ALMOST TIGHT REGRET BOUNDS FOR REINFORCEMENT LEARNING WITH ABSTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study reinforcement learning (RL) with abstention, where an agent in a finite-horizon Markov decision process (MDP) may terminate an episode at any time and receive a fallback reward. We formulate this setting as an augmented MDP and propose an optimism-based algorithm that incorporates abstention into both value iteration and policy execution. We establish an interaction-dependent regret upper bound whose leading term depends on the realized number of interactions $N_{\mathrm{int}}(K)$ rather than the full decision budget HK, yielding a potentially sharper upper bound when abstention reduces interactions. To further characterize the MDPs in which such improvements occur, we introduce the optimal effective horizon length $d^\star$ and the minimum sub-optimality gap $\gamma_{\min}$ over base actions. Our analysis shows that a shorter effective horizon can reduce regret, whereas a smaller gap increases the exploration cost. These effects are further illustrated through three concrete examples, including the boundary cases in which abstention is never beneficial or is uniformly optimal, as well as an intermediate setting in which the benefit of abstention varies with the problem parameters. Finally, complementary regret lower bounds show that dependence on $d^\star$ and $\gamma_{\min}$ is unavoidable in general, supporting their role in characterizing the intrinsic learning difficulty of RL with abstention.

1 INTRODUCTION

Abstention is a fundamental mechanism for reliable decision-making under uncertainty. Instead of always committing to a prediction or an action, an agent may defer, refuse, or stop acting when its decision is likely to be unreliable. This idea has a long history in supervised learning, where it is also known as the reject option or selective prediction (Chow, 1970; Bartlett & Wegkamp, 2008; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; Kalai & Kanade, 2021; Fuchs et al., 2024).

More recently, abstention has been studied in online learning and bandit settings, where an agent may abstain from a prediction at a cost (Cortes et al., 2018), decline to pull an arm, or choose a fallback decision (Yang et al., 2024; Pasteris et al., 2024), and defer treatment recommendations when available information is insufficient (Sawarni et al., 2025).

However, these existing theoretical studies of abstention primarily focus on one-step decisions. This leaves open how abstention should operate within multi-step trajectories, where actions affect subsequent states and an agent needs to stop midway. For example, a robot navigating an unknown environment may encounter unexpected obstacles or unreliable sensor readings that make further actions unsafe. It then needs to terminate the task and invoke a fallback option (Garcıa & Fernández, 2015; Alshiekh et al., 2018; Saunders et al., 2017). Similar considerations arise in other sequential decision-making settings, including large language model reasoning (Wen et al., 2025; Madhusudhan et al., 2025; Xu et al., 2026; Davidov et al., 2026).

Motivated by these applications, we study a framework covering abstention in multi-step decision: *episodic Reinforcement Learning (RL) with abstention*. In this setting, we augment a Markov Decision Process (MDP) with a terminal abstention action, allowing the agent to end an episode at any step and receive a fallback reward. The agent must therefore learn both how to act and when to stop. This introduces a new trade-off between abstention and exploration: abstention can avoid low-value interactions, but abstaining too early limits the exploration, making it hard to learn whether further interaction is worthwhile. This trade-off motivates the following question:

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.

Table 1: Comparison of regret bounds for tabular RL with and without abstention. Here, S and A denote the numbers of states and base actions, H is the horizon, and K is the number of episodes. $N_{\mathrm{int}}(K)$ counts realized interactions, $d^\star$ is the largest expected interaction length per episode among optimal policies, and $\gamma_{\min}$ is the minimum sub-optimality gap over base actions. Logarithmic factors and the additive S^2AH^3 terms in the upper bounds are omitted. Upper bounds hold with high probability, whereas lower bounds hold with constant probability.

Method	Abstention	Regret type	Regret bound	Reference
UCBVI-CH	No	Worst-case	$\widetilde{O}(H^{3/2}\sqrt{SAHK})$[1]	(Azar et al., 2017)
UCBVI-AG	Yes	Interaction-dependent	$\widetilde{O}(H^{3/2}\sqrt{SA\,N_{\mathrm{int}}(K)})$	Theorem 1
		EHL & MSG-dependent	$\widetilde{O}(\min\{H^{3/2}\sqrt{SAKd^\star} + SAH^4\gamma_{\min}^{-1}, H^{3/2}\sqrt{SAHK}\})$	Theorem 2
Lower Bound	Yes	MSG-dependent	$\widetilde{\Omega}(SAH^3\gamma_{\min}^{-1})$	Theorem 4
		EHL-dependent	$\widetilde{\Omega}(H\sqrt{SAKd^\star})$	Theorem 5

How and when does abstention benefit reinforcement learning?

To answer this question, we introduce UCBVI-AG (Algorithm 1), an optimism-based algorithm that abstains whenever the optimistic value of continued interaction does not exceed the fallback reward. Although abstention can be naturally incorporated into standard optimistic planning in this way, standard full-horizon regret bounds depend on the total decision budget HK and do not explicitly capture the interaction savings enabled by abstention. We therefore derive an interaction-dependent regret bound and relate the realized interaction count to structural properties of the MDP, characterizing when abstention yields sharper guarantees. We summarize our contributions below and compare the regret bounds in Table 1.

- **Interaction-dependent Upper Bound (Theorem 1).** We establish a regret upper bound for UCBVI-AG whose leading term depends on the realized number of real interactions $N_{\mathrm{int}}(K)$ rather than the full horizon budget HK, yielding a potentially sharper regret upper bound when abstention reduces interactions.
- **Instance-dependent Upper Bound (Theorem 2).** To characterize the MDPs in which abstention can yield potential benefits, we introduce two structural quantities: the optimal effective horizon length (EHL) $d^\star$, defined as the largest expected interaction length among optimal policies, and the minimum sub-optimality gap (MSG) $\gamma_{\min}$ over base actions. The resulting instance-dependent regret bound shows how a shorter EHL can reduce regret, while a smaller MSG increases the exploration cost.
- **Illustrative Examples (Section 6).** We provide concrete examples to illustrate our theoretical results, including boundary cases where abstention is never beneficial or uniformly optimal, and an intermediate setting in which the benefit of abstention varies with the problem parameters
- **Lower Bounds (Section 7).** We establish complementary lower bounds showing that the corresponding upper-bound are tight in the dependence on S, A, $d^\star$ and $\gamma_{\min}$, up to logarithmic factors. These results show that dependence on $d^\star$ and $\gamma_{\min}$ is unavoidable in general, and support the key role of both quantities in characterizing the intrinsic learning difficulty of RL with abstention.

2 Related Works

Machine Learning and Multi-Armed Bandits with abstention. Learning with abstention, also known as learning with a reject option or selective prediction, has been widely studied in supervised learning. In this setting, a predictor may abstain from making a prediction when its confidence is low, thereby trading off prediction accuracy and coverage (Chow, 1970; Bartlett & Wegkamp, 2008; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; Zaoui et al., 2020; Kalai & Kanade, 2021). Related abstention mechanisms have also been studied in ranking (Cheng et al., 2010; Mao et al., 2023). More recently, abstention has been investigated in online learning and bandit settings.

[1]For fair comparison, we adapt UCBVI-CH's time-homogeneous bound to our time-inhomogeneous setting, incurring an *additional* $\sqrt{H}$ factor. Such a conversion is commonly used in RL literature (Jin et al., 2018)

Cortes et al. (2018) studied online learning with an abstention option, where the agent can refuse to predict at a cost, and subsequent works consider related abstention mechanisms in multi-armed bandits and expert-advice problems (Yang et al., 2024; Pasteris et al., 2024). In addition, Sawarni et al. (2025) studied abstention in static policy learning from observational data, where a learned treatment policy may defer on uncertain covariates. These works demonstrate that abstention can improve robustness by avoiding uncertain or low-confidence decisions. However, they mostly focus on one-step prediction, and therefore do not address the key challenge in reinforcement learning: learning an unknown transition kernel over multi-step trajectories. While Davidov et al. (2026) analyzed value-based abstention in multi-step large language model reasoning, their theoretical guarantees assume access to the true value function, whereas we learn under unknown transition dynamics and explicitly account for estimation uncertainty and exploration in our regret analysis.

Regret analysis for tabular RL. Tabular reinforcement learning studies sequential decision-making problems in which both the state space and the action space are finite. A central algorithmic principle in this literature is optimism in the face of uncertainty, which dates back to UCRL-type methods (Auer & Ortner, 2006; Auer et al., 2008) and has become a standard tool for episodic finite-horizon RL. For tabular episodic MDPs, algorithms such as UCBVI construct optimistic value estimates from empirical transition kernels and confidence bonuses, leading to minimax-optimal regret bounds (Azar et al., 2017). Subsequent works developed sharper analyses and alternative optimistic algorithms, including model-based and model-free methods, Bernstein-type bonuses, and refined concentration arguments (Dann et al., 2017; Jin et al., 2018; Zanette & Brunskill, 2019; Zhang et al., 2020; Ménard et al., 2021; Agrawal & Jia, 2017; Kearns & Singh, 2002; Strehl et al., 2006). Beyond worst-case guarantees, another line of work studies *instance-dependent or gap-dependent* regret bounds, showing that the regret can be substantially smaller on benign MDP instances where suboptimal actions can be statistically separated more quickly (Simchowitz & Jamieson, 2019; Zanette & Brunskill, 2018; Xu et al., 2021).

3 PRELIMINARIES

Notations. We use $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ to suppress logarithmic factors in S, A, H, K, and $1/\delta$. Specifically, for nonnegative functions f and g, we write $f = \widetilde{O}(g)$ if $f \le Cg\log^c(SAHK/\delta)$, and $f = \widetilde{\Omega}(g)$ if $f \ge Cg/\log^c(SAHK/\delta)$, for some constants $C > 0$, $c \ge 0$, and sufficiently large S, A, H, K, and $1/\delta$. For any $H \in \mathbb{N}$, we denote $[H] = \{1, \dots, H\}$. In addition, we omit the index set $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$ when it is clear from context.

Standard MDP. We first introduce a standard episodic finite-horizon MDP, denoted by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P, r)$. Here, $\mathcal{S}$ and $\mathcal{A}$ are the finite state and action spaces with cardinalities S and A, respectively, and H is the horizon of each episode. The transition kernel is given by $P = \{P_h\}_{h=1}^H$, where each P_h is time-dependent. For any state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, $P_h(\cdot \mid s, a)$ is a probability simplex on $\mathcal{S}$ and $P_h(s' \mid s, a)$ denotes the probability of transitioning to state s' at the next step after taking action a in state s at step h. The reward function is given by $r = \{r_h\}_{h=1}^H$, where each $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a deterministic, time-dependent function.

Unlike the standard episodic MDP, where the agent must interact with the environment for all H steps once an episode starts, many real-world decision-making scenarios allow the agent to stop early and receive a fallback reward. This motivates our introduction of an MDP with abstention.

MDP with Abstention. An episodic finite-horizon MDP with abstention is denoted by the tuple $\mathcal{M}^+ = (\mathcal{S}, \mathcal{A}^+, H, P, r, G)$. Here, $\mathcal{S}$, H, and P have the same definition as standard MDPs. However, here $\mathcal{A}^+ = \mathcal{A} \cup \{\widetilde{a}\}$ denotes an augmented action set with the *base action set* $\mathcal{A}$ and *abstention action* $\widetilde{a}$. Moreover, $G = \{G_h\}_{h=1}^H$, where each $G_h : \mathcal{S} \to [0, H - h + 1]$ is a **bounded** fallback value function. For simplicity, we assume that both the reward function r and the fallback values G are known to the agent. This is a standard simplification in RL theory, as the main statistical difficulty lies in learning the unknown transition kernel (Lattimore & Hutter, 2014; Azar et al., 2017).

An agent interacts with the environment as follows. Each episode starts from the same fixed initial state $s_1 \in \mathcal{S}$. Then, at each step $h \in [H]$, if the agent has not abstained at the previous step $h - 1$, she observes the current state s_h and chooses an action from the augmented action space $\mathcal{A}^+$. If the agent

chooses an action a_h from the base action set $\mathcal{A}$, then the episode continues, and the agent receives a reward $r_h(s_h, a_h)$. The next state s_{h+1} is sampled according to the transition kernel $P(\cdot|s_h, a_h)$. If the agent chooses abstention $\widetilde{a}$, then the episode ends immediately, and the agent receives a total fallback reward $G_h(s)$.

We only consider a deterministic policy defined as a sequence of mappings $\pi = \{\pi_h\}_{h=1}^H$, with each $\pi_h : \mathcal{S} \to \mathcal{A}^+$. For any policy π, suppose $V_{H+1}^\pi = 0$, then for time-state-action triples $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$, the value function with abstention $V_h^\pi : \mathcal{S} \to [0, H]$, the value function with base action set $U_h^\pi : \mathcal{S} \to [0, H]$, and the action value function with base action set $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to [0, H]$ are defined recursively as follows:

$$\begin{aligned} Q_h^\pi(s, a) &:= r_h(s, a) + (P_h V_{h+1}^\pi)(s, a), \\ U_h^\pi(s) &:= Q_h^\pi(s, \pi_h(s)), \text{ if } \pi_h(s) \in \mathcal{A}, \\ V_h^\pi(s) &:= \mathbb{1}_{\{\pi_h(s) \in \mathcal{A}\}} U_h^\pi(s) + \mathbb{1}_{\{\pi_h(s) = \widetilde{a}\}} G_h(s), \end{aligned}$$

where for any function $V : \mathcal{S} \to [0, H]$, we define $P_h V(s, a) := \sum_{s' \in \mathcal{S}} P_h(s'|s, a) V(s')$. Denote the optimal policy with abstention by $\pi^\star = \arg\max_\pi V_1^\pi(s_1)$, and the optimal value function with abstention by $V_1^\star(s_1) = V_1^{\pi^\star}(s_1)$. Suppose $V_{H+1}^\star = 0$, then for time-state-action triples $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$, the optimal value function with abstention: $V_h^\star$, the optimal value function with only base action set: $U_h^\star$, and the optimal action value function with only base action set: $Q_h^\star$ can be expressed as follows

$$\begin{aligned} Q_h^*(s, a) &:= r_h(s, a) + (P_h V_{h+1}^\star)(s, a), \\ U_h^\star(s) &:= \max_{a \in \mathcal{A}} Q_h^*(s, a), \\ V_h^\star(s) &:= \max\{U_h^\star(s), G_h(s)\}. \end{aligned}$$

Regret. Suppose the agent interacts with the environment for a total of K episodes. For each $k \in [K]$, let π^k denote the policy actually chosen at episode k by the agent in the augmented action set $\mathcal{A}^+$. Then the regret against the optimal policy is defined as

$$\mathrm{Reg}(K) := \sum_{k=1}^K \Big(V_1^\star(s_1) - V_1^{\pi^k}(s_1) \Big). \tag{1}$$

The objective of the agent during its interaction with the environment is to minimize the regret.

