UNDERSTANDING GAP-DEPENDENT REGRET FOR OPTIMISM-BASED REINFORCEMENT LEARNING WITH LINEAR FUNCTION APPROXIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study gap-dependent regret for reinforcement learning with linear function approximation. While prior works have established gap-dependent guarantees in this setting, existing analyses do not apply to algorithms that achieve the nearly minimax-optimal worst-case regret bound $\tilde{O}(d\sqrt{H^3K})$, where d is the feature dimension, H is the horizon length, and K is the number of episodes. We bridge this gap by establishing the first gap-dependent regret bound for the nearly minimax-optimal algorithm LSVI-UCB++ (He et al., 2023), with an expected regret bound $\tilde{O}(d^2H^3/\Delta_{\min}+d^6H^5)$, improving the dependence on both d and H in the leading gap-dependent term compared with previous results. To understand the exploration cost induced by optimism, we establish a structural lower bound $\Omega(d^2H^3/\Delta_{\min})$ for a broad class of algorithms based on persistent ellipsoidal optimism. When specialized to LSVI-UCB++, this result shows that the leading dependence of our upper bound on d, H, and $\Delta_{\min}$ is tight up to logarithmic factors. Beyond this algorithmic class, we establish a general gap-dependent lower bound $\Omega(dH^3\log K/\Delta_{\min})$ for arbitrary learning algorithms, showing that the logarithmic dependence on K and the cubic dependence on H are intrinsic to gap-dependent expected regret in linear MDPs. Together, our results substantially narrow the gap between upper and lower bounds and provide a sharper characterization of gap-dependent learning and optimism-based exploration with linear function approximation.

1 INTRODUCTION

Reinforcement learning (RL) (Sutton & Barto, 2018) provides a framework for sequential decision-making, where an agent tries to maximize cumulative rewards through interaction with an unknown environment. A central challenge is designing efficient algorithms for large state and action spaces. Linear function approximation addresses this by representing value functions using known features.

Recently, a large body of literature has studied regret upper bounds for RL with linear function approximation. These works can be categorized into two main classes. The first class considers the linear Markov decision process (MDP) (Jin et al., 2020), where the transition dynamics are assumed to be linear in the known features. In particular, Jin et al. (2020) proposed the first provably efficient algorithm, LSVI-UCB, which achieves the regret bound $\tilde{O}(\sqrt{d^3H^4K})$. Here, $\tilde{O}(\cdot)$ hides logarithmic factors, d denotes the feature dimension, H is the horizon length, and K is the number of episodes. Subsequently, He et al. (2023) proposed the LSVI-UCB++ algorithm, which leverages adaptive weighted ridge regression and pessimistic value estimation and improves the regret bound to the nearly minimax-optimal rate $\tilde{O}(d\sqrt{H^3K})$ (Zhou et al., 2021a). The second class focuses on linear mixture MDPs (Ayoub et al., 2020; Zhou et al., 2021a; Zhou & Gu, 2022), where the transition probability is modeled as a linear combination of multiple base models. Ayoub et al. (2020) proposed the UCRL-VTR algorithm with a regret bound $\tilde{O}(d\sqrt{H^4K})$. Later, Zhou et al. (2021a) developed a near-optimal algorithm for time-inhomogeneous linear mixture MDPs. Furthermore, Zhou & Gu (2022) proposed a near-optimal, horizon-free algorithm for time-homogeneous linear mixture MDPs.

In practice, RL algorithms often outperform their worst-case guarantees when a positive suboptimality gap exists (i.e., the optimal action at each state is better than suboptimal actions by a non-negligible margin). Gap-dependent analysis has been well studied in the tabular setting (Simchowitz & Jamieson,

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.

2019; Dann et al., 2021; Yang et al., 2021; Xu et al., 2021; Zheng et al., 2025b; Zhang et al., 2025b; 2026; Chen et al., 2025), and it has also gained traction in RL with linear function approximation. He et al. (2021) pioneered this direction, showing that LSVI-UCB (Jin et al., 2020) achieves an expected regret bound $\tilde{O}(d^3H^5/\Delta_{\min})$ under linear MDPs, and UCRL-VTR (Ayoub et al., 2020) attains an expected regret bound $\tilde{O}(d^2H^5/\Delta_{\min})$ under linear mixture MDPs, where the minimum gap $\Delta_{\min}$ is defined as the infimum of the positive suboptimality gap $\Delta_h(s,a)$ over all state-action-step triples (s,a,h). Papini et al. (2021) established a high-probability constant-regret guarantee under the UNISOFT condition, while Zhang et al. (2024a) obtained a high-probability constant-regret guarantee without this assumption. Both results imply expected regret of order $\tilde{O}(d^3H^5/\Delta_{\min})$. Taken together, these results show that RL with linear function approximation can achieve polylogarithmic regret.

Despite these advances, existing gap-dependent analyses do not yet cover algorithms that achieve the nearly minimax-optimal worst-case regret bound $\tilde{O}(d\sqrt{H^3K})$, e.g., LSVI-UCB++ (He et al., 2023) or UCRL-VTR+ (Zhou et al., 2021a). As a result, current gap-dependent regret bounds exhibit a loose dependence on the feature dimension d and horizon length H, i.e., $\tilde{O}(d^3H^5/\Delta_{\min})$ under linear MDPs or $\tilde{O}(d^2H^5/\Delta_{\min})$ under linear mixture MDPs, failing to reflect the superior worst-case efficiency of nearly minimax-optimal algorithms. Moreover, most existing algorithms with linear function approximation, including LSVI-UCB and UCRL-VTR, rely on optimism-based exploration, which encourages exploration through uncertainty-dependent bonuses. However, it remains unclear how much gap-dependent regret is intrinsically induced by such optimism-based exploration mechanisms and whether a nearly minimax-optimal optimism-based algorithm can attain a matching upper bound. Consequently, these gaps lead to the following open question:

What gap-dependent regret is unavoidable under optimism-based exploration, and can a nearly minimax-optimal optimism-based algorithm achieve a matching upper bound?

Improving the dependence on feature dimension d and horizon length H in regret bounds is not only of theoretical interest but also has practical implications. In domains such as robotics and healthcare, agents operate in high-dimensional environments, where function approximation is essential for learning and control. For instance, in robotic manipulation and grasping tasks, the state space consists of high-dimensional continuous observations, including robot joint configurations, object positions, and environmental variables (Toner et al., 2023). Similarly, healthcare applications such as optimizing cell growth conditions or sequential treatment planning involve biological state representations and long decision sequences, making function approximation critical for deployment (Al-Hamadani et al., 2024). In these applications, improved dependence on d and H provides a sharper characterization of the sample efficiency of RL algorithms in high-dimensional, long-horizon tasks.

In this paper, we answer the question above by establishing complementary upper and lower bounds. On the upper-bound side, we establish an improved gap-dependent regret bound for the nearly minimax-optimal LSVI-UCB++. Our bound achieves the best-known dependence on d and H in the leading gap-dependent term for linear function approximation. On the lower-bound side, we first study persistent ellipsoidal optimism, where an uncertainty-dependent ellipsoidal bonus remains active throughout learning, and establish a structural lower bound for this exploration mechanism. We then specialize the result to LSVI-UCB++, showing that the leading dependence of our upper bound on d, H, and $\Delta_{\min}$ is tight up to logarithmic factors. We also establish a general lower bound characterizing the fundamental difficulty of gap-dependent learning in linear MDPs. Together, these results narrow the gap between upper and lower bounds and provide a sharper understanding of gap-dependent learning and optimism-based exploration with linear function approximation.

Our contributions. Our contributions are summarized as follows:

(i) **Improved gap-dependent regret bound and a novel random-suffix framework:** We establish the first gap-dependent regret bound for the nearly minimax-optimal LSVI-UCB++ algorithm with linear function approximation. Specifically, we prove an expected regret upper bound of $\tilde{O}(d^2H^3/\Delta_{\min} + d^6H^5)$. Compared with existing gap-dependent guarantees, our leading term $\tilde{O}(d^2H^3/\Delta_{\min})$ substantially improves the dependence on both the feature dimension d and the horizon length H (see Table 1). In particular, when $1/\Delta_{\min} = \tilde{\Omega}(d^4)$, our bound improves on all existing gap-dependent regret guarantees in Table 1. To obtain this bound, we develop a novel random-suffix framework for gap-dependent regret analysis in linear MDPs, presented in Section 4. The framework localizes regret and variance analysis to random trajectory suffixes beginning at the first realized

positive-gap step. Our numerical experiments in Appendix B suggest approximately logarithmic regret growth of LSVI-UCB++ and show lower empirical regret than the baselines considered.

(ii) **A structural lower bound for persistent ellipsoidal optimism and a matching lower bound for LSVI-UCB++:** To understand whether the leading dependence of our upper bound can be improved, we study a broad class of algorithms based on persistent ellipsoidal optimism defined in Definition H.1. We establish a structural lower bound for this algorithmic class in Theorem 3.2, showing that, for sufficiently large episode budgets, every algorithm in the class can incur expected regret of order $\Omega(d^2H^3/\Delta_{\min})$. We further verify in Appendix H.7 that LSVI-UCB++ satisfies the required structural conditions, yielding an $\Omega(d^2H^3/\Delta_{\min})$ lower bound that matches the leading $\tilde{O}(d^2H^3/\Delta_{\min})$ term of our upper bound up to logarithmic factors.

(iii) **Improved general gap-dependent regret lower bound:** We also establish an improved gap-dependent expected regret lower bound $\Omega\left(dH^3 \log K/\Delta_{\min}\right)$ for arbitrary learning algorithms in linear MDPs, as shown in Theorem 3.3. Compared with the prior lower bound $\Omega(dH/\Delta_{\min})$ (He et al., 2021), our result shows that the logarithmic dependence on the number of episodes K in the expected regret is unavoidable. More importantly, it reveals an intrinsic horizon dependence of order H^3. This provides a stronger benchmark for gap-dependent regret in linear MDPs and shows that the cubic dependence on H in our upper bound is unavoidable even beyond optimism-based algorithms. The remaining gap between this general lower bound and the leading gap-dependent term of our upper bound is therefore a factor of d, up to logarithmic factors.

Algorithm	$\mathbb{E}[\textbf{Regret}(T)]$
LSVI-UCB	$\tilde{O}(d^3H^5/\Delta_{\min})$ (He et al., 2021)
UCRL-VTR	$\tilde{O}(d^2H^5/\Delta_{\min})$ (He et al., 2021)
LSVI-UCB	$\tilde{O}(d^3H^5/\Delta_{\min})$ (Papini et al., 2021)
Cert-LSVI-UCB	$\tilde{O}(d^3H^5/\Delta_{\min})$ (Zhang et al., 2024a)
LSVI-UCB++	$\tilde{O}(d^2H^3/\Delta_{\min} + d^6H^5)$ (Ours)

Table 1: Comparison of gap-dependent expected regret bounds for different RL algorithms with linear function approximation. We use T to denote the total number of steps, d as the feature dimension, H as the horizon length, and $\Delta_{\min}$ as the minimum gap. The notation $\tilde{O}(\cdot)$ hides logarithmic factors.

(iv) **Technical novelty:** Our three main results rely on distinct technical ideas.

Upper bound. The proof of Theorem 3.1 develops a first-positive-gap and random-suffix analysis. Lemma G.1 reduces the expected regret to the first realized positive-gap step of each suboptimal episode. To analyze the resulting random suffixes, Lemma F.3 introduces a surrogate covariance sequence for partial bonus sums, while Lemmas F.4 and F.5 control the associated value-estimation errors. These tools lead to the random-suffix variance bound in Lemma G.2 and the self-bounding argument in Lemma G.3, yielding the leading $\tilde{O}(d^2H^3/\Delta_{\min})$ regret.

Structural lower bound for persistent ellipsoidal optimism. For Theorem 3.2, we construct a hard linear MDP with action-specific unknown transition directions across $\Theta(H)$ steps (Appendix H.2). Full-domain Bellman calibration bounds the total weighted squared regression error along these directions (Appendix H.3), forcing the learner to either repeatedly select suboptimal actions or accurately estimate many transition directions (Appendix H.5). A finite-family testing argument then yields the $\Omega(d^2H^3/\Delta_{\min})$ regret lower bound (Appendix H.6).

General lower bound. For Theorem 3.3, we construct a hard linear MDP with statistically difficult alternatives across feature directions and steps. In each pair, the same action is suboptimal under one alternative but optimal under the other. Sampling it sufficiently often incurs regret under the former, while insufficient sampling prevents reliable identification and incurs regret under the latter. A change-of-measure argument formalizes this tradeoff and yields the lower bound $\Omega(dH^3 \log K/\Delta_{\min})$.

2 PRELIMINARIES

Notation. For any $C \in \mathbb{N}_+$, we write $[C] := \{1, 2, \ldots, C\}$. We denote by $\mathbb{I}[x]$ the indicator function of the event x. For a vector $\mathbf{x} \in \mathbb{R}^d$ and a positive definite matrix $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$, we use $\|\mathbf{x}\|_2$ to denote

the Euclidean norm, and $\|\mathbf{x}\|_{\mathbf{\Sigma}} := \sqrt{\mathbf{x}^\top \mathbf{\Sigma} \mathbf{x}}$. For any $a, b \in \mathbb{R}$ with $a < b$, $x \in \mathbb{R}$, let $[x]_{[a,b]}$ denote the truncation function $a \cdot \mathbb{I}[x \le a] + x \cdot \mathbb{I}[a < x < b] + b \cdot \mathbb{I}[b \le x]$.

Episodic Markov decision processes. An episodic MDP is denoted by $\mathcal{M} := (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$, where $\mathcal{S}$ is a possibly infinite measurable state space, $\mathcal{A}$ is the finite set of actions, H is the number of steps in each episode, $\mathbb{P} := \{\mathbb{P}_h\}_{h=1}^H$ is the transition kernel so that $\mathbb{P}_h(\cdot \mid s, a)$ characterizes the distribution over the next state given the state-action pair (s, a) at step h, and $r := \{r_h\}_{h=1}^H$ is the collection of reward functions. Following He et al. (2023), we assume that the reward functions $r_h(s, a) \in [0, 1]$ are deterministic and known to the learner for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$.

In each episode, an adversary selects the initial state s_1. At each step $h \in [H]$, an agent observes a state $s_h \in \mathcal{S}$, picks an action $a_h \in \mathcal{A}$, receives the reward $r_h = r_h(s_h, a_h)$ and then transitions to the next state s_{h+1}. The episode terminates at the terminal state s_{H+1} after H steps.

For convenience, we denote $\mathbb{P}_{s,a,h} f = \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)} f(s')$, $\mathbb{1}_s f = f(s)$ and $\mathbb{V}_{s,a,h}(f) = \mathbb{P}_{s,a,h} f^2 - (\mathbb{P}_{s,a,h} f)^2$ for any function $f : \mathcal{S} \to \mathbb{R}$ and state-action-step triple (s, a, h).

Policy and value functions. A policy π is a collection of H functions $\{\pi_h : \mathcal{S} \to \Delta^{\mathcal{A}}\}_{h \in [H]}$, where $\Delta^{\mathcal{A}}$ is the set of probability distributions over $\mathcal{A}$. A policy is deterministic if for any $s \in \mathcal{S}$, $\pi_h(s)$ concentrates all the probability mass on an action $a \in \mathcal{A}$. In this case, we denote $\pi_h(s) = a$.

Denote state value functions by $V_h^\pi(s) := \sum_{h'=h}^H \mathbb{E}_{(s_{h'}, a_{h'}) \sim (\mathbb{P}, \pi)} [r_{h'}(s_{h'}, a_{h'}) \mid s_h = s]$ and action value functions by $Q_h^\pi(s, a) := r_h(s, a) + \sum_{h'=h+1}^H \mathbb{E}_{(s_{h'}, a_{h'}) \sim (\mathbb{P}, \pi)} [r_{h'}(s_{h'}, a_{h'}) \mid s_h = s, a_h = a]$. Since the action space and the horizon are finite, there exists an optimal policy $\pi^\star$ that achieves the optimal value $V_h^\star(s) = \sup_\pi V_h^\pi(s) = V_h^{\pi^\star}(s)$ for all $(s, h) \in \mathcal{S} \times [H]$ (Azar et al., 2017). Then for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, the Bellman equation can be expressed as ($V_{H+1}^\pi(s) = V_{H+1}^\star(s) = 0$):

$$\begin{cases} V_h^\pi(s) = \mathbb{E}_{a' \sim \pi_h(s)}[Q_h^\pi(s, a')] \\ Q_h^\pi(s, a) := r_h(s, a) + \mathbb{P}_{s,a,h} V_{h+1}^\pi \end{cases} \text{ and } \begin{cases} V_h^\star(s) = \max_{a' \in \mathcal{A}} Q_h^\star(s, a') \\ Q_h^\star(s, a) := r_h(s, a) + \mathbb{P}_{s,a,h} V_{h+1}^\star. \end{cases} \tag{1}$$

For any algorithm over K episodes, let π^k be the policy used in the k-th episode, and s_1^k be the corresponding initial state. The regret over $T = HK$ steps is $\text{Regret}(T) := \sum_{k=1}^K (V_1^\star - V_1^{\pi^k})(s_1^k)$.

Suboptimality gap. For any given MDP, we define the suboptimality gap as follows.

Definition 2.1. For any (s, a, h), the suboptimality gap is defined as $\Delta_h(s, a) := V_h^\star(s) - Q_h^\star(s, a)$.

Equation (1) implies that for any (s, a, h), $\Delta_h(s, a) \ge 0$. Then, it is natural to define the minimum gap, which is the infimum of the strictly positive suboptimality gaps.

Definition 2.2. We define the **minimum gap** as $\Delta_{\min} := \inf_{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H], \Delta_h(s,a) > 0} \Delta_h(s, a)$.

We assume that $\Delta_{\min} > 0$ in this paper. Definitions 2.1 and 2.2 and the assumption of $\Delta_{\min} > 0$ are standard in the literature (Simchowitz & Jamieson, 2019; Dann et al., 2021; Yang et al., 2021; Xu et al., 2021; He et al., 2021; Zhang et al., 2024a; Zheng et al., 2025b; Zhang et al., 2026; 2025b).

Linear Markov decision processes. In this work, we focus on the linear Markov decision process (Jin et al., 2020; He et al., 2021; 2023), which is formally defined as follows:

Definition 2.3. An episodic MDP $\mathcal{M} := (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$ is a linear MDP with a known feature mapping $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, if for any $h \in [H]$, there exist d unknown signed measures $\boldsymbol{\theta}_h = (\boldsymbol{\theta}_h^1, ..., \boldsymbol{\theta}_h^d)$ over $\mathcal{S}$, such that for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have $\mathbb{P}_h(\cdot|s, a) = \langle \boldsymbol{\phi}(s, a), \boldsymbol{\theta}_h(\cdot) \rangle$.

For simplicity, we assume that $\|\phi(s, a)\|_2 \le 1$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $\|\boldsymbol{\theta}_h(B)\|_2 \le \sqrt{d}$ for all measurable $B \subseteq \mathcal{S}$ and $h \in [H]$. For linear MDPs, we have the following property:

Proposition 2.4 (Proposition 3.3 of He et al. (2023))**.** *For any policy π, there exist $\{\mathbf{w}_h^\pi\}_{h=1}^H$ such that for any state-action-step triple $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, we have $\mathbb{P}_{s,a,h} V_{h+1}^\pi = \langle \boldsymbol{\phi}(s, a), \mathbf{w}_h^\pi \rangle$.*

3 Theoretical Guarantees

3.1 Algorithm Review

We begin by reviewing a slightly modified version of LSVI-UCB++ (He et al., 2023) (see Appendix D for the two modifications) in Algorithm 1. LSVI-UCB++ reduces the problem of learning the optimal

value functions to a sequence of linear regression problems. Based on Proposition 2.4, Algorithm 1 constructs the estimator $\hat{\mathbf{w}}_{k,h}$ by solving the following weighted ridge regression

$$\hat{\mathbf{w}}_{k,h} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \lambda\|\mathbf{w}\|_2^2 + \sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \left(\mathbf{w}^\top \boldsymbol{\phi}(s_h^i, a_h^i) - V_{k,h+1}(s_{h+1}^i)\right)^2.$$

Here, the adjusted estimated variance $\bar{\sigma}_{k,h}^2$ is set as $\bar{\sigma}_{k,h}^2 = \max\{\sigma_{k,h}^2, H, 2d^3H^2\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}\}$, where the estimated variance $\sigma_{k,h}^2$ is defined in Equation (2) below.

With exploration bonuses, Line 11 of Algorithm 1 constructs an optimistic estimate $Q_{k,h}(s,a)$ as

$$Q_{k,h}(s,a) \approx r_h(s,a) + \hat{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s,a) + \beta\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}.$$

Here, $Q_{k,h}(s,a)$ is an upper bound on $Q_h^\star(s,a)$ with high probability when the hyperparameter β is chosen as $\tilde{\Theta}(\sqrt{d})$. The remaining components in Line 11 ensure its boundedness and monotonicity.

