OPTIMAL MULTI-REWARD REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We study an unknown-transition finite-horizon Markov decision process (MDP) with a finite collection of known reward functions $\{r^1, r^2, \ldots, r^M\}$. The goal is to output an ϵ-optimal policy for every reward using online episodic interaction only. Performance is measured by the policy error $V_0^{*,m} - V_0^{\widehat{\pi}^m,m}$ where $m \in [M]$ represents the reward function and $V_0^{*,m} = \mathbb{E}_{s_1\sim\mu}[V_1^{*,m}(s_1)]$. Under this setting, we design a provably efficient algorithm to establish a minimax sample complexity bound of

$$O\left(\frac{SAH^3}{\epsilon^2}\log M \operatorname{polylog}\left(\frac{SAH\log M}{\min\{\epsilon,1\}\,\delta}\right)\right)$$

episodes, with no additional burn-in cost. This matches the information-theoretic lower bound up to a factor of $\operatorname{polylog}(SAH\log M/(\min\{\epsilon,1\}\,\delta))$. Our method combines three technical ingredients. First, we adapt MVP (Zhang et al., 2025) to reward-switching learning to construct optimistic value estimates. Second, we use fresh replay samples to conservatively evaluate the candidate policies. Third, gap-based multiplicative weights updates adjust the reward-sampling distribution using the differences between these estimates, converting weighted learning progress into simultaneous guarantees for all rewards.

1 INTRODUCTION

Reinforcement learning (RL) studies sequential decision-making problems in which an agent interacts with an environment, observes the resulting states, and chooses actions to maximize a cumulative reward. In the standard finite-horizon formulation, the environment is modeled as a Markov decision process (MDP) with finite state-action space $\mathcal{S} \times \mathcal{A}$ and horizon H. Under this setting, a substantial literature has developed optimistic and variance-sensitive algorithms with near-optimal regret and sample-complexity guarantees (Azar et al., 2017; Jin et al., 2018; Zhang et al., 2025).

However, these results typically focus on a single reward function, an assumption that is restrictive in many applications. A fixed environment may be evaluated according to several objectives, such as safety, efficiency, fairness, or user preference, while the underlying transition dynamics remain unchanged. For example, a recommender system may be required to optimize different notions of user utility, and a control system may need to support multiple task specifications without relearning the environment from scratch. The same phenomenon appears in reward-mixing MDPs and contextual decision processes, where a common environment is used to study multiple objectives or task-dependent criteria (Hallak et al., 2015). Since all reward functions share the same transition kernel, trajectories collected for one objective may contain useful information for the others. Exploiting this shared structure is therefore both statistically natural and practically important.

At the other extreme, reward-free or reward-agnostic exploration aims to collect data that supports policy learning for rewards that are unknown during exploration or specified only afterward (Jin et al., 2020; Li et al., 2026). Such formulations require broad coverage of the MDP and can be more demanding than the setting considered here. We study an intermediate but fundamental problem: the learner is given M different reward functions in advance and may use them during exploration, but it must eventually return an approximately optimal policy for every reward. A naive approach that runs a separate RL algorithm for each reward incurs a factor proportional to M, whereas an

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.

overly reward-agnostic strategy may collect substantially more data than is necessary for the given reward class.

Recent work by Zhang et al. (2025) established that a near-optimal policy for a single reward function can be learned with the minimax-optimal sample complexity $\widetilde{O}(SAH^3/\epsilon^2)$, without additional burn-in terms. This naturally raises the following fundamental question:

Can we simultaneously learn near-optimal policies for M different reward functions using $\widetilde{O}\big(\frac{SAH^3 \log M}{\epsilon^2}\big)$ episodes, while still avoiding any additional burn-in terms?

1.1 Our Contribution

We answer this question affirmatively. More precisely, we design an online algorithm such that: for any fixed failure probability $\delta \in (0,1)$ and error threshold $\epsilon \in (0, H]$, with probability at least $1-\delta$, the algorithm returns an ϵ-optimal policy for every reward function using

$$O\left(\frac{SAH^3}{\epsilon^2} \log M \,\mathrm{polylog}\left(\frac{SAH \log M}{\min\{\epsilon, 1\}\,\delta}\right)\right)$$

episodes (see the formal statement in Theorem 1). This sample-complexity bound shows that policies for multiple reward functions can be learned jointly and efficiently, with at most an additional $\log M$ factor over the minimax-optimal sample complexity for a single reward function. On the other hand, we adapt the recent lower-bound construction for reward-agnostic reinforcement learning Ridel & Cohen (2026) to show that any algorithm for simultaneously learning near-optimal policies for M reward functions requires $\Omega\big(SAH^3 \log(M)/\epsilon^2\big)$ exploration episodes; see Appendix F for the detailed construction. Consequently, when $\log M \le S$, our upper bound matches the lower bound in Corollary 1 up to a multiplicative factor of $\mathrm{polylog}(SAH \log M/(\min\{\epsilon, 1\}\,\delta))$.

1.2 Related Works

Tabular MDPs. A large body of theoretical work on online reinforcement learning considers tabular MDPs in which all episodes share the same transition kernel, reward function, and action space. Over the past several decades, substantial progress has been made toward improving the sample efficiency of both model-based and model-free algorithms. Representative results include regret and PAC guarantees based on optimism, posterior sampling, and variance-aware exploration (Zhang et al., 2022; Agrawal & Jia, 2017; Szita & Szepesvari, 2010; Dann et al., 2017; Domingues et al., 2021; Azar et al., 2017; Efroni et al., 2019; Neu & Pike-Burke, 2020), as well as sample-efficient model-free methods (Jin et al., 2018; Dong et al., 2019; Ménard et al., 2021b; Li et al., 2021; 2022; Wang et al., 2020). Another line of work develops problem-dependent and gap-dependent guarantees (Zanette & Brunskill, 2019; Simchowitz & Jamieson, 2019; Yang et al., 2021; Xu et al., 2021; Chen et al., 2025). Most closely related to our analysis, Zhang et al. (2025) established the all-range minimax regret bound $\widetilde{O}(\min\{\sqrt{SAH^3K},\, HK\})$, which implies the minimax-optimal sample complexity $\widetilde{O}(SAH^3/\epsilon^2)$ without an additional burn-in term.

Reward-agnostic reinforcement learning. In reward-agnostic RL, the learner explores the environment without access to the downstream reward functions, which belong to a fixed class of cardinality M chosen independently of the exploration data. This framework therefore naturally extends to the simultaneous learning of policies for multiple rewards. In this setting, Zhang et al. (2020) first established a sample-complexity bound of $\widetilde{O}\big(SAH^5 \log(M)/\epsilon^2\big)$. Subsequent work by Li et al. (2026) improved the dependence on the horizon, obtaining $\widetilde{O}\big(SAH^3\mathrm{polylog}(M)/\epsilon^2 + S^4A^4H^6\mathrm{polylog}(M)/\epsilon\big)$ for reward classes of polynomial cardinality. More recently, Ridel & Cohen (2026) further improved the burn-in dependence, yielding a sample-complexity bound of $\widetilde{O}\big(SAH^3\mathrm{polylog}(M)/\epsilon^2 + S^2AH^4\mathrm{polylog}(M)/\epsilon\big)$.[1] Thus, for fixed reward classes of polynomial cardinality, the improved reward-agnostic bound has a leading term of $\widetilde{O}\big(SAH^3\mathrm{polylog}(M)/\epsilon^2\big)$ and an additional burn-in term of order $\widetilde{O}\big(S^2AH^4\mathrm{polylog}(M)/\epsilon\big)$. We refer the reader to Table 1 for a detailed comparison.

[1] The guarantees of Li et al. (2026); Ridel & Cohen (2026) cover fixed reward classes of cardinality polynomial in S, A, H. We write $\mathrm{polylog}(M)$ conservatively: replacing δ by δ/M inherits the original powers of the confidence logarithms, which need not equal one.

Table 1: Comparison of online episodic sample-complexity bounds for learning near-optimal policies for M reward functions.

Reference	Sample complexity	Comment
Jin et al. (2020)	$\widetilde{O}\left(\frac{S^2AH^5}{\epsilon^2} + \frac{S^4AH^7}{\epsilon}\right)$	Reward-free exploration.
Zhang et al. (2020)	$\widetilde{O}\left(\frac{SAH^5 \log M}{\epsilon^2}\right)$	–
Ménard et al. (2021a)	$\widetilde{O}\left(\frac{S^2AH^3}{\epsilon^2}\right)$	Reward-free exploration.
Li et al. (2026)	$\widetilde{O}\left(\frac{SAH^3 \log M}{\epsilon^2} + \frac{S^4A^4H^6 \log M}{\epsilon}\right)$	–
Ridel & Cohen (2026)	$\widetilde{O}\left(\frac{SAH^3 \log M}{\epsilon^2} + \frac{S^2AH^4 \log M}{\epsilon}\right)$	–
This work	$\widetilde{O}\left(\frac{SAH^3 \log M}{\epsilon^2}\right)$	No burn-in terms.
Lower bound	$\tilde{\Omega}\left(\frac{SAH^3 \min\{\log M, S\}}{\epsilon^2}\right)$	See Appendix F.

The lower-bound construction adapts those of Jin et al. (2020); Ridel & Cohen (2026).

2 PRELIMINARIES

In this section, we introduce the basics of finite-horizon MDPs, as well as the basic episode setting.

Basics of finite-horizon MDPs. This paper focuses on time-inhomogeneous finite-horizon MDPs. Let $\mathcal{S}$ be a finite state space with $|\mathcal{S}| = S$, and let $\mathcal{A}$ be a finite action space with $|\mathcal{A}| = A$, and let H be the horizon. The notation $P = \{P_h(\cdot|s,a)\}_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}}$ denotes the transition kernel. If action a is taken, then the state at the next step $h+1$ of the environment is randomly drawn from $P_{h,s,a} := P_h(\cdot|s,a) \in \Delta(\mathcal{S})$ and receives the reward $r_h(s,a) \in [0,1]$ instantaneously. A policy $\pi = \{\pi_h\}_{h=1}^H$ consists of functions that map each state to a probability distribution over the action space. When π is deterministic, we write $\pi_h(s)$ for the action taken at state s in layer h. We also allow episode-level mixtures of such policies.

Multi-Reward Setting. We consider the multi-reward setting with M known reward functions. Specifically, the reward function is denoted by $r^m = \{r_h^m(s,a)\}_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}}$ where $r_h^m(s,a) \in [0,1]$ for all $(h,s,a,m) \in [H]\times\mathcal{S}\times\mathcal{A}\times[M]$. All rewards share the same transition kernel of the underlying finite-horizon MDP. We state our results for $M \geq 2$; the single-reward case is covered by Zhang et al. (2025).

Value function and Q-function. Define the reward-specific value function $V^{\pi,m}$ and the Q-function $Q^{\pi,m}$ as

$$V_h^{\pi,m}(s) := \mathbb{E}^{\pi,m}\left[\sum_{h'=h}^{H} r_{h'}^m(s_{h'},a_{h'})|s_h = s\right], \qquad \forall (h,s) \in [H]\times\mathcal{S}$$

$$Q_h^{\pi,m}(s,a) := \mathbb{E}^{\pi,m}\left[\sum_{h'=h}^{H} r_{h'}^m(s_{h'},a_{h'})|s_h = s, a_h = a\right], \qquad \forall (h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}$$

where the expectation $\mathbb{E}^{\pi,m}[\cdot]$ is taken over the randomness of an episode $\{(s_h,a_h)\}_{h=1}^H$ generated under policy π and reward function r^m. Denote $V_0^{\pi,m} = \mathbb{E}_{s_1\sim\mu}[V_1^{\pi,m}(s_1)]$ where μ is the initial distribution. Accordingly, we define the optimal reward-specific value function and Q-function as

$$V_h^{*,m}(s) := \max_\pi V_h^{\pi,m}(s), \qquad \forall (h,s) \in [H]\times\mathcal{S}$$

$$Q_h^{*,m}(s,a) := \max_\pi Q_h^{\pi,m}(s,a), \qquad \forall (h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}$$

and $V_0^{*,m} = \max_\pi V_0^{\pi,m}$.

The learning problem. Let $V_0^{*,m}$ be the optimal expected value function subject to the reward function r^m and initial distribution μ, and $V_0^{\pi,m}$ be the value of π in r^m. In the multi-reward setting, each reward function r^m induces its own optimal policy. The PAC sample complexity quantifies the number of episodes required to learn an ϵ-optimal policy collection $\{\widehat{\pi}^m\}_{m\in[M]}$ satisfying

$$V_0^{*,m} - V_0^{\widehat{\pi}^m,m} \le \epsilon, \quad \forall m \in [M]$$

simultaneously with high probability.

Notations. We denote $\mathbb{V}(P, X) := \langle P, X^2\rangle - \langle P, X\rangle^2$ as the variance on random variable X on distribution P, where X^2 is element-wise square of X. We use $[N]$ to denote the set $\{1, 2, \ldots, N\}$.

3 ALGORITHM

We introduce `Online-Multi-Reward-Learning-with-Exponential-Weight` (Algorithm 1), which adapts the `MVP` framework in Zhang et al. (2021) to the multi-reward MDP setting by incorporating three novel techniques: *shared-model multi-reward learning*, *sealed replay*, and *multiplicative weight update*. In each phase ℓ, we generate candidate policies $\{(\pi^{\ell,k,m})_{k\in[K]}\}_{m\in[M]}$ and optimistic estimate $\hat{v}^+_{\ell,m}$ using `Multi-Reward-Switching-MVP` (Algorithm 2). Then, `Sealed-Replay` (Algorithm 3) collects fresh samples to construct the retained empirical kernel $\widetilde{P}$. Finally, we evaluate the candidate policies under $\widetilde{P}$ using `Conservative-Evaluation` (Algorithm 4) to derive the conservative estimate $\hat{v}^-_{\ell,m}$, and use the estimate gap $\hat{v}^+_{\ell,m} - \hat{v}^-_{\ell,m}$ to update the reward sample distribution p_ℓ with multiplicative weight.

Shared-model multi-reward learning. At the beginning of phase ℓ, we fix the reward-sampling distribution p_ℓ and run `Multi-Reward-Switching-MVP`, the multi-reward MVP subroutine (Algorithm 2), for K episodes. At the beginning of each episode k, following prior work Zhang et al. (2025), the subroutine first computes, for every reward index $m \in [M]$, an optimistic value function $\overline{V}_1^{\ell,k,m}$ and the corresponding greedy counterfactual policy $\pi^{\ell,k,m}$, using the shared empirical transition kernel $\widehat{P}$ and the reward-specific exploration bonuses $b_h^{\ell,k,m}(s,a)$. The agent then samples a reward index $m^{\ell,k} \sim p_\ell$ and executes the roll-out policy $\pi^{\ell,k} := \pi^{\ell,k,m^{\ell,k}}$.

The empirical transition model and the statistics maintained under the doubling schedule are carried over across phases, allowing learning for all rewards to benefit from previously collected transitions. The subroutine records all counterfactual policies $\{\pi^{\ell,k,m}\}_{k\in[K],\, m\in[M]}$, the roll-out policies $(\pi^{\ell,k})_{k\in[K]}$, and the initial states $(s_1^{\ell,k})_{k\in[K]}$. It returns the average optimistic estimates $\hat{v}^+_{\ell,m}$ for all $m \in [M]$ (line 21), together with the visit counts $N_{\ell,h}(s,a)$ accumulated only during the current phase.

Sealed replay. In line 7, we invoke `Sealed-Replay` (Algorithm 3 in Appendix B) to execute each of the K roll-out policies $(\pi^{\ell,k})_{k\in[K]}$ collected during the learning stage for a constant 8 times, yielding $8K$ replay episodes in total. Only tuples (h,s,a) whose visit counts $N_{\ell,h}(s,a)$, accumulated during that learning stage (Algorithm 2), meet or exceed the threshold specified in line 6 are retained, denoted as $\mathcal{R}_\ell = \left\{(h,s,a) : N_{\ell,h}(s,a) \ge 8\log\frac{1}{\delta''}\right\}$. For each retained tuple (h,s,a), `Sealed-Replay` stores the first $\widetilde{N}_{\ell,h}(s,a)$ replay samples, capped at $N_{\ell,h}(s,a)$. It then constructs a fresh empirical transition model $\widetilde{P}_\ell$ using only these replay samples. Replay samples are not used to update the learning model and are discarded at the end of phase ℓ.

Conservative evaluation. Using the sealed replay samples collected in phase ℓ, we invoke `Conservative-Evaluation` (Algorithm 4 in Appendix B) to compute a conservative value estimate for each counterfactual policy $\pi^{\ell,k,m}$, $k \in [K]$, $m \in [M]$, under the corresponding reward function r^m. For each retained row $(h,s,a) \in \mathcal{R}_\ell$, starting from $\underline{V}_{H+1}^{\ell,k,m} \equiv 0$, we compute the evaluation backup

$$\underline{Q}_h^{\ell,k,m}(s,a) = r_h^m(s,a) + \frac{\widetilde{N}_{\ell,h}(s,a)}{\widetilde{N}_{\ell,h}(s,a)+1}\left\langle \widetilde{P}_{\ell,h,s,a}, \underline{V}_{h+1}^{\ell,k,m}\right\rangle. \tag{1}$$

For all other triples not in $\mathcal{R}_\ell$, we set $\underline{Q}_h^{\ell,k,m}(s,a) = r_h^m(s,a)$. The value recursion follows the corresponding counterfactual policy, with $\underline{V}_h^{\ell,k,m}(s) = \sum_{a\in\mathcal{A}} \pi_h^{\ell,k,m}(a \mid s)\underline{Q}_h^{\ell,k,m}(s,a)$. Averaging the resulting values at the recorded initial states yields the conservative estimate $\hat{v}_{\ell,m}^- = \frac{1}{K}\sum_{k=1}^K \underline{V}_1^{\ell,k,m}(s_1^{\ell,k})$.

Remark 1. *A key feature of this conservative evaluation is that it does not subtract an explicit pessimistic bonus. Instead, it shrinks the continuation value in each backward backup according to the visit count of the corresponding triple (h,s,a). This design offers two advantages: (i) it keeps the mean underestimation bias small; and (ii) it controls upward fluctuations even when $N_{\ell,h}(s,a)$ is small, so that $\hat{v}_{\ell,m}^-$ cannot exceed the true value very often.*

Gap-based multiplicative weights update. Finally, we update the reward-sampling distribution using the gap between the optimistic and conservative estimates:

$$p_{\ell+1}(m) \propto p_\ell(m)\exp\left(\frac{\hat{v}_{\ell,m}^+ - \hat{v}_{\ell,m}^-}{H}\right). \tag{2}$$

Thus, rewards with larger certificate gaps receive larger multiplicative weight increases relative to rewards with smaller gaps.

Output. After L phases, we return, for each $m\in[M]$, the uniform mixture of the counterfactual policies indexed by $[L]\times[K]$ (line 16) $\overline{\pi}^m = \mathrm{Unif}\big((\pi^{\ell,k,m})_{\ell\in[L],\,k\in[K]}\big)$. Each phase uses K learning episodes and $8K$ replay episodes, for a total of $9KL$ episodes of environment interaction.

Algorithm 1 `Online-Multi-Reward-Learning-with-Exponential-Weight` $(\mathcal{S},\mathcal{A},H,K,L,\{r^m\}_{m\in[M]},\epsilon,\delta)$

1: **input:** state space $\mathcal{S}$, action space $\mathcal{A}$, horizon H, learning episodes per phase K, phases L, reward functions $\{r^m\}_{m\in[M]}$, accuracy $\epsilon\in(0,H]$, confidence parameter $\delta\in(0,1)$
2: Set $\delta' = \frac{\delta}{64SAHKLM}$, $\delta'' = \frac{\delta}{64SAHL}$, and $p_1(m)\leftarrow 1/M$ for all $m\in[M]$.
3: For all (h,s,a,s'), set $N_h^{\mathsf{all}}(s,a), N_h(s,a,s'), N_h(s,a)\leftarrow 0$; initialize $\widehat{P}_{h,s,a}\in\Delta(\mathcal{S})$.
4: **for** phase $\ell = 1,2,\dots,L$ **do**
5: $\quad N_\ell, \{\hat{v}_{\ell,m}^+\}_{m\in[M]}, (\pi^{\ell,k})_{k\in[K]}, \{(\pi^{\ell,k,m})_{k\in[K]}\}_{m\in[M]}, (s_1^{\ell,k})_{k\in[K]} \leftarrow$ `Multi-Reward-Switching-MVP`(p_ℓ);
6: $\quad \mathcal{R}_\ell \leftarrow \{(h,s,a) : N_{\ell,h}(s,a)\ge 8\log\frac{1}{\delta''}\}$ ▷ get the retained rows
7: $\quad \widetilde{P}_\ell, \widetilde{N}_\ell \leftarrow$ `Sealed-Replay`$(\mathcal{R}_\ell, N_\ell, (\pi^{\ell,k})_{k\in[K]})$; ▷ fresh replay model
8: $\quad$ **for** $m\in[M]$ **do**
9: $\quad\quad$ **for** $k = 1,2,\dots,K$ **do**
10: $\quad\quad\quad \underline{V}_1^{\ell,k,m} \leftarrow$ `Conservative-Evaluation`$(\widetilde{P}_\ell, \widetilde{N}_\ell, \mathcal{R}_\ell, \pi^{\ell,k,m}, r^m)$;
11: $\quad\quad$ **end for**
12: $\quad\quad \hat{v}_{\ell,m}^- \leftarrow \frac{1}{K}\sum_{k=1}^K \underline{V}_1^{\ell,k,m}(s_1^{\ell,k})$
13: $\quad$ **end for**
14: $\quad$ For all $m\in[M]$, update ▷ multiplicative weight update

$$p_{\ell+1}(m) \leftarrow \frac{p_\ell(m)\exp\big((\hat{v}_{\ell,m}^+ - \hat{v}_{\ell,m}^-)/H\big)}{\sum_{j=1}^M p_\ell(j)\exp\big((\hat{v}_{\ell,j}^+ - \hat{v}_{\ell,j}^-)/H\big)}.$$

15: **end for**
16: Set $\overline{\pi}^m \leftarrow \mathrm{Unif}\big((\pi^{\ell,k,m})_{\ell\in[L],\,k\in[K]}\big)$ for all $m\in[M]$.
17: **output:** $\{\overline{\pi}^m\}_{m\in[M]}$

4 Sample Complexity Analysis

In this section, we present the statement and proof sketch of the main theorem.