4 Algorithm

In this section, we present UCBVI-AG (Algorithm 1), an optimism-based algorithm that incorporates abstention into both **backward value iteration** and **policy execution**.

At the beginning of each episode k, the agent estimates the transition kernel from $\mathcal{H}$, which contains only transitions collected through real interactions before episode k.

Backward Value Iteration with Fallback reward (Lines 4–12). For each time-state-action triple $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$, the agent computes an optimistic action-value estimate $Q_h^k(s, a)$ using the empirical transition kernel and an exploration bonus. It then sets $U_h^k(s) = \max_{a \in \mathcal{A}} Q_h^k(s, a)$ and incorporates the fallback reward into the optimistic value function through $V_h^k(s) = \max\{U_h^k(s), G_h(s)\}$, accounting for both continued interaction and immediate abstention.

Policy Execution with Abstention Gate (Lines 13–20). During policy execution, UCBVI-AG is supposed to decide whether to *abstain* or *continue interacting* with the environment. At each visited state s_h^k, the agent abstains, receives the fallback reward, and terminates the episode if $G_h(s_h^k) \geq U_h^k(s_h^k)$. On the optimism event, this condition ensures that abstention is an optimal action. Otherwise, the agent executes $a_h^k \in \arg\max_{a \in \mathcal{A}} Q_h^k(s_h^k, a)$ and adds the observed transition to $\mathcal{H}$.

5 Regret Upper Bounds

In this section, we establish theoretical regret upper bounds for UCBVI-AG, as stated in Theorems 1 and 2. Since only actual interactions with the environment generate transition samples, we first define the total number of environment interactions up to episode K.

Algorithm 1 **U**pper **C**onfidence **B**ound **V**alue **I**teration with **A**bstention **G**ate (UCBVI-AG)

Input: Fallback reward G; confidence level $\delta \in (0,1)$; fixed episode budget $K \geq 1$; $L_1 = \log(6HSAK/\delta)$; fixed bonus coefficient $c \geq 1$; real-interaction dataset $\mathcal{H} = \varnothing$.
for $k = 1, 2, \ldots, K$ **do**
  From $\mathcal{H}$, for all $h \in [H]$ and $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, compute,

$$N_h^k(s,a,s') = \sum\nolimits_{(\bar{s},\bar{a},\bar{s}',\bar{h})\in\mathcal{H}} \mathbb{1}\{\bar{h}=h, \bar{s}=s, \bar{a}=a, \bar{s}'=s'\}.$$

$$N_h^k(s,a) = \sum\nolimits_{s'\in\mathcal{S}} N_h^k(s,a,s'), \quad \widehat{P}_h^k(s' \mid s,a) = \frac{N_h^k(s,a,s')}{N_h^k(s,a) \vee 1}.$$

  Set $V_{H+1}^k(s) = 0$ for all $s \in \mathcal{S}$.
  for $h = H, H-1, \ldots, 1$ **do**
    for each $s \in \mathcal{S}$, $a \in \mathcal{A}$ **do**
      $b_h^k(s,a) \leftarrow cH\sqrt{L_1/(N_h^k(s,a) \vee 1)}$.
      $Q_h^k(s,a) \leftarrow \min\{H-h+1,\ r_h(s,a) + (\widehat{P}_h^k V_{h+1}^k)(s,a) + b_h^k(s,a)\}$.
    end for
    $U_h^k(s) \leftarrow \max_{a\in\mathcal{A}} Q_h^k(s,a)$ for all $s \in \mathcal{S}$.
    $V_h^k(s) \leftarrow \max\{U_h^k(s), G_h(s)\}$ for all $s \in \mathcal{S}$. `//Abstention-augmented Value`
  end for
  for $h = 1, 2, \ldots, H$ **do**
    if $G_h(s_h^k) \geq U_h^k(s_h^k)$ **then** `//Abstention Gate`
      Choose abstention action $\widetilde{a}$, receive fallback reward $G_h(s_h^k)$. **Break**
    else
      Execute $a_h^k \in \arg\max_{a\in\mathcal{A}} Q_h^k(s_h^k, a)$ in the environment and observe s_{h+1}^k.
      Update the data set by

$$\mathcal{H} \leftarrow \mathcal{H} \cup \{(s_h^k, a_h^k, s_{h+1}^k, h)\}.$$

    end if
  end for
end for

Definition 1 (Number of Interactions)**.** Suppose Algorithm 1 has total K episodes. We say that the agent *interacts with* the environment at time step h if she does not abstain and chooses some base action $a \in \mathcal{A}$. The total number of real interactions by time K is defined

$$N_{\text{int}}(K) := \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{1}\{\text{the agent interacts with environment at time step } h \text{ and episode } k\}.$$

Interaction-dependent Upper Bound. We first provide an interaction-dependent regret upper bound for UCBVI-AG as follows. The detailed proof is provided in Appendix B.

Theorem 1 (Interaction-dependent Regret Upper Bound)**.** *For any fixed $\delta \in (0,1)$, let $L_1 = \log(6HSAK/\delta)$, and $c = 1$ be absolute constants in UCBVI-AG. Define $L_2 := \log(12S^2AHK/\delta)$. With probability at least $1-\delta$, the regret of UCBVI-AG can be upper bounded as*

$$\text{Reg}(K) \leq 10(H^{3/2}\sqrt{SA\,N_{\text{int}}(K)L_1} + S^2AH^3L_2\log(K+1)). \tag{2}$$

The resulting bound is $\widetilde{O}(H^{3/2}\sqrt{SA\,N_{\text{int}}(K)} + S^2AH^3)$. For comparison, the standard UCBVI-CH analysis yields an upper bound $\widetilde{O}(H^{3/2}\sqrt{SAHK} + S^2AH^3)$. Thus, our bound replaces the total number of decision steps HK in the first term with the realized number of real interactions $N_{\text{int}}(K)$, yielding a potentially sharper guarantee when abstention substantially reduces interactions. Intuitively, this improvement arises because, in UCBVI-AG, the agent abstains whenever the best optimistic base value function cannot exceed the cumulative fallback reward. In such cases, uninformative or sub-optimal interaction steps are skipped, and regret is accumulated only on steps where the agent actually interacts with the environment.

Remark 1 (Comparison with the minimax lower bound)**.** *(Domingues et al., 2021, Theorem 9) establish an expected regret lower bound of $\Omega(\sqrt{H^3SAK})$ for standard finite-horizon MDPs with*

time-dependent transitions, characterizing the intrinsic difficulty of learning without abstention. Our interaction-dependent upper bound does not have such dependence on HK, highlighting how abstention can reduce this difficulty by allowing the learner to avoid unnecessary interactions.

Moreover, this lower bound extends to our setting with $G_h(s) = 0$, since abstention yields zero payoff and cannot improve upon continuing with nonnegative base rewards in this case. In the full-interaction regime $N_{\mathrm{int}}(K) = HK$, the bound becomes $\Omega(H\sqrt{SA\,N_{\mathrm{int}}(K)})$, showing that the N_{int} dependence is unavoidable. The first term of our upper bound is larger by a factor of $\sqrt{H}$, up to logarithmic factors, reflecting a limitation of the Hoeffding-based analysis. A Bernstein-based analysis may improve this horizon dependence; we leave this refinement to future work.

Instance-dependent Upper Bound. Theorem 1 provides a potentially sharper regret upper bound for Algorithm 1 (UCBVI-AG). The improvement, however, depends on how much smaller the number of real interactions $N_{\mathrm{int}}(K)$ is than the total number of decision steps HK. Since $N_{\mathrm{int}}(K)$ is random and not known a priori, the above bound does not immediately reveal when abstention leads to a substantial improvement. We therefore relate it to intrinsic properties of the augmented MDP.

To this end, we first introduce the *effective horizon length of optimal policies*, which measures the largest expected number of real interactions per episode among optimal policies.

Definition 2 (**E**ffective **H**orizon **L**ength (**EHL**))**.** Let $\Pi^\star$ denote the set of deterministic Markov policies that select an optimal augmented action at every state-time pair. For a fixed MDP $\mathcal{M}$, define

$$d^\star := \max_{\pi^\star \in \Pi^\star} \mathbb{E}_{\mathcal{M},\pi^\star}[\tau_{\pi^\star} - 1],$$

where τ_π is the first abstention time under π, with $\tau_\pi = H + 1$ if no abstention occurs.

The quantity $d^\star$ measures the largest expected number of interactions per episode among optimal policies and provides an algorithm-independent reference value for the agent's per-episode interaction count. The lower bound in Theorem 5 further demonstrates its relevance to intrinsic learning difficulty.

During learning, however, suboptimal base actions may lead to additional interactions beyond this baseline. To control these excess interactions, we introduce the minimum positive sub-optimality gap, a standard quantity in gap-dependent RL (Simchowitz & Jamieson, 2019; Yang et al., 2021).

Definition 3 (**M**inimum **S**ub-optimality **G**ap (**MSG**))**.** We define the minimum positive sub-optimality gap over base actions as

$$\gamma_{\min} := \min_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}: Q_h^\star(s,a) < V_h^\star(s)} \big(V_h^\star(s) - Q_h^\star(s,a)\big).$$

If the indexing set is empty, we set $\gamma_{\min} = +\infty$ and adopt $1/\gamma_{\min} = 0$.

The quantity $\gamma_{\min}$ captures an intrinsic aspect of learning difficulty: distinguishing nearly optimal base actions from optimal behavior. A smaller gap may require more accurate estimation of the unknown transition kernel and thus more exploration, potentially increasing cumulative regret. The lower bound in Theorem 4 substantiates this interpretation, showing that an inverse dependence on $\gamma_{\min}$ in regret is unavoidable in general.

Together, EHL and MSG connect the agent's interaction count to intrinsic properties of the augmented MDP. The following proposition bounds $N_{\mathrm{int}}(K)$ by the optimal interaction baseline $Kd^\star$, plus gap-dependent terms that grow only polylogarithmically with K. The proof is provided in Appendix C.1.

Proposition 1 (Bound of Interaction)**.** *Fix $K \geq 1$ and $\delta \in (0,1)$. Run UCBVI-AG (Algorithm 1) with confidence parameter $\delta/2$ and bonus coefficient $c = 1$. Then, with probability at least $1-\delta$,*

$$N_{\mathrm{int}}(K) \leq \min\Big\{4Kd^\star + \frac{42S^2AH^4L_2\log(K+1)}{\gamma_{\min}} + \frac{400SAH^5L_1}{\gamma_{\min}^2}, HK\Big\}.$$

Combining Proposition 1 with the interaction-dependent regret bound yields a guarantee in terms of the optimal effective horizon length and the minimum positive sub-optimality gap as follows.

Theorem 2 (Instance-dependent Regret Upper Bound)**.** *Under the same algorithm settings as Proposition 1, with probability at least $1-\delta$,*

$$\mathrm{Reg}(K) = \widetilde{O}(\min\{H^{3/2}\sqrt{SAKd^\star} + SAH^4{\gamma_{\min}}^{-1}, H^{3/2}\sqrt{SAHK}\} + S^2AH^3).$$

Theorem 2 characterizes how the benefit of abstention depends on the optimal effective horizon length $d^\star$ and the minimum sub-optimality gap $\gamma_{\min}$, while ensuring a regret bound no worse in order than the standard case $\widetilde{O}(H^2\sqrt{SAK} + S^2AH^3)$.

The gap $\gamma_{\min}$ controls the gap-dependent exploration cost. When $\gamma_{\min} \lesssim H^2\sqrt{SA/K}$, suboptimal actions are difficult to distinguish from optimal ones, and the term $SAH^4/\gamma_{\min}$ can exceed the standard term $H^2\sqrt{SAK}$. Consequently, even if $d^\star = o(H)$, a shorter effective horizon alone does not guarantee an order-wise improvement at the given episode budget K.

$\gamma_{\min}$ \ $d^\star$	$\Theta(H)$	$0 < d^\star = o(H)$
$\lesssim H^2\sqrt{SA/K}$	$\widetilde{O}(H^2\sqrt{SAK})$	$\widetilde{O}(H^2\sqrt{SAK})$
$\gg H^{5/2}\sqrt{SA/Kd^\star}$	$\widetilde{O}(H^2\sqrt{SAK})$	$\widetilde{O}\Big(H^{3/2}\sqrt{SAKd^\star}\Big)$

Table 2: Order-wise comparison across EHL and MSG regimes. The additive term S^2AH^3 is omitted.

By contrast, when $\gamma_{\min}$ is bounded away from zero independently of K, the gap-dependent term grows only polylogarithmically with K. For sufficiently large K, the regret bound is therefore dominated by $d^\star$. In particular, if $d^\star = o(H)$, the leading term is order-wise smaller than the standard full-horizon regret bound. Table 2 summarizes these regimes. In the next section, focusing on the case where $\gamma_{\min}$ is bounded away from zero, we illustrate the effect of abstention through three concrete examples corresponding to $d^\star = H$, $d^\star = 0$, and $0 < d^\star < H$, respectively.

6 Illustrative Examples

Throughout this section, we consider a simplified fallback value: $G_h(s) = (H - h + 1)r_{\text{abs}},\ r_{\text{abs}} \in [0, 1]$. Detailed proofs are provided in Appendix D. We begin with two simple extreme boundary examples, where $d^\star = H$ and $d^\star = 0$.

Example 1 (No Beneficial Abstention)**.** Consider an augmented MDP in which abstention is never beneficial. Mathematically, we assume that $r_{\text{abs}} < r_h(s, a)$ for all $h \in [H]$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Since $V^\star_{h+1}(s') \geq G_{h+1}(s')$, every base action satisfies $Q^\star_h(s, a) \geq r_h(s, a) + (H - h)r_{\text{abs}} > G_h(s)$. Consequently, every optimal policy interacts throughout the episode, and $d^\star = H$. Substituting this value into Theorem 2 gives $\text{Reg}(K) = \widetilde{O}(H^2\sqrt{SAK} + S^2AH^3)$.

This example shows that if abstention is never beneficial, then our regret bound reduces to the baseline bound, indicating that our algorithm *recovers* the standard tabular RL setting as a *special* case.