The algorithm also uses a pessimistic estimate $\check{Q}_{k,h}(s,a)$, which serves as a lower bound for $Q_h^\star(s,a)$ with high probability. By Equation (1) and Line 17 of Algorithm 1, this estimate allows us to control the value estimation error $V_{k,h}(s) - V_h^\star(s)$ via the error $V_{k,h}(s) - \check{V}_{k,h}(s)$. The pessimistic estimate $\check{Q}_{k,h}(s,a)$ is constructed in a manner analogous to the optimistic one. We first obtain the vector $\check{\mathbf{w}}_{k,h}$ by solving the following weighted ridge regression:

$$\check{\mathbf{w}}_{k,h} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \lambda\|\mathbf{w}\|_2^2 + \sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \left(\mathbf{w}^\top \boldsymbol{\phi}(s_h^i, a_h^i) - \check{V}_{k,h+1}(s_{h+1}^i)\right)^2.$$

We then compute $\check{Q}_{k,h}$ as follows: $\check{Q}_{k,h}(s,a) \approx r_h(s,a) + \check{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s,a) - \bar{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}$, where the hyperparameter $\bar{\beta}$ can be chosen as $\tilde{\Theta}(\sqrt{d^3H^2})$. Finally, LSVI-UCB++ constructs $\sigma_{k,h}^2$ as follows:

$$\sigma_{k,h}^2 = \bar{\mathbb{V}}_{s_h^k, a_h^k, h} V_{k,h+1} + E_{k,h} + D_{k,h} + H. \tag{2}$$

In (2), $\bar{\mathbb{V}}_{s_h^k, a_h^k, h} V_{k,h+1}$ denotes the estimated variance of the value function and is defined as

$$\left[\tilde{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s_h^k, a_h^k)\right]_{[0,H^2]} - \left(\left[\hat{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s_h^k, a_h^k)\right]_{[0,H]}\right)^2.$$

Here,

$$\tilde{\mathbf{w}}_{k,h} := \arg\min_{\mathbf{w}\in\mathbb{R}^d} \lambda\|\mathbf{w}\|_2^2 + \sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \left(\mathbf{w}^\top \boldsymbol{\phi}(s_h^i, a_h^i) - V_{k,h+1}^2(s_{h+1}^i)\right)^2$$

is the solution to the weighted ridge regression problem for the squared value function. In addition,

$$E_{k,h} = \min\left\{\tilde{\beta}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}, H^2\right\} + \min\left\{2H\bar{\beta}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}, H^2\right\},$$
$$D_{k,h} = \min\left\{4d^3H^2\left(\hat{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s_h^k, a_h^k) - \check{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s_h^k, a_h^k) + 2\bar{\beta}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}\right), d^3H^3\right\}.$$

Here, $E_{k,h}$ bounds the error between the estimated variance $\bar{\mathbb{V}}_{s_h^k, a_h^k, h} V_{k,h+1}$ and the true variance $\mathbb{V}_{s_h^k, a_h^k, h} V_{k,h+1}$ of $V_{k,h+1}$, and $D_{k,h}$ bounds the error between $\bar{\mathbb{V}}_{s_h^k, a_h^k, h} V_{k,h+1}$ and $\mathbb{V}_{s_h^k, a_h^k, h} V_{h+1}^\star$.

3.2 Gap-Dependent Regret Upper Bound

We now present our main result: the first gap-dependent regret upper bound for a nearly minimax-optimal algorithm in RL with linear function approximation. In Algorithm 1, for any confidence parameter $\delta \in (0,1)$, define the confidence radii $\beta = \beta(\delta)$, $\bar{\beta} = \bar{\beta}(\delta)$, and $\tilde{\beta} = \tilde{\beta}(\delta)$ as

$$\beta(\delta) = \Theta\Big(H\sqrt{d\lambda} + \sqrt{d\log^2\left(1 + dT/(\delta\lambda)\right)}\Big), \bar{\beta}(\delta) = \Theta\Big(H\sqrt{d\lambda} + \sqrt{d^3H^2\log^2\left(dT/(\delta\lambda)\right)}\Big),$$
$$\tilde{\beta}(\delta) = \Theta\Big(H^2\sqrt{d\lambda} + \sqrt{d^3H^4\log^2\left(dT/(\delta\lambda)\right)}\Big), \tag{3}$$

where the hidden constants are chosen to be sufficiently large universal constants.

Theorem 3.1. *For any linear MDP, if we set $\lambda = 1/H^2$ and the confidence radii as specified in Equation* (3) *with $\delta = 1/(19T)$, then the expected regret of Algorithm 1 satisfies*

$$\mathbb{E}\left[\mathrm{Regret}(T)\right] \le O\left(d^2H^3\iota_2^3/\Delta_{\min} + d^6H^5\iota_2^3\right),$$

where $\iota_2 = \log\left(1 + dHK\right)$.

Algorithm 1 LSVI-UCB++

Require: Regularization parameter $\lambda > 0$, confidence radii $\beta, \bar{\beta}, \tilde{\beta} > 0$. Episode number $K \in \mathbb{N}_+$.

Initialize $k_{\text{mid}}, k_{\text{last}} \leftarrow 0$, and for each step $h \in [H]$, set $\mathbf{\Sigma}_{0,h}, \mathbf{\Sigma}_{1,h} = \lambda \mathbf{I}_d$.
For each step $h \in [H]$ and state-action $(s,a) \in \mathcal{S} \times \mathcal{A}$, set $Q_{0,h}(s,a) = H, \check{Q}_{0,h}(s,a) = 0$.
for episode $k = 1, \ldots, K$ **do**
  Receive the initial state s_1^k.
  Set $V_{k,H+1}(s) = \check{V}_{k,H+1}(s) = 0$ for all $s \in \mathcal{S}$.
  for step $h = H, \ldots, 1$ **do**
    $\hat{\mathbf{w}}_{k,h} = \mathbf{\Sigma}_{k,h}^{-1} \sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) V_{k,h+1}(s_{h+1}^i)$.
    $\check{\mathbf{w}}_{k,h} = \mathbf{\Sigma}_{k,h}^{-1} \sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) \check{V}_{k,h+1}(s_{h+1}^i)$.
    $\widetilde{\mathbf{w}}_{k,h} = \mathbf{\Sigma}_{k,h}^{-1} \sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) V_{k,h+1}^2(s_{h+1}^i)$.
    if there exists $h' \in [H]$ such that $\det(\mathbf{\Sigma}_{k,h'}) \geq 2 \det(\mathbf{\Sigma}_{k_{\text{last}},h'})$, **then** for any $(s,a) \in \mathcal{S} \times \mathcal{A}$,
      $Q_{k,h}(s,a) = \min\left\{r_h(s,a) + \hat{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s,a) + \beta \|\boldsymbol{\phi}(s,a)\|_{\mathbf{\Sigma}_{k,h}^{-1}}, Q_{k-1,h}(s,a), H\right\}$,
      $\check{Q}_{k,h}(s,a) = \max\left\{r_h(s,a) + \check{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s,a) - \bar{\beta} \|\boldsymbol{\phi}(s,a)\|_{\mathbf{\Sigma}_{k,h}^{-1}}, \check{Q}_{k-1,h}(s,a), 0\right\}$.
      Set $k_{\text{mid}} \leftarrow k$.
    else
      $Q_{k,h}(s,a) = Q_{k-1,h}(s,a),\ \check{Q}_{k,h}(s,a) = \check{Q}_{k-1,h}(s,a), \forall (s,a) \in \mathcal{S} \times \mathcal{A}$.
    end if
    $V_{k,h}(s) = \max_a Q_{k,h}(s,a),\ \check{V}_{k,h}(s) = \max_a \check{Q}_{k,h}(s,a), \forall s \in \mathcal{S}$.
  end for
  Set the last updating episode $k_{\text{last}} \leftarrow k_{\text{mid}}$.
  for step $h = 1, \ldots, H$ **do**
    Take $a_h^k = \pi_h^k(s_h^k) = \arg\max_a Q_{k,h}(s_h^k, a)$, with ties broken by the smallest indexed action.
    Set the estimated variance $\sigma_{k,h}^2$ as in Equation (2).
    $\bar{\sigma}_{k,h}^2 = \max\left\{\sigma_{k,h}^2, H, 2d^3H^2 \|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\mathbf{\Sigma}_{k,h}^{-1}}\right\}$.
    $\mathbf{\Sigma}_{k+1,h} = \mathbf{\Sigma}_{k,h} + \bar{\sigma}_{k,h}^{-2} \boldsymbol{\phi}(s_h^k, a_h^k) \boldsymbol{\phi}(s_h^k, a_h^k)^\top$.
    Receive the next state s_{h+1}^k.
  end for
end for

The full proof is deferred to Appendix G. Compared with prior gap-dependent expected regret upper bounds for linear function approximation (He et al., 2021; Papini et al., 2021; Zhang et al., 2024a), our result improves the dependence on both the feature dimension d and the horizon length H in the leading gap-dependent term. In particular, while existing bounds scale as $\tilde{O}(d^3H^5/\Delta_{\min})$ for LSVI-UCB or $\tilde{O}(d^2H^5/\Delta_{\min})$ for UCRL-VTR (see Table 1), our bound improves the leading gap-dependent term to $\tilde{O}(d^2H^3/\Delta_{\min})$. The experiments in Appendix B further show lower empirical regret than other baselines and suggest approximately logarithmic regret growth.

3.3 A Structural Lower Bound for Persistent Ellipsoidal Optimism

We next study the exploration cost induced by persistent ellipsoidal optimism. We establish a structural lower bound for this exploration mechanism and then specialize the result to LSVI-UCB++, obtaining a lower bound that matches the leading gap-dependent term in Theorem 3.1 up to logarithmic factors.

We consider persistent ellipsoidal optimistic algorithms defined in Definition H.1. These algorithms use predictably weighted ridge regression to estimate Bellman targets and form optimistic candidates by adding an ellipsoidal bonus $\beta \|\phi(s,a)\|_{\Lambda_{u,h}^{-1}}$ whenever updating value estimates at episode u. At such an episode, the step-h covariance matrix is $\Lambda_{u,h} = \lambda I_d + \sum_{\ell<u} \alpha_{\ell,h} \phi(s_h^\ell, a_h^\ell) \phi(s_h^\ell, a_h^\ell)^\top$, where $\alpha_{\ell,h} > 0$ is the weight of the step-h sample from episode ℓ. Episode k uses the value estimates from the latest update, allowing for both every-episode and rare-switching updates.

The class captures a broad family of LSVI-style optimistic algorithms, allowing either unweighted or predictably weighted regression and either every-episode or rare-switching updates. Representative examples include LSVI-UCB and LSVI-UCB++, corresponding respectively to unweighted regression

with every-episode updates and predictable variance-based weighting with rare switching. More generally, the framework covers other variants satisfying the same structural conditions, while the underlying combination of ridge regression and ellipsoidal optimism also appears in classical optimistic linear-bandit methods such as OFUL (Abbasi-Yadkori et al., 2011).

We now quantify the exploration cost shared by this algorithmic class. Our structural lower bound shows that, for every algorithm in the class and a sufficiently large episode budget, there exists a linear MDP on which the algorithm incurs expected regret of at least order $d^2H^3/\Delta_{\min}$.

Theorem 3.2 (Informal)**.** *For any feature dimension $d \geq 15$, horizon $H \geq 16$, minimum gap $\Delta_{\min} \in (0, H/16]$, and any persistent ellipsoidal optimistic algorithm satisfying Definition H.1, if the episode number K satisfies the sufficient-budget condition in Theorem H.3, then there exists a d-dimensional linear MDP with minimum gap $\Delta_{\min}$ such that*

$$\mathbb{E}[\mathrm{Regret}(T)] \geq \Omega\left(\frac{d^2H^3}{\Delta_{\min}}\right).$$

The formal statement and complete proof are given in Theorem H.3 and Appendix H, respectively. The theorem requires the episode budget K to exceed an explicit threshold that depends polynomially on the problem parameters and on the algorithmic quantities governing the regression weights, regularization, and confidence radius. The proof constructs a hard linear MDP with multiple action-specific unknown transition directions across $\Theta(H)$ steps (Appendix H.2). At each such step, the learner must either continue selecting sufficiently many suboptimal actions or estimate many unknown transition directions to sufficiently high accuracy (Appendix H.5); a finite-family testing argument then converts this exploration–estimation tradeoff into the stated lower bound (Appendix H.6).

Theorem 3.2 identifies the $d^2H^3/\Delta_{\min}$ dependence as an unavoidable exploration cost for the considered class of persistent ellipsoidal optimistic algorithms. The quadratic dependence on d is consistent with gap-dependent results for ellipsoidal optimism in linear bandits, such as OFUL (Abbasi-Yadkori et al., 2011), whose leading inverse-gap term is also quadratic in d up to logarithmic factors. In linear MDPs, this exploration cost propagates across the horizon, yielding the additional H^3 dependence. We further verify in Section H.7 that LSVI-UCB++ satisfies the required conditions, implying an $\Omega(d^2H^3/\Delta_{\min})$ lower bound for LSVI-UCB++ itself. This matches the leading $d^2H^3/\Delta_{\min}$ term in our upper bound up to logarithmic factors.

3.4 A General Gap-Dependent Regret Lower Bound

Beyond the algorithm class considered above, we next establish a gap-dependent regret lower bound that applies to arbitrary learning algorithms. This result characterizes an intrinsic difficulty of gap-dependent learning in linear MDPs, independently of the particular exploration mechanism.

Theorem 3.3. *For any feature dimension $d \geq 7$, horizon $H \geq 8$, minimum gap $\Delta_{\min} \in (0, H/4]$, episode number $K \geq (dH^3/\Delta_{\min}^2)^2$, and any learning algorithm, there exists a d-dimensional linear MDP with minimum gap $\Delta_{\min}$ such that*

$$\mathbb{E}[\mathrm{Regret}(T)] \geq cdH^3 \log K/\Delta_{\min},$$

where $c > 0$ is a universal constant.

The proof is deferred to Appendix I. Compared with the prior lower bound $\Omega(dH/\Delta_{\min})$ (He et al., 2021), our result provides two substantial improvements. First, it establishes at finite K that the logarithmic dependence on the number of episodes is unavoidable in gap-dependent expected regret. Second, it reveals an intrinsic cubic dependence on the horizon H. In particular, the H^3 dependence matches that of our upper bound in Theorem 3.1, showing that the improved horizon dependence achieved by LSVI-UCB++ is tight even beyond the class of optimism-based algorithms. The remaining gap between this general lower bound and the leading gap-dependent term of our upper bound is therefore a factor of d in the feature dimension, up to logarithmic factors.

We also note that some prior works (Papini et al., 2021; Zhang et al., 2024a) establish high-probability gap-dependent regret bounds of order $\tilde{O}(d^3H^5 \log(1/\delta)/\Delta_{\min})$ that do not explicitly depend on $\log K$. This does not contradict Theorem 3.3, since our result concerns expected regret. Converting such high-probability guarantees into expected regret typically requires choosing a confidence level δ that depends on K, thereby introducing an additional $\log K$ factor. For example, setting $\delta = 1/K$ (see Remark 5.2 of Zhang et al. (2024a)) yields $\mathbb{E}[\mathrm{Regret}(T)] \leq \tilde{O}(d^3H^5 \log K/\Delta_{\min})$.

4 PROOF SKETCH OF THEOREM 3.1

In this section, we present the random-suffix framework underlying the proof of Theorem 3.1. The use of the first positive-gap step to localize regret has precedents in tabular RL (Dann et al., 2021). For linear MDPs, we develop a novel random-suffix analysis that jointly controls cumulative exploration bonuses and variance. For each episode in which a positive gap is realized, we take the first such step as the suffix boundary. We express the expected regret in terms of value differences at these boundaries and use the bonus and variance bounds to obtain the improved dependence on H.

For each episode k, define $\tau_k := \inf\{h \in [H] : \Delta_h(s_h^k, a_h^k) > 0\}$, with the convention $\tau_k = H + 1$ if no positive gap is encountered. Let $\mathcal{K}_\tau := \{k \in [K] : \tau_k \le H\}$ and $L := |\mathcal{K}_\tau|$. We further define $\mathcal{J}_\tau := \{(k, h) : k \in \mathcal{K}_\tau, \tau_k \le h \le H\}$ and, for each $h \in [H]$, $\mathcal{K}_h := \{k \in [K] : \tau_k \le h\}$.

We begin with a first-positive-gap decomposition that expresses the expected regret in terms of value differences at the random suffix boundaries.

Lemma 4.1. *For LSVI-UCB++ with K episodes and $T = HK$ steps, the expected regret satisfies*

$$\mathbb{E}[\mathrm{Regret}(T)] = \mathbb{E}\left[\sum\nolimits_{k\in\mathcal{K}_\tau} (V_{\tau_k}^\star - V_{\tau_k}^{\pi^k})(s_{\tau_k}^k)\right].$$

The proof is given in Appendix G. Since all realized gaps before τ_k are zero, the performance-difference identity implies that the expected cumulative gap over the remaining suffix equals $(V_{\tau_k}^\star - V_{\tau_k}^{\pi^k})(s_{\tau_k}^k)$. Thus, each episode with a realized positive-gap action is charged only once at its first such step, which forms the basis of our random-suffix analysis.

The following lemma controls the cumulative optimistic error at these first positive-gap steps.

Lemma 4.2 (Informal)**.** *Let $\iota_1 := \log(1 + dHK/\delta)$. Then with high probability,*

$$\sum\nolimits_{k\in\mathcal{K}_\tau} \left(Q_{k,\tau_k}(s_{\tau_k}^k, a_{\tau_k}^k) - Q_{\tau_k}^{\pi^k}(s_{\tau_k}^k, a_{\tau_k}^k)\right) \le O\left(d^2H^3\iota_1^3/\Delta_{\min} + d^6H^5\iota_1^3\right). \tag{4}$$

The formal statement is given in Lemma G.3. By optimism, the left-hand side upper bounds $\sum_{k\in\mathcal{K}_\tau}(V_{\tau_k}^\star - V_{\tau_k}^{\pi^k})(s_{\tau_k}^k)$. Therefore, combining Lemmas 4.1 and 4.2, together with the trivial bound $\mathrm{Regret}(T) \le T$, gives Theorem 3.1. We next explain the key ideas behind Lemma 4.2.

4.1 PROOF SKETCH OF LEMMA 4.2

For simplicity, denote the left-hand side of Equation (4) by X. For every $k \in \mathcal{K}_\tau$, the definition of τ_k gives $\Delta_{\tau_k}(s_{\tau_k}^k, a_{\tau_k}^k) \ge \Delta_{\min}$. Since π^k is greedy with respect to $Q_{k,h}$ and $Q_{k,h}$ is optimistic, we have $Q_{k,\tau_k}(s_{\tau_k}^k, a_{\tau_k}^k) - Q_{\tau_k}^{\pi^k}(s_{\tau_k}^k, a_{\tau_k}^k) \ge \Delta_{\min}$. Summing over $k \in \mathcal{K}_\tau$ therefore gives

$$\Delta_{\min} L \le X. \tag{5}$$

We next derive an upper bound on X. By the Bellman equation and the confidence bound for the optimistic estimate, for every (k, h) we have (here we define $Q_{k,H+1}(s, a) := Q_{H+1}^{\pi^k}(s, a) := 0$)

$$\begin{aligned}(Q_{k,h} - Q_h^{\pi^k})(s_h^k, a_h^k) &\le (Q_{k,h+1} - Q_{h+1}^{\pi^k})(s_{h+1}^k, a_{h+1}^k) \\ &\quad + \left(\mathbb{P}_{s_h^k, a_h^k, h} - \mathbb{1}_{s_{h+1}^k}\right)(V_{k,h+1} - V_{h+1}^{\pi^k}) + 4\min\{\beta\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}, H\}. \end{aligned} \tag{6}$$

Unlike the usual full-trajectory analysis (He et al., 2023), we sum Equation (6) only from the random starting point τ_k. Summing over $h = \tau_k, \ldots, H$ and $k \in \mathcal{K}_\tau$ gives $X \le M_\tau + B_\tau$, where $M_\tau := \sum_{(k,h)\in\mathcal{J}_\tau}(\mathbb{P}_{s_h^k, a_h^k, h} - \mathbb{1}_{s_{h+1}^k})(V_{k,h+1} - V_{h+1}^{\pi^k})$ is a martingale term and $B_\tau := \sum_{h=1}^H \sum_{k\in\mathcal{K}_h} 4\min\{\beta\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}, H\}$ is the corresponding partial sum of bonuses. The martingale term can be bounded by $M_\tau \le O(\sqrt{H^3 L \log(KH/\delta)})$ with high probability. The main difficulty is therefore to control B_τ.

The next lemma provides a partial-sum bonus bound that is specifically designed for the episode subsets $\mathcal{K}_h$ generated by the first-positive-gap decomposition.