Algorithm 2 `Multi-Reward-Switching-MVP`(p)

input: reward-sampling distribution p over $[M]$, $c_1 = \frac{460}{9}$, $c_2 = \frac{544}{9}$.
shared state: N^{all}, $N_h(s,a,s')$, $N_h(s,a)$, and $\widehat{P}$, initialized once in Algorithm 1 and updated in place.
Set all entries of $N_\ell = \{N_{\ell,h}(s,a)\}_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}}$ to zero.
for $k = 1, 2, \cdots, K$ **do**
  Set $\overline{V}^{k,m}_{H+1}(s) \leftarrow 0$ for all $s \in \mathcal{S}, m \in [M]$;
  for $h = H, H-1, \cdots, 1$ **do**
    for $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $m \in [M]$ **do**

$$b^{k,m}_h(s,a) \leftarrow c_1\sqrt{\frac{\mathbb{V}(\widehat{P}_{h,s,a}, \overline{V}^{k,m}_{h+1})\log\frac{1}{\delta'}}{\max\{N_h(s,a),1\}}} + c_2\frac{H\log\frac{1}{\delta'}}{\max\{N_h(s,a),1\}}, \tag{3}$$

$$\overline{Q}^{k,m}_h(s,a) \leftarrow \min\left\{r^m_h(s,a) + \langle\widehat{P}_{h,s,a}, \overline{V}^{k,m}_{h+1}\rangle + b^{k,m}_h(s,a), H\right\}. \tag{4}$$

    end for
    Set $\overline{V}^{k,m}_h(s) \leftarrow \max_{a\in\mathcal{A}} \overline{Q}^{k,m}_h(s,a)$ and $\pi^{k,m}_h(s) \in \arg\max_{a\in\mathcal{A}} \overline{Q}^{k,m}_h(s,a)$ for all $(s,m) \in \mathcal{S}\times[M]$.
  end for
  Draw $m^k \sim p$, set $\pi^k \leftarrow \pi^{k,m^k}$; start a fresh episode with $s^k_1 \sim \mu$. ▷ roll-out policy
  for $h = 1, 2, \cdots, H$ **do**
    Observe s^k_h, execute $a^k_h = \pi^k_h(s^k_h)$, receive r^k_h, observe s^k_{h+1}. $(s,a,s') \leftarrow (s^k_h, a^k_h, s^k_{h+1})$;
    Update $N^{\mathsf{all}}_h(s,a) \leftarrow N^{\mathsf{all}}_h(s,a)+1$, $N_h(s,a,s') \leftarrow N_h(s,a,s')+1$, $N_{\ell,h}(s,a) \leftarrow N_{\ell,h}(s,a)+1$.
    if $N^{\mathsf{all}}_h(s,a) \in \{2^j : j = 0, 1, \ldots, \lfloor\log_2(KL)\rfloor\}$ **then**
      $N_h(s,a) \leftarrow \sum_{\widetilde{s}} N_h(s,a,\widetilde{s})$; ▷ number of visits to (h,s,a) in this epoch.
      Set $\widehat{P}_{h,s,a,\widetilde{s}} \leftarrow \frac{N_h(s,a,\widetilde{s})}{N_h(s,a)}$ and $N_h(s,a,\widetilde{s}) \leftarrow 0$ for all $\widetilde{s} \in \mathcal{S}$.
    end if
  end for
end for
Set $\hat{v}^+_m \leftarrow \frac{1}{K}\sum_{k=1}^K \overline{V}^{k,m}_1(s^k_1)$ for all $m \in [M]$; ▷ optimistic estimate
return: N_ℓ, $\{\hat{v}^+_m\}_{m\in[M]}$, $(\pi^k)_{k\in[K]}$, $\{(\pi^{k,m})_{k\in[K]}\}_{m\in[M]}$, $(s^k_1)_{k\in[K]}$

Theorem 1 (Sample Complexity of Algorithm 1)**.** *There exists a universal constant $C > 0$ such that, for any $M \geq 2$, $SAH \geq 2$, $\delta \in (0,1)$, and $\epsilon \in (0,H]$, setting $\iota = \log^4\frac{SAH\log M}{\min\{\epsilon,1\}\delta}$, $K = \lceil C\frac{SAH^2\iota^{3/4}}{\epsilon}\rceil$ learning episodes per phase, and $L = \lceil C\frac{H\iota^{1/4}}{\epsilon}\log M\rceil$ phases in Algorithm 1 ensures that, with probability at least $1-\delta$, the algorithm returns a policy collection $\{\overline{\pi}^m\}_{m\in[M]}$ such that $\overline{\pi}^m$ is ϵ-optimal for every $m \in [M]$. The total number of episodes, including learning and replay, is at most*

$$9KL = O\left(\frac{SAH^3\iota\log M}{\epsilon^2}\right) = O\left(\frac{SAH^3\log M}{\epsilon^2}\log^4\left(\frac{SAH\log M}{\min\{\epsilon,1\}\delta}\right)\right).$$

Theorem 1 extends the burn-in-free sample-complexity guarantee of MVP (Zhang et al., 2025) to the multi-reward setting, with optimal logarithmic dependence on M up to additional logarithmic factors.[2]

Additional notations. For phase ℓ, we write $\overline{V}^{\ell,k,m}$ and $\underline{V}^{\ell,k,m}$ for the optimistic value function $\overline{V}^{k,m}$ and conservative value function $\underline{V}^{\ell,k,m}$ when evaluating $\pi^{\ell,k,m}$ in phase ℓ under r^m respectively. We use the same convention for other phase-dependent quantities, including $\overline{Q}^{\ell,k,m}$, $\underline{Q}^{\ell,k,m}$, $m^{\ell,k}$, $s^{\ell,k}_1$ and let $v_{\ell,m} := \frac{1}{K}\sum_{k=1}^K V_1^{\pi^{\ell,k,m},m}(s^{\ell,k}_1)$. Recall that $\hat{v}^+_{\ell,m}$ is the optimistic estimate computed in line 5 of Algorithm 1, whereas $\hat{v}^-_{\ell,m}$ is the conservative estimate computed in line 12 of Algorithm 1.

High-level idea. To establish ϵ-optimality for all M rewards, it suffices, up to an initial-state concentration term, to control the worst-reward average optimistic surplus $\max_{m\in[M]}\frac{1}{L}\sum_{\ell=1}^L(\hat{v}^+_{\ell,m} - v_{\ell,m})$.

[2]This dependence is optimal in the regime $M \lesssim \exp(S)$. For $M \gtrsim \exp(S)$, the reward-free algorithm of Ménard et al. (2021a) achieves an optimal sample complexity of $\widetilde{O}(S^2AH^3/\epsilon^2)$ episodes.

For this high-level discussion, we consider the high-probability optimism event, on which $\hat{v}^+_{\ell,m} \geq v_{\ell,m}$ for all ℓ, m. Our strategy is to estimate these surpluses using the gaps $\hat{v}^+_{\ell,m} - \hat{v}^-_{\ell,m}$ and apply multiplicative weights updates to convert gap-weighted progress into simultaneous guarantees for every reward. Two difficulties arise along this approach.

First, accurate phase-wise pessimism is expensive. If $\hat{v}^-_{\ell,m} \leq v_{\ell,m}$ held in every phase, each estimated gap would directly upper-bound the corresponding optimistic surplus. A conventional approach would enforce this inequality through pessimistic Bellman backups with confidence bonuses. However, keeping the resulting confidence widths sufficiently small in every phase can require substantially more replay samples than our intended per-phase budget of $K = \widetilde{O}(SAH^2/\epsilon)$ episodes.

Second, a generic multiplicative-weights regret bound is too loose. For arbitrary signed gaps in $[-H, H]$, the standard worst-case analysis, with an appropriately tuned learning rate, gives

$$\max_{m\in[M]} \frac{1}{L}\sum_{\ell=1}^{L}\left(\hat{v}^+_{\ell,m} - \hat{v}^-_{\ell,m}\right) - \frac{1}{L}\sum_{\ell=1}^{L}\sum_{m=1}^{M} p_\ell(m)\left(\hat{v}^+_{\ell,m} - \hat{v}^-_{\ell,m}\right) \lesssim H\sqrt{\frac{\log M}{L}}.$$

This bound does not yield an $O(\epsilon)$ overhead with $L = \widetilde{O}(H \log M/\epsilon)$ phases. Instead, we seek a constant-factor comparison with an additive $O(H\log(M/\delta)/L)$ term:

$$\max_{m\in[M]} \sum_{\ell=1}^{L}\frac{1}{L}\left(\hat{v}^+_{\ell,m} - \hat{v}^-_{\ell,m}\right) \lesssim \frac{4}{L}\sum_{\ell=1}^{L}\sum_{m=1}^{M} p_\ell(m)\left(\hat{v}^+_{\ell,m} - \hat{v}^-_{\ell,m}\right) + \frac{H}{L}\log\frac{64M}{\delta}. \tag{5}$$

Such a comparison follows directly in the case $(\hat{v}^+_{\ell,m} - \hat{v}^-_{\ell,m}) \geq 0$ for all (ℓ, m), but our evaluation gaps can be negative because phase-wise pessimism is not guaranteed.

We address both difficulties by designing $\hat{v}^-_{\ell,m}$ as a *conservative evaluation* based on fresh replay samples, rather than as a phase-wise lower confidence bound. The evaluator is conservative in conditional mean with a small bias. It also satisfies a conditional exponential-moment guarantee for $\hat{v}^-_{\ell,m} - v_{\ell,m}$, which controls cumulative overestimation. These modifications allow us to establish an approximate version of equation 5.

The key is therefore not to eliminate individual negative gaps, but to control their effect through conditional moments and cumulative evaluation error, adding only an $O(H\log(M/\delta)/L)$ overhead.

Given eq. (5), it remains to control the weighted gap $\frac{1}{L}\sum_{\ell=1}^{L}\sum_{m\in[M]} p_\ell(m)\left(\hat{v}^+_{\ell,m} - \hat{v}^-_{\ell,m}\right)$. We bound this quantity by combining the regret guarantee of reward-switching MVP. Together with the cumulative overestimation guarantee, we obtain that, with high probability, $\max_{m\in[M]} \frac{1}{L}\sum_{\ell=1}^{L}\left(\hat{v}^+_{\ell,m} - v_{\ell,m}\right) = O(\epsilon)$.

The full proof of Theorem 1 is deferred to Appendix C. We sketch the key arguments below.

4.1 Multi-Reward Learning Guarantee.

Since the learning state persists across phases, optimism and the MVP analysis (Zhang et al., 2025) apply over all KL learning episodes.

Lemma 1 (Optimism). *With probability at least* $1 - 4SAHKLM\delta'$, $V_1^{*,m}(s) \leq \overline{V}_1^{\ell,k,m}(s)$ *for all* $(s, \ell, k, m) \in \mathcal{S} \times [L] \times [K] \times [M]$.

The difference between the optimistic estimate and the policy value is then nonnegative and bounds suboptimality. Along the executed policies, the global learning analysis gives

$$\sum_{\ell=1}^{L}\sum_{k=1}^{K}\left(\overline{V}_1^{\ell,k,m^{\ell,k}}(s_1^{\ell,k}) - V_1^{\pi^{\ell,k,m^{\ell,k}},m^{\ell,k}}(s_1^{\ell,k})\right) \lesssim \min\left\{\sqrt{SAH^3KL\log_2^3(2KL)\log\frac{M}{\delta'}}, HKL\right\}$$

with probability at least $1 - 5\delta'$. Applying conditional-mean concentration to the surpluses and then intersecting with optimism transfers the selected surplus to its weighted average.

Lemma 2. *With probability at least* $1-(4SAHKLM+5)\delta'-\delta/64$,

$$\frac{1}{L}\sum_{\ell=1}^{L}\sum_{m=1}^{M} p_\ell(m)\left(\hat{v}^{+}_{\ell,m}-v_{\ell,m}\right) \lesssim \sqrt{\frac{SAH^3}{KL}\log_2^3(2KL)\log\frac{M}{\delta'}}+\frac{H}{KL}\log\frac{64}{\delta}. \tag{6}$$

We postpone the proofs to Appendices C.1.1 and C.1.2.

4.2 Conservative Evaluation with Sealed Replay

In each phase, we replay each of the K recorded roll-out policies eight times. With high probability, this yields enough fresh samples to match the learning-phase visit counts for every retained tuple (h,s,a).

Lemma 3. *With probability at least* $1-2SAHL\delta''$*, the collected replay counts satisfy* $\widetilde{N}_{\ell,h}(s,a)=N_{\ell,h}(s,a)$ *for any* $(h,s,a)\in\mathcal{R}_\ell$ *and any* $\ell\in[L]$.

Let $\mathcal{F}_\ell$ denote the history immediately before the replay stage of phase ℓ. For a fixed policy π, $\underline{V}^{\pi}$ provides a conservative estimate of its value function V^{π}. Since the replay samples are independent of the candidate policies $\{\pi^{k,m}\}_{k\in[K],\,m\in[M]}$ conditional on $\mathcal{F}_\ell$, we obtain the following guarantee:

Lemma 4. *For every* $\ell\in[L]$*, it holds that* $\mathbb{E}\left[\exp\left(\frac{\hat{v}^{-}_{\ell,m}-v_{\ell,m}}{H}\right)|\mathcal{F}_\ell\right]\le 1$*. Moreover, with probability at least* $1-\delta/64$*, it holds that* $\sum_{\ell=1}^{L}(\hat{v}^{-}_{\ell,m}-v_{\ell,m})\le H\log\frac{64M}{\delta}$ *simultaneously for all* $m\in[M]$,

We next bound the mean underestimation error.

Lemma 5 (Conservative weighted bound)**.** *With probability at least* $1-SAHL\delta''$*, simultaneously for all* $\ell\in[L]$, $0\le\sum_{m=1}^{M}p_\ell(m)\left(v_{\ell,m}-\mathbb{E}[\hat{v}^{-}_{\ell,m}\mid\mathcal{F}_\ell]\right)\lesssim\frac{SAH^2}{K}\log\frac{1}{\delta''}$.

We defer the proofs to Appendices C.2.1, C.2.2, and C.3.

4.3 Gap-based Multiplicative Weights Update

Define $W_\ell=\frac{1}{M}\sum_{m=1}^{M}\exp\left(\frac{1}{H}\sum_{\ell'=1}^{\ell-1}\left(\hat{v}^{+}_{\ell',m}-\hat{v}^{-}_{\ell',m}\right)\right)$ for each $1\le\ell\le L+1$. By the multiplicative weights update rule eq. (2), we have that $W_{\ell+1}/W_\ell=\sum_{m=1}^{M}p_\ell(m)\exp\left((\hat{v}^{+}_{\ell,m}-\hat{v}^{-}_{\ell,m})/H\right)$.

Define $(x)_+=\max\{x,0\}$. With Lemma 4, we can show that

$$\mathbb{E}\left[(W_{\ell+1}/W_\ell)|\mathcal{F}_\ell\right]\le 1+\frac{4}{H}\cdot\sum_{m=1}^{M}p_\ell(m)\mathbb{E}\left[(\hat{v}^{+}_{\ell,m}-v_{\ell,m})_+ + v_{\ell,m}-\hat{v}^{-}_{\ell,m}|\mathcal{F}_\ell\right], \tag{7}$$

which further implies

$$\mathbb{E}\left[W_{L+1}\exp\left(-\frac{4}{H}\sum_{\ell=1}^{L}\sum_{m=1}^{M}p_\ell(m)\left[(\hat{v}^{+}_{\ell,m}-v_{\ell,m})_+ + v_{\ell,m}-\mathbb{E}[\hat{v}^{-}_{\ell,m}|\mathcal{F}_\ell]\right]\right)\right]\le 1. \tag{8}$$

As a result, with high probability, we have that

$$\max_{m\in[M]}\frac{1}{L}\sum_{\ell=1}^{L}\left(\hat{v}^{+}_{\ell,m}-\hat{v}^{-}_{\ell,m}\right)\le\frac{4}{L}\sum_{\ell=1}^{L}\sum_{m=1}^{M}p_\ell(m)\left[(\hat{v}^{+}_{\ell,m}-v_{\ell,m})_+ + v_{\ell,m}-\mathbb{E}[\hat{v}^{-}_{\ell,m}|\mathcal{F}_\ell]\right]+O(\epsilon).$$

Finally, on the intersection of the optimism and concentration events, Lemmas 2 and 5 imply, respectively,

$$\frac{1}{L}\sum_{\ell=1}^{L}\sum_{m=1}^{M}p_\ell(m)\left(\hat{v}^{+}_{\ell,m}-v_{\ell,m}\right)_+=O(\epsilon),\quad \frac{1}{L}\sum_{\ell=1}^{L}\sum_{m=1}^{M}p_\ell(m)\left(v_{\ell,m}-\mathbb{E}\left[\hat{v}^{-}_{\ell,m}\mid\mathcal{F}_\ell\right]\right)=O(\epsilon),$$

under the choices of K and L. Combining these bounds with the preceding inequalities finishes the proof.

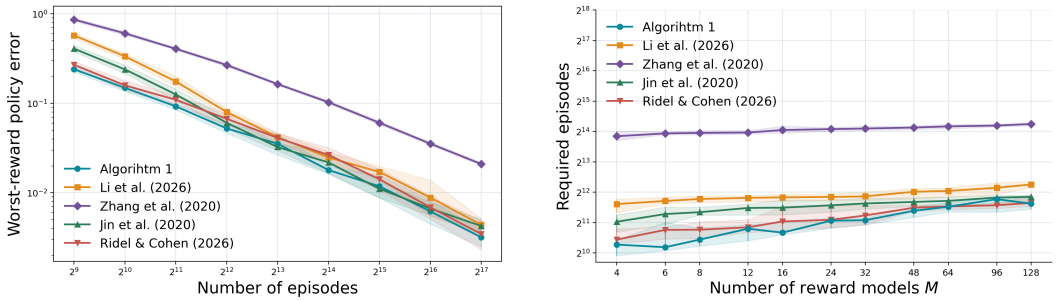
Remark 2. *We retain the term $(\hat{v}^+_{\ell,m} - v_{\ell,m})_+$ rather than $\hat{v}^+_{\ell,m} - v_{\ell,m}$ to ensure that the exponential-moment bound remains valid even when optimism fails. Although $\hat{v}^+_{\ell,m} - v_{\ell,m} \geq 0$ holds with high probability, we must still account for the failure event when taking expectations.*

5 Numerical Experiments

We conduct two experiments to evaluate the episode efficiency of Algorithm 1 and four reward-agnostic baseline algorithms in Jin et al. (2020); Zhang et al. (2020); Ridel & Cohen (2026); Li et al. (2026) on random nonstationary finite-horizon MDPs with $S = 10$, $A = 5$, $H = 6$; the complete transition kernel, reward functions, and experimental settings are provided in Appendix E. For each run t, total episode budget B, and M reward models, we measure the worst-reward policy error and report its average over T independent runs:

$$\mathrm{Error}_t(B, M) := \max_{m \in [M]} \left(V_{0,t}^{\star,m} - V_{0,t}^{\overline{\pi}^m_{B,t},m} \right), \quad \overline{\mathrm{Error}}_T(B, M) := \frac{1}{T} \sum_{t=1}^{T} \mathrm{Error}_t(B, M),$$

where $\overline{\pi}^m_{B,t}$ is the policy obtained in the t-th run with an episode budget of B. The first experiment fixes $M = 16$ and examines the mean worst-reward error as the episode budget increases. The second varies M and estimates the budget required to achieve $\overline{\mathrm{Error}}_8(B, M) \leq 0.1$.