Example 2 (Uniformly Optimal Abstention)**.** Let $r_{\max} := \max_{h,s,a} r_h(s, a)$ and suppose that $r_{\text{abs}} > r_{\max}$. Backward induction gives $V^\star_h(s) = G_h(s), Q^\star_h(s, a) = r_h(s, a) + (H - h)r_{\text{abs}}$. Thus, abstention is uniquely optimal at every state-time pair, and $d^\star = 0, \gamma_{\min} = r_{\text{abs}} - r_{\max} > 0$. Theorem 2 yields $\text{Reg}(K) = \widetilde{O}(SAH^4/(r_{\text{abs}} - r_{\max}) + S^2AH^3)$.

The example above illustrates the case where abstention is uniformly optimal. The agent only needs a short initial exploration phase with no polynomial dependence on K to estimate the value of continuing with the base action set. Once the agent identifies that abstention is optimal, it will abstain in subsequent episodes and incur no further regret. As a result, the number of real interactions N_{int} admits an upper bound with no polynomial dependence on K, and the regret bound no longer scales with the full interaction budget HK.

Next, we consider a more refined example (as illustrated in Figure 1) to better understand how the structure of an MDP shapes the benefit of abstention.

Example 3 (Good-State Survival MDP)**.** Consider an augmented MDP with $S = 2$ and $A = 2$. The state space $\mathcal{S} = \{s_{\text{g}}, s_{\text{b}}\}$ consists of two states: the good state s_{g}, the bad state s_{b}. The base action space $\mathcal{A} = \{a_{\text{g}}, a_{\text{b}}\}$ consists of two actions: the good action a_{g} and the bad action a_{b}. The MDP starts with the good state s_{g}.

For all $h \in [H]$ and an $\eta \in (0, 1)$, the transition kernel is given as in Table 3. In other words, once the agent enters the bad state s_{b}, it is absorbed there regardless of whether it chooses a_{g} or a_{b}. If the agent is in the good state s_{g}, then choosing the bad action a_{b} moves it to the bad state with probability

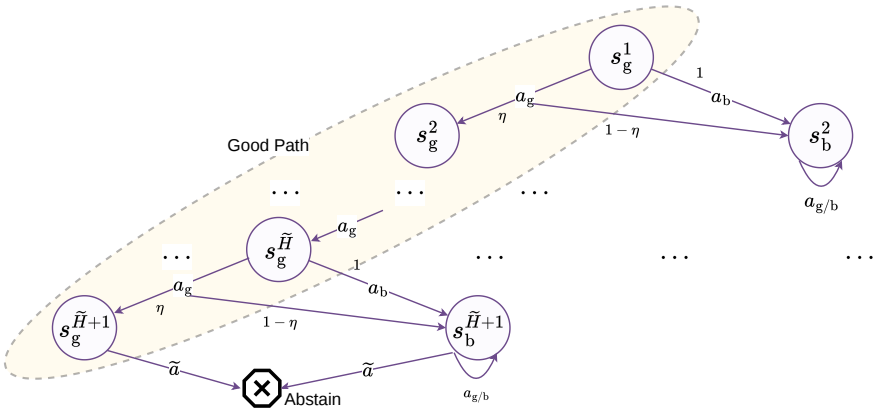


Figure 1: Illustration of Example 3. The optimal policy follows the *good path* (shaded dashed region) by choosing a_{g} at s_{g} and $\widetilde{a}$ at s_{b} during the early phase $h \le \widetilde{H}$, and abstains at both states afterwards.

one. In contrast, choosing the good action a_{g} keeps the agent in the good state with probability η and moves it to the bad state with probability $1-\eta$.

The reward function has a two-phase structure. Fix a threshold $\widetilde{H} < H$. In the early stages $h \le \widetilde{H}$, interaction at the good state is valuable: $r_h(s_{\mathrm{g}}, a_{\mathrm{g}}) = r_1 > 0$, while all other state-action pairs yield reward 0. In the later stages $h > \widetilde{H}$, the same good state-action pair yields only a smaller reward $r_h(s_{\mathrm{g}}, a_{\mathrm{g}}) = r_2$, with $0 < r_2 < r_1$, and all other rewards remain 0. On the other hand, the one step fallback reward satisfies that $r_2 < r_{\mathrm{abs}} < r_1$. If the agent abstains at step h in any state $s \in \mathcal{S}$, the episode terminates and it receives $G_h(s) = (H-h+1)r_{\mathrm{abs}}$. □

Remark 2. *The two-phase reward structure is natural in sequential decision-making problems where rewards decrease over time. For example, early moves in a game may create a decisive advantage, early reasoning steps in an LLM may gather useful evidence, and early corrective actions in control may keep the system recoverable. In contrast, once the trajectory enters a late or low-value phase, further interaction may yield only marginal reward and can be dominated by the fallback option.*

In Example 3, the optimal policy follows the *good path* (see Figure 1): it chooses a_{g} at s_{g} and abstains at s_{b} during the early phase $h \le \widetilde{H}$, and abstains at both states afterwards. Thus, optimal interaction continues only while the agent remains in the good state during the early phase. The probability of interacting at step $h \le \widetilde{H}$ is η^{h-1}, giving $d^\star = \sum_{h=1}^{\widetilde{H}} \eta^{h-1} = (1-\eta^{\widetilde{H}})/(1-\eta)$. This EHL can be much smaller than H. Moreover, $\gamma_{\min} = r_{\mathrm{abs}} - r_2$. Substituting them into Theorem 2 gives the following Theorem 3.

s	a	$P_h(s_{\mathrm{g}} \mid s,a)$	$P_h(s_{\mathrm{b}} \mid s,a)$
s_{g}	a_{g}	η	$1-\eta$
s_{g}	a_{b}	0	1
s_{b}	a_{g}	0	1
s_{b}	a_{b}	0	1

Table 3: Transition Kernel of Example 3.

Remark 3. *Although Example 3 is presented for the simplest setting with two states and two base actions, the construction extends naturally to arbitrary numbers of states S and actions A by enlarging the bad-state and bad-action groups (to sizes $S-1$ and $A-1$ respectively) while preserving the same underlying transition structure. Hence, the advantages of abstention captured by the bound below in Theorem 3 are not artifacts of the two-state two-action setting, which was adopted purely for expositional clarity.*

Theorem 3. *Under the augmented MDP in Example 3 with a general number of states and actions ($S, A \ge 2$), and the same conditions as Theorem 2, with probability at least $1-\delta$, the regret of UCBVI-AG can be upper bounded as*

$$\mathrm{Reg}(K) \le \widetilde{O}\left(H^{3/2}\sqrt{SAK\frac{1-\eta^{\widetilde{H}}}{1-\eta}} + \frac{SAH^4}{r_{\mathrm{abs}}-r_2} + S^2AH^3 \right).$$

This bound shows how the effective horizon length $d^\star$ of the good path determines the potential benefit of abstention. The last two terms account for gap-dependent correction. For fixed instance

parameters, they grow only polylogarithmically with K and become lower order for sufficiently large K. As a result, the bound is **dominated** by the first term.

The first term improves the full horizon H in the standard UCBVI-CH guarantee with $d^\star = (1 - \eta^{\widetilde{H}})/(1-\eta) \le \min\{\widetilde{H}, 1/1-\eta\}$. Thus, when the early phase $\widetilde{H}$ is short or the probability of staying in the good state η is sufficiently small, $d^\star$ can be much smaller than H.

7 Regret Lower Bounds

The preceding theorems and examples illustrate how abstention can yield sharper regret upper bounds. We now establish complementary regret lower bounds that hold with constant probability, showing that our upper bounds are almost tight and the dependence on $d^\star$ and $\gamma_{\min}$ is unavoidable in general. These results support the role of both quantities in characterizing the intrinsic learning difficulty of RL with abstention. Detailed proofs are provided in Appendices E and F.

We first establish a lower bound for instances with a MSG $\gamma_{\min} = \gamma$.

Theorem 4 (MSG-dependent Lower Bound)**.** *Let $S \ge 7$, $A \ge 2$, and define $m = 2^{\lfloor \log_2((S-2)/2) \rfloor}, d_0 = \log_2 m$. Assume $H \ge \max\{16, 4(d_0+1)\}$, and let K and γ satisfy $\sqrt{SAH^3/K} < \gamma \le 1$. There exists a finite family $\mathfrak{M}_\gamma$ of augmented MDPs, with common known rewards and fallbacks, such that $\gamma_{\min}(\mathcal{M}) = \gamma$ for every $\mathcal{M} \in \mathfrak{M}_\gamma$. For an arbitrary learning algorithm* Alg*, there exists $\mathcal{M} \in \mathfrak{M}_\gamma$ such that*

$$\mathbb{P}_{\mathcal{M},\mathrm{Alg}}\left(\mathrm{Reg}_{\mathcal{M}}(K) \ge \frac{SAH^3}{196608\gamma}\right) \ge \frac{1}{4}. \tag{3}$$

This result shows that small positive gaps can impose a substantial exploration cost even when abstention is available. Under the stated sample-size condition, it matches the $1/\gamma_{\min}$ dependence of the corresponding upper-bound term in Theorem 2, up to a factor of H and logarithmic factors.

We next establish a lower bound showing how EHL $d^\star$ contributes to the intrinsic learning difficulty.

Theorem 5 (EHL-dependent Lower Bound)**.** *Let $S \ge 7$, $A \ge 2$, and let m, d_0 be defined as in Theorem 4. Fix an integer ℓ with $2(d_0+1) \le \ell \le H/2$, let $M = m(A-1)\lceil \ell/2 \rceil \ge 8$, and suppose $K \ge MH^2/\ell^2$. There exists a finite family $\mathfrak{M}_{\ell,K}$ of augmented MDPs, with common known rewards and fallbacks, such that for every instance $\mathcal{M} \in \mathfrak{M}_{\ell,K}$, the EHL $d^\star(\mathcal{M})$ satisfies $\ell/2 < d^\star(\mathcal{M}) \le \ell$. For an arbitrary learning algorithm* Alg*, there exists $\mathcal{M} \in \mathfrak{M}_{\ell,K}$ such that*

$$\mathbb{P}_{\mathcal{M},\mathrm{Alg}}\left(\mathrm{Reg}(K) \ge \frac{1}{192} H\sqrt{SAK\, d^\star(\mathcal{M})}\right) \ge \frac{1}{7}. \tag{4}$$

This result shows that even when optimal policies terminate early, learning their decisions can incur regret of order $H\sqrt{SAKd^\star}$. It matches the $\sqrt{Kd^\star}$ dependence of the corresponding upper-bound term in Theorem 2, up to a factor of $\sqrt{H}$ and logarithmic factors.

8 Conclusion

We studied episodic RL with abstention, where the agent may stop at an intermediate step and receive a fallback reward. We proposed UCBVI-AG, an optimistic value-iteration algorithm with an abstention gate, and showed its regret upper bound scales with the realized number of interactions rather than the full horizon budget. We further introduced EHL $d^\star$ and MSG $\gamma_{\min}$ to characterize when abstention yields sharper regret upper bounds. Complementary lower bounds support the role of both quantities in characterizing the intrinsic learning difficulty.

Several directions remain open. First, our upper bounds retain additional factors of $\sqrt{H}$ and H compared to the corresponding EHL-dependent and MSG-dependent lower bounds, respectively, apart from logarithmic factors. A Bernstein-Concentration based analysis may reduce these gaps in horizon dependence. Second, our analysis focuses on tabular MDPs. Extending the framework to linear MDPs and more general function approximation settings would broaden its applicability to large-scale sequential decision-making problems. We leave these to future work.

AI USE STATEMENT

We used generative AI tools to assist with writing and language polishing, to identify potentially relevant literature, and to review mathematical derivations and proofs for potential errors or gaps. AI-generated suggestions were used as supplementary feedback rather than as independent verification of correctness. We checked the cited sources and independently reviewed AI-assisted revisions to the mathematical arguments. The authors take full responsibility for the final content of the paper, including the accuracy of citations and the validity of all mathematical claims and proofs.

REPRODUCIBILITY STATEMENT

This work is purely theoretical. All assumptions and conditions are explicitly stated alongside the theoretical results, and complete proofs are provided in the appendix. We also provide pseudocode for UCBVI-AG in Algorithm 1 and detailed descriptions of the hard-instance constructions used to establish our upper and lower bounds in the appendix.

REFERENCES

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in neural information processing systems*, 30, 2017.

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32-1, 2018.

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in neural information processing systems*, 19, 2006.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

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.

Peter L Bartlett and Marten H Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(8), 2008.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Summer school on machine learning*, pp. 208–240. Springer, 2003.

Weiwei Cheng, Michaël Rademaker, Bernard De Baets, and Eyke Hüllermeier. Predicting partial orders: ranking with abstention. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 215–230. Springer, 2010.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Scott Yang. Online learning with abstention. In *International Conference on Machine Learning*, pp. 1059–1067. PMLR, 2018.

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.

Hen Davidov, Nachshon Cohen, Oren Kalinsky, Yaron Fairstein, Guy Kushilevitz, Ram Yazdi, and Patrick Rebeschini. Knowing when to quit: A principled framework for dynamic abstention in llm reasoning. *arXiv preprint arXiv:2604.18419*, 2026.

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.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.

Tobias Fuchs, Florian Kalinke, and Klemens Böhm. Partial-label learning with a reject option. *arXiv preprint arXiv:2402.00592*, 2024.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Adam Kalai and Varun Kanade. Towards optimally abstaining from prediction with ood test examples. *Advances in Neural Information Processing Systems*, 34:12774–12785, 2021.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.

Tor Lattimore and Marcus Hutter. Near-optimal pac bounds for discounted mdps. *Theoretical Computer Science*, 558:125–143, 2014.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. Do llms know when to not answer? investigating abstention abilities of large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9329–9345, 2025.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Ranking with abstention. *arXiv preprint arXiv:2307.02035*, 2023.

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, 2021.

Stephen Pasteris, Alberto Rumi, Maximilian Thiessen, Shota Saito, Atsushi Miyauchi, Fabio Vitale, and Mark Herbster. Bandits with abstention under expert advice. *Advances in Neural Information Processing Systems*, 37:126059–126087, 2024.

William Saunders, Girish Sastry, Andreas Stuhlmueller, and Owain Evans. Trial without error: Towards safe reinforcement learning via human intervention. *arXiv preprint arXiv:1707.05173*, 2017.

Ayush Sawarni, Jikai Jin, Justin Whitehouse, and Vasilis Syrgkanis. Policy learning with abstention. *arXiv preprint arXiv:2510.19672*, 2025.

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

Alexander L Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L Littman. Pac model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pp. 881–888, 2006.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556, 2025.