Lemma 4.3 (Informal)**.** *Set $\lambda = 1/H^2$ and $\iota = \log(1 + K/(d\lambda H))$. For any $h \in [H]$ and $\beta', C \ge 1$, with high probability,*

$$\sum\nolimits_{k\in\mathcal{K}_h} \min\{\beta'\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}, C\} \le 10d^4H^3C\iota + 10\beta' d^4H^2\iota + 2\beta'\sqrt{d\iota \sum\nolimits_{k\in\mathcal{K}_h}(\sigma_{k,h}^2 + H)}.$$

The formal statement and proof are given in Lemma F.3 in the appendix. This lemma requires a new argument because, after restricting the episodes to $\mathcal{K}_h$, the subsequence of covariance matrices $\{\mathbf{\Sigma}_{k,h} : k \in \mathcal{K}_h\}$ no longer satisfies the one-step recursive update needed for a standard elliptical-potential argument. To recover this structure, write $\mathcal{K}_h = \{k_1 < \cdots < k_{|\mathcal{K}_h|}\}$ and construct a surrogate covariance sequence $\{\mathbf{\Sigma}'_i\}$ using only these selected episodes and recursively defined surrogate variance weights. We show by induction that $\mathbf{\Sigma}'_i \preceq \mathbf{\Sigma}_{k_i,h}$ and $\bar{\sigma}'_{i,h} \geq \bar{\sigma}_{k_i,h}$. Consequently, each original bonus can be upper bounded by the corresponding surrogate bonus. Since the surrogate sequence again admits a one-step recursive update, the standard elliptical-potential argument becomes applicable. This surrogate construction is what allows the variance-weighted bonus analysis of LSVI-UCB++ to be localized to the random subsets $\mathcal{K}_h$.

Applying Lemma 4.3 with $\beta' = \beta$ and $C = H$ at every step and using Cauchy–Schwarz gives $B_\tau \leq O\left(d^4H^5\iota + \beta d^4H^3\iota + \beta\sqrt{dH\iota \sum_{(k,h)\in\mathcal{J}_\tau}(\sigma^2_{k,h} + H)}\right)$. It therefore remains to control the estimated variances over the same random suffix set $\mathcal{J}_\tau$.

Our next key lemma develops a random-suffix analogue of the variance analysis used in the worst-case regret bound of LSVI-UCB++.

Lemma 4.4 (Informal)**.** *With high probability,*

$$\sum\nolimits_{(k,h)\in\mathcal{J}_\tau} \sigma^2_{k,h} \leq O\left(H^2L + d^{10}H^9\iota_1^3\right). \tag{7}$$

The formal statement is given in Lemma G.2, with the complete proof provided in Appendix G. Here, we briefly outline the main ideas of the proof. According to the definition of $\sigma^2_{k,h}$, we need to handle components involving the estimated variance $\bar{\mathbb{V}}_{s^k_h,a^k_h,h}V_{k,h+1}$, the variance-estimation error $E_{k,h}$, the correction term $D_{k,h}$, and the additive term H.

For the estimated variance $\bar{\mathbb{V}}_{s^k_h,a^k_h,h}V_{k,h+1}$, we approximate it by the true variance $\mathbb{V}_{s^k_h,a^k_h,h}V_{k,h+1}$, where the approximation error is controlled by $E_{k,h}$. We then compare $\mathbb{V}_{s^k_h,a^k_h,h}V_{k,h+1}$ with $\mathbb{V}_{s^k_h,a^k_h,h}V^{\pi^k}_{h+1}$. The partial sums of the resulting errors can be bounded using Lemma F.4. In particular, the partial sum of $\mathbb{V}_{s^k_h,a^k_h,h}V^{\pi^k}_{h+1}$ over the random suffixes is bounded by $\tilde{O}\big(H^2L + H^3\big)$ through a suffix version of the total-variance argument combined with martingale concentration. The partial sums of $E_{k,h}$ can be controlled directly using Lemma 4.3. For $D_{k,h}$, we approximate its main component by $\mathbb{P}_{s^k_h,a^k_h,h}(V_{k,h+1} - \check{V}_{k,h+1})$, where the remaining error can again be bounded by the bonus. We further establish a recursive bound for the partial sum of $\mathbb{P}_{s^k_h,a^k_h,h}(V_{k,h+1} - \check{V}_{k,h+1})$ over $(k,h) \in \mathcal{J}_\tau$ in Lemma F.5. Combining these bounds gives a self-bounding inequality for $\sum_{(k,h)\in\mathcal{J}_\tau}\sigma^2_{k,h}$, whose solution yields Equation (7).

Using Lemma 4.4 in the bonus bound, together with the martingale bound and $\beta = O(\sqrt{d}\,\iota_1)$, gives

$$X \leq O(d^6H^5\iota_1^3 + d\sqrt{H^3\iota_1^3L}). \tag{8}$$

Finally, Equation (5) implies $L \leq X/\Delta_{\min}$. Substituting this relation into Equation (8) gives $X \leq O(d^6H^5\iota_1^3 + d\sqrt{H^3\iota_1^3X/\Delta_{\min}})$. Solving this inequality proves Lemma 4.2.

The improved H^3 dependence therefore follows from combining the first-positive-gap decomposition with the random-suffix variance analysis: the former reduces the analysis to L relevant suffixes, while the latter controls the sum of the true policy variances by $\tilde{O}(H^2L + H^3)$. Together with $\Delta_{\min}L \leq X$, this yields the leading term $\tilde{O}(d^2H^3/\Delta_{\min})$ and completes the proof sketch.

5 Conclusion

We establish the first gap-dependent regret guarantee for the nearly minimax-optimal LSVI-UCB++, with a leading term $\tilde{O}(d^2H^3/\Delta_{\min})$. We also develop a structural lower bound for persistent ellipsoidal optimism that, when specialized to LSVI-UCB++, matches the leading term of our upper bound up to logarithmic factors. Beyond this class, we prove a general lower bound $\Omega(dH^3\log K/\Delta_{\min})$ for arbitrary learning algorithms, showing that the H^3 dependence and logarithmic dependence on K are intrinsic. Resolving the remaining gap between the leading gap-dependent term of our upper bound and the general lower bound is an important direction for future work.

AI USE STATEMENT

AI tools were used for language polishing, presentation refinement, mathematical proofreading, and assistance with code debugging. All scientific ideas, problem formulation, mathematical claims, proof strategies and key arguments, experimental design, implementation decisions, and interpretation of results were independently developed and verified by the authors.

ETHICS STATEMENT

This work is primarily theoretical and does not involve human subjects, personal data, or real-world experiments requiring ethical approval. The empirical evaluation is conducted entirely on synthetic linear MDPs. We have followed the guidelines outlined in the ICLR Code of Ethics and have aimed to ensure transparency, integrity, and fairness throughout the research process. We do not foresee any direct ethical concerns or potential harms arising from this study.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide the formal problem setup and assumptions in Section 2 and complete proofs of all theoretical results in the appendix. Appendix B describes the synthetic linear MDP construction, algorithmic settings, and evaluation protocol used in our experiments. The code for method implementation is included in the supplementary materials.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pp. 67–83. PMLR, 2020.

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in Neural Information Processing Systems*, 30, 2017.

Mokhaled NA Al-Hamadani, Mohammed A Fadhel, Laith Alzubaidi, and Balazs Harangi. Reinforcement learning algorithms and applications in healthcare and robotics: A comprehensive and systematic review. *Sensors*, 24(8):2461, 2024.

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 49–56. MIT Press, 2007.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in Neural Information Processing Systems*, 21, 2008.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020.

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

Shulun Chen, Runlong Zhou, Zihan Zhang, Maryam Fazel, and Simon Du. Sharp gap-dependent variance-aware regret bounds for tabular mdps. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 112169–112205. Curran Associates, Inc., 2025. doi: 10.52202/085713-3747. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a2fb8d64706276cc6c49a52f143e0e5c-Paper-Conference.pdf.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pp. 1507–1516. PMLR, 2019.

Christoph Dann, Teodor Vanislavov Marinov, Mehryar Mohri, and Julian Zimmert. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 1–12, 2021.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pp. 100–118, 1975.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pp. 4171–4180. PMLR, 2021.

Jiafan He, Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for linear markov decision processes. In *International Conference on Machine Learning*, pp. 12790–12822. PMLR, 2023.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

Zeyu Jia, Lin Yang, Csaba Szepesvari, and Mengdi Wang. Model-based reinforcement learning with value-targeted regression. In *Learning for Dynamics and Control*, pp. 666–686. PMLR, 2020.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in Neural Information Processing Systems*, 31, 2018.

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

Safwan Labbi, Daniil Tiapkin, Lorenzo Mancini, Paul Mangold, and Eric Moulines. Federated ucbvi: Communication-efficient federated regret minimization with heterogeneous agents. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1315–1323. PMLR, 03–05 May 2025. URL `https://proceedings.mlr.press/v258/labbi25a.html`.

Gen Li, Laixi Shi, Yuxin Chen, and Yuejie Chi. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Information and Inference: A Journal of the IMA*, 12 (2):969–1043, 2023.

Zihang Liang, Haochen Zhang, and Lingzhou Xue. Provably efficient federated reinforcement learning with linear function approximation and logarithmic communication cost. *arXiv preprint arXiv:2609.00193*, 2026.

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.

Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pp. 2010–2020. PMLR, 2020.

Thanh Nguyen-Tang, Ming Yin, Sunil Gupta, Svetha Venkatesh, and Raman Arora. On instance-dependent bounds for offline reinforcement learning with linear function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 9310–9318, 2023.

Jungseul Ok, Alexandre Proutiere, and Damianos Tranos. Exploration in structured reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Matteo Papini, Andrea Tirinzoni, Aldo Pacchiano, Marcello Restelli, Alessandro Lazaric, and Matteo Pirotta. Reinforcement learning in linear mdps: Constant regret and representation selection. *Advances in Neural Information Processing Systems*, 34:16371–16383, 2021.

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

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

Ambuj Tewari and Peter Bartlett. Optimistic linear programming gives logarithmic regret for irreducible mdps. In *Advances in Neural Information Processing Systems*, pp. 1505–1512, 2008.

Tyler Toner, Miguel Saez, Dawn M Tilbury, and Kira Barton. Opportunities and challenges in applying reinforcement learning to robotic manipulation: An industrial case study. *Manufacturing Letters*, 35:1019–1030, 2023.

Andrew J Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. In *International Conference on Machine Learning*, pp. 22384–22429. PMLR, 2022.

Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, and Rahul Jain. Learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3007–3015. PMLR, 2021.

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

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

Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International conference on machine learning*, pp. 6995–7004. PMLR, 2019.

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.

Haochen Zhang, Zhong Zheng, and Lingzhou Xue. Gap-dependent bounds for federated q-learning. In *Forty-second International Conference on Machine Learning*, 2025a.

Haochen Zhang, Zhong Zheng, and Lingzhou Xue. Regret-optimal q-learning with low cost for single-agent and federated reinforcement learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 128711–128799. Curran Associates, Inc., 2025b. doi: 10.52202/085713-4288. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/bb203e938836544655996d1bb94a0fd7-Paper-Conference.pdf.

Haochen Zhang, Zhong Zheng, and Lingzhou Xue. Q-learning with fine-grained gap-dependent regret. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 48118–48162, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/4f8d1e02a057beb7920618ac73788078-Paper-Conference.pdf.

Junkai Zhang, Weitong Zhang, and Quanquan Gu. Optimal horizon-free reward-free exploration for linear mixture mdps. In *International Conference on Machine Learning*, pp. 41902–41930. PMLR, 2023.

Weitong Zhang, Zhiyuan Fan, Jiafan He, and Quanquan Gu. Achieving constant regret in linear markov decision processes. *Advances in Neural Information Processing Systems*, 37:130694–130738, 2024a.

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

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

Zihan Zhang, Yuxin Chen, Jason D Lee, and Simon S Du. Settling the sample complexity of online reinforcement learning. In *Conference on Learning Theory*, pp. 5213–5219. PMLR, 2024b.

Zihan Zhang, Jason D Lee, Yuxin Chen, and Simon Shaolei Du. Horizon-free regret for linear markov decision processes. In *The Twelfth International Conference on Learning Representations*, 2024c.

Canzhe Zhao, Xiangcheng Zhang, Baoxiang Wang, and Shuai Li. Logarithmic regret for linear markov decision processes with adversarial corruptions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22759–22767, 2025.

Zhong Zheng, Fengyu Gao, Lingzhou Xue, and Jing Yang. Federated q-learning: Linear regret speedup with low communication cost. In *The Twelfth International Conference on Learning Representations*, 2024.

Zhong Zheng, Haochen Zhang, and Lingzhou Xue. Federated q-learning with reference-advantage decomposition: Almost optimal regret and logarithmic communication cost. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Zhong Zheng, Haochen Zhang, and Lingzhou Xue. Gap-dependent bounds for q-learning using reference-advantage decomposition. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Dongruo Zhou and Quanquan Gu. Computationally efficient horizon-free reinforcement learning for linear mixture mdps. *Advances in neural information processing systems*, 35:36337–36349, 2022.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pp. 4532–4576. PMLR, 2021a.

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pp. 12793–12802. PMLR, 2021b.

Runlong Zhou, Zihan Zhang, and Simon Shaolei Du. Sharp variance-dependent bounds in reinforcement learning: Best of both worlds in stochastic and deterministic environments. In *International Conference on Machine Learning*, pp. 42878–42914. PMLR, 2023.

The appendix is organized as follows. Appendix A reviews related work, Appendix B presents numerical experiments, and Appendices C–F provide technical preliminaries. Appendix G proves the upper bound in Theorem 3.1, Appendix H proves the structural lower bound in Theorem 3.2, and Appendix I proves the general lower bound in Theorem 3.3.

A Related Work

Near-Optimal RL. In tabular RL, algorithms are categorized into model-based and model-free approaches. Model-based methods estimate the transition and reward models from data and plan based on the learned models, while model-free methods directly maintain value function estimates and act greedily. A large body of work has focused on model-based algorithms (Agarwal et al., 2020; Agrawal & Jia, 2017; Auer et al., 2008; Azar et al., 2017; Dann et al., 2019; Zanette & Brunskill, 2019; Zhang et al., 2024b; 2021; Zhou et al., 2023). Notably, Zhang et al. (2024b) proposed an algorithm achieving a regret bound of $\tilde{O}(\min\{\sqrt{SAH^2T},\, T\})$, which matches the information-theoretic lower bound. Model-free approaches have also been studied (Jin et al., 2018; Li et al., 2023; Ménard et al., 2021; Yang et al., 2021; Zhang et al., 2020; 2025b), and several works (Zhang et al., 2020; Ménard et al., 2021; Li et al., 2023; Zhang et al., 2025b) achieved the near-optimal regret $\tilde{O}(\sqrt{SAH^2T})$. Several works also focus on federated RL, including Zheng et al. (2024); Labbi et al. (2025); Zheng et al. (2025a); Zhang et al. (2025b), with the last three attaining near-optimal regret bounds.

The literature on RL with linear function approximation can be divided based on the structural assumptions imposed on the MDP. One line of work considers linear MDPs (Yang & Wang, 2019; Jin et al., 2020; Wei et al., 2021; Wagenmaker et al., 2022; He et al., 2023; Zhang et al., 2024c). Yang & Wang (2019) provided the first sample-efficient algorithm under a generative model. Subsequently, Jin et al. (2020) proposed the first provably efficient algorithm, LSVI-UCB, which achieves a regret bound $\tilde{O}(\sqrt{d^3H^3T})$ without access to a generative model. More recently, He et al. (2023) developed the LSVI-UCB++ algorithm, improving the regret to the nearly minimax-optimal bound $\tilde{O}(d\sqrt{H^3K})$ (Zhou et al., 2021a). Another line of work studies linear mixture MDPs (Jia et al., 2020; Ayoub et al., 2020; Modi et al., 2020; Zhou et al., 2021a;b; Zhou & Gu, 2022). Jia et al. (2020) and Ayoub et al. (2020) proposed the UCRL-VTR algorithm for episodic MDPs, achieving a regret bound of $\tilde{O}(d\sqrt{H^4K})$. Subsequently, Zhou et al. (2021a) developed a near-optimal algorithm for time-inhomogeneous linear mixture MDPs by proposing a Bernstein-type concentration inequality for self-normalized martingales. Furthermore, Zhou & Gu (2022) proposed a near-optimal, horizon-free algorithm for time-homogeneous linear mixture MDPs. Zhang et al. (2023) provided the near-optimal horizon-free sample complexity in the reward-free time-homogeneous setting. Liang et al. (2026) provide a provably efficient analysis of federated RL with linear function approximation.

Gap-Dependent RL. In tabular RL, early works established logarithmic regret guarantees for undiscounted and irreducible MDPs, including finite-time results (Auer & Ortner, 2007; Tewari & Bartlett, 2008). Subsequent works developed sharper instance-dependent guarantees and extended gap-dependent analyses to episodic settings (Jaksch et al., 2010; Ok et al., 2018; Simchowitz & Jamieson, 2019; Dann et al., 2021; Yang et al., 2021; Xu et al., 2021; Zheng et al., 2025b; Chen et al., 2025; Zhang et al., 2026). For model-based algorithms, Simchowitz & Jamieson (2019), Dann et al. (2021), and Chen et al. (2025) obtain fine-grained gap-dependent regret bounds. For model-free algorithms, Yang et al. (2021) provided the first gap-dependent regret bound for UCB-Hoeffding (Jin et al., 2018), which was subsequently refined by the AMB algorithm of Xu et al. (2021). Later, Zheng et al. (2025b) further improved the result by reanalyzing the UCB-Advantage algorithm (Zhang et al., 2020) and the Q-EarlySettled-Advantage algorithm (Li et al., 2023). Zhang et al. (2025a;b) extend gap-dependent analysis to federated Q-learning settings. More recently, Zhang et al. (2026) provided the first fine-grained gap-dependent regret upper bound for UCB-Hoeffding.

For RL with linear function approximation, gap-dependent regret bounds have also been studied. For online RL, He et al. (2021) provided the first gap-dependent regret bound, showing that LSVI-UCB (Jin et al., 2020) achieves $\tilde{O}(d^3H^5/\Delta_{\min})$ in linear MDPs, and UCRL-VTR (Ayoub et al., 2020) achieves $\tilde{O}(d^2H^5/\Delta_{\min})$ in linear mixture MDPs. Subsequent works (Papini et al., 2021; Zhang et al., 2024a) further improved these results, establishing high-probability constant-regret guarantees for LSVI-UCB under the additional UNISOFT condition and for Cert-LSVI-UCB. Zhao et al. (2025) studied logarithmic regret in a different setting with adversarial reward corruptions. Nguyen-Tang et al. (2023) provided gap-dependent guarantees for offline RL with linear function approximation.

B NUMERICAL EXPERIMENTS

We provide synthetic simulations to illustrate the empirical behavior of a practical implementation of LSVI-UCB++, to compare it with representative gap-dependent baselines, and to examine whether the observed growth is compatible with the polylogarithmic dependence on K in Theorem 3.1.

Synthetic linear MDPs. We consider two families of finite-horizon linear MDPs. In the random family, we use four problem scales

$$(S, A, H, d) \in \{(5, 3, 2, 3), (5, 5, 5, 3), (8, 6, 5, 5), (10, 10, 10, 8)\}.$$

For each scale, the feature vector $\boldsymbol{\phi}(s, a) \in \mathbb{R}^d$ is sampled from $\mathrm{Dir}(1, \ldots, 1)$, so that $\boldsymbol{\phi}_j(s, a) \geq 0$ and $\sum_{j=1}^{d} \boldsymbol{\phi}_j(s, a) = 1$. The transition kernel is generated from base kernels $\boldsymbol{\theta}_{h,j} \in \Delta(\mathcal{S})$ via

$$\mathbb{P}_h(s' \mid s, a) = \langle \boldsymbol{\phi}(s, a), \boldsymbol{\theta}_h(s') \rangle,$$

where each $\boldsymbol{\theta}_{h,j}$ is independently sampled from $\mathrm{Dir}(1, \ldots, 1)$. Independently of the transition construction, each known reward $r_h(s, a)$ is sampled independently from $\mathrm{Unif}[0, 1]$. At the beginning of each episode, the initial state is drawn uniformly over the state space.

To isolate the effect of the minimum positive gap, we also use a controlled-gap family with $(S, A, H, d) = (10, 10, 10, 8)$ and $\Delta_{\min} \in \{0.2, 0.4, 0.6, 0.8\}$. Let s_0 be the initial state and partition the other nine states into a rewarding class $\mathcal{S}_+$ of four states and a zero-reward class $\mathcal{S}_-$ of five states. Using the standard basis $e_1, \ldots, e_8$ of $\mathbb{R}^8$, we set

$$\boldsymbol{\phi}(s_0, A-1) = e_1, \qquad \boldsymbol{\phi}(s_0, a) = e_2 \ \ (a = 0, \ldots, A-3), \qquad \boldsymbol{\phi}(s_0, A-2) = e_5,$$

and set $\boldsymbol{\phi}(s, a) = e_3$ for $s \in \mathcal{S}_+$ and $\boldsymbol{\phi}(s, a) = e_4$ for $s \in \mathcal{S}_-$. This construction uses five active one-hot coordinates and is zero-padded to dimension $d = 8$. At step $h = 1$, all state–action pairs have reward 0. At steps $h = 2, \ldots, H$, states in $\mathcal{S}_+$ have reward 1, while s_0 and states in $\mathcal{S}_-$ have reward 0. For every $h \in [H]$, states in $\mathcal{S}_+$ transition uniformly within $\mathcal{S}_+$ and states in $\mathcal{S}_-$ transition uniformly within $\mathcal{S}_-$. At steps $h = 2, \ldots, H$, every action at s_0 transitions uniformly into $\mathcal{S}_-$. Thus, for every state–action–step triple with $(s, h) \neq (s_0, 1)$, both the reward and transition kernel are fully specified, and the gap is zero.