(a) Convergence at $M = 16$

(b) Reward-count scaling at $\epsilon = 0.1$

Figure 1: Empirical episode efficiency in multi-reward reinforcement learning. (a) Mean worst-reward policy error versus total episode budget at $M = 16$, averaged over 12 independent runs. (b) Estimated episode budget required to reach a mean worst-reward error of 0.1 as M varies, using 8 independent runs. Shaded regions indicate pointwise 95% confidence intervals.

Figure 1a shows that our multi-reward algorithm achieves the lowest mean worst-reward error at 7 of the 9 tested budgets and first reaches the target error of 0.1 at 2,048 episodes, earlier than the four baseline implementations on the tested grid. All environment interactions are included in the reported budgets. Jin et al. (2020) attains lower point estimates at 8,192 and 32,768 episodes.

Figure 1b shows that our method requires the lowest estimated budget at 10 of the 11 tested reward counts, with Ridel & Cohen (2026) performing better at $M = 96$. Increasing M from 4 to 128 (32×) raises our estimated budget from approximately 1,240 to 3,153 episodes (2.54×). These results indicate favorable empirical episode efficiency and modest budget growth under the tested configurations.

6 Discussion

In this paper, we study reinforcement learning with M different rewards and propose an algorithm with optimal sample complexity of $\widetilde{O}(SAH^3 \log(M)/\epsilon^2)$. This sample complexity bound is minimax optimal and contains no burn-in term. We leave the following interesting directions for future work: (i) whether the minimax-optimal sample complexity for reward-agnostic reinforcement learning can be achieved without any burn-in terms; (ii) whether the shared structure among reward functions can be exploited to obtain sharper sample-complexity bounds that adapt to the intrinsic complexity of the reward class, rather than its cardinality alone.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the writing and implementing the numerical experiments. The theoretical results and proofs were developed by the authors. We have not used generative AI tools for the technical proof of the theoretical results, and the rest of the required disclosure tasks are not applicable to this work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This paper is mostly theoretical, and there are no significant ethical concerns.

REPRODUCIBILITY STATEMENT

To facilitate verification of our theoretical results, we specify the problem setup and assumptions in Section 2 and provide the algorithm in Algorithm 1. The main result and proof sketch are provided in Section 4 and the complete proofs of the main results and supporting lemmas are provided in the Appendix. In addition, we present experiments in Section 5 and have uploaded the code required for reproduction in the supplementary materials. Also, the detailed proof is provided in the Appendix.

REFERENCES

Shipra Agrawal and Randy Jia. Posterior sampling for reinforcement learning: worst-case regret bounds. *arXiv preprint arXiv:1705.07041*, 2017.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International conference on machine learning*, pp. 263–272. PMLR, 2017.

Shulun Chen, Runlong Zhou, Zihan Zhang, Maryam Fazel, and Simon Du. Sharp gap-dependent variance-aware regret bounds for tabular mdps. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 112169–112205. Curran Associates, Inc., 2025.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pp. 578–598. PMLR, 2021.

Kefan Dong, Yuanhao Wang, Xiaoyu Chen, and Liwei Wang. Q-learning with ucb exploration is sample efficient for infinite-horizon mdp. *arXiv preprint arXiv:1901.09311*, 2019.

Yonathan Efroni, Nadav Merlis, Mohammad Ghavamzadeh, and Shie Mannor. Tight regret bounds for model-based reinforcement learning with greedy policies. *Advances in Neural Information Processing Systems*, 32, 2019.

Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*, 2015.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I. Jordan. Is q-learning provably efficient?, 2018.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International conference on machine learning*, pp. 4870–4879. PMLR, 2020.

Gen Li, Laixi Shi, Yuxin Chen, Yuantao Gu, and Yuejie Chi. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Advances in Neural Information Processing Systems*, 34:17762–17776, 2021.

Gen Li, Yuling Yan, Yuxin Chen, and Jianqing Fan. Minimax-optimal reward-agnostic exploration in reinforcement learning. *Mathematics of Operations Research*, 2026.

Yuanzhi Li, Ruosong Wang, and Lin F Yang. Settling the horizon-dependence of sample complexity in reinforcement learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 965–976. IEEE, 2022.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, pp. 7599–7608. PMLR, 2021a.

Pierre Ménard, Omar Darwiche Domingues, Xuedong Shang, and Michal Valko. Ucb momentum q-learning: Correcting the bias without forgetting. In *International Conference on Machine Learning*, pp. 7609–7618. PMLR, 2021b.

Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1392–1403, 2020.

Oran Ridel and Alon Cohen. Improved bounds for reward-agnostic and reward-free exploration. *arXiv preprint arXiv:2602.16363*, 2026.

Max Simchowitz and Kevin Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32, 2019.

Istvan Szita and Csaba Szepesvari. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 1031–1038, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Ruosong Wang, Simon S Du, Lin F Yang, and Sham M Kakade. Is long horizon reinforcement learning more difficult than short horizon reinforcement learning? *arXiv preprint arXiv:2005.00527*, 2020.

Haike Xu, Tengyu Ma, and Simon Du. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. In *Conference on Learning Theory*, pp. 4438–4472. PMLR, 2021.

Kunhe Yang, Lin Yang, and Simon Du. Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pp. 1576–1584. PMLR, 2021.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pp. 7304–7312. PMLR, 2019.

Xuezhou Zhang, Yuzhe Ma, and Adish Singla. Task-agnostic exploration in reinforcement learning. *Advances in Neural Information Processing Systems*, 33:11734–11743, 2020.

Zihan Zhang, Xiangyang Ji, and Simon S Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *COLT*, pp. 4528–4531, 2021.

Zihan Zhang, Yuhang Jiang, Yuan Zhou, and Xiangyang Ji. Near-optimal regret bounds for multi-batch reinforcement learning. *Advances in Neural Information Processing Systems*, 35:24586–24596, 2022.

Zihan Zhang, Yuxin Chen, Jason Lee, and Simon S Du. Settling the sample complexity of online reinforcement learning. *Journal of the ACM*, 72(3):1–63, 2025.

A Notations and Technical Lemmas

A.1 Notations

Table 2: Parameters of MDP

$\mathcal{S},\ S=\lvert\mathcal{S}\rvert$	State space and its size
$\mathcal{A},\ A=\lvert\mathcal{A}\rvert$	Action space and its size
H	Horizon length
K	Number of learning episodes per phase
L	Number of phases
M	Number of reward functions
$s, s' \in \mathcal{S}$	States
$a, a' \in \mathcal{A}$	Actions
$h \in [H],\ h' \in [H]$	Stage indices
$k \in [K],\ k' \in [K]$	Episode indices
$P_{h,s,a}$	Transition kernel at stage h
$\widehat{P}^k_{h,s,a}$	Empirical transition kernel at k-th episode
$\widetilde{P}_{h,s,a}$	Sealed replay transition kernel
$r^m_h(s,a)$	Reward at triple (h,s,a) for $m \in [M]$
$\mu \in \Delta(\mathcal{S})$	Initial-state distribution
$\mathbb{V}(P,X)$	Variance on random variable X on P
$\langle f, g\rangle$	inner product of f, g
$x \wedge y$	$x \wedge y = \min\{x, y\}$
$x \vee y$	$x \vee y = \max\{x, y\}$
$b^{k,m}_h(s,a)$	Optimistic bonus at k-th episode for reward r^m.
$V^{\pi,m}_h(s),\ Q^{\pi,m}_h(s,a)$	Value and action-value under policy π for reward r^m
$V^{*,m}_h(s),\ Q^{*,m}_h(s,a)$	Optimal value and action-value for reward r^m
$\overline{V}^{\ell,k,m}_h$	$\overline{V}^{\ell,k,m}_h(s) := \max_{a\in\mathcal{A}} \overline{Q}^{\ell,k,m}_h(s,a)$
$\underline{V}^{\ell,k,m}_h$	$\underline{V}^{\ell,k,m}_h(s) := \sum_{a\in\mathcal{A}} \pi^{\ell,k,m}_h(a\lvert s)\cdot \underline{Q}^{\ell,k,m}_h(s,a)$
$\pi^{\ell,k,m}_h$	$\pi^{\ell,k,m}_h(s) \in \arg\max_{a\in\mathcal{A}} \overline{Q}^{\ell,k,m}_h(s,a)$, counterfactual policy at ℓ-th phase, k-th episode
$\pi^{\ell,k}$	$\pi^{\ell,k} = \pi^{\ell,k,m^{\ell,k}}$, roll-out policy at ℓ-th phase, k-th episode
$N^{\ell,k}_h(s,a)$	Number of samples in the most recently completed learning stage before episode k of phase ℓ, used to construct $\widehat{P}^{\ell,k}_{h,s,a}$.
$N_{\ell,h}(s,a)$	Cumulative number of samples collected in learning stage for (h,s,a) within phase ℓ.
$\widetilde{N}_{\ell,h}(s,a)$	Number of samples collected for (h,s,a) during the phase ℓ.
$N^k_h(s,a)$	Number of samples used to construct $\widehat{P}^k_{h,s,a}$ before k-th episode
$\mathcal{H}^\ell_{k-1}$	The σ-field generated by the history prior to sampling $m^{\ell,k}$ and $s^{\ell,k}_1$ in episode k of phase ℓ.
$\mathcal{A}^\ell_k$	$\mathcal{H}^\ell_{k-1} \vee \sigma(s^{\ell,k}_1)$
$\mathcal{B}^\ell_k$	$\mathcal{H}^\ell_{k-1} \vee \sigma(m^{\ell,k})$
$\mathcal{F}_\ell$	The filtration before the sealed replay stage starts at phase ℓ

A.2 TECHNICAL LEMMAS

Lemma 6 (Bennett's inequality)**.** *Let $Z, Z_1, ..., Z_n$ be i.i.d. random variables with values in $[0,1]$ and let $0 < \delta < 1$. Define $\mathbb{V}Z = \mathbb{E}\left[(Z - \mathbb{E}Z)^2\right]$. Then one has*

$$\mathbb{P}\left[\left|\mathbb{E}\left[Z\right] - \frac{1}{n}\sum_{i=1}^{n} Z_i\right| > \sqrt{\frac{2\mathbb{V}Z\log(2/\delta)}{n}} + \frac{\log(2/\delta)}{n}\right] \leq \delta.$$

Lemma 7 (Theorem 4 in Maurer & Pontil (2009))**.** *Consider any $0 < \delta < 1$ and any integer $n \geq 2$. Let $Z, Z_1, ..., Z_n$ be a collection of i.i.d. random variables falling within $[0,1]$. Define the empirical mean $\overline{Z} \coloneqq \frac{1}{n}\sum_{i=1}^{n} Z_i$ and empirical variance $\widehat{V}_n \coloneqq \frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})^2$. Then we have*

$$\mathbb{P}\left[\left|\mathbb{E}\left[Z\right] - \frac{1}{n}\sum_{i=1}^{n} Z_i\right| > \sqrt{\frac{2\widehat{V}_n\log(4/\delta)}{n-1}} + \frac{7\log(4/\delta)}{3(n-1)}\right] \leq \delta.$$

The claim follows by applying the one-sided bound to Z and $1 - Z$, each with failure probability $\delta/2$, and expressing the unbiased sample variance as $n\widehat{V}_n/(n-1)$.

Lemma 8. *Let $X_1, X_2, \ldots$ be a sequence of random variables taking values in $[0, l]$. For any $k \geq 1$, let $\mathcal{F}_k$ be the σ-algebra generated by $(X_1, X_2, \ldots, X_k)$, and define $Y_k \coloneqq \mathbb{E}[X_k \mid \mathcal{F}_{k-1}]$. Then for any $\delta > 0$, we have*

$$\mathbb{P}\left[\exists n, \sum_{k=1}^{n} X_k \geq 2\sum_{k=1}^{n} Y_k + 2l\log\frac{1}{\delta}\right] \leq \delta$$
$$\mathbb{P}\left[\exists n, \sum_{k=1}^{n} Y_k \geq 2\sum_{k=1}^{n} X_k + 2l\log\frac{1}{\delta}\right] \leq \delta.$$

Lemma 9 (Freedman's inequality)**.** *Let $(M_n)_{n\geq 0}$ be a martingale with respect to a filtration $(\mathcal{F}_n)_{n\geq 0}$ such that $M_0 = 0$ and $|M_n - M_{n-1}| \leq c$ $(\forall n \geq 1)$ hold for some quantity $c > 0$. Define $\mathsf{Var}_n \coloneqq \sum_{k=1}^{n}\mathbb{E}\left[(M_k - M_{k-1})^2 \mid \mathcal{F}_{k-1}\right]$ for every $n \geq 0$. Then for any integer $n \geq 1$, $0 < \epsilon \leq nc^2$, and $0 < \delta < 1$, one has*

$$\mathbb{P}\left[|M_n| \geq 2\sqrt{2}\sqrt{\mathsf{Var}_n\log\frac{1}{\delta}} + 2\sqrt{\epsilon\log\frac{1}{\delta}} + 2c\log\frac{1}{\delta}\right] \leq 2\left(\log_2\left(\frac{nc^2}{\epsilon}\right) + 1\right)\delta.$$

Lemma 10 (Lemma 14 in Zhang et al. (2021))**.** *Let $f : \Delta^S \times \mathbb{R}^S \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ with $f(p, v, n, \iota) = pv + \max\left\{\bar{c}_1\sqrt{\frac{\mathbb{V}(p,v)\iota}{n}}, \bar{c}_2\frac{\iota}{n}\right\}$ with $\bar{c}_1 = \frac{20}{3}$ and $\bar{c}_2 = \frac{400}{9}$. Then f satisfies*

1. *$f(p, v, n, \iota)$ is non-decreasing in $v(s)$ for all $p \in \Delta^S$,$v \in [0,1]^S$ and $n, \iota > 0$;*
2. *$f(p, v, n, \iota) \geq pv + 2\sqrt{\frac{\mathbb{V}(p,v)\iota}{n}} + \frac{14\iota}{3n}$ for all p, v and $n, \iota > 0$.*

Lemma 11. *Let $N_h^k(s, a)$ denote the size of the most recently completed batch in Algorithm 2 before episode k, with value zero if no batch has been completed. For $K \geq 1$, it holds that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\max\{N_h^k(s_h^k, a_h^k), 1\}} \leq 2SAH\log_2(2K) \tag{9}$$

Lemma 12 (Ville's inequality)**.** *Let $X_0, X_1, \cdots$ be a non-negative supermartingale. Then for any real number $a > 0$,*

$$\mathbb{P}\left[\sup_{n\geq 0} X_n \geq a\right] \leq \frac{\mathbb{E}[X_0]}{a} \tag{10}$$

B AUXILIARY ALGORITHMS

In this section, we include Algorithm 3: `Sealed-Replay` and 4: `Conservative-Evaluation` for completeness.

Algorithm 3 `Sealed-Replay`$(\mathcal{R}, N, \mathcal{P})$

input: tuple set $\mathcal{R}$, fixed sample size $N = \{N_h(s,a)\}$, indexed policy sequence $\mathcal{P} = (\pi^1, \dots, \pi^K)$
initialization: set $\widetilde{N}_h(s,a), \widetilde{N}_h(s,a,\cdot) \leftarrow 0$, initialize $\widetilde{P}_{h,s,a}$ to the fixed uniform distribution $\mathrm{Unif}(\mathcal{S})$ for all $(h,s,a) \in \mathcal{R}$; and set $L_{\mathrm{rep}} = 8$.
for $l = 1, 2, \cdots, L_{\mathrm{rep}}$ **do**
  for $k = 1, 2, \cdots, K$ **do**
    Start a fresh episode with $\widetilde{s}_1 \sim \mu$.
    for $h = 1, 2, \cdots, H$ **do**
      Observe $\widetilde{s}_h$, execute $\widetilde{a}_h = \pi_h^k(\widetilde{s}_h)$, observe $\widetilde{s}_{h+1}$.
      if $(h, \widetilde{s}_h, \widetilde{a}_h) \in \mathcal{R}$ and $\widetilde{N}_h(\widetilde{s}_h, \widetilde{a}_h) < N_h(\widetilde{s}_h, \widetilde{a}_h)$ **then**
        Update $\widetilde{N}_h(\widetilde{s}_h, \widetilde{a}_h, \widetilde{s}_{h+1}) \leftarrow \widetilde{N}_h(\widetilde{s}_h, \widetilde{a}_h, \widetilde{s}_{h+1}) + 1$, $\widetilde{N}_h(\widetilde{s}_h, \widetilde{a}_h) \leftarrow \widetilde{N}_h(\widetilde{s}_h, \widetilde{a}_h) + 1$
        Update retained empirical kernel with $\widetilde{P}_{h,\widetilde{s}_h,\widetilde{a}_h,s'} = \frac{\widetilde{N}_h(\widetilde{s}_h,\widetilde{a}_h,s')}{\widetilde{N}_h(\widetilde{s}_h,\widetilde{a}_h)}$ for all $s' \in \mathcal{S}$
      end if
    end for
  end for
end for
return: $\widetilde{P}$, $\widetilde{N}$

Algorithm 4 `Conservative-Evaluation`$(P, N, \mathcal{R}, \pi, r)$

input: Evaluation kernel P, count function N, retained tuple set $\mathcal{R}$, policy π, reward function r
initialization: Set $\underline{V}_{H+1}(s) \leftarrow 0$ for all $s \in \mathcal{S}$
for $h = H, H-1, \dots, 1$ **do** ▷ evaluate π under r
  for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**

$$\underline{Q}_h(s,a) \leftarrow \begin{cases} \max\left\{ r_h(s,a) + \dfrac{N_h(s,a)}{N_h(s,a)+1} \langle P_{h,s,a}, \underline{V}_{h+1} \rangle, 0 \right\}, & (h,s,a) \in \mathcal{R}, \\ r_h(s,a), & \text{otherwise.} \end{cases}$$

  end for
  For all $s \in \mathcal{S}$, set $\underline{V}_h(s) \leftarrow \sum_{a \in \mathcal{A}} \pi_h(a|s) \underline{Q}_h(s,a)$.
end for
return: $\underline{V}_1$

C MISSING PROOFS IN SECTION 4

In this section, we provide the full proof of Theorem 1.

Theorem 2 (Restatement of Theorem 1)**.** *There exists a universal constant $C > 0$ such that the following holds. Under the multi-reward MDP setting with $M \geq 2$ and $SAH \geq 2$, fix any $\delta \in (0,1)$ and $\epsilon \in (0, H]$. Set*

$$\iota = \log^4 \frac{SAH \log M}{\min\{\epsilon, 1\}\, \delta}, \qquad K = \left\lceil C \frac{SAH^2 \iota^{3/4}}{\epsilon} \right\rceil, \qquad L = \left\lceil C \frac{H \iota^{1/4}}{\epsilon} \log M \right\rceil.$$

Run Algorithm 1 for L phases with K learning episodes per phase. With probability at least $1 - \delta$, the algorithm returns episode-level mixture policies $\{\overline{\pi}^m\}_{m \in [M]}$ such that

$$V_0^{*,m} - V_0^{\overline{\pi}^m, m} \leq \epsilon \qquad \textit{for every } m \in [M],$$

using a total of $9KL$ episodes, which is at most

$$O\left(\frac{SAH^3}{\epsilon^2} (\log M) \log^4 \left(\frac{SAH \log M}{\min\{\epsilon, 1\}\, \delta} \right) \right),$$

including all learning and replay episodes.

Proof outline: We first bound the weighted average optimistic error over all KL learning episodes. Then, we control the cumulative overestimation and conditional mean bias of conservative evaluation. The reward-weight update combines these bounds to control each reward's cumulative certificate across the phases $\sum_{\ell=1}^{L}(\hat{v}^{+}_{\ell,m} - \hat{v}^{-}_{\ell,m})$.

Finally, we use initial-state concentration to transfer the guarantee to the output mixture $\overline{\pi}^m$.

We retain δ' and δ'' as the local confidence parameters, with $0 < \delta' \le \min\{e^{-1}, (2KL)^{-1}\}$ and $0 < \delta'' \le e^{-1}$. These ranges hold for the choices in Algorithm 1.