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.

Rui Xu, Yi Chen, Sihong Xie, and Hui Xiong. Geometry-calibrated conformal abstention for language models. *arXiv preprint arXiv:2604.27914*, 2026.

Junwen Yang, Tianyuan Jin, and Vincent YF Tan. Multi-armed bandits with abstention. *arXiv preprint arXiv:2402.15127*, 2024.

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. Problem dependent reinforcement learning bounds which can identify bandit structure in mdps. In *International Conference on Machine Learning*, pp. 5747–5755. PMLR, 2018.

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.

Ahmed Zaoui, Christophe Denis, and Mohamed Hebiri. Regression with reject option and application to knn. *Advances in Neural Information Processing Systems*, 33:20073–20082, 2020.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learningvia reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33: 15198–15207, 2020.

A SUMMARY OF NOTATIONS

We summarize the notation used in the proofs below as in Table 4.

Notations	Descriptions
$b_h^k(s,a)$	Exploration bonus for (h,s,a) in episode k.
$q_h^k(s,a)$	Transition estimation correction term: $2SH^2L_2/(N_h^k(s,a)\vee 1)$ (Lemma 6).
L_1, L_2	Logarithmic confidence factors defined in the corresponding theorem or proposition.
$\widetilde{\rho}_h^k(s)$	Optimistic value gap: $V_h^k(s) - V_h^{\pi_k}(s)$ (Lemma 8).
$\widehat{\mathrm{Reg}}(K)$	Cumulative optimistic value gap: $\sum_{k=1}^K \widetilde{\rho}_1^k(s_1)$ (Lemma 1).
α	Recursion coefficient: $1 + 1/(2H)$.
ε_h^k	Transition martingale difference defined in Lemma 8.
$\mathcal{E}_H$	Hoeffding confidence event defined in Lemma 5.
$\mathcal{E}_B$	Bernstein confidence event defined in Lemma 6.
$\mathcal{E}_O$	Optimism event: $U_h^k(s) \geq U_h^\star(s)$ and $V_h^k(s) \geq V_h^\star(s)$ for all k,h,s (Lemma 7).
$\mathcal{G}$	Joint confidence event: $\mathcal{E}_H \cap \mathcal{E}_B$.
τ_k	First abstention stage in episode k; $\tau_k = H+1$ if no abstention occurs.
I_h^k	Indicator of a base-action interaction: $I_h^k = \mathbf{1}\{h < \tau_k\}$, for $h \in [H]$.
$N_{\mathrm{int}}(K)$	Total number of base-action interactions: $\sum_{k=1}^K \sum_{h=1}^H I_h^k = \sum_{k=1}^K (\tau_k - 1)$.

Table 4: Summary of frequently used notations in proof.

In the following sections, we prove Theorem 1 in Appendix B, Proposition 1 and theorem 2 in Appendix C, Theorem 3 in Appendix D, Theorem 4 in Appendix E, and Theorem 5 in Appendix F.

B PROOF OF THEOREM 1

In this section, we prove Theorem 1. The statement and proof of key lemmas are provided in Appendix B.1 and B.2. The auxiliary lemmas are detailed in Appendix B.3.

Proof of Theorem 1. By Lemmas 4 to 6, the intersection of $\mathcal{G}$ and the martingale event in Lemma 4 has probability at least $1-\delta$. On this intersection, Lemma 7 gives $V_h^k \geq V_h^\star$ for all k,h, and hence

$$\begin{aligned}
\mathrm{Reg}(K) &\leq \widehat{\mathrm{Reg}}(K) \\
&:= \sum_{k=1}^K \big(V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)\big) \\
&\leq e^{1/2}\Big(2\underbrace{\sum_{k,h} I_h^k b_h^k(s_h^k,a_h^k)}_{J_1} + \underbrace{\sum_{k,h} I_h^k q_h^k(s_h^k,a_h^k)}_{J_2}\Big) + \underbrace{\sum_{k,h} I_h^k \alpha^h \varepsilon_h^k}_{J_3},
\end{aligned}$$

where the last inequality follows from Lemma 1 and the nonnegativity of b_h^k and q_h^k. Applying Lemmas 2 to 4 yields

$$\begin{aligned}
\mathrm{Reg}(K) \leq\ & 4e^{1/2}cH\sqrt{HSA\,N_{\mathrm{int}}(K)L_1} + 2e^{1/2}S^2AH^3L_2(2+\log K) \\
&+ 2e^{1/2}H\sqrt{2N_{\mathrm{int}}(K)\log\frac{3HK}{\delta}}.
\end{aligned}$$

Since $\log(3HK/\delta) \leq L_1$, $HSA \geq 1$, $c=1$, and $2+\log K \leq 3\log(K+1)$ for $K \geq 1$, this proves Eqn. (2). □

B.1 Decomposition of Pseudo-Regret

Lemma 1 (Decomposition of Pseudo-Regret)**.** *Suppose the event $\mathcal{G}$ holds. The pseudo-regret $\widehat{\mathrm{Reg}}(K) := \sum_{k=1}^{K}\big(V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)\big)$ can be decomposed as follows.*

$$\widehat{\mathrm{Reg}}(K) \le e^{1/2}(2J_1 + J_2) + J_3,$$

where

$$J_1 := \sum_{k,h} I_h^k b_h^k(s_h^k, a_h^k), \qquad J_2 := \sum_{k,h} I_h^k q_h^k(s_h^k, a_h^k), \qquad J_3 := \sum_{k,h} I_h^k \alpha^h \varepsilon_h^k.$$

Proof. On every interaction step, Lemma 8 and the definition of ε_h^k give

$$\widetilde{\rho}_h^k(s_h^k) \le 2b_h^k(s_h^k, a_h^k) + q_h^k(s_h^k, a_h^k) + \alpha\widetilde{\rho}_{h+1}^k(s_{h+1}^k) + \alpha\varepsilon_h^k.$$

At the first abstention step, $\widetilde{\rho}_h^k(s_h^k) = 0$. If the episode never abstains, the terminal value gap is zero at $H+1$. Unrolling along the interacting prefix therefore gives

$$\widetilde{\rho}_1^k(s_1^k) \le \sum_{h=1}^{H} I_h^k \alpha^{h-1}\big(2b_h^k(s_h^k, a_h^k) + q_h^k(s_h^k, a_h^k)\big) + \sum_{h=1}^{H} I_h^k \alpha^h \varepsilon_h^k.$$

Summing over episodes and using $\widehat{\mathrm{Reg}}(K) = \sum_{k=1}^{K} \widetilde{\rho}_1^k(s_1^k)$, we obtain

$$\begin{aligned}
\widehat{\mathrm{Reg}}(K) &\le \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \alpha^{h-1}\big(2b_h^k(s_h^k, a_h^k) + q_h^k(s_h^k, a_h^k)\big) + J_3 \\
&\le e^{1/2}\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k\big(2b_h^k(s_h^k, a_h^k) + q_h^k(s_h^k, a_h^k)\big) + J_3 \\
&= e^{1/2}(2J_1 + J_2) + J_3.
\end{aligned}$$

The second inequality follows from $\alpha^{h-1} \le \alpha^H = (1 + 1/(2H))^H \le e^{1/2}$ and $b_h^k, q_h^k \ge 0$. The weights in J_3 are retained because ε_h^k may have either positive or negative sign. Hence we prove the lemma. □

B.2 Bounding the Decomposition Terms

In this section, we bound the three sums introduced in Lemma 1: the bonus sum J_1 in Lemma 2, the correction sum J_2 in Lemma 3, and the weighted martingale sum J_3 in Lemma 4.

Lemma 2 (Bound on Bonus Accumulation)**.** *Recall that $b_h^k(s,a) := cH\sqrt{L_1/(N_h^k(s,a) \vee 1)}$ for some constant $c \ge 1$ and $L_1 = \log(6HSAK/\delta)$. Then we have*

$$J_1 \le 2cH\sqrt{L}\sum_{h,s,a}\sqrt{N_h^{K+1}(s,a)} \le 2cH\sqrt{HSA\,N_{\mathrm{int}}(K)L}.$$

Proof. Fix (h,s,a) and let $n = N_h^{K+1}(s,a)$. Because a stage-dependent triple is visited at most once per episode, the pre-episode counts on its successive visits are $0, 1, \ldots, n-1$. For $n \ge 2$,

$$\sum_{j=0}^{n-1}\frac{1}{\sqrt{j \vee 1}} = 2 + \sum_{j=2}^{n-1} j^{-1/2} \le 2 + \int_1^{n-1} x^{-1/2}\,dx = 2\sqrt{n-1} \le 2\sqrt{n}.$$

The bound also holds for $n = 0, 1$, with an empty sum when $n = 0$. Thus, by the definition of b_h^k,

$$J_1 = cH\sqrt{L}\sum_{h,s,a}\sum_{j=0}^{N_h^{K+1}(s,a)-1}\frac{1}{\sqrt{j \vee 1}} \le 2cH\sqrt{L}\sum_{h,s,a}\sqrt{N_h^{K+1}(s,a)} \le 2cH\sqrt{HSA\,N_{\mathrm{int}}(K)L_1},$$

where the last inequality follows from Cauchy–Schwarz inequality and $\sum_{h,s,a} N_h^{K+1}(s,a) = N_{\mathrm{int}}(K)$. □

Lemma 3 (Bound on Correction Accumulation)**.** *For every realization of the interaction history,*

$$J_2 \le 2S^2AH^3L_2(2+\log K).$$

Proof. For a triple with $n = N_h^{K+1}(s,a) \ge 1$ visits,

$$\sum_{j=0}^{n-1} \frac{1}{j \vee 1} = 1 + \sum_{j=1}^{n-1} \frac{1}{j} \le 2 + \log K,$$

where $n \le K$; the contribution is zero when $n = 0$. Grouping by the HSA triples gives

$$J_2 = 2SH^2L_2 \sum_{h,s,a} \sum_{j=0}^{N_h^{K+1}(s,a)-1} \frac{1}{j \vee 1} \le 2S^2AH^3L_2(2+\log K).$$

□

The next lemma controls the transition sampling errors accumulated along the agent's trajectory. Since these errors vanish on inactive steps, the bound depends on the number of real interactions rather than HK.

Lemma 4 (Bound on the Martingale Sum)**.** *Recall that* $J_3 := \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \alpha^h \varepsilon_h^k$*, where* $\alpha = 1 + 1/(2H)$*. With probability at least* $1 - \delta/3$*,*

$$J_3 \le 2e^{1/2} H \sqrt{2N_{\text{int}}(K) \log \frac{3HK}{\delta}}.$$

Proof. We first index the $T := HK$ episode-stage slots chronologically by $t = (k-1)H + h$, and set

$$X_t := I_h^k \alpha^h \varepsilon_h^k, \qquad A_t := I_h^k.$$

After abstention, all remaining slots of that episode have $X_t = A_t = 0$. Let $\mathcal{F}_{t-1}$ denote the history before the transition in slot t, including the interaction decision and chosen action. In particular, A_t is $\mathcal{F}_{t-1}$-measurable.

On an active step with $h < H$, recall that

$$\varepsilon_h^k = (P_h \widetilde{\rho}_{h+1}^k)(s_h^k, a_h^k) - \widetilde{\rho}_{h+1}^k(s_{h+1}^k).$$

Hence $|\varepsilon_h^k| \le 2H$. Using $\alpha^h \le e^{1/2}$ and defining $B := 2e^{1/2}H$, we obtain

$$\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = 0, \qquad |X_t| \le BA_t.$$

For any fixed $\lambda > 0$, conditional Hoeffding's lemma (Lattimore & Szepesvári, 2020) gives

$$\mathbb{E}[e^{\lambda X_t} \mid \mathcal{F}_{t-1}] \le \exp\left(\frac{\lambda^2 B^2}{2} A_t\right).$$

Indeed, when $A_t = 1$, the increment is centered and lies in $[-B, B]$; when $A_t = 0$, both sides equal one.

Define the partial sums

$$M_t := \sum_{j=1}^{t} X_j, \qquad C_t := \sum_{j=1}^{t} A_j, \qquad M_0 = C_0 = 0,$$

and the exponential process

$$Z_t(\lambda) := \exp\left(\lambda M_t - \frac{\lambda^2 B^2}{2} C_t\right).$$

Since A_t is $\mathcal{F}_{t-1}$-measurable,

$$\mathbb{E}[Z_t(\lambda) \mid \mathcal{F}_{t-1}] = Z_{t-1}(\lambda) e^{-\lambda^2 B^2 A_t/2} \mathbb{E}[e^{\lambda X_t} \mid \mathcal{F}_{t-1}] \le Z_{t-1}(\lambda).$$

Thus, $(Z_t(\lambda))_{t=0}^T$ is a nonnegative supermartingale with $Z_0(\lambda) = 1$, and hence

$$\mathbb{E}[Z_T(\lambda)] \le 1. \tag{5}$$

By definition, we have $M_T = J_3, C_T = N_{\text{int}}(K)$. Because C_T is random, we first bound the probability of a large deviation for each possible deterministic count $n \in \{1, \dots, T\}$.

Set $\ell = \log(3T/\delta)$ and, for each such n, choose $\lambda_n = \sqrt{2\ell/(B^2 n)}$. On the event

$$\{M_T \ge B\sqrt{2n\ell},\ C_T = n\},$$

we have

$$\lambda_n M_T - \frac{\lambda_n^2 B^2}{2} C_T \ge \lambda_n B\sqrt{2n\ell} - \frac{\lambda_n^2 B^2 n}{2} = 2\ell - \ell = \ell.$$

Therefore, by Markov's inequality and Eqn. (5),

$$\mathbb{P}\Big(M_T \ge B\sqrt{2n\ell},\ C_T = n\Big) \le \mathbb{P}\big(Z_T(\lambda_n) \ge e^\ell\big) \le e^{-\ell}.$$

If $C_T = 0$, then every increment is zero and $M_T = 0$. Taking a union bound over the remaining possible counts gives

$$\mathbb{P}\Big(M_T > B\sqrt{2C_T\ell}\Big) \le \sum_{n=1}^{T} \mathbb{P}\Big(M_T > B\sqrt{2n\ell},\ C_T = n\Big) \le Te^{-\ell} = \frac{\delta}{3}.$$

Substituting $B = 2e^{1/2}H$, $T = HK$, and $C_T = N_{\text{int}}(K)$ proves the lemma. □

B.3 Auxiliary Lemmas

We first show that the bonus $b_h^k(s,a)$ uniformly bounds the transition estimation error applied to the optimal value function with high probability.