Let U_+ and U_- denote the uniform distributions over $\mathcal{S}_+$ and $\mathcal{S}_-$, respectively, and let $p_\star = 0.60 + \varepsilon$, where $\varepsilon \sim \mathrm{Unif}[-0.025, 0.025]$ is sampled independently for each MDP instance. At step $h = 1$, action $A-1$ at s_0 transitions into $\mathcal{S}_+$ with total probability $p_\star$, actions $0, \ldots, A-3$ do so with total probability $p_\star - \Delta_{\min}/(H-1)$, and action $A-2$ does so with total probability $p_\star - \Delta_{\mathrm{bad}}/(H-1)$, where $\Delta_{\mathrm{bad}} = 1.2$; conditional on either class, the next state is uniform within that class.

To make the linear-MDP structure explicit, we set

$$\boldsymbol{\theta}_{1,1} = p_\star U_+ + (1 - p_\star) U_-, \qquad \boldsymbol{\theta}_{1,2} = \left(p_\star - \frac{\Delta_{\min}}{H-1}\right) U_+ + \left(1 - p_\star + \frac{\Delta_{\min}}{H-1}\right) U_-,$$

$$\boldsymbol{\theta}_{1,5} = \left(p_\star - \frac{\Delta_{\mathrm{bad}}}{H-1}\right) U_+ + \left(1 - p_\star + \frac{\Delta_{\mathrm{bad}}}{H-1}\right) U_-, \qquad \boldsymbol{\theta}_{1,3} = U_+, \qquad \boldsymbol{\theta}_{1,4} = U_-.$$

For $h \geq 2$, we set

$$\boldsymbol{\theta}_{h,1} = \boldsymbol{\theta}_{h,2} = \boldsymbol{\theta}_{h,4} = \boldsymbol{\theta}_{h,5} = U_-, \qquad \boldsymbol{\theta}_{h,3} = U_+,$$

with all remaining coordinates equal to the zero measure. Hence $\mathbb{P}_h(\cdot \mid s, a) = \langle \boldsymbol{\phi}(s, a), \boldsymbol{\theta}_h(\cdot) \rangle$ for every (s, a, h). Moreover, for every measurable $B \subseteq \mathcal{S}$, at most five coordinates of $\boldsymbol{\theta}_h(B)$ are nonzero and each has magnitude at most one, so $\|\boldsymbol{\theta}_h(B)\|_2 \leq \sqrt{5} \leq \sqrt{d}$.

Consequently, each of the nine noninitial states has positive probability of being reached after the first action. Since $H = 10$, $p_\star \in [0.575, 0.625]$, and $\Delta_{\min} \leq 0.8 < \Delta_{\mathrm{bad}}$, all specified transition probabilities lie in $[0, 1]$. The value gaps at $(s_0, h = 1)$ are exactly $\Delta_{\min}$ for actions $0, \ldots, A-3$ and Δ_{bad} for action $A-2$. All other state–action–step triples have zero gap, so the minimum positive gap is exactly $\Delta_{\min}$ for every generated instance. The random common shift changes the MDP instance without changing any action gap.

For each panel, we run each algorithm for $K = 20{,}000$ episodes, average over ten independently generated MDP instances, and use paired rollout seeds across algorithms on each instance. Since the

MDP is known in simulation, regret is evaluated exactly by dynamic programming as

$$\sum_{k=1}^{K} \big(V_1^\star - V_1^{\pi^k}\big)(s_1^k),$$

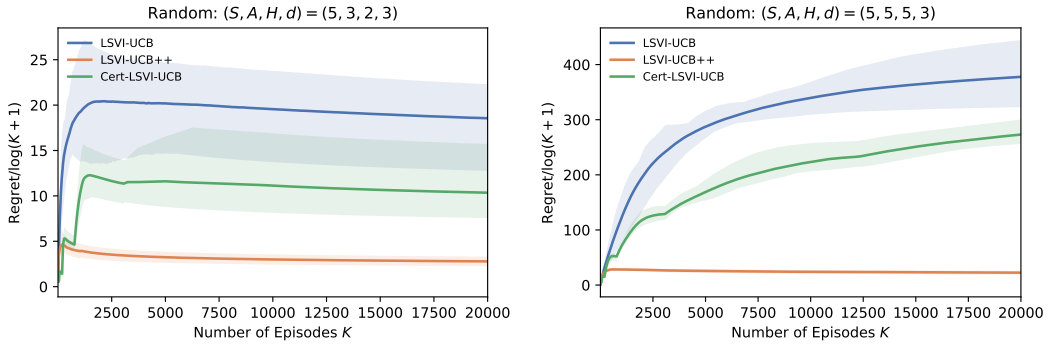
which matches the regret definition in Section 2. The figures report the mean curve together with the 10th–90th percentile interval.

Algorithms and parameters. We compare LSVI-UCB++ with LSVI-UCB (Jin et al., 2020) and Cert-LSVI-UCB (Zhang et al., 2024a). To ensure that all methods use the same reward information, all three algorithms use the known reward function and estimate only the transition-dependent continuation-value term. Thus, for LSVI-UCB, the regression target is $V_{k,h+1}(s_{h+1}^i)$ and the optimistic value is computed as

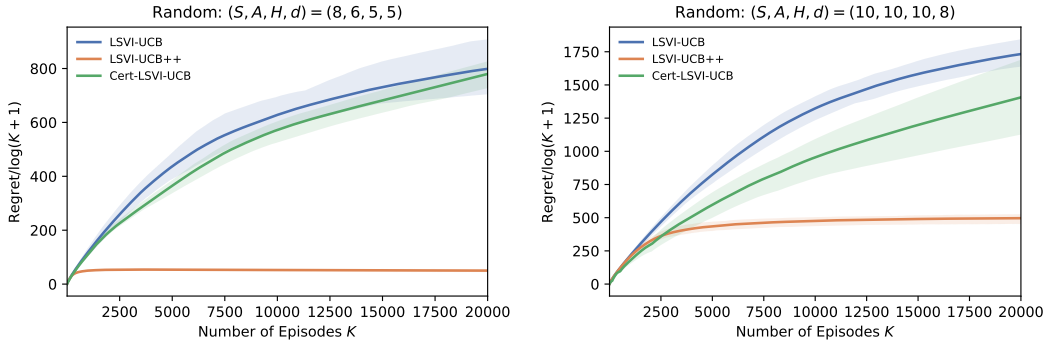
$$\Lambda_{k,h} = \lambda I + \sum_{i<k} \boldsymbol{\phi}(s_h^i, a_h^i)\boldsymbol{\phi}(s_h^i, a_h^i)^\top, \qquad w_{k,h} = \Lambda_{k,h}^{-1} \sum_{i<k} \boldsymbol{\phi}(s_h^i, a_h^i) V_{k,h+1}(s_{h+1}^i),$$

$$Q_{k,h}(s,a) = \min\big\{r_h(s,a) + \langle w_{k,h}, \boldsymbol{\phi}(s,a)\rangle + \beta\|\boldsymbol{\phi}(s,a)\|_{\Lambda_{k,h}^{-1}},\, H\big\}.$$

We use $\lambda = 1$ and $\beta = Hd$ for LSVI-UCB. For the practical implementation of LSVI-UCB++, we set $\lambda = H^{-2}$, $\beta = \bar{\beta} = \sqrt{d}$, $\widetilde{\beta} = \sqrt{dH}$, and use determinant-triggered rare switching with threshold 2. For Cert-LSVI-UCB, we use $\lambda = 1$ and the confidence-radius form of Zhang et al. (2024a) with multiplicative coefficient 4×10^{-4}.



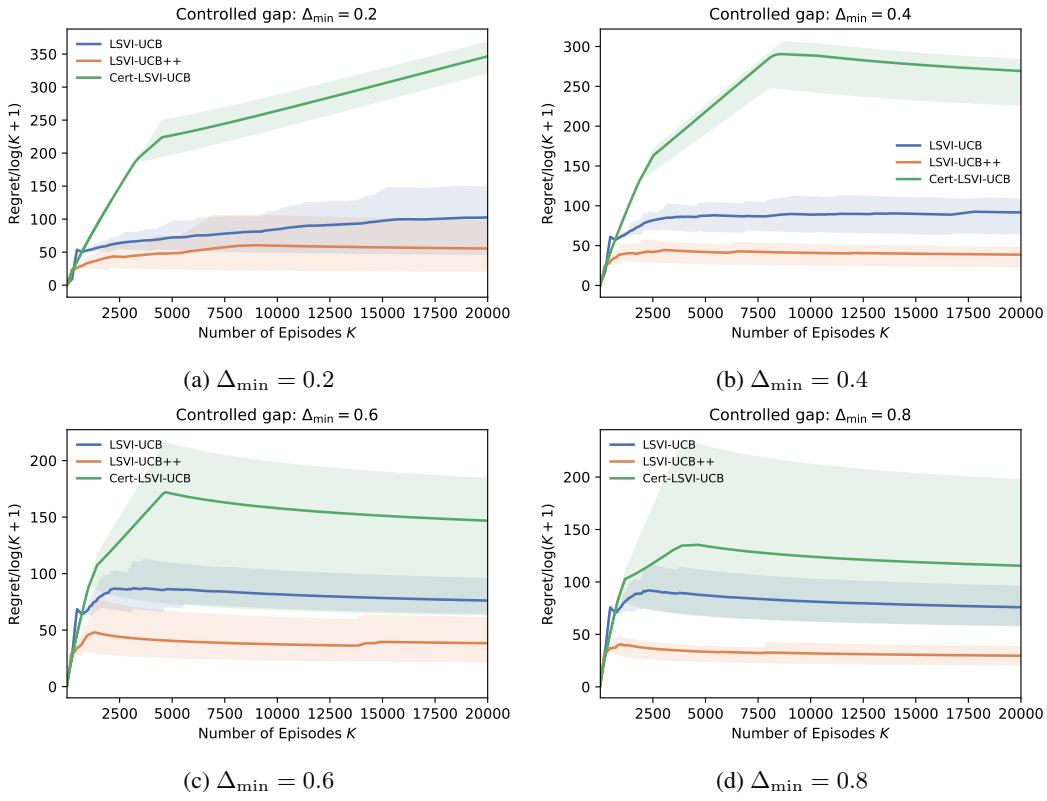
(a) $(S, A, H, d) = (5, 3, 2, 3)$ (b) $(S, A, H, d) = (5, 5, 5, 3)$



(c) $(S, A, H, d) = (8, 6, 5, 5)$ (d) $(S, A, H, d) = (10, 10, 10, 8)$

Figure 1: Log-normalized regret $\mathrm{Regret}(T)/\log(K+1)$ on random synthetic linear MDPs. Each curve is averaged over ten MDP instances, with shaded 10th–90th percentile intervals.

Results. Figure 1 shows that LSVI-UCB++ obtains lower empirical regret than both baselines across the four random linear MDP scales. Its normalized curves $\mathrm{Regret}(T)/\log(K+1)$ become nearly flat toward the end of the tested range, suggesting approximately logarithmic regret growth over this finite range. This trend is qualitatively compatible with the polylogarithmic dependence on K



(a) $\Delta_{\min} = 0.2$ (b) $\Delta_{\min} = 0.4$

(c) $\Delta_{\min} = 0.6$ (d) $\Delta_{\min} = 0.8$

Figure 2: Log-normalized regret $\mathrm{Regret}(T)/\log(K+1)$ on controlled-gap linear MDPs with $(S, A, H, d) = (10, 10, 10, 8)$. Each curve is averaged over ten MDP instances, with shaded 10th–90th percentile intervals.

established in Theorem 3.1; it does not imply that the higher-order logarithmic factors in the upper bound are absent.

Figure 2 further examines the dependence on the minimum positive gap. LSVI-UCB++ achieves lower empirical regret than both baselines under the reported parameter settings. Over the later part of the tested range, its regret decreases as $\Delta_{\min}$ increases from 0.2 to 0.8, consistent with the gap-dependent intuition that larger action gaps make suboptimal actions easier to eliminate.

C Auxiliary Lemmas

In this section, we introduce several auxiliary lemmas that will be used to support our proof.

Lemma C.1 (Azuma-Hoeffding inequality)**.** *Suppose $\{X_k\}_{k=0}^{\infty}$ is a martingale and $|X_k - X_{k-1}| \leq c_k$, $\forall k \in \mathbb{N}_+$, almost surely. Then for any $N \in \mathbb{N}_+$ and $\epsilon > 0$, it holds that:*

$$\mathbb{P}\left(X_N - X_0 \geq \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{2\sum_{k=1}^{N} c_k^2}\right),$$

and

$$\mathbb{P}\left(|X_N - X_0| \geq \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2}{2\sum_{k=1}^{N} c_k^2}\right).$$

Lemma C.2 (Freedman's inequality (Freedman, 1975, Theorem 1.6))**.** *Suppose $\{X_k\}_{k=0}^{\infty}$ is a martingale with respect to the filtration $\{\mathcal{F}_k\}_{k=0}^{\infty}$ and $|X_k - X_{k-1}| \leq c$, $\forall k \in \mathbb{N}_+$, almost surely. Define*

$$\mathrm{Var}_N := \sum_{k=1}^{N} \mathbb{E}\left[(X_k - X_{k-1})^2 \mid \mathcal{F}_{k-1}\right].$$

Then, for any $\epsilon, v > 0$, it holds that

$$\mathbb{P}\left(\exists N \in \mathbb{N}_+ : X_N - X_0 \geq \epsilon \text{ and } \mathrm{Var}_N \leq v\right) \leq \exp\left(-\frac{\epsilon^2}{2(v + c\epsilon)}\right).$$

Consequently, for any $N \in \mathbb{N}_+$, $v > 0$, and $\delta \in (0, 1)$, it holds that

$$\mathbb{P}\left(X_N - X_0 > 2\sqrt{v \log\left(\frac{1}{\delta}\right)} + 2c \log\left(\frac{1}{\delta}\right),\ \mathrm{Var}_N \leq v\right) \leq \delta.$$

Proof. The first claim follows from Theorem 1.6 of Freedman (1975). To prove the consequence, let

$$q := \log\left(\frac{1}{\delta}\right), \qquad \epsilon := 2\sqrt{vq} + 2cq.$$

By Freedman's inequality,

$$\begin{aligned}\mathbb{P}\left(X_N - X_0 > \epsilon, \mathrm{Var}_N \leq v\right) &\leq \mathbb{P}\left(\exists n \in \mathbb{N}_+ : X_n - X_0 > \epsilon \text{ and } \mathrm{Var}_n \leq v\right) \\ &\leq \exp\left(-\frac{\epsilon^2}{2(v + c\epsilon)}\right).\end{aligned}$$

Moreover,

$$\epsilon^2 - 2q(v + c\epsilon) = 2vq + 4cq\sqrt{vq} \geq 0,$$

which implies

$$\frac{\epsilon^2}{2(v + c\epsilon)} \geq q.$$

Therefore,

$$\mathbb{P}\left(X_N - X_0 > \epsilon,\ \mathrm{Var}_N \leq v\right) \leq e^{-q} = \delta.$$

□

Lemma C.3 (Lemma 12 of Abbasi-Yadkori et al. (2011))**.** *If $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$ are two positive definite matrices such that $\mathbf{A} \succeq \mathbf{B}$, then for any $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}\|_{\mathbf{A}} \leq \|\mathbf{x}\|_{\mathbf{B}} \cdot \sqrt{\det(\mathbf{A})/\det(\mathbf{B})}$.*

Lemma C.4 (Lemma 11 of Abbasi-Yadkori et al. (2011))**.** *Let $\{\mathbf{x}_k\}_{k=1}^K$ be a sequence of vectors in $\mathbb{R}^d$. Let matrix $\mathbf{\Sigma}_0$ be a $d \times d$ positive definite matrix and define $\mathbf{\Sigma}_k = \mathbf{\Sigma}_0 + \sum_{i=1}^k \mathbf{x}_i \mathbf{x}_i^\top$, then we have*

$$\sum_{i=1}^k \min\left\{1, \mathbf{x}_i^\top \mathbf{\Sigma}_{i-1}^{-1} \mathbf{x}_i\right\} \leq 2 \log\left(\frac{\det \mathbf{\Sigma}_k}{\det \mathbf{\Sigma}_0}\right).$$

In addition, if $\|\mathbf{x}_i\|_2 \leq L$ holds for all $i \in [K]$, then

$$\sum_{i=1}^k \min\left\{1, \mathbf{x}_i^\top \mathbf{\Sigma}_{i-1}^{-1} \mathbf{x}_i\right\} \leq 2 \log\left(\frac{\det \mathbf{\Sigma}_k}{\det \mathbf{\Sigma}_0}\right) \leq 2\left(d \log\left((\mathit{trace}(\mathbf{\Sigma}_0) + kL^2)/d\right) - \log\det \mathbf{\Sigma}_0\right).$$

Lemma C.5 (Lemma 4.4 of Zhou & Gu (2022))**.** *Let $\{\sigma_k\}_{k\geq 1}$ be a sequence of non-negative numbers, $\alpha, \gamma, \lambda > 0$, $\{\mathbf{a}_k\}_{k\geq 1} \subset \mathbb{R}^d$ and $\|\mathbf{a}_k\|_2 \leq A$. Let $\{\bar{\sigma}_k\}_{k\geq 1}$ and $\{\hat{\mathbf{\Sigma}}_k\}_{k\geq 1}$ be (recursively) defined as follows: $\hat{\mathbf{\Sigma}}_1 = \lambda \mathbf{I}_d$,*

$$\forall k \geq 1,\ \bar{\sigma}_k = \max\{\sigma_k, \alpha, \gamma \|\mathbf{a}_k\|_{\hat{\mathbf{\Sigma}}_k^{-1}}^{1/2}\},\ \hat{\mathbf{\Sigma}}_{k+1} = \hat{\mathbf{\Sigma}}_k + \mathbf{a}_k \mathbf{a}_k^\top / \bar{\sigma}_k^2.$$

Let $\iota_0 = \log(1 + KA^2/(d\lambda\alpha^2))$. Then we have

$$\sum_{k=1}^K \min\left\{1, \|\mathbf{a}_k\|_{\hat{\mathbf{\Sigma}}_k^{-1}}\right\} \leq 2d\iota_0 + 2\gamma^2 d\iota_0 + 2\sqrt{d\iota_0} \sqrt{\sum_{k=1}^K (\sigma_k^2 + \alpha^2)}.$$

D MINOR MODIFICATIONS TO LSVI-UCB++

Our implementation of LSVI-UCB++ makes two minor modifications to the pseudocode of He et al. (2023). The first clarifies the timing of the rare-switching update, while the second relaxes the adjusted variance normalization. We describe both modifications below and explain why the results from He et al. (2023) invoked in our analysis remain valid.

Rare-switching update schedule. In the original pseudocode of He et al. (2023), the last updating episode k_{last} is reset to the current episode k immediately after the determinant-doubling criterion is triggered inside the backward loop over $h = H, \dots, 1$. Taken literally, this changes the reference covariance matrices within the same backward pass: after the first step is updated, the criterion is subsequently compared against $\Sigma_{k,h'}$ itself and therefore cannot remain triggered. To implement the episode-level rare-switching update intended in He et al. (2023), Algorithm 1 introduces an auxiliary variable k_{mid}. Whenever the updating criterion is triggered, we set $k_{\mathrm{mid}} \leftarrow k$, while keeping k_{last} fixed throughout the entire backward pass. Only after all steps have been processed do we set $k_{\mathrm{last}} \leftarrow k_{\mathrm{mid}}$. Consequently, when an episode triggers the determinant-doubling criterion, the optimistic and pessimistic value functions are updated consistently at all steps using the same previous updating episode k_{last} as reference. This modification only makes explicit the episode-level update schedule used in the rare-switching analysis of He et al. (2023). Hence, the rare-switching and covering-number arguments used in their analysis continue to apply.