At phase ℓ, recall that $\mathcal{F}^{\ell}_{k-1}$ contains all the information before the initial state $s_1^{\ell,k}$. I.e., $\mathcal{F}^{\ell}_{k-1}$ includes all completed learning and replay history, reward-sampling distribution p_ℓ, and the sampled reward index $m^{\ell,k}$, before observing $s_1^{\ell,k}$. Let $\mathcal{F}_\ell$ contain the history immediately before phase ℓ's sealed replay stage, and recall

$$v_{\ell,m} := \frac{1}{K}\sum_{k=1}^{K} V_1^{\pi^{\ell,k,m},m}(s_1^{\ell,k}), \qquad \overline{\pi}^m = \mathrm{Unif}\big((\pi^{\ell,k,m})_{\ell\in[L],\,k\in[K]}\big).$$

Within a fixed phase, we suppress ℓ when there is no confusion. Every candidate $\pi^{\ell,k,m}$ is fixed before the independent draw $s_1^{\ell,k} \sim \mu$. Define

$$\begin{aligned}\Delta^m_{\mathrm{init}} := \frac{1}{KL}\sum_{\ell=1}^{L}\sum_{k=1}^{K}\mathbb{E}_{s_1\sim\mu}\left[V_1^{*,m}(s_1) - V_1^{\pi^{\ell,k,m},m}(s_1)\right] \\ - \frac{1}{KL}\sum_{\ell=1}^{L}\sum_{k=1}^{K}\left(V_1^{*,m}(s_1^{\ell,k}) - V_1^{\pi^{\ell,k,m},m}(s_1^{\ell,k})\right).\end{aligned}$$

For each $m \in [M]$, the output policy error decomposes as

$$V_0^{*,m} - V_0^{\overline{\pi}^m,m} = \underbrace{\Delta^m_{\mathrm{init}}}_{\mathrm{Term}_1} + \underbrace{\frac{1}{KL}\sum_{\ell=1}^{L}\sum_{k=1}^{K}\left(V_1^{*,m}(s_1^{\ell,k}) - V_1^{\pi^{\ell,k,m},m}(s_1^{\ell,k})\right)}_{\mathrm{Term}_2}. \tag{11}$$

Bound of Term_1. Since the counterfactual policy $\pi^{\ell,k,m}$ is fixed before drawing $s_1^{\ell,k}$, the functions $V_1^{*,m}$ and $V_1^{\pi^{\ell,k,m},m}$ are $\mathcal{F}^{\ell}_{k-1}$-measurable and

$$0 \le V_1^{*,m}(s) - V_1^{\pi^{\ell,k,m},m}(s) \le H \implies 0 \le \frac{V_1^{*,m}(s) - V_1^{\pi^{\ell,k,m},m}(s)}{H} \le 1$$

Since for $x \in [0,1]$, the convexity of exponential function gives

$$e^{-x} \le (1-x)e^0 + xe^{-1} = 1 - (1-e^{-1})x.$$

Hence

$$\exp\left(-\frac{V_1^{*,m}(s_1^{\ell,k}) - V_1^{\pi^{\ell,k,m},m}(s_1^{\ell,k})}{H}\right) \le 1 - \frac{1-e^{-1}}{H}\left(V_1^{*,m}(s_1^{\ell,k}) - V_1^{\pi^{\ell,k,m},m}(s_1^{\ell,k})\right).$$

Taking the conditional expectation on both sides with $\mathcal{F}^{\ell}_{k-1}$, recall that $\pi^{\ell,k,m}$ is fixed before the draw of $s_1^{\ell,k}$ and $V_1^{*,m}$ and $V_1^{\pi^{\ell,k,m},m}$ are $\mathcal{F}^{\ell}_{k-1}$-measurable, then

$$\begin{aligned}&\mathbb{E}\left[\exp\left(-\frac{V_1^{*,m}(s_1^{\ell,k}) - V_1^{\pi^{\ell,k,m},m}(s_1^{\ell,k})}{H}\right) \mid \mathcal{F}^{\ell}_{k-1}\right] \\ \le& 1 - \frac{1-e^{-1}}{H}\left(V_0^{*,m} - V_0^{\pi^{\ell,k,m},m}\right) \\ \le& \exp\left(-\frac{1-e^{-1}}{H}\left(V_0^{*,m} - V_0^{\pi^{\ell,k,m},m}\right)\right).\end{aligned} \tag{12}$$

where the last inequality follows $1 - x \le e^{-x}$ for all $x \in \mathbb{R}$. Multiplying both sides of eq. (12) by the positive, $\mathcal{F}_{k-1}^{\ell}$-measurable factor

$$\exp\left(\frac{1-e^{-1}}{H}\left(V_0^{*,m} - V_0^{\pi^{\ell,k,m},m}\right)\right)$$

and bringing this factor inside the conditional expectation yields:

$$\mathbb{E}\left[\exp\left\{\frac{1}{H}\left[(1-e^{-1})\left(V_0^{*,m} - V_0^{\pi^{\ell,k,m},m}\right) - \left(V_1^{*,m}(s_1^{\ell,k}) - V_1^{\pi^{\ell,k,m},m}(s_1^{\ell,k})\right)\right]\right\} \mid \mathcal{F}_{k-1}^{\ell}\right] \le 1 \tag{13}$$

holds for all $\ell, k \in [L] \times [K]$. Write the product formula

$$\begin{aligned}
&\exp\left\{\frac{1}{H}\sum_{\ell=1}^{L}\sum_{k=1}^{K}\left[(1-e^{-1})\left(V_0^{*,m} - V_0^{\pi^{\ell,k,m},m}\right) - \left(V_1^{*,m}(s_1^{\ell,k}) - V_1^{\pi^{\ell,k,m},m}(s_1^{\ell,k})\right)\right]\right\} \\
&= \prod_{\ell=1}^{L}\prod_{k=1}^{K}\exp\left\{\frac{1}{H}\left[(1-e^{-1})\left(V_0^{*,m} - V_0^{\pi^{\ell,k,m},m}\right) - \left(V_1^{*,m}(s_1^{\ell,k}) - V_1^{\pi^{\ell,k,m},m}(s_1^{\ell,k})\right)\right]\right\}.
\end{aligned} \tag{14}$$

Denote

$$\begin{aligned}
z_{\ell,k} &= \exp\left\{\frac{1}{H}\left[(1-e^{-1})\left(V_0^{*,m} - V_0^{\pi^{\ell,k,m},m}\right) - \left(V_1^{*,m}(s_1^{\ell,k}) - V_1^{\pi^{\ell,k,m},m}(s_1^{\ell,k})\right)\right]\right\} \\
Z_{\ell,k} &= \left(\prod_{\ell'=1}^{\ell-1}\prod_{k'=1}^{K} z_{\ell',k'}\right)\left(\prod_{k'=1}^{k} z_{\ell,k'}\right)
\end{aligned}$$

i.e., $Z_{\ell,k}$ is the product of $z_{\ell',k'}$ till ℓ, k. By the tower rule and eq. (13)

$$\begin{aligned}
\mathbb{E}[Z_{L,K}] &= \mathbb{E}[Z_{L,K-1}z_{L,K}] = \mathbb{E}\left[\mathbb{E}\left[Z_{L,K-1}z_{L,K} \mid \mathcal{F}_{K-1}^{L}\right]\right] \\
&= \mathbb{E}\left[Z_{L,K-1}\mathbb{E}\left[z_{L,K} \mid \mathcal{F}_{K-1}^{L}\right]\right] \le \mathbb{E}[Z_{L,K-1}] \le 1
\end{aligned}$$

Then, iteratively apply the above iteration and the definition of the mixture policy $\overline{\pi}^m = \mathrm{Unif}((\pi^{\ell,k,m})_{\ell\in[L],k\in[K]})$, $V_0^{\overline{\pi}^m,m} = \frac{1}{KL}\sum_{\ell,k} V_0^{\pi^{\ell,k,m},m}$

$$\begin{aligned}
&\mathbb{E}[Z_{L,K}] \\
=&\mathbb{E}\left[\exp\left\{\frac{1}{H}\sum_{\ell=1}^{L}\sum_{k=1}^{K}\left[(1-e^{-1})\left(V_0^{*,m} - V_0^{\pi^{\ell,k,m},m}\right) - \left(V_1^{*,m}(s_1^{\ell,k}) - V_1^{\pi^{\ell,k,m},m}(s_1^{\ell,k})\right)\right]\right\}\right] \\
=&\mathbb{E}\left[\exp\left\{\frac{KL}{H}\left[(1-e^{-1})\left(V_0^{*,m} - V_0^{\overline{\pi}^m,m}\right) - \mathrm{Term}_2\right]\right\}\right] \le 1
\end{aligned}$$

Since the exponential term is a nonnegative random variable with expectation at most 1, Markov's inequality implies

$$\mathbb{P}\left(\exp\left\{\frac{KL}{H}\left[(1-e^{-1})\left(V_0^{*,m} - V_0^{\overline{\pi}^m,m}\right) - \mathrm{Term}_2\right]\right\} \ge \frac{1}{\delta'}\right) \le \delta'$$

Then for any fixed m, with probability at least $1 - \delta'$

$$(1-e^{-1})\left(V_0^{*,m} - V_0^{\overline{\pi}^m,m}\right) \le \mathrm{Term}_2 + \frac{H}{KL}\log\frac{1}{\delta'}$$

combine $(1-e^{-1})^{-1} \le 2$, it further implies

$$\left(V_0^{*,m} - V_0^{\overline{\pi}^m,m}\right) \le 2\mathrm{Term}_2 + \frac{2H}{KL}\log\frac{1}{\delta'}$$

simultaneously for all rewards with probability at least $1 - M\delta'$.

Bound of Term_2**.** On the optimism event in Lemma 1,

$$\mathrm{Term}_2 \le \frac{1}{L}\sum_{\ell=1}^{L}(\hat{v}^{+}_{\ell,m} - v_{\ell,m}) = \underbrace{\frac{1}{L}\sum_{\ell=1}^{L}(\hat{v}^{+}_{\ell,m} - \hat{v}^{-}_{\ell,m})}_{\text{Cumulative certificate}} + \underbrace{\frac{1}{L}\sum_{\ell=1}^{L}(\hat{v}^{-}_{\ell,m} - v_{\ell,m})}_{\text{Conservative evaluation error}}. \tag{15}$$

The conservative evaluation error is signed. Iterating the conditional exponential-moment bound in Lemma 4 and applying Markov's inequality gives, with probability at least $1 - M\delta'$, simultaneously for all $m \in [M]$,

$$\sum_{\ell=1}^{L}(\hat{v}^{-}_{\ell,m} - v_{\ell,m}) \le H \log\frac{1}{\delta'}.$$

Intersecting with the initial-state and optimism events therefore yields

$$V_0^{*,m} - V_0^{\overline{\pi}^m,m} \le \frac{2}{L}\sum_{\ell=1}^{L}(\hat{v}^{+}_{\ell,m} - \hat{v}^{-}_{\ell,m}) + \frac{2H}{L}\log\frac{1}{\delta'} + \frac{2H}{KL}\log\frac{1}{\delta'}.$$

It remains to control the cumulative certificates through the reward-weight update. The required weighted optimistic error and conditional mean evaluation bias are analyzed in Appendix C.1 and Appendices C.2–C.3, respectively, and combined in Appendix C.4.

C.1 Optimistic Evaluation Part for Multi-Reward MDPs

We first show the key Lemmas in the multi-reward MDP setting with respect to the optimistic evaluation in the shared-model multi-reward learning.

C.1.1 Proof of Lemma 1

The proof follows the same structure as the optimism analysis in Zhang et al. (2025), with additional concentration events for the M reward functions over all KL learning episodes. Write $\widehat{P}^{\ell,k}_{h,s,a}$ for the empirical transition and let $N_h^{\ell,k}(s,a)$ denote the value of $N_h(s,a)$ in Algorithm 2 before episode k of phase ℓ, i.e., the number of samples used to construct the current empirical transition tuple $\widehat{P}^{\ell,k}_{h,s,a}$. Note that the shared empirical transition kernel $\widehat{P}^{\ell,k}_{h,s,a}$ is not reinitialized at phase boundaries, as well as $N_h^{\ell,k}(s,a)$, distinct from the sealed replay model $\widetilde{P}_\ell = \{\widetilde{P}_{\ell,h,s,a}\}_{h,s,a}$, and count $\widetilde{N}_\ell = \{\widetilde{N}_{\ell,h}(s,a)\}_{h,s,a}$ in line 7.

Lemma 13 (Formal statement of Lemma 1)**.** *With probability at least* $1 - 4SAHKLM\delta'$*, one has*

$$Q_h^{*,m}(s,a) \le \overline{Q}_h^{\ell,k,m}(s,a) \quad \textit{and} \quad V_h^{*,m}(s) \le \overline{V}_h^{\ell,k,m}(s).$$

for all $(h,s,a,\ell,k,m) \in [H]\times\mathcal{S}\times\mathcal{A}\times[L]\times[K]\times[M]$.

Proof. For each (h,s,a), represent its successive learning transitions over all phases by an independent i.i.d. sequence with law $P_{h,s,a}$, independent of the replay streams. Since the learning state persists across phases, the potential batches are fixed blocks of this sequence, of sizes $1,1,2,4,\ldots$, whether or not they are actually completed. Apply Lemma 7, with confidence parameter $4\delta'$, to $V_{h+1}^{*,m}/H$ on each block of size at least two and each $m \in [M]$. There are at most $1 + \lfloor \log_2(KL) \rfloor \le KL$ potential batches per row. Thus, with probability at least $1 - 4SAHM(1 + \lfloor \log_2(KL) \rfloor)\delta'$, every empirical row used by the algorithm with $N_h^{\ell,k}(s,a) \ge 2$ satisfies

$$\begin{aligned}\left|\left\langle \widehat{P}^{\ell,k}_{h,s,a} - P_{h,s,a}, V_{h+1}^{*,m}\right\rangle\right| &\le \sqrt{\frac{2\mathbb{V}(\widehat{P}^{\ell,k}_{h,s,a}, V_{h+1}^{*,m})\log(1/\delta')}{N_h^{\ell,k}(s,a)-1}} + \frac{7H\log(1/\delta')}{3(N_h^{\ell,k}(s,a)-1)} \\ &\le 2\sqrt{\frac{\mathbb{V}(\widehat{P}^{\ell,k}_{h,s,a}, V_{h+1}^{*,m})\log(1/\delta')}{N_h^{\ell,k}(s,a)}} + \frac{14H\log(1/\delta')}{3N_h^{\ell,k}(s,a)}.\end{aligned} \tag{16}$$

This event is uniform over all KL learning episodes, including adaptively selected completed batches.

We prove optimism by backward induction on h. The terminal values are zero. If $N_h^{\ell,k}(s,a) < 2$, the bonus and $\log(1/\delta') \geq 1$ imply $\overline{Q}_h^{\ell,k,m}(s,a) = H \geq Q_h^{*,m}(s,a)$. Otherwise, apply Lemma 10 to $\overline{V}_{h+1}^{\ell,k,m}/H$ and multiply by H. The algorithm's constants satisfy $c_1 \geq 20/3$ and $c_2 \geq 400/9$. Whenever the Q-update is not clipped at H, the monotonicity of f and the induction hypothesis give

$$\begin{aligned}\overline{Q}_h^{\ell,k,m}(s,a) &\geq r_h^m(s,a) + Hf\left(\widehat{P}_{h,s,a}^{\ell,k}, \frac{\overline{V}_{h+1}^{\ell,k,m}}{H}, N_h^{\ell,k}(s,a), \log\frac{1}{\delta'}\right) \\ &\geq r_h^m(s,a) + Hf\left(\widehat{P}_{h,s,a}^{\ell,k}, \frac{V_{h+1}^{*,m}}{H}, N_h^{\ell,k}(s,a), \log\frac{1}{\delta'}\right) \\ &\geq r_h^m(s,a) + \langle P_{h,s,a}, V_{h+1}^{*,m}\rangle = Q_h^{*,m}(s,a), \end{aligned} \tag{17}$$

where the last inequality uses the second conclusion of Lemma 10 and equation 16. Clipping also preserves this bound since $Q_h^{*,m}(s,a) \leq H$. Taking the maximum over a proves $\overline{V}_h^{\ell,k,m}(s) \geq V_h^{*,m}(s)$ on the same event. □

C.1.2 Proof of Lemma 2

The following cumulative regret bound guarantee extends the analysis in Zhang et al. (2025), which controls the cumulative optimism surplus along the roll-out policies.

Lemma 14. *Over all KL learning episodes in Algorithm 2, with probability at least $1 - 5\delta'$,*

$$\sum_{\ell=1}^{L}\sum_{k=1}^{K}\left(\overline{V}_1^{\ell,k,m^{\ell,k}}(s_1^{\ell,k}) - V_1^{\pi^{\ell,k,m^{\ell,k}},m^{\ell,k}}(s_1^{\ell,k})\right) \lesssim \min\left\{\sqrt{SAH^3KL\log_2^3(2KL)\log\frac{M}{\delta'}}, HKL\right\}.$$

The proof of Lemma 14 is given in Appendix D.1 and its supporting profile and bonus arguments, which apply once from initialization to all KL episodes of the persistent learning process. For each fixed global profile, each optimistic continuation vector at stage $h+1$ depends only on the known rewards and learning streams at stages strictly greater than h. The profile concentration events are uniform over all profiles and candidate choices, so they also cover the reward sampling distribution p_ℓ determined by previous replay outcomes. The forward martingale arguments use histories containing these outcomes; the next learning transition still has conditional law $P_{h,s,a}$. Thus the same $5\delta'$ failure bound applies to the entire run.

In proving Lemma 2, denote the optimism gap and weighted gap

$$\Delta^{\ell,k,m} := \overline{V}_1^{\ell,k,m}(s_1^{\ell,k}) - V_1^{\pi^{\ell,k,m},m}(s_1^{\ell,k}), \qquad \Delta^{\ell,k} := \sum_{m=1}^{M} p_\ell(m)\Delta^{\ell,k,m},$$

and the nonnegative counterfactual differences and their weighted average by

$$\Delta_+^{\ell,k,m} := \left[\Delta^{\ell,k,m}\right]_+, \qquad \Delta_+^{\ell,k} := \sum_{m=1}^{M} p_\ell(m)\Delta_+^{\ell,k,m}.$$

Let $\mathcal{H}_{k-1}^{\ell}$ contain all completed learning and replay history, including prior auxiliary evaluations and weight updates, before drawing $m^{\ell,k}$ and $s_1^{\ell,k}$. The distribution p_ℓ, current counterfactual policies $\{\pi^{\ell,k,m}\}_{m\in[M]}$, and optimistic values are measurable with respect to this history. Conditional on $\mathcal{H}_{k-1}^{\ell}$, the fresh draws $s_1^{\ell,k} \sim \mu$ and $m^{\ell,k} \sim p_\ell$ are independent. Use the conditioning fields

$$\mathcal{A}_k^{\ell} := \mathcal{H}_{k-1}^{\ell} \vee \sigma(s_1^{\ell,k}), \qquad \mathcal{B}_k^{\ell} := \mathcal{H}_{k-1}^{\ell} \vee \sigma(m^{\ell,k}). \tag{18}$$

Thus $\mathcal{A}_k^{\ell}$ reveals the initial state but not the reward index, whereas $\mathcal{B}_k^{\ell}$ reveals the reward index but not the initial state. In particular,

$$\mathbb{E}\left[\Delta_+^{\ell,k,m^{\ell,k}} \mid \mathcal{A}_k^{\ell}\right] = \sum_{m=1}^{M} p_\ell(m)\Delta_+^{\ell,k,m} = \Delta_+^{\ell,k}.$$

The selected counterfactual difference $\Delta_+^{\ell,k,m^{\ell,k}}$ is an unbiased random probe of the weighted average $\Delta_+^{\ell,k}$. Consider the random process with $M_0^1 = 1$, $M_0^\ell = M_K^{\ell-1}$ for $\ell \geq 2$, and

$$M_k^\ell := M_{k-1}^\ell \exp\left((1-e^{-1})\frac{\Delta_+^{\ell,k}}{H} - \frac{\Delta_+^{\ell,k,m^{\ell,k}}}{H}\right), \qquad k \in [K].$$

Since $\Delta_+^{\ell,k}$ is $\mathcal{A}_k^\ell$-measurable,

$$\mathbb{E}\left[M_k^\ell \mid \mathcal{A}_k^\ell\right] = M_{k-1}^\ell \exp\left((1-e^{-1})\frac{\Delta_+^{\ell,k}}{H}\right)\mathbb{E}\left[\exp\left(-\frac{\Delta_+^{\ell,k,m^{\ell,k}}}{H}\right) \mid \mathcal{A}_k^\ell\right].$$

For all $y \in [0,1]$, $e^{-y} \leq 1-(1-e^{-1})y$. Since $0 \leq \Delta_+^{\ell,k,m} \leq H$, we have

$$\mathbb{E}\left[\exp\left(-\frac{\Delta_+^{\ell,k,m^{\ell,k}}}{H}\right) \mid \mathcal{A}_k^\ell\right] \leq 1-(1-e^{-1})\frac{\Delta_+^{\ell,k}}{H} \leq \exp\left(-(1-e^{-1})\frac{\Delta_+^{\ell,k}}{H}\right).$$

Therefore, $\mathbb{E}[M_k^\ell \mid \mathcal{A}_k^\ell] \leq M_{k-1}^\ell$. Taking conditional expectations over the initial distribution again gives $\mathbb{E}[M_k^\ell \mid \mathcal{H}_{k-1}^\ell] \leq M_{k-1}^\ell$. By Ville's inequality in Lemma 12, for any concentration failure level $\delta'' \in (0,1)$,

$$\mathbb{P}\left(\max_{\ell\in[L],\, k\in[K]} M_k^\ell \geq \frac{1}{\delta''}\right) \leq \delta''.$$

Consequently, with probability at least $1-\delta''$,

$$\sum_{\ell=1}^{L}\sum_{k=1}^{K}\Delta_+^{\ell,k} \leq \frac{1}{1-e^{-1}}\left(\sum_{\ell=1}^{L}\sum_{k=1}^{K}\Delta_+^{\ell,k,m^{\ell,k}} + H\log\frac{1}{\delta''}\right) \leq 2\left(\sum_{\ell=1}^{L}\sum_{k=1}^{K}\Delta_+^{\ell,k,m^{\ell,k}} + H\log\frac{1}{\delta''}\right).$$

The preceding concentration event is unconditional. At the same time, on its intersection with the optimism event in Lemma 13, the optimism implies

$$\Delta_+^{\ell,k,m} = \Delta^{\ell,k,m}, \qquad \Delta_+^{\ell,k} = \Delta^{\ell,k}$$

Intersecting further with Lemma 14 yields, with probability at least $1-(4SAHKLM+5)\delta' - \delta''$,

$$\begin{aligned}
\frac{1}{L}\sum_{\ell=1}^{L}\sum_{m=1}^{M} p_\ell(m)(\hat{v}_{\ell,m}^+ - v_{\ell,m}) &= \frac{1}{KL}\sum_{\ell=1}^{L}\sum_{k=1}^{K}\Delta^{\ell,k} \\
&\leq \frac{2}{KL}\left(\sum_{\ell=1}^{L}\sum_{k=1}^{K}\Delta^{\ell,k,m^{\ell,k}} + H\log\frac{1}{\delta''}\right) \\
&\lesssim \sqrt{\frac{SAH^3}{KL}\log_2^3(2KL)\log\frac{M}{\delta'}} + \frac{H}{KL}\log\frac{1}{\delta''}. \qquad (19)
\end{aligned}$$

The concentration failure level in this argument is free. Applying Ville's inequality at level $\delta/64$ gives the probability and bound stated in Lemma 2, with the algorithm's replay confidence parameter unchanged. The quantity $v_{\ell,m}$ already uses the recorded initial states $(s_1^{\ell,k})_{k\in[K]}$, so this step requires no additional initial-state correction. This completes the proof.