Lemma 5 (Hoeffding-type Confidence Bound for Optimal Value Function)**.** *Recall that* $b_h^k(s,a) = cH\sqrt{L_1/(N_h^k(s,a) \vee 1)}$*, where* $L_1 = \log(6HSAK/\delta)$ *and* $c = 1$*. With probability at least* $1 - \delta/3$*, the event*

$$\mathcal{E}_{\text{H}} := \left\{ \left| ((\widehat{P}_h^k - P_h)V_{h+1}^\star)(s,a) \right| \le b_h^k(s,a), \quad \forall k, h, s, a \right\}$$

holds.

Proof. For the analysis, we represent transition samples using independent sequences indexed by (h,s,a). For each triple, let $(X_{h,s,a,i})_{i \ge 1}$ be i.i.d. with distribution $P_h(\cdot \mid s,a)$. On the ith active visit to (h,s,a), the next state is $X_{h,s,a,i}$.

For each deterministic $n \in [K]$, Hoeffding's inequality (Hoeffding, 1963) gives

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} V_{h+1}^\star(X_{h,s,a,i}) - (P_h V_{h+1}^\star)(s,a) \right| > H\sqrt{\frac{L_1}{2n}} \right) \le 2e^{-L_1}.$$

Taking a union bound over all (h,s,a,n) gives failure probability at most $2HSAKe^{-L_1} = \delta/3$.

On this event, we can substitute $n = N_h^k(s,a)$ whenever the count $N_h^k(s,a)$ is positive. Then $\widehat{P}_h^k V_{h+1}^\star = (N_h^k(s,a))^{-1} \sum_{i=1}^{N_h^k(s,a)} V_{h+1}^\star(X_{h,s,a,i})$. The resulting bound is at most $b_h^k(s,a)$.

If the count $N_h^k(s,a)$ is zero, the algorithm sets $\widehat{P}_h^k(\cdot \mid s,a) = 0$, so the error is at most $H \le cH\sqrt{L_1} = b_h^k(s,a)$. In conclusion, we prove this lemma. □

The preceding lemma controls the transition estimation error only when applied to $V_{h+1}^\star$. To handle data-dependent value functions, we next establish an entrywise Bernstein confidence bound and derive a correction bound that holds uniformly over all $f : \mathcal{S} \to [0, H]$.

Lemma 6 (Bernstein-type Confidence and Transition Correction Bounds)**.** *Define* $q_h^k(s,a) := 2SH^2L_2/(N_h^k(s,a)\vee 1)$ *as a transition estimation correction term, where* $L_2 := \log(12S^2AHK/\delta)$. *With probability at least* $1-\delta/3$*, the Bernstein confidence event*

$$\mathcal{E}_{\mathrm{B}} := \left\{ \left| \widehat{P}_h^k(s' \mid s,a) - P_h(s' \mid s,a) \right| \le \sqrt{\frac{2P_h(s' \mid s,a)L_2}{N_h^k(s,a)\vee 1}} + \frac{2L_2}{3(N_h^k(s,a)\vee 1)}, \quad \forall k,h,s,a,s' \right\}$$

holds. Moreover, on $\mathcal{E}_{\mathrm{B}}$*, simultaneously for all* k,h,s,a *and all functions* $f:\mathcal{S}\to[0,H]$*, including data-dependent functions,*

$$\left| ((\widehat{P}_h^k - P_h)f)(s,a) \right| \le \frac{(P_hf)(s,a)}{2H} + q_h^k(s,a).$$

Proof. In this proof, we use the same independent sample sequence $(X_{h,s,a,i})_{i\ge 1}$ introduced in the proof of Lemma 5.

For fixed (h,s,a,s') and deterministic $n\in[K]$, the variables $\mathbf{1}\{X_{h,s,a,i}=s'\}$ are i.i.d. Bernoulli with mean $p := P_h(s' \mid s,a)$ and variance $p(1-p)\le p$. Bernstein's inequality (Boucheron et al., 2003) therefore gives

$$\mathbb{P}\left(\left| \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_{h,s,a,i}=s'\} - p \right| > \sqrt{\frac{2pL_2}{n}} + \frac{2L_2}{3n} \right) \le 2e^{-L_2}.$$

Taking a union bound over all S^2AHK choices of (h,s,a,s',n), the corresponding inequalities hold simultaneously with probability at least $1-2S^2AHKe^{-L_2} = 1-\delta/6$.

On this event, whenever $N_h^k(s,a)>0$, substituting $n = N_h^k(s,a)$ and using

$$\widehat{P}_h^k(s' \mid s,a) = \frac{1}{N_h^k(s,a)} \sum_{i=1}^{N_h^k(s,a)} \mathbf{1}\{X_{h,s,a,i}=s'\}$$

yields the desired entrywise bound. When $N_h^k(s,a)=0$, the absolute error is at most one, while $2L_2/3\ge 1$, so the bound also holds. Consequently, $\mathbb{P}(\mathcal{E}_{\mathrm{B}})\ge 1-\delta/6\ge 1-\delta/3$.

For the second statement, we work on $\mathcal{E}_{\mathrm{B}}$ and fix any (k,h,s,a). Write $N := N_h^k(s,a)\vee 1$ and $p_{s'} := P_h(s' \mid s,a)$. For any $f:\mathcal{S}\to[0,H]$, the triangle inequality and the entrywise bound established above give

$$\left| ((\widehat{P}_h^k - P_h)f)(s,a) \right| \le \sum_{s'\in\mathcal{S}} \left| \widehat{P}_h^k(s' \mid s,a) - p_{s'} \right| f(s') \le \sum_{s'\in\mathcal{S}} \left(\sqrt{\frac{2p_{s'}L_2}{N}} f(s') + \frac{2L_2 f(s')}{3N} \right).$$

For each $s'\in\mathcal{S}$, AM–GM inequality yields

$$\sqrt{\frac{2p_{s'}L_2}{N}} f(s') = \sqrt{\frac{p_{s'}f(s')}{H} \cdot \frac{2Hf(s')L_2}{N}} \le \frac{p_{s'}f(s')}{2H} + \frac{Hf(s')L_2}{N}.$$

Summing over s' and using $\sum_{s'} f(s') \le SH$ and $H\ge 1$, we obtain

$$\begin{aligned} \left| ((\widehat{P}_h^k - P_h)f)(s,a) \right| &\le \frac{(P_hf)(s,a)}{2H} + \frac{L_2}{N}\left(H+\frac{2}{3}\right)\sum_{s'\in\mathcal{S}} f(s') \\ &\le \frac{(P_hf)(s,a)}{2H} + \frac{2SH^2L_2}{N} \\ &= \frac{(P_hf)(s,a)}{2H} + q_h^k(s,a). \end{aligned}$$

This argument holds on $\mathcal{E}_{\mathrm{B}}$ for every (k,h,s,a) and every $f:\mathcal{S}\to[0,H]$. The bound therefore also applies to data-dependent functions, without any additional union bound or independence assumption. □

As a direct application of Lemma 5, we have the following lemma on optimistic estimation.

Lemma 7 (Optimistic Base Value Estimation). *On $\mathcal{E}_{\mathrm{H}}$, the event*

$$\mathcal{E}_{\mathrm{O}} := \{U_h^k(s) \geq U_h^\star(s), \quad V_h^k(s) \geq V_h^\star(s), \quad \forall k, h, s\}$$

holds. In particular, $\mathbb{P}(\mathcal{E}_{\mathrm{O}}) \geq 1 - \delta/3$.

Proof. On $\mathcal{E}_{\mathrm{H}}$,

$$(\widehat{P}_h^k V_{h+1}^\star)(s,a) + b_h^k(s,a) \geq (P_h V_{h+1}^\star)(s,a). \tag{6}$$

We use backward induction, starting from $V_{H+1}^k = V_{H+1}^\star = 0$. If $V_{h+1}^k \geq V_{h+1}^\star$, the nonnegative entries of $\widehat{P}_h^k$ imply

$$r_h(s,a) + (\widehat{P}_h^k V_{h+1}^k)(s,a) + b_h^k(s,a) \geq r_h(s,a) + (P_h V_{h+1}^\star)(s,a) = Q_h^\star(s,a).$$

Since $Q_h^\star(s,a) \leq H - h + 1$, clipping preserves this inequality: $Q_h^k(s,a) \geq Q_h^\star(s,a)$. Maximizing over base actions gives $U_h^k(s) \geq U_h^\star(s)$, and

$$V_h^k(s) = \max\{U_h^k(s), G_h(s)\} \geq \max\{U_h^\star(s), G_h(s)\} = V_h^\star(s). \tag{7}$$

This closes the induction. Nonnegativity of $\widehat{P}_h^k$ is sufficient, including when it is the zero measure at an unvisited triple. □

We next combine the confidence bounds above with optimism to derive a one-step recursion for the difference between the optimistic value and the value of the executed policy. The recursion separates the bonus and correction terms from the next-step value difference and a transition sampling error, and terminates when the agent abstains.

Lemma 8 (One-step Decomposition under Abstention). *Define $\widetilde{\rho}_h^k(s) := V_h^k(s) - V_h^{\pi^k}(s), \alpha := 1 + 1/(2H)$, with the terminal convention $V_{H+1}^k = V_{H+1}^{\pi^k} = V_{H+1}^\star = 0$. For each interaction step $h < \tau_k$ with $h < H$, define $\varepsilon_h^k := (P_h \widetilde{\rho}_{h+1}^k)(s_h^k, a_h^k) - \widetilde{\rho}_{h+1}^k(s_{h+1}^k)$. If stage H is an interaction step, set $\varepsilon_H^k = 0$. On $\mathcal{G} := \mathcal{E}_{\mathrm{H}} \cap \mathcal{E}_{\mathrm{B}}$, for all episodes k and interaction steps $h < \tau_k$,*

$$\widetilde{\rho}_h^k(s_h^k) \leq 2b_h^k(s_h^k, a_h^k) + q_h^k(s_h^k, a_h^k) + \alpha \widetilde{\rho}_{h+1}^k(s_{h+1}^k) + \alpha \varepsilon_h^k, \tag{8}$$

where $\widetilde{\rho}_{H+1}^k(s_{H+1}^k)$ is interpreted as zero, without requiring a terminal-state observation. If $\tau_k \leq H$, then $\widetilde{\rho}_{\tau_k}^k(s_{\tau_k}^k) = 0$.

Proof. Work on $\mathcal{G}$ and fix an episode k. If the algorithm abstains at stage $h = \tau_k \leq H$, the gate condition gives

$$G_h(s_h^k) \geq U_h^k(s_h^k).$$

Thus $V_h^k(s_h^k) = G_h(s_h^k)$. Since the executed policy π^k also abstains immediately at this state,

$$V_h^{\pi^k}(s_h^k) = G_h(s_h^k),$$

and consequently $\widetilde{\rho}_h^k(s_h^k) = 0$.

Now consider an interaction step $h < \tau_k$ and abbreviate $(s,a) := (s_h^k, a_h^k)$. The gate and action-selection rule imply

$$U_h^k(s) > G_h(s), \qquad a \in \arg\max_{a' \in \mathcal{A}} Q_h^k(s,a').$$

Hence $V_h^k(s) = Q_h^k(s,a)$. The clipped update and the Bellman equation for π^k give, respectively,

$$Q_h^k(s,a) \leq r_h(s,a) + (\widehat{P}_h^k V_{h+1}^k)(s,a) + b_h^k(s,a)$$

and

$$V_h^{\pi^k}(s) = r_h(s,a) + (P_h V_{h+1}^{\pi^k})(s,a).$$

Subtracting and decomposing the transition terms yields

$$\begin{aligned}
\widetilde{\rho}_h^k(s) &\leq b_h^k(s,a) + (\widehat{P}_h^k V_{h+1}^k - P_h V_{h+1}^{\pi^k})(s,a) \\
&= b_h^k(s,a) + ((\widehat{P}_h^k - P_h) V_{h+1}^\star)(s,a) \\
&\quad + ((\widehat{P}_h^k - P_h)(V_{h+1}^k - V_{h+1}^\star))(s,a) + (P_h \widetilde{\rho}_{h+1}^k)(s,a).
\end{aligned} \tag{9}$$

We next bound the RHS of Eqn. (9).

First, by the definition of $\mathcal{E}_{\mathrm{H}}$,

$$\left|((\widehat{P}_h^k - P_h)V_{h+1}^\star)(s,a)\right| \le b_h^k(s,a). \tag{10}$$

Moreover, define $f := V_{h+1}^k - V_{h+1}^\star$. By optimism on $\mathcal{E}_{\mathrm{H}}$ in Lemma 7, optimality of $V^\star$, and the boundedness of V^k, we have

$$0 \le f \le V_{h+1}^k - V_{h+1}^{\pi^k} = \widetilde{\rho}_{h+1}^k \le H.$$

Applying Lemma 6 on $\mathcal{E}_{\mathrm{B}}$, we obtain

$$\begin{aligned}((\widehat{P}_h^k - P_h)f)(s,a) &\le \left|((\widehat{P}_h^k - P_h)f)(s,a)\right| \\ &\le \frac{(P_h f)(s,a)}{2H} + q_h^k(s,a) \\ &\le \frac{(P_h \widetilde{\rho}_{h+1}^k)(s,a)}{2H} + q_h^k(s,a),\end{aligned} \tag{11}$$

where the last inequality follows from $f \le \widetilde{\rho}_{h+1}^k$ and the nonnegativity of the transition probabilities. Plugging Eqns. (10) and (11) into Eqn. (9) gives

$$\begin{aligned}\widetilde{\rho}_h^k(s) &\le 2b_h^k(s,a) + q_h^k(s,a) + \left(1 + \frac{1}{2H}\right)(P_h \widetilde{\rho}_{h+1}^k)(s,a) \\ &\le 2b_h^k(s_h^k, a_h^k) + q_h^k(s_h^k, a_h^k) + \alpha \widetilde{\rho}_{h+1}^k(s_{h+1}^k) + \alpha \varepsilon_h^k,\end{aligned}$$

where the last inequality follows from the definition of ε_h^k and hence proves this lemma. □

C Proof of Theorem 2

C.1 Proof of Proposition 1

Lemma 9 (Regret-controlled interaction count)**.** *Suppose that, before episode k, the agent selects a deterministic Markov policy π^k based on the past history. For any fixed $K \ge 1$ and $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$N_{\mathrm{int}}(K) \le 2Kd^\star + \frac{2H}{\gamma_{\min}}\mathrm{Reg}(K) + H\log\frac{1}{\delta}.$$

Proof. If $\gamma_{\min} = +\infty$, every base action is optimal, so $d^\star = H$ and the claim follows from $N_{\mathrm{int}}(K) \le KH$. Henceforth, assume $\gamma_{\min} \in (0,\infty)$.