Relaxed adjusted variance normalization. The second modification concerns the adjusted variance used in the weighted ridge regression. The original algorithm defines

$$\bar{\sigma}_{k,h} = \max\left\{\sigma_{k,h}, H, 2d^3H^2\|\phi(s_h^k, a_h^k)\|^{1/2}_{\Sigma_{k,h}^{-1}}\right\},$$

whereas Algorithm 1 uses

$$\bar{\sigma}^2_{k,h} = \max\left\{\sigma^2_{k,h}, H, 2d^3H^2\|\phi(s_h^k, a_h^k)\|_{\Sigma_{k,h}^{-1}}\right\}.$$

All other components of the weighted regression and variance estimator remain unchanged. Let $u_{k,h} := \|\phi(s_h^k, a_h^k)\|_{\Sigma_{k,h}^{-1}}$. Our normalization guarantees

$$\bar{\sigma}^2_{k,h} \geq \sigma^2_{k,h}, \qquad \bar{\sigma}^2_{k,h} \geq H, \qquad \bar{\sigma}^2_{k,h} \geq 2d^3H^2u_{k,h}.$$

Consequently,

$$\bar{\sigma}^{-1}_{k,h} \leq H^{-1/2}, \qquad \frac{2Hu_{k,h}}{\bar{\sigma}^2_{k,h}} \leq \frac{1}{d^3H}.$$

These are the properties of the adjusted variance needed in the self-normalized concentration arguments used by He et al. (2023). Moreover, $\bar{\sigma}^2_{k,h} \geq \sigma^2_{k,h}$ preserves the required conditional-variance domination. Therefore, the arguments underlying Lemmas B.1 and B.5 of He et al. (2023) remain valid under our normalization with the same orders of the confidence radii.

The relaxed normalization is useful for our gap-dependent analysis because the corresponding weighted elliptical-potential argument can be applied with $\alpha = \sqrt{H}$ and $\gamma = \sqrt{2d^3H^2}$, which yields the sharper partial bonus bound developed below. The rare-switching and covering-number arguments used in our analysis are also unaffected by the modified update schedule. Moreover, the concentration arguments inherited from He et al. (2023) are compatible with the normalization properties verified above and continue to apply under our modified definition of $\bar{\sigma}_{k,h}$. Consequently, Algorithm 1 preserves the nearly minimax-optimal leading-order regret established by He et al. (2023).

E PROBABILITY EVENTS

In this section, we introduce several high-probability events for LSVI-UCB++. For every $(k, h) \in [K] \times [H + 1]$, define the clipped optimistic and pessimistic value functions as

$$V^{\mathrm{clip}}_{k,h}(s) := \max\{V_{k,h}(s), 0\}, \qquad s \in \mathcal{S},$$

and

$$\check{V}^{\mathrm{clip}}_{k,h}(s) := \min\{\check{V}_{k,h}(s), H\}, \qquad s \in \mathcal{S}.$$

Since $V_{k,h}(s) \le H$ and $\check{V}_{k,h}(s) \ge 0$ by construction, we have $0 \le V_{k,h}^{\mathrm{clip}}(s), \check{V}_{k,h}^{\mathrm{clip}}(s) \le H$ for any $s \in \mathcal{S}$.

Lemma E.1. *For LSVI-UCB++, we have the following high-probability events.*

(a) Define $\mathcal{E}_1$ as the event that the following inequalities hold for any $(s,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ simultaneously:

$$\left|\hat{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s,a) - \mathbb{P}_{s,a,h} V_{k,h+1}\right| \le \bar{\beta}\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s,a)},$$
$$\left|\tilde{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s,a) - \mathbb{P}_{s,a,h} V_{k,h+1}^2\right| \le \tilde{\beta}\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s,a)},$$
$$\left|\check{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s,a) - \mathbb{P}_{s,a,h} \check{V}_{k,h+1}\right| \le \bar{\beta}\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s,a)}.$$

Here, for sufficiently large constants $\bar{c}_\beta, \tilde{c}_\beta \ge 1$,

$$\bar{\beta} = \bar{c}_\beta \left(H\sqrt{d\lambda} + \sqrt{d^3 H^2 \log^2\left(dHK/(\delta\lambda)\right)} \right)$$

and

$$\tilde{\beta} = \tilde{c}_\beta \left(H^2\sqrt{d\lambda} + \sqrt{d^3 H^4 \log^2\left(dHK/(\delta\lambda)\right)} \right).$$

Then the event $\mathcal{E}_1$ holds with probability at least $1 - 7\delta$.

(b) Define $\mathcal{E}_2$ as the event such that, for any $(s,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, the weight vector $\hat{\mathbf{w}}_{k,h}$ satisfies

$$\left|\hat{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s,a) - \mathbb{P}_{s,a,h} V_{k,h+1}\right| \le \beta\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s,a)},$$

where for sufficiently large constant $c_\beta \ge 1$,

$$\beta = c_\beta \left(H\sqrt{d\lambda} + \sqrt{d \log^2\left(1 + dKH/(\delta\lambda)\right)} \right).$$

Then the event $\mathcal{E}_2$ holds with probability at least $1 - 8\delta$.

(c) Define $\mathcal{E}_3$ as the event

$$\mathcal{E}_3 := \left\{ \sum_{(k,h)\in\mathcal{J}_\tau} \left(\mathbb{P}_{s_h^k, a_h^k, h} - \mathbb{1}_{s_{h+1}^k}\right)\left(V_{k,h+1}^{\mathrm{clip}} - V_{h+1}^{\pi^k}\right) \le 2\sqrt{2H^3 L \log\left(\frac{KH}{\delta}\right)} \right\}.$$

Then the event $\mathcal{E}_3$ holds with probability at least $1 - \delta$.

(d) Define $\mathcal{E}_4$ as the event

$$\mathcal{E}_4 := \left\{ \sum_{(k,h)\in\mathcal{J}_\tau} (h - \tau_k + 1)\left(\mathbb{P}_{s_h^k, a_h^k, h} - \mathbb{1}_{s_{h+1}^k}\right)\left(V_{k,h+1}^{\mathrm{clip}} - V_{h+1}^{\pi^k}\right) \le 2\sqrt{2H^5 L \log\left(\frac{KH}{\delta}\right)} \right\}.$$

Then the event $\mathcal{E}_4$ holds with probability at least $1 - \delta$.

(e) Define $\mathcal{E}_5$ as the event

$$\mathcal{E}_5 := \left\{ \sum_{(k,h)\in\mathcal{J}_\tau} (h - \tau_k + 1)\left(\mathbb{P}_{s_h^k, a_h^k, h} - \mathbb{1}_{s_{h+1}^k}\right)\left(V_{k,h+1}^{\mathrm{clip}} - \check{V}_{k,h+1}^{\mathrm{clip}}\right) \le 2\sqrt{2H^5 L \log\left(\frac{KH}{\delta}\right)} \right\}.$$

Then the event $\mathcal{E}_5$ holds with probability at least $1 - \delta$.

(f) Define $\mathcal{E}_6$ as the event

$$\mathcal{E}_6 := \left\{ \sum_{(k,h)\in\mathcal{J}_\tau} \mathbb{V}_{s_h^k,a_h^k,h} V_{h+1}^{\pi^k} \le 2H^2L + 3H^3 \log\left(\frac{K}{\delta}\right) \right\}.$$

Then the event $\mathcal{E}_6$ holds with probability at least $1-\delta$.

Proof of Lemma E.1. Parts (a) and (b) follow from the arguments of Lemmas B.1 and B.5 in He et al. (2023), respectively. As discussed in Section D, these arguments remain valid under the adjusted-variance normalization used in Algorithm 1.

Retained sequence forms a martingale-difference sequence. We establish a common retained-sequence construction used in the proofs of parts (c), (d) and (e). Let $\mathcal{F}_{k,h}$ denote the σ-field generated by all samples before episode k, together with the trajectory in episode k up to step h after the action a_h^k is selected and before the next state s_{h+1}^k is observed. By the definition of τ_k, for every $(k,h)\in[K]\times[H]$,

$$\{\tau_k \le h\} = \bigcup_{h'=1}^{h} \{\Delta_{h'}(s_{h'}^k, a_{h'}^k) > 0\} \in \mathcal{F}_{k,h},$$

since $s_{h'}^k$ and $a_{h'}^k$ are $\mathcal{F}_{k,h}$-measurable for every $h' \le h$ and the true gap functions are fixed by the underlying MDP.

Order all pairs $(k,h)\in[K]\times[H]$ lexicographically, and let $(k(n),h(n))$ denote the n-th pair in this order. Let $\{\mathcal{H}_n\}_{n=0}^{KH}$ be the corresponding filtration, where

$$\mathcal{H}_{n-1} := \mathcal{F}_{k(n),h(n)}, \qquad n\in[KH],$$

and $\mathcal{H}_{KH}$ is the terminal sigma-field containing all information revealed after the last transition increment. Thus, $\mathcal{H}_{n-1}$ contains the information available immediately before $s_{h(n)+1}^{k(n)}$ is observed, whereas $\mathcal{H}_n$ contains the information available after this transition. For convenience, set $\mathcal{H}_{KH+1} := \mathcal{H}_{KH}$. For any stopping time γ taking values in $\{0,\dots,KH+1\}$, let

$$\mathcal{H}_\gamma := \{A\in\mathcal{H}_{KH} : A\cap\{\gamma=n\}\in\mathcal{H}_n \text{ for every } n\in\{0,\dots,KH+1\}\}$$

denote the corresponding stopped sigma-field.

Define

$$Z_n := \mathbb{I}[\tau_{k(n)} \le h(n)], \qquad n\in[KH].$$

Then Z_n is $\mathcal{H}_{n-1}$-measurable, and

$$M := \sum_{n=1}^{KH} Z_n = \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{I}[\tau_k \le h] = \sum_{k\in\mathcal{K}_\tau} (H-\tau_k+1) \le HL.$$

Set $N_0 := 0$, and for each $j\in[KH]$, define

$$N_j := \inf\left\{ n\in[KH] : \sum_{i=1}^{n} Z_i = j \right\},$$

with the convention $N_j = KH+1$ if the set is empty. Thus, N_j is the position of the j-th retained pair when $j\le M$, whereas $N_j = KH+1$ when $j>M$. Moreover, for every $n\in[KH]$,

$$\{N_j = n\} = \left\{ \sum_{i=1}^{n-1} Z_i = j-1 \right\} \cap \{Z_n = 1\} \in \mathcal{H}_{n-1},$$

so N_j is a predictable stopping time.

Let $\{\xi_n\}_{n=1}^{KH}$ be any sequence such that ξ_n is $\mathcal{H}_n$-measurable and $\mathbb{E}[\xi_n \mid \mathcal{H}_{n-1}] = 0$. Define

$$\eta_j := \sum_{n=1}^{KH} \mathbb{I}[N_j = n]\xi_n, \qquad j\in[KH].$$

Thus, $\eta_j = \xi_{N_j}$ when $j \le M$, while $\eta_j = 0$ when $j > M$.

Fix any $A \in \mathcal{H}_{N_{j-1}}$. Since $N_{j-1} \le n-1$ on $\{N_j = n\}$ and $\{N_j = n\} \in \mathcal{H}_{n-1}$, it follows that

$$A \cap \{N_j = n\} = A \cap \{N_{j-1} \le n-1\} \cap \{N_j = n\} \in \mathcal{H}_{n-1}.$$

Therefore,

$$\mathbb{E}[\mathbb{I}_A \eta_j] = \sum_{n=1}^{KH} \mathbb{E}\left[\mathbb{I}_{A \cap \{N_j = n\}} \xi_n\right] = \sum_{n=1}^{KH} \mathbb{E}\left[\mathbb{I}_{A \cap \{N_j = n\}} \mathbb{E}[\xi_n \mid \mathcal{H}_{n-1}]\right] = 0.$$

Since this holds for every $A \in \mathcal{H}_{N_{j-1}}$,

$$\mathbb{E}[\eta_j \mid \mathcal{H}_{N_{j-1}}] = 0.$$

Moreover, η_j is $\mathcal{H}_{N_j}$-measurable. Therefore, $\{\eta_j\}_{j=1}^{KH}$ is a martingale-difference sequence with respect to the filtration $\{\mathcal{H}_{N_j}\}_{j=0}^{KH}$.

If $|\eta_j| \le B$ for every $j \in [KH]$, then for any deterministic $m \in [KH]$, Lemma C.1 gives

$$\mathbb{P}\left(\sum_{j=1}^{m} \eta_j > B\sqrt{2m \log\left(\frac{KH}{\delta}\right)}\right) \le \frac{\delta}{KH}.$$

Taking a union bound over $m \in [KH]$, with probability at least $1-\delta$, we have simultaneously for all $m \in [KH]$,

$$\sum_{j=1}^{m} \eta_j \le B\sqrt{2m \log\left(\frac{KH}{\delta}\right)}.$$

Hence, on this event, if $M \ge 1$, evaluating the uniform bound at $m = M$ gives

$$\sum_{j=1}^{M} \eta_j \le B\sqrt{2M \log\left(\frac{KH}{\delta}\right)},$$

while the case $M = 0$ is immediate.

(c) Set

$$\xi_n := \left(\mathbb{P}_{s^{k(n)}_{h(n)}, a^{k(n)}_{h(n)}, h(n)} - \mathbb{1}_{s^{k(n)}_{h(n)+1}}\right)\left(V^{\mathrm{clip}}_{k(n),h(n)+1} - V^{\pi^{k(n)}}_{h(n)+1}\right), \qquad n \in [KH].$$

By the Markov property,

$$\mathbb{E}[\xi_n \mid \mathcal{H}_{n-1}] = 0.$$

Let $\{\eta_j\}_{j=1}^{KH}$ be the corresponding retained sequence. Since both $V^{\mathrm{clip}}_{k,h+1}$ and $V^{\pi^k}_{h+1}$ take values in $[0, H]$, we have $|\eta_j| \le 2H$. Applying the uniform-prefix bound above with $B = 2H$ and using $M \le HL$, with probability at least $1-\delta$,

$$\sum_{(k,h) \in \mathcal{J}_\tau} \left(\mathbb{P}_{s^k_h, a^k_h, h} - \mathbb{1}_{s^k_{h+1}}\right)\left(V^{\mathrm{clip}}_{k,h+1} - V^{\pi^k}_{h+1}\right) = \sum_{j=1}^{M} \eta_j \le 2\sqrt{2H^3 L \log\left(\frac{KH}{\delta}\right)}.$$

This proves part (c).

(d) Set

$$\xi_n := Z_n\big(h(n) - \tau_{k(n)} + 1\big)\left(\mathbb{P}_{s^{k(n)}_{h(n)}, a^{k(n)}_{h(n)}, h(n)} - \mathbb{1}_{s^{k(n)}_{h(n)+1}}\right)\left(V^{\mathrm{clip}}_{k(n),h(n)+1} - V^{\pi^{k(n)}}_{h(n)+1}\right),\ n \in [KH].$$

To verify predictability, observe that

$$Z_n\big(h(n) - \tau_{k(n)} + 1\big) = \sum_{t=1}^{h(n)} \big(h(n) - t + 1\big)\,\mathbb{1}[\tau_{k(n)} = t].$$

For every $t \le h(n)$, the event $\{\tau_{k(n)} = t\}$ is $\mathcal{H}_{n-1}$-measurable. Therefore, $Z_n\big(h(n) - \tau_{k(n)} + 1\big)$ is $\mathcal{H}_{n-1}$-measurable, and the Markov property yields

$$\mathbb{E}[\xi_n \mid \mathcal{H}_{n-1}] = 0.$$

Let $\{\eta_j\}_{j=1}^{KH}$ be the corresponding retained sequence. Since $0 \le Z_n\big(h(n) - \tau_{k(n)} + 1\big) \le H$, we have $|\eta_j| \le 2H^2$. Applying the uniform-prefix bound above with $B = 2H^2$ and using $M \le HL$, with probability at least $1 - \delta$,

$$\sum_{(k,h)\in\mathcal{J}_\tau} (h - \tau_k + 1)\Big(\mathbb{P}_{s_h^k, a_h^k, h} - \mathbb{1}_{s_{h+1}^k}\Big)\Big(V_{k,h+1}^{\mathrm{clip}} - V_{h+1}^{\pi^k}\Big) = \sum_{j=1}^{M} \eta_j \le 2\sqrt{2H^5 L \log\left(\frac{KH}{\delta}\right)}.$$

This proves part (d).

(e) For any $n \in [KH]$, set

$$\xi_n := Z_n\big(h(n) - \tau_{k(n)} + 1\big)\left(\mathbb{P}_{s_{h(n)}^{k(n)}, a_{h(n)}^{k(n)}, h(n)} - \mathbb{1}_{s_{h(n)+1}^{k(n)}}\right)\left(V_{k(n),h(n)+1}^{\mathrm{clip}} - \check{V}_{k(n),h(n)+1}^{\mathrm{clip}}\right),$$

and let $\{\eta_j\}_{j=1}^{KH}$ be the corresponding retained sequence. By the same argument as in part (d), $\mathbb{E}[\xi_n \mid \mathcal{H}_{n-1}] = 0$ and $|\eta_j| \le 2H^2$. Therefore, with probability at least $1 - \delta$,

$$\sum_{(k,h)\in\mathcal{J}_\tau} (h - \tau_k + 1)\left(\mathbb{P}_{s_h^k, a_h^k, h} - \mathbb{1}_{s_{h+1}^k}\right)\left(V_{k,h+1}^{\mathrm{clip}} - \check{V}_{k,h+1}^{\mathrm{clip}}\right) = \sum_{j=1}^{M} \eta_j \le 2\sqrt{2H^5 L \log\left(\frac{KH}{\delta}\right)}.$$

This proves part (e).

(f) If $L = 0$, then $\mathcal{J}_\tau$ is empty and the conclusion holds immediately. We therefore consider the case $L \ge 1$ and write the elements of $\mathcal{K}_\tau$ in increasing order as

$$k_1 < k_2 < \cdots < k_L.$$

For each $i \in [L]$, define

$$U_i := \sum_{h=\tau_{k_i}}^{H} \mathbb{V}_{s_h^{k_i}, a_h^{k_i}, h} V_{h+1}^{\pi^{k_i}},$$

and set $U_i := 0$ for $L < i \le K$.

Let $\ell(k,h) := (k-1)H + h$ denote the lexicographic position of (k,h). Set $\rho_0 := 0$, and for every $i \in [K]$, define

$$\gamma_i := \begin{cases} \ell(k_i, \tau_{k_i}) - 1, & i \le L, \\ KH, & i > L, \end{cases} \qquad \rho_i := \begin{cases} k_i H - 1, & i \le L, \\ KH, & i > L. \end{cases}$$

For every $(k,t) \in [K] \times [H]$, we have

$$\{i \le L, k_i = k, \tau_k = t\} \in \mathcal{F}_{k,t} = \mathcal{H}_{\ell(k,t)-1}.$$

Moreover,

$$\{i \le L, k_i = k\} = \left\{\sum_{k'=1}^{k-1} \mathbb{1}[\tau_{k'} \le H] = i - 1\right\} \cap \{\tau_k \le H\} \in \mathcal{F}_{k,H} = \mathcal{H}_{kH-1}.$$

Therefore, γ_i and ρ_i are stopping times with respect to $\{\mathcal{H}_n\}_{n=0}^{KH}$. Moreover,

$$\mathcal{H}_{\rho_{i-1}} \subseteq \mathcal{H}_{\gamma_i} \subseteq \mathcal{H}_{\rho_i}, \qquad i \in [K].$$

We first bound the conditional expectation of U_i. Fix any $(k,t) \in [K] \times [H]$ and, for every $h \in \{t, \dots, H\}$, define

$$\zeta_{k,h} := r_h(s_h^k, a_h^k) + V_{h+1}^{\pi^k}(s_{h+1}^k) - V_h^{\pi^k}(s_h^k).$$

By the Bellman equation and the Markov property,

$$\mathbb{E}\left[\zeta_{k,h}|\mathcal{F}_{k,h}\right]=0, \qquad \mathbb{E}\left[\zeta_{k,h}^2\middle|\mathcal{F}_{k,h}\right]=\mathbb{V}_{s_h^k,a_h^k,h}V_{h+1}^{\pi^k}.$$

Consequently, conditional orthogonality of the martingale differences gives

$$\begin{aligned}\mathbb{E}\left[\sum_{h=t}^{H}\mathbb{V}_{s_h^k,a_h^k,h}V_{h+1}^{\pi^k}\middle|\mathcal{F}_{k,t}\right]&=\mathbb{E}\left[\left(\sum_{h=t}^{H}\zeta_{k,h}\right)^2\middle|\mathcal{F}_{k,t}\right]\\&=\mathbb{E}\left[\left(\sum_{h=t}^{H}r_h(s_h^k,a_h^k)-V_t^{\pi^k}(s_t^k)\right)^2\middle|\mathcal{F}_{k,t}\right]\\&\le H^2,\end{aligned}$$

where the second equality follows by telescoping and using $V_{H+1}^{\pi^k}=0$. Applying this bound on each event $\{i\le L, k_i=k, \tau_k=t\}$, and noting that $U_i=0$ for $i>L$, gives

$$\mathbb{E}\left[U_i|\mathcal{H}_{\gamma_i}\right]\le H^2, \qquad i\in[K].$$

This is the suffix version of the total-variance argument used in Lemma C.5 of Jin et al. (2018).

