha\le\alpha_0} 2H\alpha p_n(\alpha), \qquad L_H(n) = \max_{0<\alpha\le\alpha_0} \frac{27}{32} H\alpha e^{-64n\alpha^2/3}.$$

For fixed n, both objectives extend continuously to $\alpha = 0$ with value zero. To compute each maximum, we evaluate the objective on a logarithmic grid of α, refine the search near the best grid point by bounded one-dimensional optimization, and retain the largest value found. The bounds satisfy $L_H(n) \le \mathcal{R}_{\mathrm{out}}^{\le\alpha_0}(n) \le U_H(n)$.

The theoretical curves use 45 logarithmically spaced values of the scaled sample size nb^2/H^2 between 0.001 and 0.3, with n rounded to an integer. We run the experiment at scaled sample sizes 0.005, 0.015, 0.03, and 0.08. At each sample size, we find the value of α that maximizes expected regret, then generate 256 independent training datasets at that value. Each horizon uses a fixed true sign vector, and seed 20260912 determines the task instances and training data. Table 4 reports the first sample size at which $U_H(n)$ falls below b.

H	k	Training trajectories
8	1	473
16	2	1,893
32	4	7,573
64	8	30,293

Table 4: Training trajectories needed for the maximum expected regret of the moment estimator to fall below b, with $b = 1/16$ and $k/H = 1/8$.

I.1.3 Learning with Unknown Dynamics

We use the $H = 2m$ family from Section 3.2 with $m = 8$, $k = 1$, $\gamma = 1/4$, and $\alpha = 1/256$. The vectors θ, ω, and z are drawn once using seed 20260913 and then kept fixed. The process learner receives z and estimates θ from the observed transitions. The outcome learner uses transitions and terminal outcomes to estimate θ and ω. Both learners use the same trajectories, generated with uniform actions and independent transition draws at every block.

We perform 512 independent runs at 60 sample sizes, obtained by rounding 65 logarithmically spaced values from 1 to $65{,}536$ and removing duplicates. The transition estimator uses $\operatorname{sign}\sum_i x_{ij}v_{ij}$, and the reward estimator uses $\operatorname{sign}\sum_i v_{ij}(2Y_i - 1)$. Ties are broken independently and uniformly at random. We compute the regret of each learned policy from the known task parameters. The limiting process regret is $b_\alpha = 1/256$. We report the sample standard deviation of regret across the runs. The estimated sample requirement is the point at which the mean difference between outcome and process regret becomes negative, with interpolation in log sample size. Figure 2(d) compares the experimental means with the expected regrets in (32) and (41).

To examine how transition strength affects learning, we fix $\alpha = 1/1024$ and vary $\gamma \in \{1/16, 1/8, 1/4\}$. For each value, we perform 512 independent runs at logarithmically spaced sample sizes up to $4/\gamma^2$, using the same fixed sign vectors. Figure 3 plots process regret divided by b_α against $n\gamma^2$ and compares it with the expected regret formula. As the transition estimates improve, process regret approaches the limit set by the PRM errors.

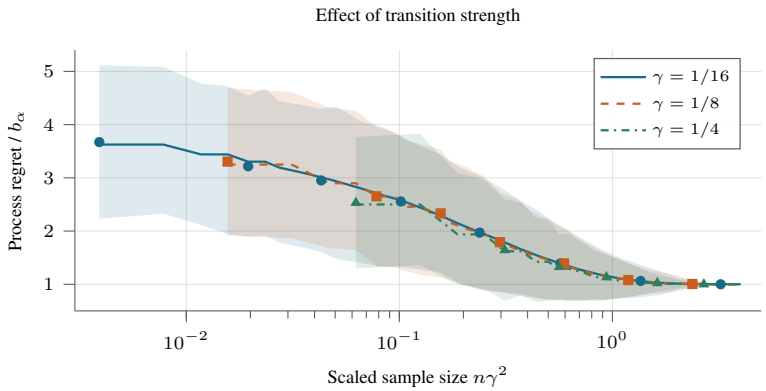

Figure 3: Effect of transition strength on process regret. Lines show expected regret, markers show means across 512 independent runs, and shading shows the mean $\pm$ one standard deviation across runs.

I.1.4 Hybrid Reward Learning

We evaluate the hybrid estimator in (101) on the same fixed task as Appendix I.1.3, with $m = 8$, $k = 1$, $\alpha = 1/256$, and $\gamma = 1/4$. We generate 2,048 independent training runs using seed 20260925 and evaluate budgets $n \in \{4096, 8192, 16384, 32768, 65536\}$. Within each run, the three methods use the same trajectory prefixes. Each method uses n rollouts in total, and the hybrid method receives both the fixed PRM and the observed outcomes.