Fix a deterministic Markov policy π. Construct $\pi^\sharp \in \Pi^\star$ by retaining π's action whenever it is optimal and replacing it with an optimal augmented action otherwise. Couple their trajectories so that they use the same transition sample whenever their states and actions coincide, while preserving their respective marginal laws.

Let

$$E_\pi := \{\exists h < \tau_\pi : V_h^\star(s_h) > Q_h^\star(s_h, a_h)\}.$$

On E_π^c, all base actions taken by π are optimal and therefore retained by $\pi^\sharp$. The trajectories consequently coincide until π stops, at which point $\pi^\sharp$ either also stops or continues. Thus, $\tau_\pi - 1 \le \tau_{\pi^\sharp} - 1$ on E_π^c. Since each trajectory contains at most H interactions, we have, trajectory-wise under the coupling,

$$\tau_\pi - 1 \le \tau_{\pi^\sharp} - 1 + H\mathbf{1}_{E_\pi}.$$

Taking expectation under the joint coupling law and using its marginals yields

$$\mathbb{E}_\pi[\tau_\pi - 1] \le \mathbb{E}_{\pi^\sharp}[\tau_{\pi^\sharp} - 1] + H\mathbb{P}_\pi(E_\pi) \le d^\star + H\mathbb{P}_\pi(E_\pi). \tag{12}$$

By the Bellman equations and telescoping,

$$V_1^\star(s_1) - V_1^\pi(s_1) = \mathbb{E}_\pi\left[\sum_{h<\tau_\pi} \big(V_h^\star(s_h) - Q_h^\star(s_h,a_h)\big) + \mathbf{1}\{\tau_\pi \le H\}\big(V_{\tau_\pi}^\star(s_{\tau_\pi}) - G_{\tau_\pi}(s_{\tau_\pi})\big)\right],$$

where the final term is interpreted as zero if no abstention occurs. All terms are nonnegative, and on the event E_π the first sum is at least $\gamma_{\min}$. Hence

$$\mathbb{P}_\pi(E_\pi) \le \frac{V_1^\star(s_1) - V_1^\pi(s_1)}{\gamma_{\min}}.$$

Plugging this into Eqn. (12) gives

$$\mathbb{E}_\pi[\tau_\pi - 1] \le d^\star + \frac{H}{\gamma_{\min}}\big(V_1^\star(s_1) - V_1^\pi(s_1)\big). \tag{13}$$

Let $\mathcal{F}_{k-1}$ denote the history available when π^k is selected, and define

$$L_k := \tau_k - 1, \qquad \mu_k := \mathbb{E}[L_k \mid \mathcal{F}_{k-1}].$$

Recall $N_{\mathrm{int}}(K) = \sum_{k=1}^K L_k$, we next bound $N_{\mathrm{int}}(K)$ by concentration.

By convexity, $e^x \le (1-x)e^0 + xe^1 = 1 + (e-1)x$ for all $x \in [0,1]$. Since $0 \le L_k/H \le 1$, taking conditional expectations yields

$$\mathbb{E}[e^{L_k/H} \mid \mathcal{F}_{k-1}] \le 1 + (e-1)\frac{\mu_k}{H} \le \exp\left(\frac{2\mu_k}{H}\right),$$

where the last inequality follows from $1 + u \le e^u$, $e - 1 < 2$, and $\mu_k \ge 0$.

Therefore,

$$M_t := \exp\left(\frac{1}{H}\sum_{k=1}^{t}(L_k - 2\mu_k)\right), \qquad M_0 = 1,$$

is a nonnegative supermartingale. By Markov's inequality, with probability at least $1-\delta$,

$$N_{\mathrm{int}}(K) = \sum_{k=1}^K L_k \le 2\sum_{k=1}^K \mu_k + H\log\frac{1}{\delta}.$$

Substituting the bound on $\sum_{k=1}^K \mu_k$ proves the claim. □

Proof of Proposition 1. We apply Lemma 9 and Theorem 1, each with failure probability $\delta/2$. In this proof, set

$$L_1 = \log\frac{12HSAK}{\delta}, \qquad L_2 = \log\frac{24S^2AHK}{\delta}.$$

By a union bound, with probability at least $1-\delta$,

$$N_{\mathrm{int}}(K) \le 2Kd^\star + \frac{2H}{\gamma_{\min}}\mathrm{Reg}(K) + H\log\frac{2}{\delta}$$

and

$$\mathrm{Reg}(K) \le 10\left(H^{3/2}\sqrt{SA\,N_{\mathrm{int}}(K)L_1} + S^2AH^3L_2\log(K+1)\right)$$

hold simultaneously. Substituting the regret bound gives

$$N_{\mathrm{int}}(K) \le 2Kd^\star + H\log\frac{2}{\delta} + \frac{20S^2AH^4L_2\log(K+1)}{\gamma_{\min}} + \frac{20H^{5/2}\sqrt{SAL_1}}{\gamma_{\min}}\sqrt{N_{\mathrm{int}}(K)}.$$

Using AM-GM inequality,

$$\frac{20H^{5/2}\sqrt{SAL_1}}{\gamma_{\min}}\sqrt{N_{\mathrm{int}}(K)} \le \frac{N_{\mathrm{int}}(K)}{2} + \frac{200SAH^5L_1}{\gamma_{\min}^2}.$$

Rearranging yields

$$\begin{aligned} N_{\mathrm{int}}(K) &\le 4Kd^\star + 2H\log\frac{2}{\delta} + \frac{40S^2AH^4L_2\log(K+1)}{\gamma_{\min}} + \frac{400SAH^5L_1}{\gamma_{\min}^2} \\ &\le 4Kd^\star + \frac{42S^2AH^4L_2\log(K+1)}{\gamma_{\min}} + \frac{400SAH^5L_1}{\gamma_{\min}^2}, \end{aligned}$$

which proves the proposition. □

C.2 Proof of Theorem 2

Proof of Theorem 2. We work on the event of probability at least $1-\delta$ constructed in the proof of Proposition 1, on which both the interaction bound and Eqn. (2) hold.

Substituting the interaction bound into Eqn. (2) and using $\sqrt{x+y+z} \le \sqrt{x}+\sqrt{y}+\sqrt{z}$ gives

$$\mathrm{Reg}(K) \le 20H^{3/2}\sqrt{SAKd^\star L_1} + \frac{200SAH^4L_1}{\gamma_{\min}} + 10S^2AH^3L_2\log(K+1) + 10\sqrt{43\left(\frac{SAH^4L_1}{\gamma_{\min}}\right)(S^2AH^3L_2\log(K+1))}.$$

By AM-GM inequality, the last term is at most

$$\frac{200SAH^4L_1}{\gamma_{\min}} + \frac{43}{8}S^2AH^3L_2\log(K+1).$$

Combining these bounds and using $10 + 43/8 \le 16$ yields

$$\mathrm{Reg}(K) \le 20H^{3/2}\sqrt{SAKd^\star L_1} + \frac{400SAH^4L_1}{\gamma_{\min}} + 16S^2AH^3L_2\log(K+1),$$

as claimed. □

D Proofs for the Illustrative Examples in Section 6

Throughout the proofs, set $V^\star_{H+1} = G_{H+1} = 0$.

D.1 Proof for Example 1

Proof. For every state-time pair, $V^\star_{h+1} \ge G_{h+1}$ because abstention is an available action. Since G_{h+1} is constant across states,

$$\begin{aligned} Q^\star_h(s,a) &= r_h(s,a) + (P_hV^\star_{h+1})(s,a) \\ &\ge r_h(s,a) + (H-h)r_{\mathrm{abs}} \\ &> (H-h+1)r_{\mathrm{abs}} = G_h(s). \end{aligned}$$

Abstention is therefore strictly suboptimal at every state-time pair. Every optimal policy takes H base actions, which gives $d^\star = H$. Consequently,

$$\min\left\{H^{3/2}\sqrt{SAKd^\star} + \frac{SAH^4}{\gamma_{\min}}, H^2\sqrt{SAK}\right\} = H^2\sqrt{SAK},$$

including the case $\gamma_{\min} = +\infty$. The claimed regret bound follows from Theorem 2.

On the optimism event, $Q^k_h(s,a) \ge Q^\star_h(s,a) > G_h(s)$. Thus, the abstention gate never triggers and $N_{\mathrm{int}}(K) = HK$. □

D.2 Proof for Example 2

Proof. We show by backward induction that $V^\star_h(s) = G_h(s)$ for every h, s. The identity holds at $H+1$. Suppose it holds at $h+1$. For every base action,

$$\begin{aligned} Q^\star_h(s,a) &= r_h(s,a) + (P_hG_{h+1})(s,a) \\ &= r_h(s,a) + (H-h)r_{\mathrm{abs}} \\ &\le r_{\max} + (H-h)r_{\mathrm{abs}} \\ &< (H-h+1)r_{\mathrm{abs}} = G_h(s). \end{aligned}$$

Hence abstention is uniquely optimal and $V^\star_h(s) = G_h(s)$, completing the induction. Every optimal policy abstains at stage one, so $d^\star = 0$. Moreover, all base-action gaps are positive:

$$V^\star_h(s) - Q^\star_h(s,a) = r_{\mathrm{abs}} - r_h(s,a).$$

Taking the minimum over (h,s,a) yields $\gamma_{\min} = r_{\mathrm{abs}} - r_{\max}$. Substituting these quantities into Proposition 1 and Theorem 2 proves the stated interaction and regret bounds, respectively. □

D.3 Proof of Theorem 3

Proof. We first compute the value function. **Optimal values and actions.** At s_b, every base action gives zero reward and remains in s_b. Backward induction therefore gives

$$V_h^\star(s_b) = G_h(s_b), \qquad Q_h^\star(s_b, a) = (H-h)r_{\mathrm{abs}}.$$

Since $r_{\mathrm{abs}} > 0$, abstention is uniquely optimal at s_b. A bad action at s_g likewise satisfies

$$Q_h^\star(s_g, a_b) = (H-h)r_{\mathrm{abs}}.$$

For $h > \widetilde{H}$, backward induction gives $V_h^\star(s_g) = G_h(s_g)$: provided that the identity holds at $h+1$,

$$Q_h^\star(s_g, a_g) = r_2 + (H-h)r_{\mathrm{abs}} < G_h(s_g).$$

Thus abstention is uniquely optimal in the late phase.

For the early phase, define $u_h := V_h^\star(s_g) - G_h(s_g)$ and note that $u_{\widetilde{H}+1} = 0$. Using the Bellman equation,

$$\begin{aligned} Q_h^\star(s_g, a_g) - G_h(s_g) &= r_1 + \eta V_{h+1}^\star(s_g) + (1-\eta)V_{h+1}^\star(s_b) - G_h(s_g) \\ &= r_1 - r_{\mathrm{abs}} + \eta u_{h+1} > 0. \end{aligned}$$

Since $Q_h^\star(s_g, a_b) < G_h(s_g)$, action a_g is uniquely optimal for $h \le \widetilde{H}$. Therefore,

$$u_h = r_1 - r_{\mathrm{abs}} + \eta u_{h+1} = (r_1 - r_{\mathrm{abs}})\frac{1-\eta^{\widetilde{H}-h+1}}{1-\eta}.$$

Optimal effective horizon length. An optimal policy interacts at stage h exactly when $h \le \widetilde{H}$ and the process has remained at s_g through the first $h-1$ transitions. This event has probability η^{h-1}. Thus, every optimal policy has the same expected interaction length, and

$$d^\star = \sum_{h=1}^{H} \mathbb{P}_{\pi^\star}(h < \tau_{\pi^\star}) = \sum_{h=1}^{\widetilde{H}} \eta^{h-1} = \frac{1-\eta^{\widetilde{H}}}{1-\eta}.$$

Minimum sub-optimality gap. The base-action gaps at every state-time pair are

State, action, and phase	$V_h^\star(s) - Q_h^\star(s,a)$
$s_b,\ a \in \mathcal{A},\ h \in [H]$	r_{abs}
$s_g,\ a_b,\ h \le \widetilde{H}$	$r_{\mathrm{abs}} + u_h$
$s_g,\ a_b,\ h > \widetilde{H}$	r_{abs}
$s_g,\ a_g,\ h \le \widetilde{H}$	0
$s_g,\ a_g,\ h > \widetilde{H}$	$r_{\mathrm{abs}} - r_2$

Since $0 < r_2 < r_{\mathrm{abs}}$ and $\widetilde{H} < H$, the smallest positive gap is attained in the last row. Hence,

$$\gamma_{\min} = r_{\mathrm{abs}} - r_2.$$

The same values and gaps hold in the extension with additional bad states and bad actions: every bad state has value G_h, and every bad action at s_g has value $(H-h)r_{\mathrm{abs}}$.

Substituting the above expressions for $d^\star$ and $\gamma_{\min}$ into Theorem 2 proves Theorem 3.

□

E Proof of the MSG Lower Bound

We first construct a family of hard MDPs with $\gamma_{\min} = \gamma$ in Appendix E.1, then establish the auxiliary information-theoretic inequalities in Appendix E.2, and finally combine these ingredients with a truncation argument to prove the constant-probability regret lower bound in Appendix E.3.

E.1 Construction of Hard MDP Instance

In this section, we define the augmented MDPs $\{\mathcal{M}_0\} \cup \{\mathcal{M}_c : c \in \mathcal{C}\}$ used in the lower bound, forming $\mathfrak{M}_\gamma$, where $\mathcal{M}_0$ is a reference MDP and each $\mathcal{M}_c$ differs from $\mathcal{M}_0$ only at a single time-state-action triple c. An illustration of a hard instance can be found in Figure 2. Use the parameters in Theorem 4: $S \geq 7$, $A \geq 2$,

$$m = 2^{\lfloor \log_2((S-2)/2) \rfloor}, \qquad d_0 = \log_2 m, \qquad H_e = \lfloor H/2 \rfloor, \qquad L = H - H_e.$$

Assume $H \geq \max\{16, 4(d_0+1)\}$ and $\sqrt{SAH^3/K} < \gamma \leq 1$. Then $m \geq 2$, $m \geq S/8$, $L \geq 8$, and $\gamma \leq L/8$.

Definition of the reference MDP $\mathcal{M}_0$. First, we define the reference MDP $\mathcal{M}_0$. The state space $\mathcal{S}$ and action space $\mathcal{A}$ are constructed as follows.