C.2 Conservative Evaluation and Multiplicative Weight Update Parts for Multi-Reward MDPs

In this section, we prove the replay and exponential evaluation guarantees. Retain the auxiliary run and histories defined above, and write

$$d_{\ell,h}^{k,m}(s,a) := \mathbb{P}_{\pi^{\ell,k,m}}\left((s_h,a_h) = (s,a) \mid s_1 = s_1^{\ell,k}\right), \qquad (20)$$

$$D_{\ell,h,s,a}^{\mathrm{avg}} := \sum_{m=1}^{M} p_\ell(m)\sum_{k=1}^{K} d_{\ell,h}^{k,m}(s,a). \qquad (21)$$

For the recorded roll-out policies $\pi^{\ell,k} := \pi^{\ell,k,m^{\ell,k}}$, also define

$$d_h^{\pi,\mu}(s,a) := \mathbb{P}_{\pi,s_1\sim\mu}\left((s_h,a_h)=(s,a)\right) \tag{22}$$

$$D^{\mu}_{\ell,h,s,a} := \sum_{k=1}^{K} d_h^{\pi^{\ell,k},\mu}(s,a). \tag{23}$$

Here $N_{\ell,h}(s,a)$ counts visits in phase ℓ's K learning episodes.

C.2.1 Retained row receives sufficiently many fresh replay samples.

We first bound the learning counts and weighted occupancies, using the histories $\mathcal{H}^{\ell}_{k-1}$, $\mathcal{A}^{\ell}_{k}$, and $\mathcal{B}^{\ell}_{k}$ from Appendix C.1.2.

Lemma 15. *For each fixed phase ℓ and tuple (h,s,a), conditional on $\mathcal{H}^{\ell}_{0}$,*

$$D^{\mathrm{avg}}_{\ell,h,s,a} \le 2N_{\ell,h}(s,a) + 2\log\frac{1}{\delta''}, \tag{24}$$

$$N_{\ell,h}(s,a) \le 2D^{\mu}_{\ell,h,s,a} + 2\log\frac{1}{\delta''}, \tag{25}$$

with failure probabilities at most $(\delta'')^{3/2}$ and $(\delta'')^{2}$, respectively.

Proof. Let

$$X^{k}_{\ell,h,s,a} := \mathbb{1}_{\{(s_h^{\ell,k},a_h^{\ell,k})=(s,a)\}}$$

so the sample count can be expressed as $N_{\ell,h}(s,a)=\sum_{k=1}^{K} X^{k}_{\ell,h,s,a}$. Recall the completed-history filtration $(\mathcal{H}^{\ell}_{k})$ and the two intermediate sigma-fields $\mathcal{A}^{\ell}_{k} = \mathcal{H}^{\ell}_{k-1}\vee\sigma(s_1^{\ell,k})$ and $\mathcal{B}^{\ell}_{k} = \mathcal{H}^{\ell}_{k-1}\vee\sigma(m^{\ell,k})$ defined in eq. (18). Conditional on $\mathcal{H}^{\ell}_{k-1}$, all counterfactual policies are fixed and $m^{\ell,k}$ and $s_1^{\ell,k}$ are independent. The conditional independence of the reward and initial-state draws gives

$$\mathbb{E}[X^{k}_{\ell,h,s,a}\mid\mathcal{A}^{\ell}_{k}] = \sum_{m=1}^{M} p_\ell(m)d^{k,m}_{\ell,h}(s,a), \qquad \mathbb{E}[X^{k}_{\ell,h,s,a}\mid\mathcal{B}^{\ell}_{k}] = d_h^{\pi^{\ell,k},\mu}(s,a).$$

Proof of equation 24. Denote the weighted occupancy as

$$\overline{d}^{k}_{\ell,h}(s,a) := \sum_{m=1}^{M} p_\ell(m)d^{k,m}_{\ell,h}(s,a)$$

Random variable $X^{k}_{\ell,h,s,a}$ satisfies

$$\mathbb{E}\left[X^{k}_{\ell,h,s,a}|\mathcal{A}^{\ell}_{k}\right] = \sum_{m=1}^{M} p_\ell(m)d^{k,m}_{\ell,h}(s,a) = \overline{d}^{k}_{\ell,h}(s,a)$$

For a Bernoulli variable X with a conditional mean $\mathbb{E}[X|\mathcal{F}]=u$ for some sigma-field $\mathcal{F}$

$$\mathbb{E}\left[e^{3u/4-(\log 4)X}|\mathcal{F}\right] = e^{3u/4}(1-u) + \frac{ue^{3u/4}}{4} = e^{3u/4}\left(1-\frac{3u}{4}\right) \le 1 \tag{26}$$

Consider the random process

$$Y_k := \exp\left(\frac{3}{4}\sum_{k'=1}^{k}\overline{d}^{k'}_{\ell,h}(s,a) - (\log 4)\sum_{k'=1}^{k}X^{k'}_{\ell,h,s,a}\right)$$

Then by equation 26

$$\mathbb{E}\left[Y_k\mid\mathcal{A}^{\ell}_{k}\right] = Y_{k-1}\mathbb{E}\left[e^{3\overline{d}^{k}_{\ell,h}(s,a)/4-(\log 4)X^{k}_{\ell,h,s,a}}\mid\mathcal{A}^{\ell}_{k}\right] \le Y_{k-1}.$$

Since $\mathcal{H}_{k-1}^{\ell} \subseteq \mathcal{A}_k^{\ell} \subseteq \mathcal{H}_k^{\ell}$, the tower property shows that (Y_k) is a nonnegative supermartingale with respect to $(\mathcal{H}_k^{\ell})$, with $Y_0 = 1$. By Ville's inequality, conditional on $\mathcal{H}_0^{\ell}$, with probability at least $1 - (\delta'')^{3/2}$,

$$Y_K \le (\delta'')^{-3/2} \implies \frac{3}{4} D_{\ell,h,s,a}^{\mathrm{avg}} - (\log 4) N_{\ell,h}(s,a) \le \frac{3}{2} \log \frac{1}{\delta''}.$$

Since $\frac{4}{3} \log 4 \le 2$, this implies equation 24 with failure probability at most $(\delta'')^{3/2}$, i.e.

$$D_{\ell,h,s,a}^{\mathrm{avg}} \le 2 N_{\ell,h}(s,a) + 2 \log \frac{1}{\delta''}$$

with probability at least $1 - (\delta'')^{3/2}$.

Proof of equation 25 Conditional on $\mathcal{B}_k^{\ell}$, the roll-out policy $\pi^{\ell,k} = \pi^{\ell,k,m^{\ell,k}}$ is fixed, while the initial state $s_1^{\ell,k} \sim \mu$ has not yet been observed. Hence,

$$\mathbb{E}\big[X_{\ell,h,s,a}^k \mid \mathcal{B}_k^{\ell}\big] = d_h^{\pi^{\ell,k},\mu}(s,a).$$

For a Bernoulli variable X with conditional mean $\mathbb{E}[X|\mathcal{F}] = u$ for some sigma-field $\mathcal{F}$

$$\mathbb{E}\left[e^{X-(e-1)u} | \mathcal{F}\right] = e^{-(e-1)u}\,(eu + 1 - u) = e^{-(e-1)u}\,(1 + (e-1)u) \le 1 \tag{27}$$

Define the random process

$$Z_k := \exp\left(\sum_{k'=1}^{k} X_{\ell,h,s,a}^{k'} - (e-1) \sum_{k'=1}^{k} d_h^{\pi^{\ell,k'},\mu}(s,a) \right), \qquad Z_0 = 1.$$

Since Z_{k-1} is $\mathcal{H}_{k-1}^{\ell}$-measurable and hence $\mathcal{B}_k^{\ell}$-measurable,

$$\mathbb{E}\big[Z_k \mid \mathcal{B}_k^{\ell}\big] = Z_{k-1} \mathbb{E}\Big[\exp\Big(X_{\ell,h,s,a}^k - (e-1) d_h^{\pi^{\ell,k},\mu}(s,a)\Big) \Big| \mathcal{B}_k^{\ell}\Big] \le Z_{k-1}.$$

Taking conditional expectations again given $\mathcal{H}_{k-1}^{\ell}$ shows that (Z_k) is a nonnegative supermartingale with respect to $(\mathcal{H}_k^{\ell})$.

By Ville's inequality, conditional on $\mathcal{H}_0^{\ell}$, with probability at least $1 - (\delta'')^2$,

$$\sum_{k=1}^{K} X_{\ell,h,s,a}^k - (e-1) \sum_{k=1}^{K} d_h^{\pi^{\ell,k},\mu}(s,a) \le 2 \log \frac{1}{\delta''}.$$

Using $N_{\ell,h}(s,a) = \sum_{k=1}^{K} X_{\ell,h,s,a}^k$, the definition of $D_{\ell,h,s,a}^{\mu}$, and $e - 1 < 2$, we obtain

$$N_{\ell,h}(s,a) \le 2 D_{\ell,h,s,a}^{\mu} + 2 \log \frac{1}{\delta''}.$$

This proves equation 25 with conditional failure probability at most $(\delta'')^2$. Integrating over the phase-start history and taking a union bound over all phases and rows gives failure probabilities at most $SAHL(\delta'')^{3/2}$ and $SAHL(\delta'')^2$ for equation 24 and equation 25, respectively. In particular, the first event alone has failure probability at most $SAHL\delta''$, as used in Appendix C.3. Their joint failure probability is at most $SAHL\left((\delta'')^{3/2} + (\delta'')^2\right) \le 2SAHL\delta''$. □

Lemma 16 (Formal statement of Lemma 3)**.** *With probability at least* $1 - 2SAHL\delta''$*, the number of the replay samples collected in the sealed replay stage satisfies*

$$\widetilde{N}_{\ell,h}(s,a) = N_{\ell,h}(s,a), \qquad \ell \in [L], \quad (h,s,a) \in \mathcal{R}_{\ell}.$$

Proof. On the learning-count event equation 25, every retained row satisfies

$$D_{\ell,h,s,a}^{\mu} \ge \frac{N_{\ell,h}(s,a) - 2\log(1/\delta'')}{2} \ge \frac{3}{8} N_{\ell,h}(s,a), \tag{28}$$

since $N_{\ell,h}(s,a) \geq 8\log(1/\delta'')$. Conditional on the history $\mathcal{F}_\ell$ immediately before replay, all rollout policies, retained rows, and quotas are fixed. Let $X^{k,l}_{\ell,h,s,a}$ indicate a visit to (h,s,a) in the l-th fresh replay of the indexed policy $\pi^{\ell,k}$, and let

$$T_{\ell,h,s,a} := \sum_{k=1}^{K}\sum_{l=1}^{L_{\mathrm{rep}}} X^{k,l}_{\ell,h,s,a}.$$

The replay episodes are conditionally independent, including episodes that replay identical policies. For $L_{\mathrm{rep}} = 8$, on the learning-count event,

$$\mathbb{E}[T_{\ell,h,s,a} \mid \mathcal{F}_\ell] = 8D^{\mu}_{\ell,h,s,a} \geq 3N_{\ell,h}(s,a).$$

Since the sealed replay phase only records the first $N_{\ell,h}(s,a)$ samples,

$$\widetilde{N}_{\ell,h}(s,a) = \min\{T_{\ell,h,s,a}, N_{\ell,h}(s,a)\}.$$

Then, the Chernoff's inequality gives,

$$\begin{aligned}\mathbb{P}\left(\widetilde{N}_{\ell,h}(s,a) < N_{\ell,h}(s,a) \mid \mathcal{F}_\ell\right) \leq& \mathbb{P}\left(T_{\ell,h,s,a} \leq \tfrac{1}{3}\mathbb{E}[T_{\ell,h,s,a} \mid \mathcal{F}_\ell] \mid \mathcal{F}_\ell\right)\\ \leq& \exp\left(-\tfrac{2}{3}N_{\ell,h}(s,a)\right)\\ \leq& (\delta'')^{16/3}.\end{aligned}$$

Using the failure bound $(\delta'')^2$ for equation 25, a union bound over all phases and tuples bounds the probability of any unfilled $\widetilde{N}_{\ell,h}(s,a) < N_{\ell,h}(s,a)$ by

$$SAHL\left((\delta'')^2 + (\delta'')^{16/3}\right) \leq 2SAHL\delta''.$$

This demonstrates that, with probability at least $1 - 2SAHL\delta''$, across all L phases, samples drawn via sealed replay always include as many new samples as were collected during the learning phase. Then we complete the proof. □

C.2.2 Proof of Lemma 4

Lemma 17 (Formal statement of Lemma 4)**.** *For every $\ell \in [L]$, it holds that $\mathbb{E}\left[\exp\left(\frac{\hat{v}^-_{\ell,m} - v_{\ell,m}}{H}\right)|\mathcal{F}_\ell\right] \leq 1$. Moreover, with probability at least $1 - \delta/64$, it holds that $\sum_{\ell=1}^{L}(\hat{v}^-_{\ell,m} - v_{\ell,m}) \leq H\log\frac{64M}{\delta}$ simultaneously for all $m \in [M]$,*

Proof. We will prove that $\mathbb{E}\left[\exp\left(\frac{\hat{v}^-_{\ell,m} - v_{\ell,m}}{H}\right)|\mathcal{F}_\ell\right] \leq 1$. The final tail probability bound follows by Markov's inequality and the bound $\mathbb{E}\left[\exp\left(\sum_{\ell=1}^{L}\frac{\hat{v}^-_{\ell,m} - v_{\ell,m}}{H}\right)\right] \leq 1$.

Conditional on $\mathcal{F}_\ell$, the counterfactual policies $\{(\pi^{\ell,k,m})_{k\in[K]}\}$, initial states $(s_1^{\ell,k})_{k\in[K]}$, retained set $\mathcal{R}_\ell$, and phase learning tuple count $N_{\ell,h}(s,a)$ are fixed. Represent the fresh replay transitions by arrays $Z^{(i)}_{\ell,h,s,a} \overset{\text{i.i.d.}}{\sim} P_{h,s,a}$ for $i \geq 1$. The auxiliary run uses the fixed-prefix model

$$\widetilde{P}_{\ell,h,s,a} = \frac{1}{N_{\ell,h}(s,a)}\sum_{i=1}^{N_{\ell,h}(s,a)} \delta_{Z^{(i)}_{\ell,h,s,a}}, \qquad (h,s,a) \in \mathcal{R}_\ell,$$

Under the success event of Lemma 3, $\widetilde{P}_{\ell,h,s,a}$ is constructed via $N_{\ell,h}(s,a)$ fresh replayed samples. The recursion gives $0 \leq \underline{V}_h \leq H - h + 1$, and its clipping at zero is inactive because rewards and continuation values are nonnegative. For a retained row, put $n = N_{\ell,h}(s,a)$, so

$$\underline{Q}^{\ell,k,m}_h(s,a) = r^m_h(s,a) + \frac{1}{n+1}\sum_{i=1}^{n}\underline{V}^{\ell,k,m}_{h+1}(Z^{(i)}_{\ell,h,s,a}). \tag{29}$$

We first record the scalar inequality for i.i.d. variables $Y_1, \dots, Y_n \in [0, H]$:

$$\mathbb{E}\left[\exp\left(\frac{\sum_{i=1}^n Y_i}{H(n+1)}\right)\right] \le \left[1 + \frac{\mathbb{E}[Y_1]}{H}\left(e^{1/(n+1)} - 1\right)\right]^n \le \exp\left(\frac{\mathbb{E}[Y_1]}{H}\right). \tag{30}$$

The first inequality follows from convexity of the exponential on $[0, 1]$. The second uses $1 + x \le e^x$ and $n(e^{1/(n+1)} - 1) \le 1$ holds for all $n \ge 0$; the latter follows from $e^x \le (1-x)^{-1}$ for $0 \le x < 1$.

We prove by backward induction that, for every h and s,

$$\mathbb{E}\left[\exp\left(\frac{\underline{V}_h^{\ell,k,m}(s) - V_h^{\pi,m}(s)}{H}\right) \mid \mathcal{F}_\ell\right] \le 1. \tag{31}$$

Base case. The terminal values are zero since $\underline{V}_{H+1}^{\ell,k,m}(s) = V_{H+1}^{\pi^{\ell,k,m},m} = 0$, the base case holds.

Inductive Hypothesis. Suppose

$$\mathbb{E}\left[\exp\left(\frac{\underline{V}_{h+1}^{\ell,k,m}(s) - V_{h+1}^{\pi,m}(s)}{H}\right) \mid \mathcal{F}_\ell\right] \le 1. \tag{32}$$

Inductive Step. Conditional on $\mathcal{F}_\ell$ and $\underline{V}_{h+1}^{\ell,k,m}$, the current tuple samples remain i.i.d., since $\underline{V}_{h+1}^{\ell,k,m}$ depends only on replay samples at stages strictly greater than h.