For every $i\in[L]$, since $0\le V_{h+1}^{\pi^{k_i}}\le H$,

$$0\le\mathbb{V}_{s_h^{k_i},a_h^{k_i},h}V_{h+1}^{\pi^{k_i}}\le H^2.$$

Since the suffix contains at most H steps and $U_i=0$ for $i>L$,

$$0\le U_i\le H^3, \qquad i\in[K].$$

Since $V_{H+1}^{\pi^{k_i}}=0$ for every $i\in[L]$, the variance term corresponding to $h=H$ vanishes. Hence, for every $i\in[L]$, U_i is determined by the states and actions observed up to step H of episode k_i and is therefore $\mathcal{H}_{\rho_i}$-measurable. Moreover, for $i>L$, $U_i=0$ and is thus trivially $\mathcal{H}_{\rho_i}$-measurable. Therefore, U_i is $\mathcal{H}_{\rho_i}$-measurable for every $i\in[K]$. For each $i\in[K]$, define

$$D_i:=U_i-\mathbb{E}\left[U_i|\mathcal{H}_{\gamma_i}\right].$$

Since $\mathcal{H}_{\gamma_i}\subseteq\mathcal{H}_{\rho_i}$, the conditional expectation $\mathbb{E}[U_i\mid\mathcal{H}_{\gamma_i}]$ is also $\mathcal{H}_{\rho_i}$-measurable. Therefore, D_i is $\mathcal{H}_{\rho_i}$-measurable and satisfies

$$\mathbb{E}[D_i\mid\mathcal{H}_{\gamma_i}]=0.$$

Since $\mathcal{H}_{\rho_{i-1}}\subseteq\mathcal{H}_{\gamma_i}$, the tower property gives

$$\mathbb{E}[D_i\mid\mathcal{H}_{\rho_{i-1}}]=0.$$

Thus, $\{D_i\}_{i=1}^K$ is a martingale-difference sequence with respect to the filtration $\{\mathcal{H}_{\rho_i}\}_{i=0}^K$. Furthermore, $|D_i|\le H^3$ and

$$\mathbb{E}\left[D_i^2\middle|\mathcal{H}_{\gamma_i}\right]\le\mathbb{E}\left[U_i^2\middle|\mathcal{H}_{\gamma_i}\right]\le H^3\mathbb{E}\left[U_i|\mathcal{H}_{\gamma_i}\right]\le H^5.$$

The tower property therefore yields

$$\mathbb{E}[D_i^2\mid\mathcal{H}_{\rho_{i-1}}]\le H^5.$$

For any deterministic $m\in[K]$, setting $v=H^5m$ and $c=H^3$ in Lemma C.2 gives

$$\mathbb{P}\left(\sum_{i=1}^{m}D_i>2\sqrt{H^5m\log\left(\frac{K}{\delta}\right)}+2H^3\log\left(\frac{K}{\delta}\right)\right)\le\frac{\delta}{K}.$$

Taking a union bound over all deterministic $m\in[K]$, with probability at least $1-\delta$, the preceding inequality holds simultaneously for every $m\in[K]$. Evaluating it at $m=L$ gives

$$\begin{aligned}\sum_{i=1}^{L}U_i&\le\sum_{i=1}^{L}\mathbb{E}\left[U_i|\mathcal{H}_{\gamma_i}\right]+2\sqrt{H^5L\log\left(\frac{K}{\delta}\right)}+2H^3\log\left(\frac{K}{\delta}\right)\\&\le H^2L+2\sqrt{H^5L\log\left(\frac{K}{\delta}\right)}+2H^3\log\left(\frac{K}{\delta}\right)\\&\le 2H^2L+3H^3\log\left(\frac{K}{\delta}\right),\end{aligned}$$

where the last inequality uses $2\sqrt{ab} \le a + b$ with $a = H^2 L$ and $b = H^3 \log\left(\frac{K}{\delta}\right)$. Finally,

$$\sum_{i=1}^{L} U_i = \sum_{k \in \mathcal{K}_\tau} \sum_{h=\tau_k}^{H} \mathbb{V}_{s_h^k, a_h^k, h} V_{h+1}^{\pi^k} = \sum_{(k,h) \in \mathcal{J}_\tau} \mathbb{V}_{s_h^k, a_h^k, h} V_{h+1}^{\pi^k}.$$

This proves the result. □

F Properties of Value Function Estimates

Following Lemma B.4 and Lemma B.2 in He et al. (2023), we obtain the following two lemmas.

Lemma F.1. *On the event $\mathcal{E}_1 \cap \mathcal{E}_2$, for any $(s, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, we have*

$$Q_{k,h}(s,a) \ge Q_h^\star(s,a) \ge \check{Q}_{k,h}(s,a), \qquad V_{k,h}(s) \ge V_h^\star(s) \ge \check{V}_{k,h}(s).$$

In particular, on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, Lemma F.1, together with the nonnegativity of $\check{V}_{k,h}$ and the upper bound $V_{k,h} \le H$, implies that

$$0 \le \check{V}_{k,h}(s) \le V_h^\star(s) \le V_{k,h}(s) \le H$$

for every $(s, h, k) \in \mathcal{S} \times [H] \times [K]$. Together with the terminal identities $\check{V}_{k,H+1} = V_{H+1}^\star = V_{k,H+1} = 0$, this implies that, on $\mathcal{E}_1 \cap \mathcal{E}_2$,

$$\check{V}_{k,h}^{\mathrm{clip}}(s) = \check{V}_{k,h}(s), \qquad V_{k,h}^{\mathrm{clip}}(s) = V_{k,h}(s) \tag{9}$$

for every $(s, h, k) \in \mathcal{S} \times [H+1] \times [K]$.

Lemma F.2. *On the event $\mathcal{E}_1 \cap \mathcal{E}_2$, for any $(h, k) \in [H] \times [K]$, the estimated variance satisfies*

$$\left| \bar{\mathbb{V}}_{s_h^k, a_h^k, h} V_{k,h+1} - \mathbb{V}_{s_h^k, a_h^k, h} V_{k,h+1} \right| \le E_{k,h}.$$

The following lemma provides a bound on the sum of the bonuses over the episodes in $\mathcal{K}_h$.

Lemma F.3 (Formal Statement of Lemma 4.3)**.** *On the event $\mathcal{E}_1 \cap \mathcal{E}_2$, suppose that $\lambda = 1/H^2$ and let $\iota = \log(1 + K/(d\lambda H)) = \log(1 + T/d)$. For any $h \in [H]$ and parameters $\beta' \ge 1$ and $C \ge 1$, the sum of the bonuses over the episodes in $\mathcal{K}_h$ can be bounded as*

$$\begin{aligned}
&\sum_{k \in \mathcal{K}_h} \min\left\{ \beta' \sqrt{\boldsymbol{\phi}(s_h^k, a_h^k)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^k, a_h^k)}, C \right\} \\
&\quad \le 10 d^4 H^3 C \iota + 10 \beta' d^4 H^2 \iota + 2\beta' \sqrt{d\iota \sum_{k \in \mathcal{K}_h} \left(\sigma_{k,h}^2 + H \right)}.
\end{aligned}$$

Proof of Lemma F.3. On $\mathcal{E}_1 \cap \mathcal{E}_2$, Lemma F.2 and the definition of $D_{k,h}$ imply that $\sigma_{k,h}^2 \ge H$, and hence $\sigma_{k,h}$ is well defined and nonnegative.

Fix any step $h \in [H]$. If $\mathcal{K}_h$ is empty, the conclusion holds immediately. Otherwise, write the elements of $\mathcal{K}_h$ in increasing order as

$$k_1 < k_2 < \cdots < k_{|\mathcal{K}_h|}.$$

Since $\beta' \ge 1$, the sum of the bonuses is bounded by

$$\begin{aligned}
&\sum_{i=1}^{|\mathcal{K}_h|} \min\left\{ \beta' \sqrt{\boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i})^\top \boldsymbol{\Sigma}_{k_i,h}^{-1} \boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i})}, C \right\} \\
&\le \sum_{i=1}^{|\mathcal{K}_h|} \beta' \min\left\{ \sqrt{\boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i})^\top \boldsymbol{\Sigma}_{k_i,h}^{-1} \boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i})}, 1 \right\} \\
&\quad + C \sum_{i=1}^{|\mathcal{K}_h|} \mathbb{I}\left[\sqrt{\boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i})^\top \boldsymbol{\Sigma}_{k_i,h}^{-1} \boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i})} \ge 1 \right].
\end{aligned} \tag{10}$$

For the first term in Equation (10), for any $i \in [|\mathcal{K}_h|]$, define

$$\mathbf{\Sigma}'_i = \lambda \mathbf{I}_d + \sum_{j=1}^{i-1} (\bar{\sigma}'_{j,h})^{-2} \boldsymbol{\phi}(s_h^{k_j}, a_h^{k_j}) \boldsymbol{\phi}(s_h^{k_j}, a_h^{k_j})^\top$$

and

$$\bar{\sigma}'_{i,h} = \max \left\{ \sigma_{k_i,h}, \sqrt{H}, \sqrt{2d^3H^2} \|\boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i})\|^{1/2}_{(\mathbf{\Sigma}'_i)^{-1}} \right\}.$$

We use induction to prove that, for any $i \in [|\mathcal{K}_h|]$, $\mathbf{\Sigma}'_i \preceq \mathbf{\Sigma}_{k_i,h}$ and $\bar{\sigma}'_{i,h} \geq \bar{\sigma}_{k_i,h}$.

For $i = 1$, we have

$$\mathbf{\Sigma}'_1 = \lambda \mathbf{I}_d \preceq \mathbf{\Sigma}_{k_1,h}.$$

Therefore, $(\mathbf{\Sigma}'_1)^{-1} \succeq \mathbf{\Sigma}^{-1}_{k_1,h}$, and hence

$$\begin{aligned}
\bar{\sigma}'_{1,h} &= \max \left\{ \sigma_{k_1,h}, \sqrt{H}, \sqrt{2d^3H^2} \|\boldsymbol{\phi}(s_h^{k_1}, a_h^{k_1})\|^{1/2}_{(\mathbf{\Sigma}'_1)^{-1}} \right\} \\
&\geq \max \left\{ \sigma_{k_1,h}, \sqrt{H}, \sqrt{2d^3H^2} \|\boldsymbol{\phi}(s_h^{k_1}, a_h^{k_1})\|^{1/2}_{\mathbf{\Sigma}^{-1}_{k_1,h}} \right\} \geq \bar{\sigma}_{k_1,h}.
\end{aligned}$$

Suppose that the conclusion holds for $1, 2, \ldots, i-1$. Since $\bar{\sigma}'_{j,h} \geq \bar{\sigma}_{k_j,h}$ for every $j \in [i-1]$, we have

$$\begin{aligned}
\mathbf{\Sigma}'_i &= \lambda \mathbf{I}_d + \sum_{j=1}^{i-1} (\bar{\sigma}'_{j,h})^{-2} \boldsymbol{\phi}(s_h^{k_j}, a_h^{k_j}) \boldsymbol{\phi}(s_h^{k_j}, a_h^{k_j})^\top \\
&\preceq \lambda \mathbf{I}_d + \sum_{j=1}^{i-1} (\bar{\sigma}_{k_j,h})^{-2} \boldsymbol{\phi}(s_h^{k_j}, a_h^{k_j}) \boldsymbol{\phi}(s_h^{k_j}, a_h^{k_j})^\top \preceq \mathbf{\Sigma}_{k_i,h},
\end{aligned}$$

where the last inequality follows because $k_1 < k_2 < \cdots < k_i$ and $\mathbf{\Sigma}_{k_i,h}$ contains all covariance updates before episode k_i, including those corresponding to episodes $k_1, \ldots, k_{i-1}$. Consequently, $(\mathbf{\Sigma}'_i)^{-1} \succeq \mathbf{\Sigma}^{-1}_{k_i,h}$, and thus

$$\begin{aligned}
\bar{\sigma}'_{i,h} &= \max \left\{ \sigma_{k_i,h}, \sqrt{H}, \sqrt{2d^3H^2} \|\boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i})\|^{1/2}_{(\mathbf{\Sigma}'_i)^{-1}} \right\} \\
&\geq \max \left\{ \sigma_{k_i,h}, \sqrt{H}, \sqrt{2d^3H^2} \|\boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i})\|^{1/2}_{\mathbf{\Sigma}^{-1}_{k_i,h}} \right\} \geq \bar{\sigma}_{k_i,h}.
\end{aligned}$$

This completes the induction.

Since $\mathbf{\Sigma}'_i \preceq \mathbf{\Sigma}_{k_i,h}$, we have $(\mathbf{\Sigma}'_i)^{-1} \succeq \mathbf{\Sigma}^{-1}_{k_i,h}$ and therefore

$$\begin{aligned}
&\sum_{i=1}^{|\mathcal{K}_h|} \beta' \min \left\{ \sqrt{\boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i})^\top \mathbf{\Sigma}^{-1}_{k_i,h} \boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i})}, 1 \right\} \\
&\leq \sum_{i=1}^{|\mathcal{K}_h|} \beta' \min \left\{ \sqrt{\boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i})^\top (\mathbf{\Sigma}'_i)^{-1} \boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i})}, 1 \right\} \\
&\leq 10\beta' d^4 H^2 \iota + 2\beta' \sqrt{d\iota \sum_{i=1}^{|\mathcal{K}_h|} \left(\sigma^2_{k_i,h} + H \right)},
\end{aligned} \tag{11}$$

where the last inequality follows from Lemma C.5 with $\mathbf{a}_i = \boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i})$, $\sigma_i = \sigma_{k_i,h}$, $\alpha = \sqrt{H}$, $\gamma = \sqrt{2d^3H^2}$, and $\hat{\mathbf{\Sigma}}_i = \mathbf{\Sigma}'_i$.

We next bound the second term in Equation (10). Let

$$\{i_1, i_2, \ldots, i_m\} = \left\{ i \in [|\mathcal{K}_h|] : \sqrt{\boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i})^\top \mathbf{\Sigma}^{-1}_{k_i,h} \boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i})} \geq 1 \right\},$$

where $i_1 < i_2 < \cdots < i_m$. Set $\mathbf{\Sigma}''_0 = \lambda \mathbf{I}_d$, and for every $t \in [m]$, define

$$\mathbf{\Sigma}''_t = \lambda \mathbf{I}_d + \sum_{j=1}^{t} \bar{\sigma}^{-2}_{k_{i_j},h} \boldsymbol{\phi}(s_h^{k_{i_j}}, a_h^{k_{i_j}}) \boldsymbol{\phi}(s_h^{k_{i_j}}, a_h^{k_{i_j}})^\top.$$

Since $k_{i_1} < k_{i_2} < \cdots < k_{i_m}$, the matrix $\mathbf{\Sigma}_{k_{i_t},h}$ contains all covariance updates corresponding to $k_{i_1}, \ldots, k_{i_{t-1}}$. Therefore,

$$\mathbf{\Sigma}''_{t-1} \preceq \mathbf{\Sigma}_{k_{i_t},h}.$$

Consequently, $(\mathbf{\Sigma}''_{t-1})^{-1} \succeq \mathbf{\Sigma}^{-1}_{k_{i_t},h}$, and hence

$$\begin{aligned} &\sum_{t=1}^{m} \boldsymbol{\phi}(s_h^{k_{i_t}}, a_h^{k_{i_t}})^\top (\mathbf{\Sigma}''_{t-1})^{-1} \boldsymbol{\phi}(s_h^{k_{i_t}}, a_h^{k_{i_t}}) \\ &\geq \sum_{t=1}^{m} \boldsymbol{\phi}(s_h^{k_{i_t}}, a_h^{k_{i_t}})^\top \mathbf{\Sigma}^{-1}_{k_{i_t},h} \boldsymbol{\phi}(s_h^{k_{i_t}}, a_h^{k_{i_t}}) \geq m. \end{aligned} \tag{12}$$

For every selected index i_t, the definition of i_t gives

$$\|\boldsymbol{\phi}(s_h^{k_{i_t}}, a_h^{k_{i_t}})\|_{\mathbf{\Sigma}^{-1}_{k_{i_t},h}} \geq 1.$$

By the definition of $\bar{\sigma}_{k,h}$, it follows that

$$\bar{\sigma}^2_{k_{i_t},h} \geq 2d^3H^2.$$

Moreover, since $\mathbf{\Sigma}''_{t-1} \succeq \lambda \mathbf{I}_d$, $\lambda = 1/H^2$, and $\|\boldsymbol{\phi}(s_h^{k_{i_t}}, a_h^{k_{i_t}})\|_2 \leq 1$, we have

$$\bar{\sigma}^{-2}_{k_{i_t},h} \boldsymbol{\phi}(s_h^{k_{i_t}}, a_h^{k_{i_t}})^\top (\mathbf{\Sigma}''_{t-1})^{-1} \boldsymbol{\phi}(s_h^{k_{i_t}}, a_h^{k_{i_t}}) \leq \frac{H^2}{2d^3H^2} = \frac{1}{2d^3} \leq 1.$$

In addition, by the definition of $\bar{\sigma}^2_{k,h}$, together with $\mathbf{\Sigma}_{k,h} \succeq \lambda \mathbf{I}_d$ and $\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_2 \leq 1$, we have

$$\bar{\sigma}^2_{k,h} \leq \frac{5d^3H^2}{\sqrt{\lambda}}.$$

Moreover, since $\bar{\sigma}^2_{k_{i_t},h} \geq H$ and $\|\boldsymbol{\phi}(s_h^{k_{i_t}}, a_h^{k_{i_t}})\|_2 \leq 1$, we have

$$\left\| \bar{\sigma}^{-1}_{k_{i_t},h} \boldsymbol{\phi}(s_h^{k_{i_t}}, a_h^{k_{i_t}}) \right\|_2^2 \leq \frac{1}{H}.$$

Therefore, since $m \leq |\mathcal{K}_h| \leq K$, Lemma C.4 gives

$$\sum_{t=1}^{m} \bar{\sigma}^{-2}_{k_{i_t},h} \boldsymbol{\phi}(s_h^{k_{i_t}}, a_h^{k_{i_t}})^\top (\mathbf{\Sigma}''_{t-1})^{-1} \boldsymbol{\phi}(s_h^{k_{i_t}}, a_h^{k_{i_t}}) \leq 2d \log\left(1 + \frac{m}{d\lambda H}\right) \leq 2d\iota.$$

Combining this bound with $\bar{\sigma}^2_{k,h} \leq 5d^3H^2/\sqrt{\lambda}$ and $\lambda = H^{-2}$, we obtain

$$\begin{aligned} &\sum_{t=1}^{m} \boldsymbol{\phi}(s_h^{k_{i_t}}, a_h^{k_{i_t}})^\top (\mathbf{\Sigma}''_{t-1})^{-1} \boldsymbol{\phi}(s_h^{k_{i_t}}, a_h^{k_{i_t}}) \\ &\leq \frac{5d^3H^2}{\sqrt{\lambda}} \sum_{t=1}^{m} \bar{\sigma}^{-2}_{k_{i_t},h} \boldsymbol{\phi}(s_h^{k_{i_t}}, a_h^{k_{i_t}})^\top (\mathbf{\Sigma}''_{t-1})^{-1} \boldsymbol{\phi}(s_h^{k_{i_t}}, a_h^{k_{i_t}}) \\ &\leq 10d^4H^3\iota. \end{aligned} \tag{13}$$

Combining Equation (12) and Equation (13) gives

$$m \leq 10d^4H^3\iota.$$

Substituting this bound and Equation (11) into Equation (10), we obtain

$$\sum_{i=1}^{|\mathcal{K}_h|} \min\left\{\beta'\sqrt{\boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i})^\top \boldsymbol{\Sigma}_{k_i,h}^{-1}\boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i})}, C\right\}$$
$$\leq 10d^4H^3C\iota + 10\beta' d^4H^2\iota + 2\beta'\sqrt{d\iota\sum_{i=1}^{|\mathcal{K}_h|}\left(\sigma_{k_i,h}^2 + H\right)}.$$

Since $\{k_1, \ldots, k_{|\mathcal{K}_h|}\} = \mathcal{K}_h$, this is equivalent to

$$\sum_{k\in\mathcal{K}_h} \min\left\{\beta'\sqrt{\boldsymbol{\phi}(s_h^{k}, a_h^{k})^\top \boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s_h^{k}, a_h^{k})}, C\right\}$$
$$\leq 10d^4H^3C\iota + 10\beta' d^4H^2\iota + 2\beta'\sqrt{d\iota\sum_{k\in\mathcal{K}_h}\left(\sigma_{k,h}^2 + H\right)}.$$

This completes the proof of Lemma F.3. □

Lemma F.4. *On the events $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_4$ in Lemma E.1, we have*

$$\sum_{(k,h)\in\mathcal{J}_\tau} \mathbb{P}_{s_h^k, a_h^k, h}\left(V_{k,h+1} - V_{h+1}^{\pi^k}\right)$$
$$\leq 40d^4H^6\iota + 40\beta d^4H^4\iota + 8\beta\sqrt{dH^3\iota\sum_{(k,h)\in\mathcal{J}_\tau}\left(\sigma_{k,h}^2 + H\right)} + 2\sqrt{2H^5L\log(KH/\delta)}.$$

Proof of Lemma F.4. For any $h' \in [H]$ and $k \in [K]$, if $k_{\text{last}} \geq 1$, then we have

$$\begin{aligned}
&V_{k,h'}(s_{h'}^k) - V_{h'}^{\pi^k}(s_{h'}^k) = Q_{k,h'}(s_{h'}^k, a_{h'}^k) - Q_{h'}^{\pi^k}(s_{h'}^k, a_{h'}^k)\\
&\leq \min\left\{\hat{\mathbf{w}}_{k_{\text{last}},h'}^\top\boldsymbol{\phi}(s_{h'}^k, a_{h'}^k) + \beta\sqrt{\boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)^\top\boldsymbol{\Sigma}_{k_{\text{last}},h'}^{-1}\boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)}, H\right\} - \mathbb{P}_{s_{h'}^k, a_{h'}^k, h'}V_{k,h'+1}\\
&\quad + \mathbb{P}_{s_{h'}^k, a_{h'}^k, h'}\left(V_{k,h'+1} - V_{h'+1}^{\pi^k}\right)\\
&\leq \mathbb{P}_{s_{h'}^k, a_{h'}^k, h'}\left(V_{k,h'+1} - V_{h'+1}^{\pi^k}\right) + 2\min\left\{\beta\sqrt{\boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)^\top\boldsymbol{\Sigma}_{k_{\text{last}},h'}^{-1}\boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)}, H\right\}\\
&\leq \mathbb{P}_{s_{h'}^k, a_{h'}^k, h'}\left(V_{k,h'+1} - V_{h'+1}^{\pi^k}\right) + 4\min\left\{\beta\sqrt{\boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)^\top\boldsymbol{\Sigma}_{k,h'}^{-1}\boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)}, H\right\}\\
&= V_{k,h'+1}(s_{h'+1}^k) - V_{h'+1}^{\pi^k}(s_{h'+1}^k) + \left(\mathbb{P}_{s_{h'}^k, a_{h'}^k, h'} - \mathbb{1}_{s_{h'+1}^k}\right)\left(V_{k,h'+1} - V_{h'+1}^{\pi^k}\right)\\
&\quad + 4\min\left\{\beta\sqrt{\boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)^\top\boldsymbol{\Sigma}_{k,h'}^{-1}\boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)}, H\right\}. \qquad (14)
\end{aligned}$$

Here, the first inequality holds due to the update rule of $Q_{k,h}(s_h^k, a_h^k)$ in Line 11 of Algorithm 1 and the Bellman equation in Equation (1), the second inequality holds due to the event $\mathcal{E}_2$ in Lemma E.1 and the fact that $V_{k,h'+1} = V_{k_{\text{last}},h'+1}$, and the last inequality holds due to the update condition in Line 10 of Algorithm 1 and Lemma C.3.