- The state space consists of a complete binary routing tree with m leaf states $\{x_1, \ldots, x_m\}$, together with two absorbing states s_g and s_b. We refer to $x_1, \ldots, x_m$ as the *waiting states*, and to s_g, s_b as the *good state* and *bad state*, respectively. The root of the routing tree is the initial state s_1. If necessary, dummy states are added so that the total number of states is exactly S.
- At each waiting state x_i, one base action a_w is reserved as the *waiting action*, while the remaining $A-1$ base actions are called *gambling actions*. We have an additional $\tilde{a}$ *abstention action*.

The transition kernel of $\mathcal{M}_0$ is defined as follows.

- At each internal node of the tree, two base actions $a_{\mathrm{L}}, a_{\mathrm{R}}$ lead to the left and right children, respectively, while all remaining base actions lead to the left child. All transitions are deterministic. The tree and the two outcome states use $2m+1 \leq S$ states. Dummy states are absorbing, unreachable from the root, and have zero rewards and fallbacks. Thus every waiting state x_i can be reached from s_1 within exactly d_0 steps.
- For each waiting state x_i and each $h \leq H_e$, the waiting action keeps the agent at the same state:
$$\mathbb{P}_h[x_i \mid x_i, a_w] = 1.$$
For every gambling action $a \neq a_w$, taking a leads to s_g or s_b:
$$\mathbb{P}_h[s_g \mid x_i, a] = \frac{1}{2}, \qquad \mathbb{P}_h[s_b \mid x_i, a] = \frac{1}{2}.$$
- After the active window, i.e. for $h > H_e$, every base action from a waiting state deterministically leads to s_b.
- The states s_g and s_b are absorbing.

Next, we define the reward and abstention value.

- All ordinary rewards are zero.
$$r_h(s,a) = 0, \qquad \forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}.$$
- The fallback rewards are defined by, for the good and bad states,
$$G_h(s_g) := \min\{L, H-h+1\}, \qquad G_h(s_b) := 0.$$
For every routing-tree state s (including each waiting state), let
$$G_h(s) := \begin{cases} \dfrac{L}{2} + \gamma, & h \leq H_e, \\ 0, & h > H_e. \end{cases}$$

Definition of the hard MDP $\mathcal{M}_c$. Define the set of hard time-state-action triples as

$$\mathcal{C} := \{(h, x_i, a) : d_0 < h \leq H_e,\ i \in [m],\ a \neq a_w\}.$$

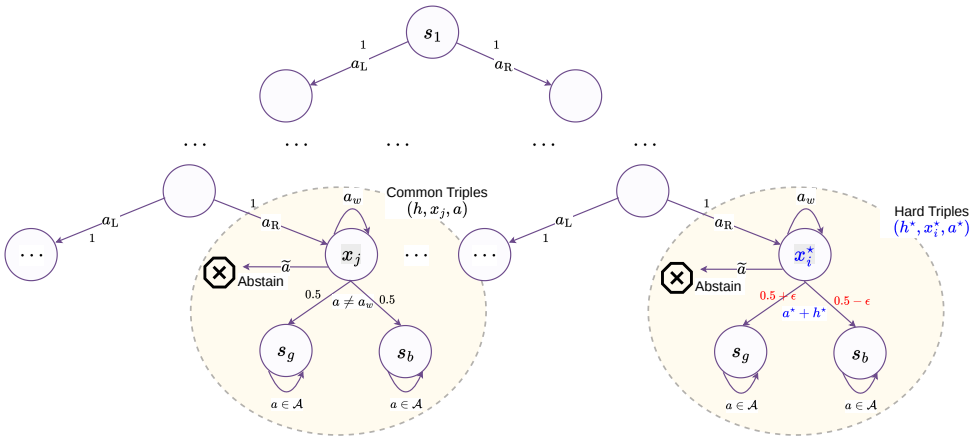


Figure 2: Illustration of a hard instance

Here m is the number of waiting-state leaves, and the number of candidate time steps is $H_e - d_0$. For each candidate cell $c \in \mathcal{C}$, let

$$T_c := \sum_{k=1}^{K} \mathbf{1}\{\text{the agent interacts at } c \text{ in episode } k\}$$

denote the total number of times the agent samples c over the K episodes. Since c contains a fixed stage index, it can be sampled at most once per episode, and hence $0 \le T_c \le K$.

For every $c = (h^*, x_{i^*}, a^*) \in \mathcal{C}$, we define an alternative MDP $\mathcal{M}_c$ through $\mathcal{M}_0$. The only difference between $\mathcal{M}_c$ and $\mathcal{M}_0$ is the transition kernel at the triple c. Specifically,

$$\mathbb{P}_{h^*}[s_g \mid x_{i^*}, a^*] = \frac{1}{2} + \epsilon, \quad \text{and} \quad \mathbb{P}_{h^*}[s_b \mid x_{i^*}, a^*] = \frac{1}{2} - \epsilon,$$

where $\epsilon := 2\gamma/L$. Since $\gamma \le L/8$, we have $\epsilon \le 1/4$.

All other transition probabilities, rewards, and fallback rewards remain identical to those of $\mathcal{M}_0$.

Properties of the hard instances. Write $b = L/2 + \gamma$. For every candidate $c = (h, x_i, a)$, $h \le H_e$ implies $H - h \ge L$, so

$$V^\star_{h+1}(s_g) = L, \qquad V^\star_{h+1}(s_b) = 0.$$

Under $\mathcal{M}_0$, every active-window tree state has optimal value b, and

$$Q^\star_{h,\mathcal{M}_0}(x_i, a) = L/2, \qquad V^\star_{h,\mathcal{M}_0}(x_i) - Q^\star_{h,\mathcal{M}_0}(x_i, a) = \gamma.$$

Under $\mathcal{M}_c$, gambling at the planted triple has value

$$Q^\star_{h^*,\mathcal{M}_c}(x_{i^*}, a^*) = L/2 + 2\gamma = b + \gamma.$$

An active-window tree state has value $b + \gamma$ if the planted triple is still reachable from that state-time pair, and value b otherwise. After the window, every tree state has value zero. In particular,

$$V^\star_{1,\mathcal{M}_0}(s_1) = b, \qquad V^\star_{1,\mathcal{M}_c}(s_1) = b + \gamma.$$

These statements follow by backward induction, since waiting and routing are deterministic and all base rewards are zero.

We next verify the minimum positive gap over *all* base actions. Under $\mathcal{M}_0$, active-window gambling gaps equal γ; routing and waiting gaps are either zero or b at the window boundary. Under $\mathcal{M}_c$, ordinary gambling gaps are γ or 2γ, the planted gambling gap is zero, and deterministic routing or waiting gaps belong to $\{0, \gamma, b, b + \gamma\}$. At s_g, continuation gaps are zero or one; at s_b and dummy states they are zero. Hence every positive base-action gap is at least γ, since $\gamma \le 1$. An ordinary gamble at a leaf other than the planted leaf has gap exactly γ; such a leaf exists because $m \ge 2$. Consequently,

$$\gamma_{\min}(\mathcal{M}_0) = \gamma_{\min}(\mathcal{M}_c) = \gamma \qquad \text{for every } c \in \mathcal{C}.$$

The rewards and fallbacks are known and identical across this family. All fallbacks lie in $[0, H-h+1]$: at active tree states, $b \le L/2+1 \le L \le H-h+1$, and the outcome-state bounds hold by definition.

Finally, one episode visits at most one candidate, since each gambling action leads to an absorbing state. Thus, with $T = \sum_{c\in\mathcal{C}} T_c$,

$$T \le K, \qquad M := |\mathcal{C}| = m(A-1)(H_e - d_0) \ge SAH/64, \qquad M \ge 8.$$

Here we used $m \ge S/8$, $A-1 \ge A/2$, and $H_e - d_0 \ge H/4$.

E.2 Auxiliary Lemmas

For $p, q \in (0,1)$, let

$$\mathrm{kl}(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$$

denote the Bernoulli KL divergence. Then we have the following bound

Lemma 10 (Bernoulli KL upper bound)**.** *For every $p, q \in (0,1)$, $\mathrm{kl}(p,q) \le \frac{(p-q)^2}{q(1-q)}$. In particular, for $0 < \varepsilon \le 1/4$,*

$$\mathrm{kl}\left(\frac{1}{2}, \frac{1}{2}+\varepsilon\right) \le 3\varepsilon^2.$$

Proof. Using $\log x \le x - 1$ on the two terms of the Bernoulli KL divergence gives

$$\mathrm{kl}(p,q) \le p\left(\frac{p}{q} - 1\right) + (1-p)\left(\frac{1-p}{1-q} - 1\right) = \frac{(p-q)^2}{q(1-q)}.$$

For $p = 1/2$ and $q = 1/2 + \varepsilon$,

$$\mathrm{kl}\left(\frac{1}{2}, \frac{1}{2}+\varepsilon\right) = -\frac{1}{2}\log(1-4\varepsilon^2).$$

Since $-\log(1-x) \le x/(1-x)$ and $1 - 4\varepsilon^2 \ge 3/4$ for $\varepsilon \le 1/4$,

$$\mathrm{kl}\left(\frac{1}{2}, \frac{1}{2}+\varepsilon\right) \le \frac{2\varepsilon^2}{1-4\varepsilon^2} \le 3\varepsilon^2.$$

□

We will also use the following Bretagnolle–Huber inequality (Lattimore & Szepesvári, 2020).

Lemma 11 (Bretagnolle–Huber)**.** *For probability measures $\mathbb{P}_0, \mathbb{P}_1$ on a common space and any event $\mathcal{A}$, $\mathbb{P}_0(\mathcal{A}) + \mathbb{P}_1(\mathcal{A}^c) \ge \frac{1}{2}\exp(-D_{\mathrm{KL}}(\mathbb{P}_0, \mathbb{P}_1))$.*

The next is the standard KL decomposition for adaptive experiments.

Lemma 12 (Adaptive KL decomposition)**.** *Consider two MDPs $\mathcal{M}_0$ and $\mathcal{M}_c$ that differ only in the transition kernel at a single state-action-time triple c. Suppose that one interaction at c reveals a Bernoulli observation with parameter p_c^0 under $\mathcal{M}_0$ and parameter p_c^1 under $\mathcal{M}_c$. Let $\mathbb{P}_0$ and $\mathbb{P}_c$ denote the laws of the complete K-episode interaction transcript, and let T_c be the number of times the agent interacts at c. Then*

$$D_{\mathrm{KL}}(\mathbb{P}_0, \mathbb{P}_c) = \mathbb{E}_0[T_c]\,\mathrm{kl}(p_c^0, p_c^1).$$

Proof. Apply the chain rule for KL divergence along the complete interaction transcript. Conditional on the same past history, the agent uses the same decision rule under the two environments, and all rewards, fallbacks, and transition kernels other than the one at c coincide. Hence all conditional KL terms vanish except when c is actually sampled. Each such interaction contributes $\mathrm{kl}(p_c^0, p_c^1)$. Taking expectation under $\mathcal{M}_0$ gives the result. □

Lemma 13 (Pinsker's inequality)**.** *For any two probability measures $\mathbb{P}_1, \mathbb{P}_2$, and event E, we have*

$$\mathrm{TV}(\mathbb{P}_1, \mathbb{P}_2) := \sup_E |\mathbb{P}_1(E) - \mathbb{P}_2(E)| \le \sqrt{D_{\mathrm{KL}}(\mathbb{P}_1, \mathbb{P}_2)/2}.$$

E.3 Proof of Theorem 4

Proof of Theorem 4. Fix an algorithm. The assumed sample-size condition implies $K \geq 8$. Use the construction and candidate counts above, with $T = \sum_c T_c$. Conditional on the deterministic policy selected for an episode, its path before the first gambling action is deterministic. Thus whether it samples c is determined by that policy, before observing any random outcome. Under $\mathcal{M}_0$, any policy that gambles has value at most $L/2 = b - \gamma$. Under $\mathcal{M}_c$, any policy that does not sample c has value at most b, whereas the optimal value is $b + \gamma$. Hence the following inequalities hold pathwise under the respective environments:

$$\mathrm{Reg}_{\mathcal{M}_0}(K) \geq \gamma T, \qquad \mathrm{Reg}_{\mathcal{M}_c}(K) \geq \gamma(K - T_c). \tag{14}$$

Write $\mathbb{P}_0, \mathbb{P}_c$ for the respective transcript laws and suppress the instance subscript on regret when the probability measure is specified.

Step 1: Define an exploration budget and split into two cases. Let

$$n := \left\lfloor \frac{ML^2}{384\gamma^2} \right\rfloor.$$

Since $M \geq 8$, $L \geq 8$, and $\gamma \leq 1$, the expression inside the floor is at least one. Also $M \leq SAH$ and $L \leq H$, so the assumed sample-size condition gives

$$1 \leq n \leq \frac{ML^2}{384\gamma^2} \leq \frac{SAH^3}{384\gamma^2} < \frac{K}{384} \leq \frac{K}{8}.$$

If $\mathbb{P}_0(T \geq n) \geq 1/4$, the first inequality in Eqn. (14) immediately gives

$$\mathbb{P}_0(\mathrm{Reg}(K) \geq \gamma n) \geq 1/4,$$

which proves the theorem. It remains to consider what happens if $\mathbb{P}_0(T \geq n) < 1/4$.

Define an auxiliary algorithm that follows the original algorithm until it has sampled n candidates in total. After completing that episode, it abstains at the root in every remaining episode. Write $\overline{\mathbb{P}}_0, \overline{\mathbb{P}}_c$ for its corresponding laws, and $\overline{T}_c, \overline{T}$ for its counts. The original and auxiliary algorithms can be coupled to agree until the budget is reached. In particular,

$$\overline{T} \leq n, \qquad \overline{\mathbb{P}}_0(\overline{T} \geq n) = \mathbb{P}_0(T \geq n) < 1/4.$$

Since $\mathbb{E}_{\overline{\mathbb{P}}_0} \overline{T} = \sum_c \mathbb{E}_{\overline{\mathbb{P}}_0} \overline{T}_c \leq n$, there exists c with $\mathbb{E}_{\overline{\mathbb{P}}_0} \overline{T}_c \leq n/M$. Applying Lemmas 10 and 12 to the auxiliary algorithm,

$$D_{\mathrm{KL}}(\overline{\mathbb{P}}_0, \overline{\mathbb{P}}_c) \leq \frac{12\gamma^2}{L^2} \mathbb{E}_{\overline{\mathbb{P}}_0} \overline{T}_c \leq \frac{12\gamma^2 n}{ML^2} \leq \frac{1}{32}.$$

By Lemma 13, $\mathrm{TV}(\overline{\mathbb{P}}_0, \overline{\mathbb{P}}_c) \leq 1/8$.