The process baseline estimates transition signs and follows the PRM preferences. The outcome baseline estimates both transition signs and reward preferences from its observations. The hybrid estimator uses the same transition estimate and applies the threshold $t_{\mathrm{prm}} = 124.5357157611$, which follows directly from the public task parameters. We select each policy using only its permitted observations and compute true regret from the fixed environment after training.

Figure 4 compares the empirical means with the exact expected regrets in (32), (41), and (105). Each error bar shows one sample standard deviation across independent runs. We also compute the regret difference between the hybrid method and each baseline within every run. These paired differences have negative means at all five budgets. At $n = 32{,}768$, the hybrid mean regret is 0.001404, compared with 0.003906 for process and 0.002413 for outcome.

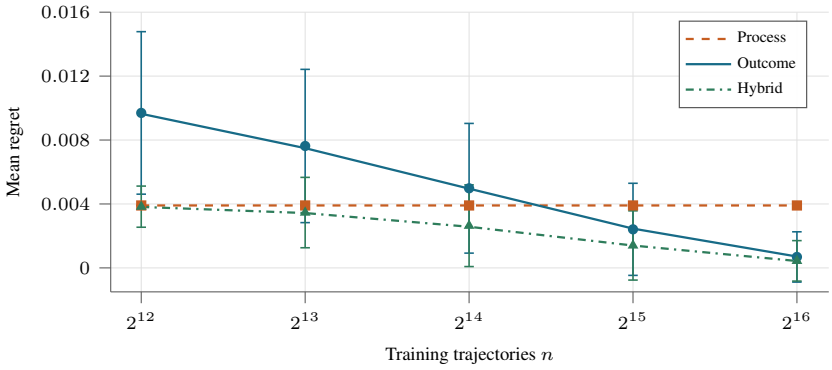

Figure 4: Hybrid learning on the fixed unknown-dynamics task. All methods receive the same total training budget. Lines give exact expected regret, and markers give means across 2,048 independent runs with error bars of one standard deviation. The hybrid rule uses the fixed PRM and outcome statistics from the same trajectories.

I.2 Effect of Potential Shaping

We add the same reward to both actions at each step and measure the change in reward error and regret.

The action-independent chain has $H = 16$, true rewards $(1.5e_0, 0)$, and base errors $(-e_0, +e_0)$, with $e_0 = 0.001$. The PRM rewards $(0.5e_0, e_0)$ favor the suboptimal action. We add $c_\Phi e_0$ to both actions for $c_\Phi \in \{0, 1, 2, 4, 8, 16, 32\}$ through the potential $\Phi_h = -(H - h + 1)c_\Phi e_0$. The difference between the action errors stays at $2e_0$, so the shaping-corrected residual remains e_0 per step. All true and PRM trajectory returns lie in $[0, 1]$; their maxima are 0.024 and 0.528, respectively.

Raw reward error increases from 0.016 to 0.528, while the residual and regret remain at 0.016 and 0.024, respectively (Figure 5). Removing the base error gives zero regret at every shaping multiplier. The common reward shift changes the raw error but preserves the preferred action, so both regret and the residual remain constant. Appendix I.3 compares the regret bounds on randomly generated MDPs.

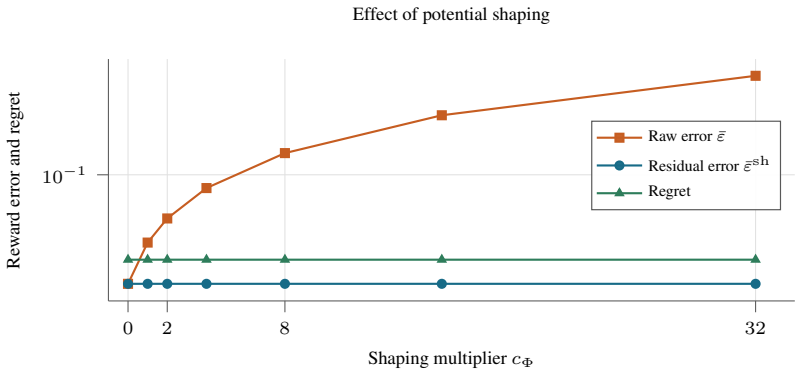

Figure 5: Effect of a common potential shaping shift on reward error and policy regret. The raw error increases, while the shaping-corrected residual and regret remain unchanged.

I.3 Regret Bounds on Random MDPs

We compare the regret bounds on 512 randomly generated finite-horizon MDPs with $H = 5$, $|\mathcal{S}| = 6$, and $|\mathcal{A}| = 3$, using seed 20260925. For each instance, the initial distribution and the transition rows at stages $h < H$ are independent $\mathrm{Dirichlet}(1)$ draws. The final transition leads to a fixed terminal state. We generate each true reward entry as $(0.45 + 0.10X)/H$, with independent $X \sim \mathrm{Beta}(2, 2)$.