Equation (14) also holds when $k_{\text{last}} = 0$, i.e., before the first policy update. Define $\hat{\mathbf{w}}_{0,h} = 0$ for every $h \in [H]$. By the definition of β and $\lambda = H^{-2}$, we have $\beta \geq \sqrt{d}$. Moreover, $\langle\boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h(\mathcal{S})\rangle = 1$ and $\|\boldsymbol{\theta}_h(\mathcal{S})\|_2 \leq \sqrt{d}$ imply $\|\boldsymbol{\phi}(s,a)\|_2 \geq 1/\sqrt{d}$, and hence

$$\beta\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{0,h}^{-1}} = \frac{\beta}{\sqrt{\lambda}}\|\boldsymbol{\phi}(s,a)\|_2 \geq H.$$

Therefore, before the first policy update,

$$Q_{k,h}(s,a) = Q_{0,h}(s,a) = H = \min\left\{r_h(s,a) + \hat{\mathbf{w}}_{0,h}^\top\boldsymbol{\phi}(s,a) + \beta\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{0,h}^{-1}}, H\right\}.$$

Thus, the first inequality still holds. For the second inequality, since $\min\{\beta\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{0,h}^{-1}}, H\} = H$ and $\mathbb{P}_{s,a,h}V_{k,h+1} \geq 0$, we have

$$\min\left\{\hat{\mathbf{w}}_{0,h}^\top \boldsymbol{\phi}(s,a) + \beta\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{0,h}^{-1}}, H\right\} - \mathbb{P}_{s,a,h}V_{k,h+1} = H - \mathbb{P}_{s,a,h}V_{k,h+1} \leq 2\min\left\{\beta\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{0,h}^{-1}}, H\right\}.$$

Furthermore, since the update condition has not been triggered, $\boldsymbol{\Sigma}_{k,h} \succeq \boldsymbol{\Sigma}_{0,h}$ and $\det(\boldsymbol{\Sigma}_{k,h}) < 2\det(\boldsymbol{\Sigma}_{0,h})$ for every $h \in [H]$. Lemma C.3 therefore gives

$$\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{0,h}^{-1}} \leq \sqrt{2}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{k,h}^{-1}} \leq 2\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{k,h}^{-1}},$$

which proves the third inequality. Hence, all steps in Equation (14) remain valid when $k_{\text{last}} = 0$.

For notational convenience, define

$$\epsilon_{k,h} := \left(\mathbb{P}_{s_h^k,a_h^k,h} - \mathbb{1}_{s_{h+1}^k}\right)\left(V_{k,h+1} - V_{h+1}^{\pi^k}\right)$$

and

$$b_{k,h} := \min\left\{\beta\sqrt{\boldsymbol{\phi}(s_h^k,a_h^k)^\top \boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s_h^k,a_h^k)}, H\right\}.$$

For every $k \in \mathcal{K}_\tau$ and $h \in \{\tau_k, \ldots, H\}$, recursively applying Equation (14) from step $h+1$ to step H and using $V_{k,H+1} = V_{H+1}^{\pi^k} = 0$ gives

$$V_{k,h+1}(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k) \leq \sum_{h'=h+1}^{H} \epsilon_{k,h'} + 4\sum_{h'=h+1}^{H} b_{k,h'}.$$

Consequently,

$$\mathbb{P}_{s_h^k,a_h^k,h}\left(V_{k,h+1} - V_{h+1}^{\pi^k}\right) = V_{k,h+1}(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k) + \epsilon_{k,h} \leq \sum_{h'=h}^{H} \epsilon_{k,h'} + 4\sum_{h'=h}^{H} b_{k,h'}.$$

Summing the preceding inequality over $h = \tau_k, \ldots, H$ and exchanging the order of summation gives

$$\sum_{h=\tau_k}^{H} \mathbb{P}_{s_h^k,a_h^k,h}\left(V_{k,h+1} - V_{h+1}^{\pi^k}\right) \leq \sum_{h=\tau_k}^{H}(h-\tau_k+1)\epsilon_{k,h} + 4\sum_{h=\tau_k}^{H}(h-\tau_k+1)b_{k,h}. \tag{15}$$

On $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_4$, by Equation (9), we have $V_{k,h}^{\text{clip}} = V_{k,h}$. Then the definition of $\mathcal{E}_4$ yields

$$\sum_{(k,h)\in\mathcal{J}_\tau}(h-\tau_k+1)\left(\mathbb{P}_{s_h^k,a_h^k,h} - \mathbb{1}_{s_{h+1}^k}\right)\left(V_{k,h+1} - V_{h+1}^{\pi^k}\right) \leq 2\sqrt{2H^5 L\log\left(\frac{KH}{\delta}\right)}. \tag{16}$$

Summing Equation (15) over all $k \in \mathcal{K}_\tau$ and applying Equation (16), we obtain

$$\sum_{(k,h)\in\mathcal{J}_\tau} \mathbb{P}_{s_h^k,a_h^k,h}\left(V_{k,h+1} - V_{h+1}^{\pi^k}\right) \leq 2\sqrt{2H^5 L\log(KH/\delta)} + \sum_{h=1}^{H}\sum_{k\in\mathcal{K}_h} 4(h-\tau_k+1)b_{k,h}. \tag{17}$$

For every $h \in [H]$, since $h - \tau_k + 1 \leq H$ for every $k \in \mathcal{K}_h$, Lemma F.3 gives

$$\sum_{k\in\mathcal{K}_h} 4(h-\tau_k+1)b_{k,h} \leq 4H\sum_{k\in\mathcal{K}_h} b_{k,h} \leq 40d^4H^5\iota + 40\beta d^4H^3\iota + 8\beta H\sqrt{d\iota\sum_{k\in\mathcal{K}_h}\left(\sigma_{k,h}^2 + H\right)}.$$

Summing the preceding inequality over $h \in [H]$ and applying the Cauchy–Schwarz inequality gives

$$\begin{aligned}
&\sum_{h=1}^{H} \sum_{k \in \mathcal{K}_h} 4(h - \tau_k + 1) \min\left\{\beta \sqrt{\boldsymbol{\phi}(s_h^k, a_h^k)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^k, a_h^k)}, H\right\} \\
&\le 40 d^4 H^6 \iota + 40 \beta d^4 H^4 \iota + 8 \beta H \sum_{h=1}^{H} \sqrt{d\iota \sum_{k \in \mathcal{K}_h} \left(\sigma_{k,h}^2 + H\right)} \\
&\le 40 d^4 H^6 \iota + 40 \beta d^4 H^4 \iota + 8 \beta \sqrt{d H^3 \iota \sum_{(k,h) \in \mathcal{J}_\tau} \left(\sigma_{k,h}^2 + H\right)}.
\end{aligned}$$

Substituting this bound into Equation (17) completes the proof. □

In addition, for the cumulative difference between the optimistic estimate $V_{k,h}(s)$ and the pessimistic estimate $\check{V}_{k,h}(s)$ over the random suffixes, we have the following lemma.

Lemma F.5. *On the events $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_5$ in Lemma E.1, we have*

$$\begin{aligned}
\sum_{(k,h) \in \mathcal{J}_\tau} \mathbb{P}_{s_h^k, a_h^k, h} \left(V_{k,h+1} - \check{V}_{k,h+1}\right) &\le 80 d^4 H^6 \iota + 40 \left(\beta + \bar{\beta}\right) d^4 H^4 \iota \\
&+ 8 \left(\beta + \bar{\beta}\right) \sqrt{d H^3 \iota \sum_{(k,h) \in \mathcal{J}_\tau} \left(\sigma_{k,h}^2 + H\right)} + 2\sqrt{2 H^5 L \log(KH/\delta)}.
\end{aligned}$$

Proof of Lemma F.5. For each step $h' \in [H]$ and episode $k \in [K]$, if $k_{\text{last}} \ge 1$, we have

$$\begin{aligned}
&V_{k,h'}(s_{h'}^k) - \check{V}_{k,h'}(s_{h'}^k) \\
&\le Q_{k,h'}(s_{h'}^k, a_{h'}^k) - \check{Q}_{k,h'}(s_{h'}^k, a_{h'}^k) \\
&\le \min\left\{r_{h'}(s_{h'}^k, a_{h'}^k) + \hat{\mathbf{w}}_{k_{\text{last}},h'}^\top \boldsymbol{\phi}(s_{h'}^k, a_{h'}^k) + \beta \sqrt{\boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)^\top \boldsymbol{\Sigma}_{k_{\text{last}},h'}^{-1} \boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)}, H\right\} \\
&\quad - \max\left\{r_{h'}(s_{h'}^k, a_{h'}^k) + \check{\mathbf{w}}_{k_{\text{last}},h'}^\top \boldsymbol{\phi}(s_{h'}^k, a_{h'}^k) - \bar{\beta} \sqrt{\boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)^\top \boldsymbol{\Sigma}_{k_{\text{last}},h'}^{-1} \boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)}, 0\right\} \\
&\le \min\left\{r_{h'}(s_{h'}^k, a_{h'}^k) + \mathbb{P}_{s_{h'}^k, a_{h'}^k, h'} V_{k,h'+1} + 2\beta \sqrt{\boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)^\top \boldsymbol{\Sigma}_{k_{\text{last}},h'}^{-1} \boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)}, H\right\} \\
&\quad - \max\left\{r_{h'}(s_{h'}^k, a_{h'}^k) + \mathbb{P}_{s_{h'}^k, a_{h'}^k, h'} \check{V}_{k,h'+1} - 2\bar{\beta} \sqrt{\boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)^\top \boldsymbol{\Sigma}_{k_{\text{last}},h'}^{-1} \boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)}, 0\right\} \\
&\le \mathbb{P}_{s_{h'}^k, a_{h'}^k, h'} \left(V_{k,h'+1} - \check{V}_{k,h'+1}\right) + 2 \min\left\{\beta \sqrt{\boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)^\top \boldsymbol{\Sigma}_{k_{\text{last}},h'}^{-1} \boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)}, H\right\} \\
&\quad + 2 \min\left\{\bar{\beta} \sqrt{\boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)^\top \boldsymbol{\Sigma}_{k_{\text{last}},h'}^{-1} \boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)}, H\right\} \\
&\le V_{k,h'+1}(s_{h'+1}^k) - \check{V}_{k,h'+1}(s_{h'+1}^k) + \left(\mathbb{P}_{s_{h'}^k, a_{h'}^k, h'} - \mathbb{1}_{s_{h'+1}^k}\right) \left(V_{k,h'+1} - \check{V}_{k,h'+1}\right) + \\
&4 \min\left\{\beta \sqrt{\boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)^\top \boldsymbol{\Sigma}_{k,h'}^{-1} \boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)}, H\right\} + 4 \min\left\{\bar{\beta} \sqrt{\boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)^\top \boldsymbol{\Sigma}_{k,h'}^{-1} \boldsymbol{\phi}(s_{h'}^k, a_{h'}^k)}, H\right\}.
\end{aligned} \tag{18}$$

Here, the first inequality follows from

$$\check{V}_{k,h'}(s_{h'}^k) = \max_a \check{Q}_{k,h'}(s_{h'}^k, a) \ge \check{Q}_{k,h'}(s_{h'}^k, a_{h'}^k).$$

The second inequality follows from the update rules of $Q_{k,h}$ and $\check{Q}_{k,h}$ in Lines 11–12 of Algorithm 1. The third inequality follows from the events $\mathcal{E}_1$ and $\mathcal{E}_2$ in Lemma E.1 and the fact that $V_{k,h'+1} = V_{k_{\text{last}},h'+1}$ and $\check{V}_{k,h'+1} = \check{V}_{k_{\text{last}},h'+1}$. The fourth inequality uses $0 \le r_{h'}(s_{h'}^k, a_{h'}^k) \le 1$, $0 \le V_{k,h'+1}, \check{V}_{k,h'+1} \le H$, and the clipping inequalities

$$\min\{x + 2y, H\} \le x + 2\min\{y, H\}, \qquad \max\{x - 2y, 0\} \ge x - 2\min\{y, H\},$$

which hold for every $x \in [0, 2H]$ and $y \geq 0$. The last inequality follows from the update condition in Line 10 of Algorithm 1 and Lemma C.3.

Equation (18) also holds when $k_{\text{last}} = 0$, i.e., before the first policy update. Define $\hat{\mathbf{w}}_{0,h} = \check{\mathbf{w}}_{0,h} = 0$ for every $h \in [H]$. By the definitions of β and $\bar{\beta}$ and $\lambda = H^{-2}$, we have $\beta, \bar{\beta} \geq \sqrt{d}$. Moreover, $\langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h(\mathcal{S}) \rangle = 1$ and $\|\boldsymbol{\theta}_h(\mathcal{S})\|_2 \leq \sqrt{d}$ imply $\|\boldsymbol{\phi}(s,a)\|_2 \geq 1/\sqrt{d}$, and hence

$$\beta \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{0,h}^{-1}} \geq H, \qquad \bar{\beta} \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{0,h}^{-1}} \geq H.$$

Therefore, before the first policy update,

$$Q_{k,h}(s,a) = Q_{0,h}(s,a) = \min\left\{ r_h(s,a) + \hat{\mathbf{w}}_{0,h}^\top \boldsymbol{\phi}(s,a) + \beta \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{0,h}^{-1}}, H \right\} = H$$

and

$$\check{Q}_{k,h}(s,a) = \check{Q}_{0,h}(s,a) = \max\left\{ r_h(s,a) + \check{\mathbf{w}}_{0,h}^\top \boldsymbol{\phi}(s,a) - \bar{\beta} \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{0,h}^{-1}}, 0 \right\} = 0,$$

where the last equality uses $0 \leq r_h(s,a) \leq H$. Thus, the first inequality in Equation (18) follows from the same greedy-action argument, and the second inequality holds with equality. Moreover, since $0 \leq V_{k,h+1}, \check{V}_{k,h+1} \leq H$, we have

$$\left| \hat{\mathbf{w}}_{0,h}^\top \boldsymbol{\phi}(s,a) - \mathbb{P}_{s,a,h} V_{k,h+1} \right| \leq H \leq \beta \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{0,h}^{-1}}$$

and

$$\left| \check{\mathbf{w}}_{0,h}^\top \boldsymbol{\phi}(s,a) - \mathbb{P}_{s,a,h} \check{V}_{k,h+1} \right| \leq H \leq \bar{\beta} \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{0,h}^{-1}},$$

which proves the third inequality. The fourth inequality then follows from the same clipping inequalities used above. Finally, since the update condition has not been triggered,

$$\boldsymbol{\Sigma}_{k,h} \succeq \boldsymbol{\Sigma}_{0,h}, \qquad \det(\boldsymbol{\Sigma}_{k,h}) < 2 \det(\boldsymbol{\Sigma}_{0,h}),$$

and Lemma C.3 gives

$$\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{0,h}^{-1}} \leq \sqrt{2} \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{k,h}^{-1}} \leq 2 \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}.$$

This proves the last step in Equation (18) for $k_{\text{last}} = 0$.

For notational convenience, define

$$\check{\epsilon}_{k,h} := \left(\mathbb{P}_{s_h^k, a_h^k, h} - \mathbb{1}_{s_{h+1}^k} \right) \left(V_{k,h+1} - \check{V}_{k,h+1} \right),$$

$$b_{k,h} := \min\left\{ \beta \sqrt{\boldsymbol{\phi}(s_h^k, a_h^k)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^k, a_h^k)}, H \right\},$$

and

$$\check{b}_{k,h} := \min\left\{ \bar{\beta} \sqrt{\boldsymbol{\phi}(s_h^k, a_h^k)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^k, a_h^k)}, H \right\}.$$

For every $k \in \mathcal{K}_\tau$ and $h \in \{\tau_k, \ldots, H\}$, recursively applying Equation (18) from step $h+1$ to step H and using $V_{k,H+1} = \check{V}_{k,H+1} = 0$ gives

$$V_{k,h+1}(s_{h+1}^k) - \check{V}_{k,h+1}(s_{h+1}^k) \leq \sum_{h'=h+1}^{H} \check{\epsilon}_{k,h'} + 4 \sum_{h'=h+1}^{H} \left(b_{k,h'} + \check{b}_{k,h'} \right).$$

Consequently,

$$\begin{aligned} \mathbb{P}_{s_h^k, a_h^k, h} \left(V_{k,h+1} - \check{V}_{k,h+1} \right) &= V_{k,h+1}(s_{h+1}^k) - \check{V}_{k,h+1}(s_{h+1}^k) + \check{\epsilon}_{k,h} \\ &\leq \sum_{h'=h}^{H} \check{\epsilon}_{k,h'} + 4 \sum_{h'=h}^{H} \left(b_{k,h'} + \check{b}_{k,h'} \right). \end{aligned}$$

Summing the preceding inequality over $h = \tau_k, \ldots, H$ and exchanging the order of summation gives

$$\sum_{h=\tau_k}^{H} \mathbb{P}_{s_h^k, a_h^k, h} \left(V_{k,h+1} - \check{V}_{k,h+1} \right) \leq \sum_{h=\tau_k}^{H} (h - \tau_k + 1) \check{\epsilon}_{k,h} + 4 \sum_{h=\tau_k}^{H} (h - \tau_k + 1) \left(b_{k,h} + \check{b}_{k,h} \right). \tag{19}$$

On $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_5$, by Equation (9), we have $\check{V}_{k,h}^{\text{clip}} = \check{V}_{k,h}, V_{k,h}^{\text{clip}} = V_{k,h}$. Then the definition of $\mathcal{E}_5$ yields

$$\sum_{(k,h)\in\mathcal{J}_\tau} (h-\tau_k+1)\left(\mathbb{P}_{s_h^k,a_h^k,h} - \mathbb{1}_{s_{h+1}^k}\right)\left(V_{k,h+1} - \check{V}_{k,h+1}\right) \le 2\sqrt{2H^5 L \log\left(\frac{KH}{\delta}\right)}. \quad (20)$$

Summing Equation (19) over all $k \in \mathcal{K}_\tau$ and applying Equation (20), we obtain

$$\begin{aligned}&\sum_{(k,h)\in\mathcal{J}_\tau} \mathbb{P}_{s_h^k,a_h^k,h}\left(V_{k,h+1} - \check{V}_{k,h+1}\right)\\ &\le 2\sqrt{2H^5 L \log\left(\frac{KH}{\delta}\right)} + \sum_{h=1}^{H}\sum_{k\in\mathcal{K}_h} 4(h-\tau_k+1)b_{k,h} + \sum_{h=1}^{H}\sum_{k\in\mathcal{K}_h} 4(h-\tau_k+1)\check{b}_{k,h}. \quad (21)\end{aligned}$$

For every $h \in [H]$, Lemma F.3 and $h - \tau_k + 1 \le H$ give

$$\sum_{k\in\mathcal{K}_h} 4(h-\tau_k+1)b_{k,h} \le 40d^4H^5\iota + 40\beta d^4H^3\iota + 8\beta H\sqrt{d\iota\sum_{k\in\mathcal{K}_h}\left(\sigma_{k,h}^2 + H\right)}$$

and

$$\sum_{k\in\mathcal{K}_h} 4(h-\tau_k+1)\check{b}_{k,h} \le 40d^4H^5\iota + 40\bar{\beta} d^4H^3\iota + 8\bar{\beta} H\sqrt{d\iota\sum_{k\in\mathcal{K}_h}\left(\sigma_{k,h}^2 + H\right)}.$$

Summing the preceding two inequalities over $h \in [H]$ and applying the Cauchy–Schwarz inequality gives

$$\begin{aligned}&\sum_{h=1}^{H}\sum_{k\in\mathcal{K}_h} 4(h-\tau_k+1)b_{k,h} + \sum_{h=1}^{H}\sum_{k\in\mathcal{K}_h} 4(h-\tau_k+1)\check{b}_{k,h}\\ &\le 80d^4H^6\iota + 40\left(\beta+\bar{\beta}\right)d^4H^4\iota + 8\left(\beta+\bar{\beta}\right)\sqrt{dH^3\iota\sum_{(k,h)\in\mathcal{J}_\tau}\left(\sigma_{k,h}^2 + H\right)}.\end{aligned}$$

Substituting this bound into Equation (21) completes the proof. □

G Proof of Theorem 3.1

We first prove Lemma G.1, which connects the expected regret with the value difference at the first realized positive-gap step.