On the tuple $(h, s, a) \in \mathcal{R}_\ell$, recall the definition of $\underline{Q}_h^{\ell,k,m}(s, a)$ in equation 29.

$$\begin{aligned}
\underline{Q}_h^{\ell,k,m}(s,a) =& r_h^m(s,a) + \frac{1}{n+1}\sum_{i=1}^n \underline{V}_{h+1}^{\ell,k,m}(Z_{\ell,h,s,a}^{(i)}) \\
Q_h^{\pi^{\ell,k,m},m}(s,a) =& r_h^m(s,a) + \langle P_{h,s,a}, V_{h+1}^{\pi^{\ell,k,m},m}\rangle
\end{aligned}$$

Subtract the equations we have

$$\underline{Q}_h^{\ell,k,m}(s,a) - Q_h^{\pi^{\ell,k,m},m}(s,a) = \frac{1}{n+1}\sum_{i=1}^n \underline{V}_{h+1}^{\ell,k,m}(Z_{\ell,h,s,a}^{(i)}) - \langle P_{h,s,a}, V_{h+1}^{\pi^{\ell,k,m},m}\rangle$$

Conditional on $\mathcal{F}_\ell$, $\underline{V}_{h+1}^{\ell,k,m}$

$$\mathbb{E}\left[\underline{V}_{h+1}^{\ell,k,m}(Z_{\ell,h,s,a}^{(i)}) | \mathcal{F}_\ell, \underline{V}_{h+1}^{\ell,k,m}\right] = \langle P_{h,s,a}, \underline{V}_{h+1}^{\ell,k,m}\rangle$$

Then apply eq. (30), with $Y_i = \underline{V}_{h+1}^{\ell,k,m}(Z_{\ell,h,s,a}^{(i)})$

$$\mathbb{E}\left[\exp\left(\frac{\sum_{i=1}^n \underline{V}_{h+1}^{\ell,k,m}(Z_{\ell,h,s,a}^{(i)})}{H(n+1)}\right) | \mathcal{F}_\ell, \underline{V}_{h+1}^{\ell,k,m}\right] \le \exp\left(\frac{\langle P_{h,s,a}, \underline{V}_{h+1}^{\ell,k,m}\rangle}{H}\right)$$

We derive

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{\underline{Q}_h^{\ell,k,m}(s,a) - Q_h^{\pi^{\ell,k,m},m}(s,a)}{H}\right) | \mathcal{F}_\ell, \underline{V}_{h+1}^{\ell,k,m}\right] \\
=&\mathbb{E}\left[\exp\left(\frac{\sum_{i=1}^n \underline{V}_{h+1}^{\ell,k,m}(Z_{\ell,h,s,a}^{(i)})}{H(n+1)}\right) | \mathcal{F}_\ell, \underline{V}_{h+1}^{\ell,k,m}\right] \times \exp\left(-\frac{\langle P_{h,s,a}, V_{h+1}^{\pi^{\ell,k,m},m}\rangle}{H}\right) \\
\le& \exp\left(\frac{\langle P_{h,s,a}, \underline{V}_{h+1}^{\ell,k,m}\rangle}{H}\right) \times \exp\left(-\frac{\langle P_{h,s,a}, V_{h+1}^{\pi^{\ell,k,m},m}\rangle}{H}\right) \\
\le& \exp\left(\frac{\langle P_{h,s,a}, \underline{V}_{h+1}^{\ell,k,m} - V_{h+1}^{\pi^{\ell,k,m},m}\rangle}{H}\right)
\end{aligned}$$

By the tower rule, since $\mathcal{F}_\ell \subseteq \sigma(\mathcal{F}_\ell, \underline{V}_{h+1}^{\ell,k,m})$

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{\underline{Q}_h^{\ell,k,m}(s,a) - Q_h^{\pi^{\ell,k,m},m}(s,a)}{H}\right)|\mathcal{F}_\ell\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[\exp\left(\frac{\underline{Q}_h^{\ell,k,m}(s,a) - Q_h^{\pi^{\ell,k,m},m}(s,a)}{H}\right)|\mathcal{F}_\ell, \underline{V}_{h+1}^{\ell,k,m}\right] \mid \mathcal{F}_\ell\right] \\
\leq&\mathbb{E}\left[\exp\left(\frac{\langle P_{h,s,a}, \underline{V}_{h+1}^{\ell,k,m} - V_{h+1}^{\pi^{\ell,k,m},m}\rangle}{H}\right) \mid \mathcal{F}_\ell\right] \\
=&\mathbb{E}\left[\prod_{s'\in\mathcal{S}} \exp\left(\frac{P_{h,s,a}(s')(\underline{V}_{h+1}^{\ell,k,m}(s') - V_{h+1}^{\pi^{\ell,k,m},m}(s'))}{H}\right)\middle|\mathcal{F}_\ell\right] \\
\leq&\prod_{s'\in\mathcal{S}} \mathbb{E}\left[\exp\left(\frac{\underline{V}_{h+1}^{\ell,k,m}(s') - V_{h+1}^{\pi^{\ell,k,m},m}(s')}{H}\right)\middle|\mathcal{F}_\ell\right]^{P_{h,s,a}(s')} \\
\leq&1.
\end{aligned}$$

The second inequality follows from Hölder's inequality that $\mathbb{E}[\prod_i X_i^{a_i}] \leq \prod_i (\mathbb{E}[X_i])^{a_i}$ for non-negative X_i and $a_i \in (0,1)$ for each i satisfying that $\sum_i a_i = 1$. We derive the last inequality from eq. (32).

On a tuple $(h,s,a) \notin \mathcal{R}_\ell$,

$$\underline{Q}_h^{\ell,k,m}(s,a) = r_h^m(s,a) \leq Q_h^{\pi^{\ell,k,m},m}(s,a)$$

so the same bound holds directly. Finally, Jensen's inequality for the action average yields

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{\underline{V}_h^{\ell,k,m}(s) - V_h^{\pi^{\ell,k,m},m}(s)}{H}\right) \mid \mathcal{F}_\ell\right] \\
\leq&\sum_{a\in\mathcal{A}} \pi_h^{\ell,k,m}(a|s)\mathbb{E}\left[\exp\left(\frac{\underline{Q}_h^{\ell,k,m}(s,a) - Q_h^{\pi^{\ell,k,m},m}(s,a)}{H}\right) \mid \mathcal{F}_\ell\right] \\
\leq&1,
\end{aligned}$$

completing the induction.

Since e^x is a convex function, Jensen's inequality gives, for all $\ell \geq 1$

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{\hat{v}_{\ell,m}^- - v_{\ell,m}}{H}\right) \mid \mathcal{F}_\ell\right] \\
\leq&\frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[\exp\left(\frac{\underline{V}_1^{\ell,k,m}(s_1^{\ell,k}) - V_1^{\pi^{\ell,k,m},m}(s_1^{\ell,k})}{H}\right) \mid \mathcal{F}_\ell\right] \leq 1
\end{aligned} \tag{33}$$

Since $\mathcal{F}_\ell$ contains all earlier optimistic evaluations and conservative evaluations, iterating conditional expectations using tower rule across phases gives

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{1}{H}\sum_{\ell=1}^L (\hat{v}_{\ell,m}^- - v_{\ell,m})\right)\right] \\
=&\mathbb{E}\left[\exp\left(\frac{1}{H}\sum_{\ell=1}^{L-1} (\hat{v}_{\ell,m}^- - v_{\ell,m})\right)\mathbb{E}\left[\exp\left(\frac{\hat{v}_{L,m}^- - v_{L,m}}{H}\right) \mid \mathcal{F}_L\right]\right] \\
\overset{(i)}{\leq}&\mathbb{E}\left[\exp\left(\frac{1}{H}\sum_{\ell=1}^{L-1} (\hat{v}_{\ell,m}^- - v_{\ell,m})\right)\right] \\
\leq&\cdots \leq 1
\end{aligned}$$

Where (i) is derived by the eq. (33).

□

C.3 Proof of Lemma 5

We now bound the conditional mean evaluation error using the same auxiliary fixed-prefix model.

Lemma 18 (Formal statement of Lemma 5)**.** *With probability at least* $1 - SAHL\delta''$*, for every* $\ell \in [L]$*,*

$$0 \le \sum_{m=1}^{M} p_\ell(m) \left(v_{\ell,m} - \mathbb{E}[\hat{v}^-_{\ell,m} \mid \mathcal{F}_\ell] \right) \le \frac{18SAH^2}{K} \log \frac{1}{\delta''}.$$

Proof. For a fixed phase ℓ and a counterfactual policy $\pi = \pi^{\ell,k,m}$, we suppress (ℓ, k, m) on the conservative evaluation, i.e., denote $\underline{Q}_h(s,a) = \underline{Q}_h^{\ell,k,m}(s,a)$, $\underline{V}_h(s) = \underline{V}_h^{\ell,k,m}(s)$, $r = r^m$ in this proof. Conditional on $\mathcal{F}_\ell$

$$\mathbb{E}[\underline{Q}_h(s,a) \mid \mathcal{F}_\ell] = \begin{cases} r_h^m(s,a) + \dfrac{N_{\ell,h}(s,a)}{N_{\ell,h}(s,a)+1} \langle P_{h,s,a}, \mathbb{E}[\underline{V}_{h+1} \mid \mathcal{F}_\ell] \rangle, & (h,s,a) \in \mathcal{R}_\ell, \\ r_h^m(s,a), & (h,s,a) \notin \mathcal{R}_\ell. \end{cases}$$

We first show

$$\mathbb{E}\left[\underline{V}_h(s) | \mathcal{F}_\ell\right] \le V_h^{\pi,m}(s)$$

by induction.

Base case. The base case holds trivially since $\underline{V}_{H+1}(s) = V_{H+1}^{\pi,m}(s) = 0$.

Inductive hypothesis. Suppose $\mathbb{E}[\underline{V}_{h+1}(s) | \mathcal{F}_\ell] \le V_{h+1}^{\pi,m}(s)$ for all $s \in \mathcal{S}$

Inductive step. On retained set $\mathcal{R}$, backward induction implies

$$\begin{aligned} \mathbb{E}\left[\underline{Q}_h(s,a) | \mathcal{F}_\ell\right] =& r_h^m(s,a) + \frac{N_{\ell,h}(s,a)}{N_{\ell,h}(s,a)+1} \langle P_{h,s,a}, \mathbb{E}[\underline{V}_{h+1} \mid \mathcal{F}_\ell] \rangle \\ \le& r_h^m(s,a) + \langle P_{h,s,a}, \mathbb{E}[\underline{V}_{h+1} \mid \mathcal{F}_\ell] \rangle \\ \le& r_h^m(s,a) + \langle P_{h,s,a}, V_{h+1}^{\pi,m} \rangle \\ =& Q_h^{\pi,m}(s,a) \end{aligned}$$

For an unretained tuple $(h,s,a) \notin \mathcal{R}_\ell$,

$$\mathbb{E}\left[\underline{Q}_h(s,a) | \mathcal{F}_\ell\right] = r_h^m(s,a) \le Q_h^{\pi,m}(s,a)$$

Together with $\mathbb{E}[\underline{V}_h(s) \mid \mathcal{F}_\ell] = \sum_a \pi_h(a|s) \mathbb{E}[\underline{Q}_h(s,a) \mid \mathcal{F}_\ell]$, we prove

$$0 \le \mathbb{E}[\underline{V}_h(s) \mid \mathcal{F}_\ell] \le V_h^{\pi,m}(s) \le H.$$

For a retained tuple $(h,s,a) \in \mathcal{R}_\ell$, subtracting the mean recursion from the true Bellman equation yields

$$\begin{aligned} & Q_h^{\pi,m}(s,a) - \mathbb{E}[\underline{Q}_h(s,a) \mid \mathcal{F}_\ell] \\ &= \langle P_{h,s,a}, V_{h+1}^{\pi,m} - \mathbb{E}[\underline{V}_{h+1} \mid \mathcal{F}_\ell] \rangle + \frac{\langle P_{h,s,a}, \mathbb{E}[\underline{V}_{h+1} \mid \mathcal{F}_\ell] \rangle}{N_{\ell,h}(s,a)+1} \\ &\le \langle P_{h,s,a}, V_{h+1}^{\pi,m} - \mathbb{E}[\underline{V}_{h+1} \mid \mathcal{F}_\ell] \rangle + \frac{H}{N_{\ell,h}(s,a)+1}. \end{aligned}$$

For an unretained tuple $(h,s,a) \notin \mathcal{R}_\ell$ the same upper bound holds with the last term replaced by H, because only the immediate reward is retained:

$$Q_h^{\pi,m}(s,a) - \mathbb{E}\left[\underline{Q}_h(s,a) | \mathcal{F}_\ell\right] \le \langle P_{h,s,a}, V_{h+1}^{\pi,m} - \mathbb{E}[\underline{V}_{h+1} \mid \mathcal{F}_\ell] \rangle + H$$

Averaging over actions and unrolling along the true transition kernel and policy π, starting from the recorded initial state $s_1^{\ell,k}$, recall the definition of $d_{\ell,h}^{k,m}(s,a)$ in eq. (20) and plug back the index (ℓ,k,m)

$$\begin{aligned}0 \leq & V_1^{\pi^{\ell,k,m},m}(s_1^{\ell,k}) - \mathbb{E}[\underline{V}_1^{\ell,k,m}(s_1^{\ell,k}) \mid \mathcal{F}_\ell] \\ \leq & H\left[\sum_{(h,s,a)\in\mathcal{R}_\ell} \frac{d_{\ell,h}^{k,m}(s,a)}{N_{\ell,h}(s,a)+1} + \sum_{(h,s,a)\notin\mathcal{R}_\ell} d_{\ell,h}^{k,m}(s,a)\right].\end{aligned}$$

Taking the average over $k \in [K]$ and then weighting by the reward-sampling distribution p_ℓ gives

$$0 \leq \sum_{m=1}^{M} p_\ell(m)\left(v_{\ell,m} - \mathbb{E}[\hat{v}_{\ell,m}^- \mid \mathcal{F}_\ell]\right) \leq \frac{H}{K}\left[\underbrace{\sum_{(h,s,a)\in\mathcal{R}_\ell} \frac{D_{\ell,h,s,a}^{\text{avg}}}{N_{\ell,h}(s,a)+1}}_{\text{retained part}} + \underbrace{\sum_{(h,s,a)\notin\mathcal{R}_\ell} D_{\ell,h,s,a}^{\text{avg}}}_{\text{unretained part}}\right]. \tag{34}$$

The right-hand side can be decomposed into the retained part and the unretained part.

Retained part. By Lemma 15, on the weighted occupancy event equation 24, a retained row satisfies

$$\frac{D_{\ell,h,s,a}^{\text{avg}}}{N_{\ell,h}(s,a)+1} \leq \frac{2N_{\ell,h}(s,a) + 2\log(1/\delta'')}{N_{\ell,h}(s,a)+1} \leq \frac{9}{4} \tag{35}$$

since $\log(1/\delta'') \geq 1$, every tuple contributes at most $18\log(1/\delta'')$ to the retained part. Since there are SAH tuples, then

$$\text{retained part} \leq \frac{9SAH}{4}$$

Unretained part. For an unretained triple $(h,s,a) \notin \mathcal{R}_\ell$, we have $N_{\ell,h}(s,a) < 8\log(1/\delta'')$. Hence, by eq. (24)

$$D_{\ell,h,s,a}^{\text{avg}} \leq 2N_{\ell,h}(s,a) + 2\log\frac{1}{\delta''} < 18\log\frac{1}{\delta''}.$$

Similarly, $|\mathcal{R}_\ell^c| \leq SAH$ implies

$$\text{unretained part} \leq 18SAH\log\frac{1}{\delta''}.$$

Combining the two parts together, we have

$$0 \leq \sum_{m=1}^{M} p_\ell(m)\left(v_{\ell,m} - \mathbb{E}\left[\hat{v}_{\ell,m}^- | \mathcal{F}_\ell\right]\right) \leq \frac{18SAH^2}{K}\log\frac{1}{\delta''}.$$

Lemma 15 ensures this event simultaneously over all phases and rows with failure probability at most $SAHL(\delta'')^{3/2} \leq SAHL\delta''$, proving the claim. The same event controls all rewards through their weighted occupancy; no additional union bound over rewards is required. □

C.4 Proof of Theorem 1

In this section, we complete the proof of our main Theorem 1.

Bounding the cumulative certificates. Fix $\ell \in [L]$ and $m \in [M]$. Conditional on $\mathcal{F}_\ell$, both $\hat{v}_{\ell,m}^+$ and $v_{\ell,m}$ are fixed. Denote the X as the normalized gap between optimistic evaluation and the conservative evaluation: $X = \frac{\max\{\hat{v}_{\ell,m}^+, v_{\ell,m}\} - \hat{v}_{\ell,m}^-}{H} \in [-1,1]$. Then Lemma 17 implies $\mathbb{E}[e^{-X} \mid \mathcal{F}_\ell] \leq 1$. For $x \in [-1,1]$, the elementary inequality $e^{-x} \geq 1 - x + x^2/3$ therefore gives $\mathbb{E}[X^2|\mathcal{F}_\ell] \leq 3\mathbb{E}[X|\mathcal{F}_\ell]$. Also noting that $e^x \leq 1 + x + x^2$, we then have $\mathbb{E}[\exp(X) \mid \mathcal{F}_\ell] \leq 1 + 4\mathbb{E}[X \mid \mathcal{F}_\ell]$.

Since $(\hat{v}_{\ell,m}^+ - \hat{v}_{\ell,m}^-)/H \leq X$, it follows that

$$\mathbb{E}\left[\exp\left(\frac{\hat{v}_{\ell,m}^+ - \hat{v}_{\ell,m}^-}{H}\right)\middle|\mathcal{F}_\ell\right] \leq 1 + \frac{4}{H}\left[(\hat{v}_{\ell,m}^+ - v_{\ell,m})_+ + v_{\ell,m} - \mathbb{E}[\hat{v}_{\ell,m}^- \mid \mathcal{F}_\ell]\right], \tag{36}$$

where $(x)_+ = \max\{x, 0\}$. We recall that $W_\ell = \frac{1}{M}\sum_{m=1}^{M} \exp\left(\frac{1}{H}\sum_{j=1}^{\ell-1}(\hat{v}^+_{j,m} - \hat{v}^-_{j,m})\right)$. In particular $W_1 = 1$.

The update in Algorithm 1 gives $\frac{W_{\ell+1}}{W_\ell} = \sum_{m=1}^{M} p_\ell(m) \exp\left(\frac{\hat{v}^+_{\ell,m} - \hat{v}^-_{\ell,m}}{H}\right)$.

Then, by eq. (36) it implies

$$\begin{aligned}
\mathbb{E}\left[\frac{W_{\ell+1}}{W_\ell} \mid \mathcal{F}_\ell\right] =& \mathbb{E}\left[\sum_{m\in[M]} p_\ell(m) \exp\left(\frac{\hat{v}^+_{\ell,m} - \hat{v}^-_{\ell,m}}{H}\right)\middle| \mathcal{F}_\ell\right] \\
\leq& 1 + \sum_{m\in[M]} p_\ell(m)\left(\frac{4}{H}\left[(\hat{v}^+_{\ell,m} - v_{\ell,m})_+ + v_{\ell,m} - \mathbb{E}[\hat{v}^-_{\ell,m} \mid \mathcal{F}_\ell]\right]\right) \\
\leq& \exp\left(\sum_{m\in[M]} p_\ell(m)\left(\frac{4}{H}\left[(\hat{v}^+_{\ell,m} - v_{\ell,m})_+ + v_{\ell,m} - \mathbb{E}[\hat{v}^-_{\ell,m} \mid \mathcal{F}_\ell]\right]\right)\right)
\end{aligned}$$

As a result, divide both sides by the right-hand side.

$$\mathbb{E}\left[\frac{W_{\ell+1}}{W_\ell}\exp\left\{-\frac{4}{H}\sum_{m=1}^{M} p_\ell(m)\left[(\hat{v}^+_{\ell,m} - v_{\ell,m})_+ + v_{\ell,m} - \mathbb{E}[\hat{v}^-_{\ell,m} \mid \mathcal{F}_\ell]\right]\right\} \mid \mathcal{F}_\ell\right] \leq 1 \tag{37}$$

Iterating conditional expectations over phases yields:

$$\mathbb{E}\left[W_{L+1}\exp\left\{-\frac{4}{H}\sum_{\ell=1}^{L}\sum_{m=1}^{M} p_\ell(m)\left[(\hat{v}^+_{\ell,m} - v_{\ell,m})_+ + v_{\ell,m} - \mathbb{E}[\hat{v}^-_{\ell,m} \mid \mathcal{F}_\ell]\right]\right\}\right] \leq 1.$$

Since $W_{L+1} \geq \frac{1}{M}\exp\left(\frac{1}{H}\sum_{\ell=1}^{L}(\hat{v}^+_{\ell,m} - \hat{v}^-_{\ell,m})\right)$ for every m, Markov's inequality gives, with probability at least $1 - \delta/64$,

$$\begin{aligned}
\max_{m\in[M]}\sum_{\ell=1}^{L}(\hat{v}^+_{\ell,m} - \hat{v}^-_{\ell,m}) \leq& 4\sum_{\ell=1}^{L}\sum_{m=1}^{M} p_\ell(m)\left[(\hat{v}^+_{\ell,m} - v_{\ell,m})_+ + v_{\ell,m} - \mathbb{E}[\hat{v}^-_{\ell,m} \mid \mathcal{F}_\ell]\right] \\
&+ H\log\frac{64M}{\delta}.
\end{aligned} \tag{38}$$

On the optimism event, the positive part can be removed. Combining Lemma 2 with Lemma 18 then gives

$$\begin{aligned}
\frac{1}{L}\max_{m\in[M]}\sum_{\ell=1}^{L}(\hat{v}^+_{\ell,m} - \hat{v}^-_{\ell,m}) \lesssim& \sqrt{\frac{SAH^3}{KL}\log_2^3(2KL)\log\frac{M}{\delta'}} + \frac{SAH^2}{K}\log\frac{1}{\delta''} \\
&+ \frac{H}{L}\log\frac{64M}{\delta} + \frac{H}{KL}\log\frac{64}{\delta}.
\end{aligned} \tag{39}$$

We account for these events together below.

Correctness of certification. Recall the definition of $\mathcal{F}^\ell_{k-1}$. By definition of $s_1^{\ell,k} \sim \mu$, we have that

$$\mathbb{E}\left[\frac{V_1^{*,m}(s_1^{\ell,k}) - V_1^{\pi^{\ell,k,m},m}(s_1^{\ell,k})}{H} \mid \mathcal{F}^\ell_{k-1}\right] = \frac{V_0^{*,m} - V_0^{\pi^{\ell,k,m},m}}{H}.$$

By the multiplicative concentration argument we derived in eq. (13), with probability at least $1 - \delta/64$,

$$\begin{aligned}
V_0^{*,m} - V_0^{\bar{\pi}^m,m} \leq& \frac{2}{KL}\sum_{\ell=1}^{L}\sum_{k=1}^{K}\left(V_1^{*,m}(s_1^{\ell,k}) - V_1^{\pi^{\ell,k,m},m}(s_1^{\ell,k})\right) + \frac{2H}{KL}\log\frac{64M}{\delta} \\
\leq& \frac{2}{L}\sum_{\ell=1}^{L}(\hat{v}^+_{\ell,m} - v_{\ell,m}) + \frac{2H}{KL}\log\frac{64M}{\delta}.
\end{aligned} \tag{40}$$

Here the second inequality is on the optimism event.