Step 2: Transfer a failure event to the alternative. Consider

$$\mathcal{F} := \{\overline{T} < n,\ \overline{T}_c < K/2\}.$$

With $\mathbb{E}_{\overline{\mathbb{P}}_0} \overline{T}_c \leq n/M$, Markov's inequality gives

$$\overline{\mathbb{P}}_0(\overline{T}_c \geq K/2) \leq \frac{2n}{MK} \leq \frac{1}{4M} \leq \frac{1}{32}.$$

Therefore

$$\overline{\mathbb{P}}_0(\mathcal{F}) \geq 1 - \frac{1}{4} - \frac{1}{32} = \frac{23}{32}, \qquad \overline{\mathbb{P}}_c(\mathcal{F}) \geq \frac{23}{32} - \frac{1}{8} = \frac{19}{32},$$

where the second inequality follows from $\mathrm{TV}(\overline{\mathbb{P}}_0, \overline{\mathbb{P}}_c) \leq 1/8$. On $\mathcal{F}$, truncation never occurs. The restrictions of the original and auxiliary laws to the event of never reaching the budget are identical under each environment. Hence

$$\mathbb{P}_c(T < n,\ T_c < K/2) = \overline{\mathbb{P}}_c(\mathcal{F}) \geq 19/32.$$

On this event the second inequality in Eqn. (14) gives $\mathrm{Reg}(K) \geq \gamma(K - T_c) > \gamma K/2 \geq \gamma n$. Together with the first case, this proves that some member of the family satisfies

$$\mathbb{P}_{\mathcal{M}}(\mathrm{Reg}(K) \geq \gamma n) \geq 1/4.$$

Step 3: Express the threshold in the problem parameters. For $x \geq 1$, $\lfloor x \rfloor \geq x/2$. Consequently,

$$\gamma n \geq \frac{ML^2}{768\gamma} \geq \frac{SAH^3}{196608\gamma},$$

where we used $M \geq SAH/64$ and $L \geq H/2$. This proves Eqn. (3). □

F Proof of the EHL Lower Bound

F.1 Construction of Hard MDP Instance

We reuse the state space, base action set, initial state, and binary routing tree from Appendix E.1, including the m leaves $\{x_1, \dots, x_m\}$ at depth d_0, the waiting action a_w, and the absorbing outcome states s_g, s_b. For the present construction, we replace the active-window endpoint H_e with ℓ and redefine the fallback values and transition perturbation as specified below.

Parameters. Fix an integer ℓ satisfying $2(d_0+1) \le \ell \le H/2$, and define

$$W := \frac{3}{4}(H-\ell).$$

The candidate time-state-action triples are

$$\mathcal{B}_\ell := \{(h, x_i, a) : \lfloor \ell/2 \rfloor < h \le \ell,\ i \in [m],\ a \in \mathcal{A} \setminus \{a_w\}\}.$$

Their number satisfies

$$M := |\mathcal{B}_\ell| = m(A-1)\lceil \ell/2 \rceil \ge \frac{SA\ell}{32}.$$

Assume $M \ge 8$ and $K \ge MH^2/\ell^2$, and set the MDP difference ε and fallback value coefficient θ as follows

$$\varepsilon := \frac{1}{4}\sqrt{\frac{M}{K}}, \qquad \theta := \frac{\varepsilon W}{3\ell}.$$

These choices ensure

$$0 < \varepsilon \le \frac{\ell}{4H} \le \frac{1}{4}, \qquad 0 < \theta \le \frac{1}{12}.$$

Components shared by all instances. The reference instance $\mathcal{M}_0$ and all alternative instances $\{\mathcal{M}_c : c \in \mathcal{B}_\ell\}$ share the following components.

- **Routing and waiting transitions.** At every stage, transitions at internal tree nodes follow the same deterministic routing rules as in Appendix E.1. At each leaf x_i, the waiting action a_w self-loops during $h \le \ell$, while every gambling action leads to either s_g or s_b. For $h > \ell$, every base action at a leaf leads deterministically to s_b. Both outcome states and all dummy states are absorbing; the dummy states are unreachable from s_1.
- **Base rewards.** All base rewards are zero:

$$r_h(s,a) = 0, \qquad (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}.$$

- **Fallback values.** For every tree state, including the leaves, set

$$G_h(s) = \frac{W}{2}\mathbf{1}\{h \le \ell\} + \theta(H-h+1).$$

  For the two outcome states, set

$$G_h(s_g) = \min\{W + \theta(H-h+1),\ H-h+1\}, \qquad G_h(s_b) = \theta(H-h+1).$$

  All dummy states have zero fallback values.

Thus, the instances differ only in their gambling success probabilities during the active window.

Reference instance $\mathcal{M}_0$. Under $\mathcal{M}_0$, every gambling action at every leaf succeeds with probability $1/2$ throughout the active window:

$$\mathbb{P}_{h,\mathcal{M}_0}(s_g \mid x_i, a) = \mathbb{P}_{h,\mathcal{M}_0}(s_b \mid x_i, a) = \frac{1}{2}, \qquad h \le \ell, \quad i \in [m], \quad a \ne a_w.$$

Together with the shared components above, this completely specifies $\mathcal{M}_0$.

Hard instances $\mathcal{M}_c$. For each candidate triple $c = (h_c, x_{i_c}, a_c) \in \mathcal{B}_\ell$, define $\mathcal{M}_c$ by modifying $\mathcal{M}_0$ only at c:

$$\mathbb{P}_{h_c,\mathcal{M}_c}(s_g \mid x_{i_c}, a_c) = \frac{1}{2} + \varepsilon, \qquad \mathbb{P}_{h_c,\mathcal{M}_c}(s_b \mid x_{i_c}, a_c) = \frac{1}{2} - \varepsilon.$$

All other transition probabilities, base rewards, and fallback values remain identical to those of $\mathcal{M}_0$.

The hard-instance family used in the theorem is

$$\mathfrak{M}_{\ell,K} := \{\mathcal{M}_c : c \in \mathcal{B}_\ell\}.$$

The reference instance $\mathcal{M}_0$ is not included in this family and is used only for the change-of-measure argument. The family depends on K through ε and θ. For fixed parameters, the known rewards and fallback values are identical across all instances and therefore reveal no information about the planted triple c.

Lemma 14 (Sub-optimality Gap)**.** *Fix a candidate triple $c = (h_c, x_{i_c}, a_c) \in \mathcal{B}_\ell$. Under $\mathcal{M}_c$, the optimal value at the initial state is*

$$V^\star_{1,\mathcal{M}_c}(s_1) = G_1(s_1) + \varepsilon W - \theta h_c.$$

For any episode policy π,

$$\begin{aligned} V^\star_{1,\mathcal{M}_c}(s_1) - V^\pi_{1,\mathcal{M}_c}(s_1) \geq (\varepsilon W - \theta h_c)\big(1 - \mathbb{P}^\pi_{\mathcal{M}_c}(c \textit{ is sampled})\big) \\ + \theta\left(\mathbb{E}^\pi_{\mathcal{M}_c}[\tau_\pi - 1] - h_c \mathbb{P}^\pi_{\mathcal{M}_c}(c \textit{ is sampled})\right). \end{aligned} \tag{15}$$

Both terms on the right-hand side are nonnegative. Every optimal policy from s_1 satisfies $\tau_\pi - 1 = h_c$ almost surely. Consequently, $d^\star(\mathcal{M}_c) = h_c$.

Proof. Set $G_{H+1} \equiv 0$. At every reachable base-action triple under $\mathcal{M}_c$,

$$(P_{h,\mathcal{M}_c} G_{h+1})(s,a) - G_h(s) \leq -\theta + \varepsilon W \mathbf{1}\{(h,s,a) = c\}. \tag{16}$$

Indeed, routing and waiting within the active window decrease the fallback value by θ, and a fair gamble decreases its expected value by θ. At the planted triple c, the increase in the success probability contributes an additional εW. Leaving the active window without gambling incurs an additional loss of $W/2$. At s_b, each interaction decreases the fallback value by θ; at s_g, the decrease is at least θ, including when the cap is active. These observations establish Eqn. (16).

Since all base rewards are zero, telescoping the fallback increments up to abstention gives

$$\begin{aligned} V^\pi_{1,\mathcal{M}_c}(s_1) - G_1(s_1) &= \mathbb{E}^\pi_{\mathcal{M}_c}\left[\sum_{h<\tau_\pi} \big((P_{h,\mathcal{M}_c} G_{h+1})(s_h,a_h) - G_h(s_h)\big)\right] \\ &\leq \varepsilon W \mathbb{P}^\pi_{\mathcal{M}_c}(c \text{ is sampled}) - \theta \mathbb{E}^\pi_{\mathcal{M}_c}[\tau_\pi - 1]. \end{aligned} \tag{17}$$

The convention $G_{H+1} \equiv 0$ also covers trajectories without abstention.

Sampling c requires at least h_c real interactions. Thus, pathwise,

$$\tau_\pi - 1 \geq h_c \mathbf{1}\{c \text{ is sampled}\},$$

and hence

$$\mathbb{E}^\pi_{\mathcal{M}_c}[\tau_\pi - 1] \geq h_c \mathbb{P}^\pi_{\mathcal{M}_c}(c \text{ is sampled}).$$

It follows that

$$\begin{aligned} V^\pi_{1,\mathcal{M}_c}(s_1) &\leq G_1(s_1) + (\varepsilon W - \theta h_c)\mathbb{P}^\pi_{\mathcal{M}_c}(c \text{ is sampled}) \\ &\leq G_1(s_1) + \varepsilon W - \theta h_c, \end{aligned}$$

where

$$\varepsilon W - \theta h_c \geq \varepsilon W - \theta \ell = \frac{2}{3}\varepsilon W > 0.$$

This upper bound is attained by the policy that routes to x_{i_c}, waits until stage h_c, takes a_c, and abstains at stage $h_c + 1$, whose value is

$$\left(\frac{1}{2} + \varepsilon\right) W + \theta(H - h_c) = G_1(s_1) + \varepsilon W - \theta h_c.$$

Hence, this policy is optimal from s_1, and $V^\star_{1,\mathcal{M}_c}(s_1) = G_1(s_1) + \varepsilon W - \theta h_c$. Subtracting the preceding upper bound on $V^\pi_{1,\mathcal{M}_c}(s_1)$ in Eqn. (17) from this optimal value and rearranging gives Eqn. (15).

For any optimal policy π, the left-hand side of Eqn. (15) is zero. Both terms on the right are nonnegative, and $\varepsilon W - \theta h_c > 0$ and $\theta > 0$. Therefore,

$$\mathbb{P}^\pi_{\mathcal{M}_c}(c \text{ is sampled}) = 1, \qquad \mathbb{E}^\pi_{\mathcal{M}_c}[\tau_\pi - 1] = h_c.$$

Since sampling c implies $\tau_\pi - 1 \geq h_c$, we conclude that $\tau_\pi - 1 = h_c$ almost surely. This holds for every optimal policy from s_1, and hence $d^\star(\mathcal{M}_c) = h_c$. □

F.2 Proof of Theorem 5

Proof. The characterization of $d^\star$ follows from Lemma 14. Fix an arbitrary learning algorithm. As in the preceding proof, let T_c count the episodes in which the algorithm samples c. Here $\mathbb{P}_0, \mathbb{P}_c$ denote the distributions of the complete interaction history under $\mathcal{M}_0, \mathcal{M}_c$, respectively, and $\mathbb{E}_0$ denotes expectation under $\mathbb{P}_0$.

Since each episode samples at most one candidate,

$$\sum_{c \in \mathcal{B}_\ell} T_c \leq K.$$

Consequently, there exists $c^0 \in \mathcal{B}_\ell$ such that $\mathbb{E}_0[T_{c^0}] \leq K/M$. Fix this candidate. By Lemmas 10 and 12,

$$D_{\mathrm{KL}}(\mathbb{P}_0, \mathbb{P}_{c^0}) = \mathbb{E}_0[T_{c^0}]\,\mathrm{kl}\left(\frac{1}{2}, \frac{1}{2} + \varepsilon\right) \leq 3\varepsilon^2 \frac{K}{M} = \frac{3}{16}.$$

Let $\mathcal{E} = \{T_{c^0} \geq K/2\}$. Markov's inequality gives

$$\mathbb{P}_0(\mathcal{E}) \leq \frac{2\mathbb{E}_0[T_{c^0}]}{K} \leq \frac{2}{M} \leq \frac{1}{4}.$$

Applying Lemma 11, we first obtain that the probability of $\mathcal{E}^c$ is not too small.

$$\mathbb{P}_{c^0}(\mathcal{E}^c) \geq \frac{1}{2}\exp\{-D_{\mathrm{KL}}(\mathbb{P}_0, \mathbb{P}_{c^0})\} - \mathbb{P}_0(\mathcal{E}) \geq \frac{1}{2}e^{-3/16} - \frac{1}{4} \geq \frac{1}{7}.$$

Conditional on the deterministic policy selected for an episode, the path before its first gambling outcome is deterministic. Therefore, whether the episode samples c^0 is determined by the selected policy, and

$$\mathbb{P}^{\pi^k}_{\mathcal{M}_{c^0}}(c^0 \text{ is sampled}) = \mathbf{1}\{c^0 \text{ is sampled in episode } k\}.$$

Applying Eqn. (15) in Lemma 14 to each episode, summing over k, and dropping the nonnegative second term yields the pathwise bound

$$\mathrm{Reg}(K) \geq (\varepsilon W - \theta h_{c^0})(K - T_{c^0}).$$

On $\mathcal{E}^c$, we consequently have

$$\begin{aligned}
\mathrm{Reg}(K) &\geq \frac{(\varepsilon W - \theta h_{c^0})K}{2} \\
&\geq \frac{\varepsilon W K}{3} = \frac{W}{12}\sqrt{MK} \\
&\geq \frac{H\sqrt{SAK\ell}}{32\sqrt{32}} \geq \frac{H}{192}\sqrt{SAK\ell} \geq \frac{H}{192}\sqrt{SAKd^\star(\mathcal{M}_{c^0})},
\end{aligned}$$

where we used $\varepsilon W - \theta h_{c^0} \geq 2\varepsilon W/3$, $W \geq 3H/8$, and $M \geq SA\ell/32$. Together with $\mathbb{P}_{c^0}(\mathcal{E}^c) \geq 1/7$ and $d^\star(\mathcal{M}_{c^0}) \leq \ell$, it proves the theorem. □