We draw a noise level $\nu \sim \mathrm{Unif}[0.0075/H, 0.04/H]$ and sample $\xi_h(s, a) \sim \mathrm{Unif}[-\nu, \nu]$ independently. For the shaping component, the potential scale s_Φ is log-uniform on $[0.005/H, 0.15/H]$, and $\Phi_h(s) \sim \mathrm{Unif}[-s_\Phi, s_\Phi]$ independently, with $\Phi_{H+1} = 0$. The PRM is $\hat{r} = r^\star + \xi + g^\Phi$, with the two error components sampled independently. These ranges place each true reward in $[0.45/H, 0.55/H]$ and each PRM reward in $[0.11/H, 0.89/H]$. Every trajectory therefore has a cumulative true reward in $[0.45, 0.55]$ and a cumulative PRM reward in $[0.11, 0.89]$.

For the full Markov policy class, we use backward induction to compute the optimal policies under the true and PRM rewards, together with $\mathrm{F}_\Pi(\varepsilon) = \max_\pi J^\varepsilon(\pi) + \max_\pi J^{-\varepsilon}(\pi)$. The linear program in (62) gives $\bar{\varepsilon}^{\mathrm{sh}}$. We compare regret, $L = J(\pi^\star) - J(\hat{\pi})$, with the raw-error bound $2\bar{\varepsilon}$, the shaping-corrected bound $2\bar{\varepsilon}^{\mathrm{sh}}$, and the error span $\mathrm{F}_\Pi(\varepsilon)$. All 512 instances satisfy $L \le \mathrm{F}_\Pi(\varepsilon) \le 2\bar{\varepsilon}^{\mathrm{sh}} \le 2\bar{\varepsilon}$.

Spearman correlations with regret are 0.143 for the raw-error bound, 0.848 for the shaping-corrected bound, and 0.853 for the error span (Figure 6). Removing potential shaping accounts for most of the increase in correlation; using policy occupancies gives a smaller further increase.

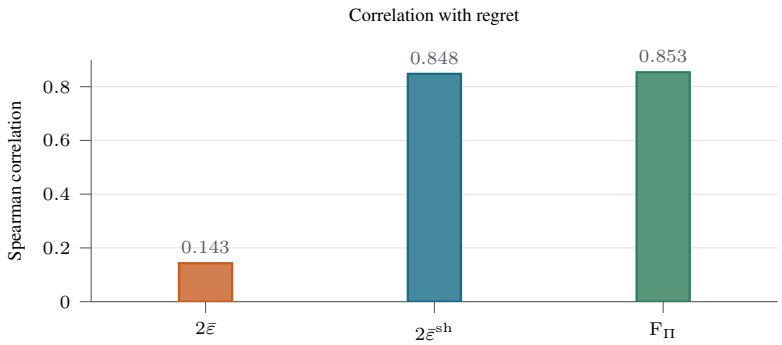

Figure 6: Spearman correlation between each regret bound and policy regret across 512 randomly generated MDPs. The bounds account for raw reward error ($2\bar{\varepsilon}$), potential shaping ($2\bar{\varepsilon}^{\mathrm{sh}}$), or policy occupancies (F_{Π}).

I.4 Exact Policy-Comparison Certificates

We use the same 512 random MDPs, full Markov policy class, and saved reward tables as above. For each instance, we solve the primal problem in (72) and the dual problem in (73) at four tolerances, $\zeta/\Delta_{\Pi}(\hat{r}) \in \{0, 0.05, 0.10, 0.20\}$. The primal LP optimizes the true-return loss over two occupancy measures subject to the PRM comparison constraint. The dual LP optimizes over the nonnegative multiplier t and the two finite-horizon value-function potentials. We include $t = 0$, so the computation also covers a minimizing sequence whose corresponding positive PRM scale tends to infinity.

For each tolerance, we compare $\mathcal{B}_{\Pi}(\zeta)$ with four bounds: $\zeta + 2\bar{\varepsilon}$, $\zeta + 2\bar{\varepsilon}^{\mathrm{sh}}$, the fixed-scale span bound $\zeta + \mathrm{F}_{\Pi}(\hat{r} - r^{\star})$, and the return-span-normalized bound. We clip every bound at the true-return span before comparing them. The best standard bound is the minimum of these four values. The exact certificate is smaller than the best standard bound on all 512 instances at every tolerance. The mean reduction is 5.9%, 6.8%, 7.8%, and 9.5% at the four tolerances, respectively. The maximum absolute difference between the primal and dual values over all 2,048 settings is 1.02×10^{-9}.