By Lemma 4 and paying an additional failure probability $\delta/64$, we have that

$$V_0^{*,m} - V_0^{\overline{\pi}^m,m} \le \frac{2}{L}\sum_{\ell=1}^{L}(\hat{v}_{\ell,m}^{+} - \hat{v}_{\ell,m}^{-}) + \frac{2H}{L}\log\frac{64M}{\delta} + \frac{2H}{KL}\log\frac{64M}{\delta}.$$

Combining this with equation 39 yields

$$\begin{aligned}\max_{m\in[M]}\left(V_0^{*,m} - V_0^{\overline{\pi}^m,m}\right) \lesssim \sqrt{\frac{SAH^3}{KL}\log_2^3(2KL)\log\frac{M}{\delta'}} + \frac{SAH^2}{K}\log\frac{1}{\delta''} \\ + \frac{H}{L}\log\frac{64M}{\delta} + \frac{H}{KL}\log\frac{64M}{\delta} = O(\epsilon).\end{aligned} \tag{41}$$

The proof of the optimality is finished.

Failure probability. The weighted optimistic bound and its underlying optimism event together cost $(4SAHKLM+5)\delta' + \delta/64$. The conditional mean evaluation bound costs $SAHL\delta''$. The weight potential, cumulative evaluation overestimation, and initial-state concentration each cost $\delta/64$. Finally, Lemma 3 costs $2SAHL\delta''$. On its success event, all retained quotas in the auxiliary run are filled, and the coupling ensures that its entire history and output agree with Algorithm 1. We intersect with this event only after deriving the preceding conditional moment bounds. Counting optimism only once, the total failure probability is at most

$$(4SAHKLM+5)\delta' + 3SAHL\delta'' + \frac{\delta}{16}. \tag{42}$$

For the algorithm's choices $\delta' = \delta/(64SAHKLM)$ and $\delta'' = \delta/(64SAHL)$, the local confidence ranges hold. Since $SAHKLM \ge 4$, equation 42 is at most $49\delta/256 < \delta$. The tail levels $\delta/64$ used above do not change either local confidence parameter.

Total episode complexity. Each of the L phases uses K learning episodes and $8K$ replay episodes. The total is therefore

$$9KL = O\left(\frac{SAH^3\log M}{\epsilon^2}\log^4\left(\frac{SAH\log M}{\min\{\epsilon,1\}\,\delta}\right)\right).$$

The ceilings are absorbed by the bounds above, and C is universal. With probability at least $1-\delta$, all returned policies are ϵ-optimal for their respective rewards. This completes the proof.

D Supplementary Proofs in Section 4

We use the auxiliary run from Appendix C. Index all learning episodes consecutively by $k \in [KL]$: a superscript k abbreviates (ℓ, j) when $k = (\ell-1)K + j$ with $j \in [K]$. Thus K still denotes the number of learning episodes per phase, and $\sum_{k,h}$ below ranges over $k \in [KL]$ and $h \in [H]$. The empirical rows $\widehat{P}_{h,s,a}^k$ and completed-batch sizes $N_h^k(s,a)$ belong to the persistent learner; replay samples do not update them.

D.1 Proof of Lemma 14

Let $e_s \in \mathbb{R}^S$ denote the s-th coordinate vector. Write $V_h^k := \overline{V}_h^{k,m^k}$, $Q_h^k := \overline{Q}_h^{k,m^k}$, $r_h^k := r_h^{m^k}$, and $b_h^k := b_h^{k,m^k}$. Throughout this proof, $0 < \delta' \le \min\{e^{-1}, (2KL)^{-1}\}$. Since the policy is greedy with respect to Q_h^k,

$$V_h^k(s_h^k) = Q_h^k(s_h^k, a_h^k) \le r_h^k(s_h^k, a_h^k) + \langle \widehat{P}_{h,s_h^k,a_h^k}^k, V_{h+1}^k\rangle + b_h^k(s_h^k, a_h^k).$$

Adding and subtracting the true transition and telescoping gives

$$\begin{aligned}
&\sum_{k=1}^{KL}\big(V_1^k(s_1^k) - V_1^{\pi^k,m^k}(s_1^k)\big)\\
&\le \sum_{k,h} b_h^k(s_h^k,a_h^k) + \sum_{k,h}\langle \widehat{P}^k_{h,s_h^k,a_h^k} - P_{h,s_h^k,a_h^k}, V_{h+1}^k\rangle\\
&\quad + \sum_{k,h}\langle P_{h,s_h^k,a_h^k} - e_{s_{h+1}^k}, V_{h+1}^k\rangle + \sum_{k=1}^{KL}\left(\sum_{h=1}^{H} r_h^k(s_h^k,a_h^k) - V_1^{\pi^k,m^k}(s_1^k)\right). \qquad (43)
\end{aligned}$$

The last two terms are martingale sums. Hoeffding–Azuma bounds them by $2\sqrt{KLH^3\log(1/\delta')}$ and $2H\sqrt{KL\log(1/\delta')}$, respectively, each with failure probability at most δ'. For the latter sum, condition on the history, m^k, and s_1^k before the episode trajectory is drawn. The forward histories also contain all completed auxiliary replay outcomes and weight updates. The martingale sums remain unchanged during replay, and the next learning transition still has conditional law P_{h,s_h^k,a_h^k}.

If KL satisfies equation 48, Lemmas 20 and 21 bound the first two terms on the same event, with failure probability at most $3\delta'$. Consequently, with probability at least $1-5\delta'$,

$$\sum_{k=1}^{KL}\big(V_1^k(s_1^k) - V_1^{\pi^k,m^k}(s_1^k)\big) \lesssim \sqrt{SAH^3KL\log_2^3(2KL)\log\frac{M}{\delta'}}.$$

If KL is below the sufficiently large universal threshold in equation 48, the deterministic bound on this surplus is HKL, which is itself at most a universal constant times the displayed square-root bound. Thus, for every $KL \ge 1$,

$$\sum_{k=1}^{KL}\big(V_1^k(s_1^k) - V_1^{\pi^k,m^k}(s_1^k)\big) \lesssim \min\left\{\sqrt{SAH^3KL\log_2^3(2KL)\log\frac{M}{\delta'}},\ HKL\right\}$$

with probability at least $1-5\delta'$. The bound HKL is used only for the value surplus, not for the sum of untruncated bonuses.

D.2 Profile-Based Concentration Lemma for Multi Reward MDPs

Following Zhang et al. (2025), couple the entire learning process from initialization to an independent stream of next-state samples for each (h,s,a): the i-th learning visit to this row uses the i-th sample in its stream. These streams are independent across rows and stages and independent of all replay streams. We establish the profile event globally, without conditioning on an intermediate phase's history; batches and unfinished batch histograms persist across phases. Let $\widehat{P}^{(j)}_{h,s,a}$ be the empirical transition of the j-th potential batch, defined even if it is never completed. Batch 1 contains the first sample; for $j \ge 2$, batch j contains samples $2^{j-2}+1,\dots,2^{j-1}$. Thus the first two batches both have size 1 but are distinct. Let $\widehat{P}^{(0)}_{h,s,a}$ be the fixed initial transition row.

Define the realized profile $\widehat{\mathcal{I}}^k_{h,s,a}$ to be 0 if no batch has been completed before episode k, and otherwise the index of the most recently completed batch. In particular,

$$\widehat{P}^k_{h,s,a} = \widehat{P}^{(\widehat{\mathcal{I}}^k_{h,s,a})}_{h,s,a}.$$

The realized profile belongs to

$$\mathcal{C} := \left\{\mathcal{I} = (\mathcal{I}^1,\dots,\mathcal{I}^{KL}) : \mathcal{I}^1 \le \cdots \le \mathcal{I}^{KL},\quad \mathcal{I}^k \in \{0,\dots,1+\lfloor\log_2(KL)\rfloor\}^{SAH}\right\}.$$

Each coordinate can be specified by the episode at which it first reaches each positive batch index, using $KL+1$ if it never does. Consequently,

$$|\mathcal{C}| \le (KL+1)^{SAH(1+\lfloor\log_2(KL)\rfloor)}.$$

Lemma 19. *Suppose $KL \geq \max\{2, SAH\}$. For every deterministic $\mathcal{I} \in \mathcal{C}$, let $\mathcal{X}_{h,\mathcal{I}} \subseteq [0,H]^S$ contain 0 and have at most $Z \geq 2$ labeled vectors. Each labeled vector must be a deterministic function of $\mathcal{I}$, the known rewards, and the transition streams at stages strictly greater than h. With probability at least $1-\delta'$, the bound*

$$\begin{aligned}
&\sum_{k,h} \langle \widehat{P}^k_{h,s^k_h,a^k_h} - P_{h,s^k_h,a^k_h}, X^k_h \rangle \\
&\lesssim \sqrt{\log_2(2KL) \sum_{k,h} \mathbb{V}(P_{h,s^k_h,a^k_h}, X^k_h) \left(SAH \log_2^2(2KL) \log(2Z) + \log \frac{1}{\delta'} \right)} \\
&\quad + H \log_2(2KL) \left(SAH \log_2^2(2KL) \log(2Z) + \log \frac{1}{\delta'} \right) \qquad (44)
\end{aligned}$$

holds uniformly for all choices $X^k_h \in \mathcal{X}_{h,\widehat{\mathcal{I}}}$.

Proof. First fix a deterministic profile $\mathcal{I}$, a batch index $j \geq 1$, and one deterministic candidate label for each (h,s,a). Write $X_{h,s,a}$ for the resulting vector. Reveal the streams backward in h, starting with $h = H$. Before revealing the batch-j samples at stage h, every $X_{h,s,a}$ is measurable because all streams at stages greater than h have already been revealed. Each centered sample contributes

$$\frac{X_{h,s,a}(s') - \langle P_{h,s,a}, X_{h,s,a} \rangle}{\max\{1, 2^{j-2}\}},$$

which has conditional mean zero, absolute value at most $H/\max\{1,2^{j-2}\}$, and conditional variance $\mathbb{V}(P_{h,s,a}, X_{h,s,a})/\max\{1,2^{j-2}\}^2$. After each stage, reveal its remaining samples with zero increments before moving to the preceding stage. This constructs a martingale with total predictable variance $\sum_{h,s,a} \mathbb{V}(P_{h,s,a}, X_{h,s,a})/\max\{1,2^{j-2}\}$. Freedman's inequality (Lemma 9), with variance peeling, therefore gives the corresponding Bernstein bound for $\sum_{h,s,a} \langle \widehat{P}^{(j)}_{h,s,a} - P_{h,s,a}, X_{h,s,a} \rangle$.

There are at most Z^{SAH} deterministic label assignments for each profile and at most $1+\lfloor \log_2(KL) \rfloor$ batches. Allocate failure probability $\delta'/[(1+\lfloor \log_2(KL) \rfloor)|\mathcal{C}|Z^{SAH}]$ to each such choice. The logarithmic cost of this union bound and the variance peeling is at most a universal constant times

$$SAH \log_2^2(2KL) \log(2Z) + \log \frac{1}{\delta'};$$

here $KL \geq SAH$ also bounds the peeling factor. Thus, simultaneously for every profile, batch, and candidate-label assignment,

$$\begin{aligned}
&\sum_{h,s,a} \langle \widehat{P}^{(j)}_{h,s,a} - P_{h,s,a}, X_{h,s,a} \rangle \\
&\lesssim \sqrt{\frac{\sum_{h,s,a} \mathbb{V}(P_{h,s,a}, X_{h,s,a})}{\max\{1,2^{j-2}\}} \left(SAH \log_2^2(2KL) \log(2Z) + \log \frac{1}{\delta'} \right)} \\
&\quad + \frac{H}{\max\{1,2^{j-2}\}} \left(SAH \log_2^2(2KL) \log(2Z) + \log \frac{1}{\delta'} \right). \qquad (45)
\end{aligned}$$

Only after establishing this uniform event do we substitute the realized profile $\widehat{\mathcal{I}}$.

A batch-j model is used on at most 2^{j-1} subsequent visits to its row. For each $i \leq 2^{j-1}$, let $\widetilde{X}^{j,i}_{h,s,a}$ be the vector used on the i-th such visit, padding with 0 if it does not occur. Although these choices are random, they are covered by the uniform event above. Visits with no completed batch contribute at most SAH^2 in total. The remaining error decomposes as

$$\sum_{j=1}^{1+\lfloor \log_2(KL) \rfloor} \sum_{i=1}^{2^{j-1}} \sum_{h,s,a} \langle \widehat{P}^{(j)}_{h,s,a} - P_{h,s,a}, \widetilde{X}^{j,i}_{h,s,a} \rangle.$$

Apply equation 45 to each summand and use Cauchy–Schwarz, together with

$$\sum_{j=1}^{1+\lfloor \log_2(KL) \rfloor} \frac{2^{j-1}}{\max\{1,2^{j-2}\}} \leq 2(1+\lfloor \log_2(KL) \rfloor) \leq 2\log_2(2KL).$$

Every observed visit appears once in the resulting variance sum. The initial contribution SAH^2 is absorbed by the linear term in equation 44, proving the claim. □

Auxiliary optimistic values for fixed profiles. For a deterministic $\mathcal{I} \in \mathcal{C}$, run the backward updates of Algorithm 2 with $\widehat{P}^k_{h,s,a}$ replaced by $\widehat{P}^{(\mathcal{I}^k_{h,s,a})}_{h,s,a}$ and with the corresponding batch size (0 for index 0, 1 for index 1, and 2^{j-2} for index $j \geq 2$). In particular, use the same bonus formula equation 3, including its $\max\{N, 1\}$ denominators, and set

$$\overline{V}^{k,m,\mathcal{I}}_{H+1}(s) = 0, \tag{46}$$

$$\overline{V}^{k,m,\mathcal{I}}_{h}(s) = \max_{a\in\mathcal{A}} \min\left\{ r^m_h(s,a) + \langle \widehat{P}^{(\mathcal{I}^k_{h,s,a})}_{h,s,a}, \overline{V}^{k,m,\mathcal{I}}_{h+1}\rangle + b^{k,m,\mathcal{I}}_h(s,a),\ H \right\}. \tag{47}$$

The family

$$\mathcal{X}_{h,\mathcal{I}} = \{\overline{V}^{k,m,\mathcal{I}}_{h+1} : k \in [KL],\ m \in [M]\} \cup \{0\}$$

has cardinality at most $Z = KLM + 1$ and satisfies the measurability assumption of Lemma 19. For the realized profile, backward induction gives $\overline{V}^{k,m,\widehat{\mathcal{I}}}_h = \overline{V}^{k,m}_h$. For fixed $\mathcal{I}$, these reward-specific backups do not depend on p_ℓ or on replay outcomes. The realized profile and candidate labels may depend on past auxiliary evaluations, but are covered by the same uniform event. Hence adaptive reward weights require no additional union bound. The same properties hold for the family of vectors $(\overline{V}^{k,m,\mathcal{I}}_{h+1})^2/H$, together with 0.

D.3 Statement and Proof of Lemma 20

Lemma 20. *Suppose* $0 < \delta' \leq \min\{e^{-1}, (2KL)^{-1}\}$ *and*

$$KL \gtrsim SAH \log_2^3(2KL) \log \frac{M}{\delta'}, \tag{48}$$

with a sufficiently large universal implied constant. With probability at least $1 - 3\delta'$*,*

$$\sum_{k,h} b^k_h(s^k_h, a^k_h) \lesssim \sqrt{SAH^3KL \log_2(2KL) \log \frac{1}{\delta'}}.$$

On the same event, $\sum_{k,h} \mathbb{V}(P_{h,s^k_h,a^k_h}, V^k_{h+1}) \lesssim KLH^2$ *and* $\sum_{k,h} \mathbb{V}(\widehat{P}^k_{h,s^k_h,a^k_h}, V^k_{h+1}) \lesssim KLH^2$.

Proof. Retain the abbreviations

$$\widehat{\texttt{Var}} := \sum_{k,h} \mathbb{V}(\widehat{P}^k_{h,s^k_h,a^k_h}, V^k_{h+1}), \qquad \texttt{Var} := \sum_{k,h} \mathbb{V}(P_{h,s^k_h,a^k_h}, V^k_{h+1}),$$

and $\texttt{BonusSum} := \sum_{k,h} b^k_h(s^k_h, a^k_h)$. By the bonus formula, Cauchy–Schwarz, and Lemma 11 applied to all KL learning episodes,

$$\begin{aligned}\texttt{BonusSum} &= \frac{460}{9} \sum_{k,h} \sqrt{\frac{\mathbb{V}(\widehat{P}^k_{h,s^k_h,a^k_h}, V^k_{h+1}) \log(1/\delta')}{\max\{N^k_h(s^k_h, a^k_h), 1\}}} + \frac{544H}{9} \sum_{k,h} \frac{\log(1/\delta')}{\max\{N^k_h(s^k_h, a^k_h), 1\}} \\ &\leq \frac{460}{9} \sqrt{2SAH \log_2(2KL) \log(1/\delta') \widehat{\texttt{Var}}} + \frac{1088}{9} SAH^2 \log_2(2KL) \log \frac{1}{\delta'}. \end{aligned} \tag{49}$$

The variance identities, telescoping over h, and $V_h^k(s_h^k) \le r_h^k(s_h^k, a_h^k) + \langle \widehat{P}^k_{h,s_h^k,a_h^k}, V_{h+1}^k \rangle + b_h^k(s_h^k, a_h^k)$ imply

$$\begin{aligned}
\widehat{\texttt{Var}} &\le \sum_{k,h} \langle \widehat{P}^k_{h,s_h^k,a_h^k} - P_{h,s_h^k,a_h^k}, (V_{h+1}^k)^2 \rangle \\
&\quad + \sum_{k,h} \langle P_{h,s_h^k,a_h^k} - e_{s_{h+1}^k}, (V_{h+1}^k)^2 \rangle + 2H\texttt{BonusSum} + 2KLH^2, \\
\texttt{Var} &\le \sum_{k,h} \langle P_{h,s_h^k,a_h^k} - e_{s_{h+1}^k}, (V_{h+1}^k)^2 \rangle \\
&\quad + 2H \sum_{k,h} \max\{\langle \widehat{P}^k_{h,s_h^k,a_h^k} - P_{h,s_h^k,a_h^k}, V_{h+1}^k \rangle, 0\} + 2H\texttt{BonusSum} + 2KLH^2.
\end{aligned}$$

Here we used $0 \le V_h^k \le H$, $\sum_{k,h} r_h^k(s_h^k, a_h^k) \le KLH$, and $x^2 - y^2 \le 2H \max\{x - y, 0\}$ for $x, y \in [0, H]$.

Apply Lemma 19 to the auxiliary value family and to its squared, normalized family. The two uniform events have total failure probability at most $2\delta'$. Inclusion of 0 allows the first event to bound the positive parts of the model errors, by selecting 0 whenever the inner product is negative; it also bounds their signed sum. For the second family use $\mathbb{V}(p, V^2/H) \le 4\mathbb{V}(p, V)$. Because $\delta' \le (2KL)^{-1}$, $\log(2(KLM+1)) \lesssim \log(M/\delta')$. The remaining squared-value martingale is at most $2H^2\sqrt{KLH\log(1/\delta')} \le KLH^2$, except on an event of probability δ'; the last inequality follows from equation 48 with a sufficiently large implied constant. Thus, on a common event of probability at least $1 - 3\delta'$,

$$\begin{aligned}
\max\{\texttt{Var}, \widehat{\texttt{Var}}\} &\lesssim H\sqrt{SAH\log_2^3(2KL)\log\frac{M}{\delta'}\,\texttt{Var}} + SAH^3\log_2^3(2KL)\log\frac{M}{\delta'} \\
&\quad + KLH^2 + H\texttt{BonusSum}.
\end{aligned}$$

Insert equation 49 and apply Young's inequality to absorb the terms involving $\sqrt{\texttt{Var}}$ and $\sqrt{\widehat{\texttt{Var}}}$ into the left-hand side. This gives

$$\max\{\texttt{Var}, \widehat{\texttt{Var}}\} \lesssim KLH^2 + SAH^3\log_2^3(2KL)\log\frac{M}{\delta'} \lesssim KLH^2,$$

where the last step uses equation 48. Finally, substitute into equation 49; the same condition absorbs its linear term into the square-root term. This proves the claim. □

D.4 Statement and Proof of Lemma 21

Lemma 21. *Under the hypotheses of Lemma 20, on the same event of probability at least* $1 - 3\delta'$*,*

$$\sum_{k,h} \langle \widehat{P}^k_{h,s_h^k,a_h^k} - P_{h,s_h^k,a_h^k}, V_{h+1}^k \rangle \lesssim \sqrt{SAH^3KL\log_2^3(2KL)\log\frac{M}{\delta'}}.$$

Proof. The value-family concentration event used in the preceding proof already gives

$$\begin{aligned}
&\sum_{k,h} \langle \widehat{P}^k_{h,s_h^k,a_h^k} - P_{h,s_h^k,a_h^k}, V_{h+1}^k \rangle \\
&\lesssim \sqrt{SAH\log_2^3(2KL)\log\frac{M}{\delta'}\,\texttt{Var}} + SAH^2\log_2^3(2KL)\log\frac{M}{\delta'}.
\end{aligned}$$

Use $\texttt{Var} \lesssim KLH^2$ from Lemma 20 and absorb the linear term by equation 48. No additional failure probability is incurred. □

E Numerical Experiments

This appendix provides the experimental setting and evaluation details for the two experiments in Section 5.

We consider random nonstationary finite-horizon MDPs with $S = 10$, $A = 5$, $H = 6$, and initial state $s_1 = 0$. For each stage $h \in [H]$ and state–action pair (s, a), we independently sample a transition row as

$$\alpha_{h,s,a} \sim \mathrm{Uniform}(0.15, 1.2), \qquad P_h(\cdot \mid s, a) \sim \mathrm{Dirichlet}(\alpha_{h,s,a}\mathbf{1}_S).$$

Thus, transitions vary with h, s, and a, have full support, and contain no absorbing states. Independently, each reward function is generated by sampling

$$u_{m,h,s} \sim \mathrm{Beta}(0.25, 0.25), \qquad r_h^m(s, a) = u_{m,h,s} \quad \text{for all } a.$$

Rewards vary across reward functions, stages, and states, but are shared across actions at each state and stage. The transition and reward tables remain fixed within each instance. Returns are undiscounted sums over H steps, with no additional terminal reward.

E.1 Experiment 1: Policy error versus episode budget

We fix $M = 16$ and test total episode budgets $B \in \{2^9, 2^{10}, \ldots, 2^{17}\}$. For each budget, we run each implementation from scratch using 12 independent environment–learner seed pairs, shared across algorithms and budgets. Optimal and learned-policy values are computed exactly by backward dynamic programming with the true kernel. We take the maximum policy error over the M rewards within each run and then average over the 12 runs, reporting $\overline{\mathrm{Error}}_{12}(B, 16)$ as defined in the main text.

For every algorithm, the number of episodes in Figure 1a counts all environment episodes actually used across all stages. We denote this total by B, consistently with the main text; no additional interaction cost is excluded. For our algorithm, this includes the initial random warmup and all subsequent learning episodes. Pointwise 95% confidence intervals are obtained from the 2.5th and 97.5th percentiles of 10,000 bootstrap means, resampling seed pairs with replacement and preserving their pairing across algorithms and budgets. The comparisons and pointwise intervals concern the mean worst-reward error under these selected configurations; the bootstrap holds configurations fixed and does not account for their selection.

E.2 Experiment 2: Required episode budget versus reward count

We vary $M \in \{4, 6, 8, 12, 16, 24, 32, 48, 64, 96, 128\}$ and use eight independent environment–learner seed pairs, with algorithm settings fixed as in Experiment 1. For each environment, we keep the transition kernel fixed and generate 128 reward functions; the instance with M rewards uses the first M functions. Each algorithm is rerun for every (B, M) setting, using the same exact policy evaluation and episode accounting as above.

For each M, we estimate the required number of episodes B to achieve $\overline{\mathrm{Error}}_8(B, M) \le 0.1$, using the same total count of episodes actually used by each algorithm. We first identify the smallest tested budget at which the mean error, and the mean errors at all larger tested budgets, are at most 0.1. We then estimate the crossing by linear interpolation of log error against log budget between that point and the preceding tested budget. Budget grids are refined near the crossing for each algorithm. Pointwise 95% confidence intervals use 10,000 paired bootstrap resamples of the eight seed pairs, recomputing the mean error curve and interpolated crossing in each resample and taking their 2.5th and 97.5th percentiles. The plotted quantity is therefore the estimated crossing of the mean error curve, rather than the average of individual runs' crossing budgets. These pointwise intervals quantify empirical variability under the selected configurations, rather than the simultaneous PAC guarantee.

F Lower Bound Construction

We establish a lower bound for learning policies for a finite collection of reward functions fixed and revealed before exploration. Our proof combines a finite-packing and decoding argument (Jin et al., 2020, Appendix D.4) with a time-dependent tree embedding (Ridel & Cohen, 2026, Appendix C.3). The construction places independent hard transitions at $\Theta(SH)$ time–leaf pairs while sharing the same physical states (see Figure 2).

Proposition 1. *There exist universal constants $c_s, c_{\mathrm{lb}} > 0$ such that the following holds. Suppose that $A \geq 2$, $L \geq 16$, $\log(AL) + 1 \leq c_s S$, $H \geq 4\big(1 + \lceil \log_2 S \rceil\big)$ and $0 < \epsilon \leq \frac{1}{192}$. There exist a class $\mathcal{P}_L$ of time-inhomogeneous transition kernels on at most S states, with A actions and horizon H, and a fixed reward class $\mathcal{R}_L \subset [0,1]^{\mathcal{S}\times\mathcal{A}\times[H]}$ satisfying $|\mathcal{R}_L| \leq SHA^2L^2$, with the following property. Any algorithm that is given $\mathcal{R}_L$ before exploration, stops after a possibly random number K of online episodes, and satisfies*

$$\mathbb{P}_P\Big(V^{\star}_{P,r} - V^{\widehat{\pi}_r}_{P,r} \leq \epsilon \quad \textit{for every } r \in \mathcal{R}_L\Big) \geq \frac{1}{2} \quad \textit{for every } P \in \mathcal{P}_L$$

must obey

$$\sup_{P\in\mathcal{P}_L} \mathbb{E}_P[K] \geq c_{\mathrm{lb}} \frac{SAH^3 \log L}{\epsilon^2}.$$

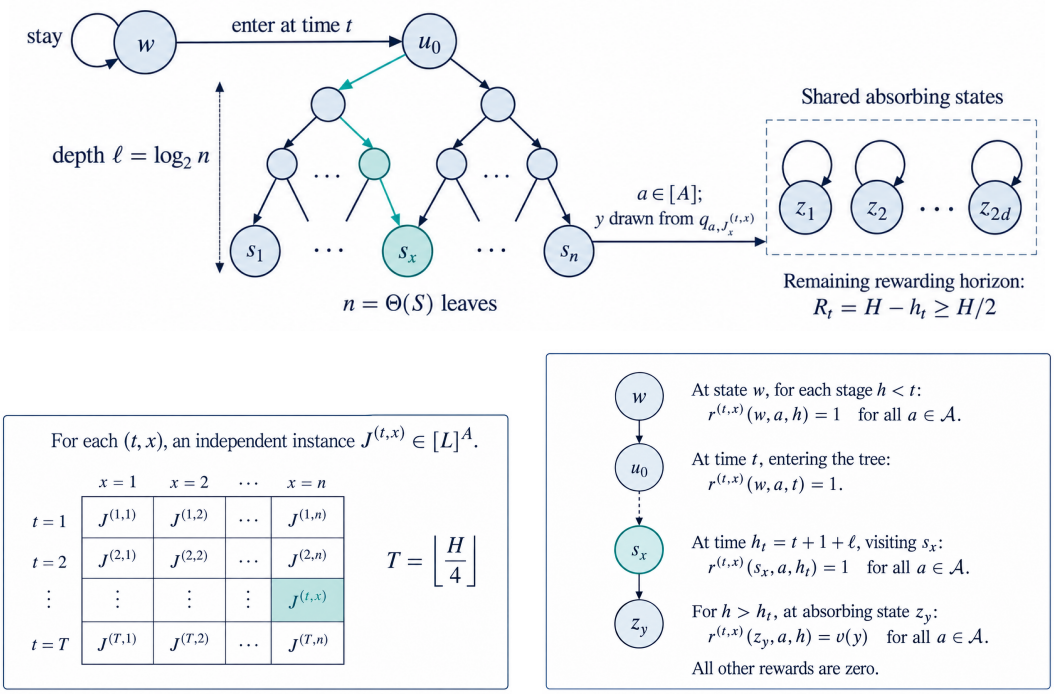


Figure 2: The lower-bound construction. **(a)** A waiting state and a deterministic binary tree lead to shared absorbing states. At stage $h_t = t + 1 + \ell$, action a at leaf s_x has transition distribution $q_{a,J_a^{(t,x)}}$. **(b)** Each time–leaf pair (t, x) has its own hidden tuple $J^{(t,x)} \in [L]^A$; different times share the same physical states. **(c)** The publicly known reward $r_{t,x,\nu}$ localizes the policy at the designated pair before evaluating the absorbing-state reward ν. The entry reward at stage t applies only to actions $a \neq 1$, whereas action 1 stays at w.

Proof. Without loss of generality, we assume $\sup_{P\in\mathcal{P}_L} \mathbb{E}_P[K]$ is finite. Set $\alpha = \frac{96\epsilon}{H}$, so $H\alpha \leq \frac{1}{2}$.

Step 1: the construction of leaf nodes. Fix $\gamma = 1/10$ and $d = \lceil 1000(\log(AL) + 1)\rceil$. The balanced packing lemma (Jin et al., 2020, Lemma D.6) provides vectors $v_{a,j} \in \{-1, 1\}^{2d}$, indexed by $(a, j) \in [A] \times [L]$, such that

$$\langle v_{a,j}, \mathbf{1}\rangle = 0, \qquad |\langle v_{a,j}, v_{a',j'}\rangle| < 2\gamma d \quad \text{whenever } (a, j) \neq (a', j'). \tag{50}$$

Our choice of d satisfies that lemma's condition $2\log L \leq \gamma^2 d - \log(4d) - 2\log A$. Taking c_s sufficiently small also ensures $d \leq S/4$. Define

$$q_0 = \frac{\mathbf{1}}{2d}, \qquad q_{a,j} = \frac{\mathbf{1} + \alpha v_{a,j}}{2d}, \tag{51}$$

and the finite collection of decoder rewards

$$\mathcal{V}_L = \left\{ \nu_{a,a',j,j'} := \frac{1}{2}\mathbf{1} + \frac{1}{3}v_{a,j} + \frac{1}{6}v_{a',j'} : a \neq a',\ j, j' \in [L] \right\}. \tag{52}$$

Then $q_{a,j} \in \Delta(2d)$, $\mathcal{V}_L \subset [0,1]^{2d}$, and $|\mathcal{V}_L| \leq A(A-1)L^2$. For a hidden tuple $J \in [L]^A$, action a produces one observation with distribution q_{a,J_a}.

The decoding argument (Jin et al., 2020, proof of Lemma D.8 and Lemma D.11) has the following property: if distributions $\{\lambda_\nu \in \Delta(A) : \nu \in \mathcal{V}_L\}$ satisfy

$$\max_a q_{a,J_a}^\top \nu - \sum_a \lambda_\nu(a) q_{a,J_a}^\top \nu \leq \frac{\alpha}{24} \quad \text{for every } \nu \in \mathcal{V}_L, \tag{53}$$

then they determine J exactly. More precisely, J_a is the unique $j \in [L]$ such that

$$\lambda_{\nu_{a,a',j,j'}}(a) > \frac{1}{2} \quad \text{for every } a' \neq a \text{ and } j' \in [L]. \tag{54}$$

Step 2: embedding at time–leaf pairs. Let n be the largest power of two not exceeding $S/4$, and write $\ell = \log_2 n$ and $T = \lfloor H/4 \rfloor$. Construct an initial waiting state w, a complete binary tree with root u_0 and leaves $s_1, \ldots, s_n$, and shared absorbing states $z_1, \ldots, z_{2d}$.

At w, action 1 stays and every other action enters u_0. At each internal tree state, action 1 moves to the left child and every other action moves to the right child. These transitions are deterministic and known. Thus, $A \geq 2$ suffices, while all A actions are available at every leaf.

For each $(t, x) \in [T] \times [n]$, assign a hidden tuple $J^{(t,x)} \in [L]^A$. A trajectory entering the tree at stage t reaches a leaf at stage $h_t = t + 1 + \ell$. At this stage, define

$$P_{h_t}(z_y \mid s_x, a) = q_{a,J_a^{(t,x)}}(y), \qquad a \in [A], \quad y \in [2d]. \tag{55}$$

At every leaf stage outside $\{h_t : t \in [T]\}$, the transition distribution is q_0. Each z_y self-loops under every action. This defines $\mathcal{P}_L$ as the hidden tuples vary. Figure 2(a)–(b) illustrates the shared physical MDP and its time–leaf indexing.

We use a separate waiting state in the tree embedding (Ridel & Cohen, 2026, Appendix C.3), so that two actions suffice for both waiting and routing. Its state count is $1 + (2n-1) + 2d = 2n + 2d \leq S$. Moreover, $n = \Theta(S)$, $T = \Theta(H)$, and the horizon condition gives $R_t := H - h_t \geq H/2$. Each episode contains at most one informative transition: after reaching a leaf, the process is absorbed. The time indices introduce no additional physical states.

Step 3: fixed rewards and localization. For each $(t, x, \nu) \in [T] \times [n] \times \mathcal{V}_L$, define the reward illustrated in Figure 2(c):

$$r_{t,x,\nu}(s,a,h) = \begin{cases} 1, & s = w,\ h < t, \\ 1, & s = w,\ h = t,\ a \neq 1, \\ 1, & s = s_x,\ h = h_t, \\ \nu(y), & s = z_y,\ h > h_t, \\ 0, & \text{otherwise.} \end{cases} \tag{56}$$

Let $\mathcal{R}_L$ contain all these rewards. They depend only on the fixed packing and known routing, not on the hidden tuples. Consequently, the whole class may be revealed before exploration, and

$$|\mathcal{R}_L| \leq TnA(A-1)L^2 \leq SHA^2L^2. \tag{57}$$

Fix $r = r_{t,x,\nu}$ and let $G = \{s_{h_t} = s_x\}$. On G, the non-absorbing rewards sum to exactly $t+1$; on G^c, they sum to at most t. Also, balance of the packing and $\|q_{a,j} - q_0\|_1 = \alpha$ imply

$$q_0^\top \nu = \frac{1}{2}, \qquad \left| q_{a,j}^\top \nu - \frac{1}{2} \right| \leq \frac{\alpha}{2}. \tag{58}$$

The event G is determined by the waiting and routing decisions, before any random leaf transition. Thus, conditioning on G^c does not select favorable absorbing-state outcomes. Since absorbing rewards are zero through stage h_t, even an early exit gives at most R_t rewarded absorbing stages.

It follows that a policy reaching the target pair and choosing its best action has value at least $t+1+\frac{R_t(1-\alpha)}{2}$, whereas the conditional value on G^c is at most $t+\frac{R_t(1+\alpha)}{2}$. Their difference is at least

$$1 - R_t\alpha \geq 1 - H\alpha \geq \frac{1}{2}. \tag{59}$$

In particular, an optimal policy reaches the designated pair with probability one, and

$$V^\star_{P,r} = t + 1 + R_t \max_a q^\top_{a,J_a^{(t,x)}}\nu.$$

For any policy π, put $p = \mathbb{P}^\pi_P(G)$ and let λ be its conditional action distribution at (s_x, h_t), arbitrary when $p = 0$. Define

$$\Delta_\nu(\lambda) = \max_a q^\top_{a,J_a^{(t,x)}}\nu - \sum_a \lambda(a) q^\top_{a,J_a^{(t,x)}}\nu.$$

The conditional expected value on G is $V^\star_{P,r} - R_t\Delta_\nu(\lambda)$; on G^c, it is at most $V^\star_{P,r} - 1/2$. Therefore,

$$V^\star_{P,r} - V^\pi_{P,r} \geq \frac{1-p}{2} + pR_t\Delta_\nu(\lambda). \tag{60}$$

If π is ϵ-optimal, then

$$p \geq 1 - 2\epsilon \geq \frac{1}{2}, \qquad \Delta_\nu(\lambda) \leq \frac{\epsilon}{pR_t} \leq \frac{4\epsilon}{H} = \frac{\alpha}{24}. \tag{61}$$

For Markov policies, $\lambda = \pi_{h_t}(\cdot \mid s_x)$; more generally, it is determined by the output policy and the known deterministic prefix. No unknown transition is needed to compute it. Applying equation 54 separately at each pair shows that simultaneous ϵ-optimality for $\mathcal{R}_L$ permits exact recovery of the entire hidden-index array.

Step 4: Putting it all together. Let $\Theta = [L]^{TnA}$. Each θ specifies the transition distributions for all TnA state–action pairs associated with the Tn time-expanded leaf nodes. Let $\mathbb{P}_\theta$ denote this transition kernel. As a comparison, let $\mathbb{P}_0$ be the reference transition kernel when every hard transition is replaced by q_0, with the same known reward class. Choose $\mathcal{P}_L = \{\mathbb{P}_\theta | \theta \in \Theta\} \cup \{\mathbb{P}_0\}$.

Let $N_{t,x,a}$ count visits to the associated hard transition before the algorithm stops. Let $\mathbb{Q}_\theta$ and $\mathbb{Q}_0$ denote the probability laws of the algorithm's stopped interaction transcript under transition kernels P_θ and P_0, respectively, with the same publicly known reward class. The transcript includes the stopping time and the algorithm's output. The adaptive KL chain rule, applied to the stopped transcript (Jin et al., 2020, proof of Lemma D.7), gives

$$\begin{aligned} D_{\mathrm{KL}}(\mathbb{Q}_\theta \,\|\, \mathbb{Q}_0) &\leq \sum_{t,x,a} \mathbb{E}_\theta[N_{t,x,a}] D_{\mathrm{KL}}\big(q_{a,\theta_{t,x,a}} \| q_0\big) \\ &\leq \alpha^2 \mathbb{E}_\theta\left[\sum_{t,x,a} N_{t,x,a}\right] \leq \alpha^2 \mathbb{E}_\theta[K]. \end{aligned} \tag{62}$$

Here $D_{\mathrm{KL}}(q_{a,j}\|q_0) \leq \alpha^2$ follows from $\log(1+u) \leq u$ and equation 50.

Denote the uniform distribution on Θ as $\mathrm{Uni}(\Theta)$ and choose θ following $\mathrm{Uni}(\Theta)$. Recall that the algorithm provides an estimator $\widehat{\theta}$ with error probability at most $1/2$ under every θ. By (Jin et al., 2020, Lemma D.10) and equation 62, we have that

$$\begin{aligned} \alpha^2 \mathbb{E}_{\theta\sim\mathrm{Uni}(\Theta)}\mathbb{E}_\theta[K] \geq& \frac{1}{|\Theta|}\sum_{\theta\in\Theta} D_{\mathrm{KL}}(\mathbb{Q}_\theta||\mathbb{Q}_0) \geq \frac{1}{2}\log|\Theta| - \log 2 \\ =& \frac{1}{2}TnA\log L - \log 2 \geq \frac{1}{4}TnA\log L. \end{aligned}$$

Consequently,

$$\sup_{P\in\mathcal{P}_L} \mathbb{E}_P[K] \geq \mathbb{E}_{\theta\sim\mathrm{Uni}(\Theta)}\mathbb{E}_\theta[K] \geq c\frac{TnA\log L}{\alpha^2} \geq c'\frac{SAH^3\log L}{\epsilon^2},$$

using $n = \Theta(S)$, $T = \Theta(H)$, and $\alpha = 96\epsilon/H$. □

Corollary 1 (Dependence on the number of rewards)**.** *For a sufficiently large universal constant C, suppose that $A \geq 2$, $S \geq C(1+\log A)$, $H \geq 4(1+\lceil \log_2 S \rceil)$, $0 < \epsilon \leq \frac{1}{192}$ and $M \geq (CSHA^2)^2$. There exists a fixed, publicly known class of exactly M reward functions for which any algorithm achieving simultaneous ϵ-optimality with probability at least $1/2$ requires*

$$\Omega\left(\frac{SAH^3}{\epsilon^2}\min\{S, \log M\}\right)$$

expected online episodes in the worst case. In particular, if $\log M \leq S$, the lower bound becomes

$$\Omega\left(\frac{SAH^3 \log M}{\epsilon^2}\right).$$

Proof. Choose

$$L = \left\lfloor \min\left\{\sqrt{\frac{M}{SA^2H}}, \exp\left(\frac{c_s S}{2}\right)\right\}\right\rfloor.$$

For sufficiently large C, the assumptions imply $L \geq 16$ and $\log A + 1 \leq c_s S/2$, so that $\log(AL) + 1 \leq c_s S$. Moreover, $|\mathcal{R}_L| \leq SHA^2L^2 \leq M$, and

$$\log L = \Omega\left(\min\left\{S, \log\frac{M}{SA^2H}\right\}\right) = \Omega(\min\{S, \log M\}),$$

where the last equality uses $M \geq (SA^2H)^2$. The conclusion follows by Proposition 1.

□