Lemma G.1 (Restatement of Lemma 4.1)**.** *For LSVI-UCB++ with K episodes and $T = HK$ steps, the expected regret satisfies*

$$\mathbb{E}\left[\text{Regret}(T)\right] = \mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\Delta_h(s_h^k, a_h^k)\right] = \mathbb{E}\left[\sum_{k\in\mathcal{K}_\tau}\left(V_{\tau_k}^\star - V_{\tau_k}^{\pi^k}\right)(s_{\tau_k}^k)\right].$$

Proof of Lemma G.1. For any fixed episode $k \in [K]$, the performance-difference identity gives

$$\left(V_1^\star - V_1^{\pi^k}\right)(s_1^k) = \mathbb{E}\left[\sum_{h=1}^{H}\Delta_h(s_h^k, a_h^k)\middle|\mathcal{F}_{k,1}\right]. \quad (22)$$

Indeed, for any $h \in [H]$, since $a_h^k = \pi_h^k(s_h^k)$, we have

$$\begin{aligned}\left(V_h^\star - V_h^{\pi^k}\right)(s_h^k) &= V_h^\star(s_h^k) - Q_h^\star(s_h^k, a_h^k) + \left(Q_h^\star - Q_h^{\pi^k}\right)(s_h^k, a_h^k)\\ &= \Delta_h(s_h^k, a_h^k) + \mathbb{E}\left[\left(V_{h+1}^\star - V_{h+1}^{\pi^k}\right)(s_{h+1}^k)\middle|\mathcal{F}_{k,h}\right],\end{aligned}$$

where the second equality follows from the Bellman equation in Equation (1). Recursively applying this equality from $h = 1$ to $h = H$ proves Equation (22).

We next express the cumulative gaps using the first realized positive-gap step τ_k. By the definition of τ_k, on the event $\{\tau_k \le H\}$, we have $\Delta_h(s_h^k, a_h^k) = 0$ for every $h < \tau_k$. On the event $\{\tau_k = H + 1\}$, we have $\Delta_h(s_h^k, a_h^k) = 0$ for every $h \in [H]$. Therefore, pathwise,

$$\sum_{h=1}^{H} \Delta_h(s_h^k, a_h^k) = \mathbb{I}[\tau_k \le H] \sum_{h=\tau_k}^{H} \Delta_h(s_h^k, a_h^k). \tag{23}$$

For any fixed $t \in [H]$, recursively applying the preceding Bellman decomposition from step t to step H gives

$$\left(V_t^\star - V_t^{\pi^k}\right)(s_t^k) = \mathbb{E}\left[\sum_{h=t}^{H} \Delta_h(s_h^k, a_h^k) \middle| \mathcal{F}_{k,t}\right].$$

Since $\{\tau_k = t\} \in \mathcal{F}_{k,t}$, multiplying the preceding equality by $\mathbb{I}[\tau_k = t]$ and taking expectations yields

$$\mathbb{E}\left[\mathbb{I}[\tau_k = t]\left(V_t^\star - V_t^{\pi^k}\right)(s_t^k)\right] = \mathbb{E}\left[\mathbb{I}[\tau_k = t] \sum_{h=t}^{H} \Delta_h(s_h^k, a_h^k)\right].$$

Summing the preceding equality over $t \in [H]$ gives

$$\mathbb{E}\left[\mathbb{I}[\tau_k \le H]\left(V_{\tau_k}^\star - V_{\tau_k}^{\pi^k}\right)(s_{\tau_k}^k)\right] = \mathbb{E}\left[\mathbb{I}[\tau_k \le H] \sum_{h=\tau_k}^{H} \Delta_h(s_h^k, a_h^k)\right].$$

Combining this equality with Equation (23), we obtain

$$\mathbb{E}\left[\sum_{h=1}^{H} \Delta_h(s_h^k, a_h^k)\right] = \mathbb{E}\left[\mathbb{I}[\tau_k \le H]\left(V_{\tau_k}^\star - V_{\tau_k}^{\pi^k}\right)(s_{\tau_k}^k)\right]. \tag{24}$$

Finally, combining Equation (22) and Equation (24) and summing over all episodes gives

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Regret}(T)\right] &= \mathbb{E}\left[\sum_{k=1}^{K}\left(V_1^\star - V_1^{\pi^k}\right)(s_1^k)\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H} \Delta_h(s_h^k, a_h^k)\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{I}[\tau_k \le H]\left(V_{\tau_k}^\star - V_{\tau_k}^{\pi^k}\right)(s_{\tau_k}^k)\right] \\
&= \mathbb{E}\left[\sum_{k \in \mathcal{K}_\tau}\left(V_{\tau_k}^\star - V_{\tau_k}^{\pi^k}\right)(s_{\tau_k}^k)\right].
\end{aligned}$$

This completes the proof. □

We then prove Lemma G.2, which bounds the sum of the estimated variances over the random suffix index set $\mathcal{J}_\tau$.

Lemma G.2. *On the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_4 \cap \mathcal{E}_5 \cap \mathcal{E}_6$, the sum of the estimated variances over $\mathcal{J}_\tau$ is upper bounded as*

$$\sum_{(k,h) \in \mathcal{J}_\tau} \sigma_{k,h}^2 \le O\left(H^2 L + d^{10} H^9 \iota_1^3\right).$$

Proof of Lemma G.2. According to the definition of $\sigma_{k,h}^2$ in Equation (2), we have

$$\begin{aligned}
&\sum_{(k,h)\in\mathcal{J}_\tau} \sigma_{k,h}^2 = \sum_{(k,h)\in\mathcal{J}_\tau} \left(\bar{\mathbb{V}}_{s_h^k,a_h^k,h} V_{k,h+1} + E_{k,h} + D_{k,h} + H\right) \\
&\le H^2 L + \underbrace{\sum_{(k,h)\in\mathcal{J}_\tau} \left(\bar{\mathbb{V}}_{s_h^k,a_h^k,h} V_{k,h+1} - \mathbb{V}_{s_h^k,a_h^k,h} V_{k,h+1}\right)}_{I_1} + \underbrace{\sum_{(k,h)\in\mathcal{J}_\tau} E_{k,h}}_{I_2} \\
&\quad + \underbrace{\sum_{(k,h)\in\mathcal{J}_\tau} D_{k,h}}_{I_3} + \underbrace{\sum_{(k,h)\in\mathcal{J}_\tau} \left(\mathbb{V}_{s_h^k,a_h^k,h} V_{k,h+1} - \mathbb{V}_{s_h^k,a_h^k,h} V_{h+1}^{\pi^k}\right)}_{I_4} + \underbrace{\sum_{(k,h)\in\mathcal{J}_\tau} \mathbb{V}_{s_h^k,a_h^k,h} V_{h+1}^{\pi^k}}_{I_5}. \quad (25)
\end{aligned}$$

Here, we used

$$\sum_{(k,h)\in\mathcal{J}_\tau} H = H|\mathcal{J}_\tau| = H \sum_{k\in\mathcal{K}_\tau} (H - \tau_k + 1) \le H^2 L.$$

For the term I_1, according to Lemma F.2, it is upper bounded by

$$I_1 = \sum_{(k,h)\in\mathcal{J}_\tau} \left(\bar{\mathbb{V}}_{s_h^k,a_h^k,h} V_{k,h+1} - \mathbb{V}_{s_h^k,a_h^k,h} V_{k,h+1}\right) \le \sum_{(k,h)\in\mathcal{J}_\tau} E_{k,h} = I_2. \quad (26)$$

For the term I_2, by the definition of $E_{k,h}$, we have

$$\begin{aligned}
I_2 &= \sum_{h=1}^{H} \sum_{k\in\mathcal{K}_h} \Bigg(\min\left\{\tilde{\beta}\sqrt{\boldsymbol{\phi}(s_h^k,a_h^k)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^k,a_h^k)}, H^2\right\} \\
&\quad + \min\left\{2H\bar{\beta}\sqrt{\boldsymbol{\phi}(s_h^k,a_h^k)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^k,a_h^k)}, H^2\right\}\Bigg) \\
&\le 20d^4H^6\iota + \left(10\tilde{\beta} + 20\bar{\beta}\right) d^4H^4\iota + \left(2\tilde{\beta} + 4\bar{\beta}\right)\sqrt{dH^3\iota \sum_{(k,h)\in\mathcal{J}_\tau} \left(\sigma_{k,h}^2 + H\right)}, \quad (27)
\end{aligned}$$

where the inequality follows by applying Lemma F.3 to $\mathcal{K}_h$ for every $h \in [H]$, summing the resulting inequalities over $h \in [H]$, and applying the Cauchy–Schwarz inequality.

For the term I_3, by the definition of $D_{k,h}$, we have

$$\begin{aligned}
I_3 &= \sum_{(k,h)\in\mathcal{J}_\tau} \min\left\{4d^3H^2\left(\hat{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s_h^k,a_h^k) - \check{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s_h^k,a_h^k) + 2\bar{\beta}\|\boldsymbol{\phi}(s_h^k,a_h^k)\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}\right), d^3H^3\right\} \\
&\le \sum_{(k,h)\in\mathcal{J}_\tau} \min\left\{4d^3H^2\left(\mathbb{P}_{s_h^k,a_h^k,h}\left(V_{k,h+1} - \check{V}_{k,h+1}\right) + 4\bar{\beta}\|\boldsymbol{\phi}(s_h^k,a_h^k)\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}\right), d^3H^3\right\} \\
&\le 4d^3H^2 \sum_{(k,h)\in\mathcal{J}_\tau} \mathbb{P}_{s_h^k,a_h^k,h}\left(V_{k,h+1} - \check{V}_{k,h+1}\right) \\
&\quad + \sum_{h=1}^{H}\sum_{k\in\mathcal{K}_h} \min\left\{16d^3H^2\bar{\beta}\sqrt{\boldsymbol{\phi}(s_h^k,a_h^k)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^k,a_h^k)}, d^3H^3\right\} \\
&\le 4d^3H^2 \sum_{(k,h)\in\mathcal{J}_\tau} \mathbb{P}_{s_h^k,a_h^k,h}\left(V_{k,h+1} - \check{V}_{k,h+1}\right) + 160d^7H^7\iota + 160\bar{\beta}d^7H^5\iota \\
&\quad + 32d^3H^2\bar{\beta}\sqrt{dH\iota \sum_{(k,h)\in\mathcal{J}_\tau} \left(\sigma_{k,h}^2 + H\right)} \\
&\le 480d^7H^8\iota + 320\left(\beta + \bar{\beta}\right) d^7H^6\iota + 64d^3H^2\left(\beta + \bar{\beta}\right)\sqrt{dH^3\iota \sum_{(k,h)\in\mathcal{J}_\tau} \left(\sigma_{k,h}^2 + H\right)} \\
&\quad + 16d^3\sqrt{H^9 L \log(KH/\delta)}. \quad (28)
\end{aligned}$$

Here, the first inequality follows from the event $\mathcal{E}_1$ in Lemma E.1. The second inequality follows from Lemma F.1, which gives $V_{k,h+1}(s) \geq V^\star_{h+1}(s) \geq \check{V}_{k,h+1}(s)$. The third inequality follows by applying Lemma F.3 to $\mathcal{K}_h$ for every $h \in [H]$ and then applying the Cauchy–Schwarz inequality. The last inequality follows from Lemma F.5 and the facts that $H \geq 1$ and $\bar{\beta} \leq \beta + \bar{\beta}$.

For the term I_4, we have

$$\begin{aligned}
I_4 &= \sum_{(k,h)\in\mathcal{J}_\tau} \left(\mathbb{P}_{s_h^k,a_h^k,h} \left(V_{k,h+1}\right)^2 - \left(\mathbb{P}_{s_h^k,a_h^k,h} V_{k,h+1}\right)^2 - \mathbb{P}_{s_h^k,a_h^k,h}(V_{h+1}^{\pi^k})^2 + \left(\mathbb{P}_{s_h^k,a_h^k,h} V_{h+1}^{\pi^k}\right)^2 \right) \\
&\leq \sum_{(k,h)\in\mathcal{J}_\tau} \left(\mathbb{P}_{s_h^k,a_h^k,h} \left(V_{k,h+1}\right)^2 - \mathbb{P}_{s_h^k,a_h^k,h} \left(V_{h+1}^{\pi^k}\right)^2 \right) \\
&\leq 2H \sum_{(k,h)\in\mathcal{J}_\tau} \mathbb{P}_{s_h^k,a_h^k,h} \left(V_{k,h+1} - V_{h+1}^{\pi^k}\right) \\
&\leq 80d^4H^7\iota + 80\beta d^4H^5\iota + 16H\beta\sqrt{dH^3\iota \sum_{(k,h)\in\mathcal{J}_\tau} \left(\sigma_{k,h}^2 + H\right)} + 8\sqrt{H^7L\log(KH/\delta)}. \quad (29)
\end{aligned}$$

Here, the first inequality follows because Lemma F.1 gives $V_{h+1}^{\pi^k}(s') \leq V^\star_{h+1}(s') \leq V_{k,h+1}(s')$. The second inequality follows from $0 \leq V_{k,h+1}(s'), V_{h+1}^{\pi^k}(s') \leq H$. The last inequality follows from Lemma F.4.

By the event $\mathcal{E}_6$ in Lemma E.1, the term I_5 satisfies

$$I_5 \leq 2H^2L + 3H^3\log(K/\delta). \quad (30)$$

Let

$$S_\tau := \sum_{(k,h)\in\mathcal{J}_\tau} \sigma_{k,h}^2, \qquad \bar{S}_\tau := \sum_{(k,h)\in\mathcal{J}_\tau} \left(\sigma_{k,h}^2 + H\right).$$

Since $|\mathcal{J}_\tau| \leq HL$, we have

$$\bar{S}_\tau = S_\tau + H|\mathcal{J}_\tau| \leq S_\tau + H^2L.$$

Substituting Equation (26), Equation (27), Equation (28), Equation (29), and Equation (30) into Equation (25), and absorbing lower-order terms, gives

$$\begin{aligned}
S_\tau \leq O\Bigg(& H^2L + d^7H^8\iota + d^7H^6(\beta+\bar{\beta})\iota + d^4H^4(\tilde{\beta}+\bar{\beta})\iota + d^3\sqrt{H^9L\log(KH/\delta)} \\
& + H^3\log(K/\delta) + \left(\tilde{\beta} + \bar{\beta} + d^3H^2(\beta+\bar{\beta}) + H\beta\right)\sqrt{dH^3\iota\left(S_\tau + H^2L\right)}\Bigg). \quad (31)
\end{aligned}$$

Since $\lambda = 1/H^2$, the definitions of β, $\tilde{\beta}$, and $\bar{\beta}$ imply

$$\beta = O\left(\sqrt{d}\,\iota_1\right), \qquad \bar{\beta} = O\left(d^{3/2}H\iota_1\right), \qquad \tilde{\beta} = O\left(d^{3/2}H^2\iota_1\right).$$

Moreover, $\iota \leq O(\iota_1)$ and $\log(KH/\delta) \leq O(\iota_1)$. Therefore,

$$(\tilde{\beta}+\bar{\beta})\sqrt{dH^3\iota} \leq O\left(d^2H^{\frac{7}{2}}\iota_1^{\frac{3}{2}}\right),$$

$$d^3H^2(\beta+\bar{\beta})\sqrt{dH^3\iota} \leq O\left(d^5H^{\frac{9}{2}}\iota_1^{\frac{3}{2}}\right),$$

and

$$H\beta\sqrt{dH^3\iota} \leq O\left(dH^{\frac{5}{2}}\iota_1^{\frac{3}{2}}\right).$$

Hence the last term in Equation (31) is bounded by

$$O\left(d^5H^{\frac{9}{2}}\iota_1^{\frac{3}{2}}\sqrt{S_\tau + H^2L}\right) \leq O\left(d^5H^{\frac{9}{2}}\iota_1^{\frac{3}{2}}\sqrt{S_\tau} + d^5H^{\frac{11}{2}}\iota_1^{\frac{3}{2}}\sqrt{L}\right).$$

Using $2\sqrt{ab} \leq a + b$, we have

$$d^3\sqrt{H^9L\iota_1} \leq O\left(H^2L + d^6H^7\iota_1\right)$$

and

$$d^5 H^{\frac{11}{2}} \iota_1^{\frac{3}{2}} \sqrt{L} \le O\left(H^2 L + d^{10} H^9 \iota_1^3\right).$$

The remaining deterministic terms in Equation (31) are also bounded by $O(d^{10}H^9\iota_1^3)$. Consequently,

$$S_\tau \le O\left(H^2 L + d^{10} H^9 \iota_1^3 + d^5 H^{\frac{9}{2}} \iota_1^{\frac{3}{2}} \sqrt{S_\tau}\right).$$

Using the fact that $x \le a\sqrt{x} + b$ implies $x \le a^2 + 2b$, we conclude that

$$S_\tau \le O\left(H^2 L + d^{10} H^9 \iota_1^3\right).$$

This completes the proof. □

With the lemma established above, we next bound the cumulative optimistic error at the first realized positive-gap steps.

Lemma G.3 (Formal Statement of Lemma 4.2)**.** *On the event $\bigcap_{i=1}^6 \mathcal{E}_i$, we have*

$$\sum_{k \in \mathcal{K}_\tau} \left(Q_{k,\tau_k}(s^k_{\tau_k}, a^k_{\tau_k}) - Q^{\pi^k}_{\tau_k}(s^k_{\tau_k}, a^k_{\tau_k})\right) \le O\left(d^6 H^5 \iota_1^3 + d\sqrt{H^3 \iota_1^3 L}\right). \tag{32}$$

Moreover,

$$\sum_{k \in \mathcal{K}_\tau} \left(Q_{k,\tau_k}(s^k_{\tau_k}, a^k_{\tau_k}) - Q^{\pi^k}_{\tau_k}(s^k_{\tau_k}, a^k_{\tau_k})\right) \le O\left(\frac{d^2 H^3 \iota_1^3}{\Delta_{\min}} + d^6 H^5 \iota_1^3\right). \tag{33}$$

Proof of Lemma G.3. For every $k \in \mathcal{K}_\tau$, the definition of τ_k gives

$$\Delta_{\tau_k}(s^k_{\tau_k}, a^k_{\tau_k}) \ge \Delta_{\min}.$$

Since π^k is greedy with respect to $Q_{k,h}$, we have

$$Q_{k,\tau_k}(s^k_{\tau_k}, a^k_{\tau_k}) = V_{k,\tau_k}(s^k_{\tau_k}).$$

Moreover, since $a^k_{\tau_k} = \pi^k_{\tau_k}(s^k_{\tau_k})$, we have

$$Q^{\pi^k}_{\tau_k}(s^k_{\tau_k}, a^k_{\tau_k}) = V^{\pi^k}_{\tau_k}(s^k_{\tau_k}).$$

Therefore, by Lemma F.1,

$$\begin{aligned} Q_{k,\tau_k}(s^k_{\tau_k}, a^k_{\tau_k}) - Q^{\pi^k}_{\tau_k}(s^k_{\tau_k}, a^k_{\tau_k}) &= V_{k,\tau_k}(s^k_{\tau_k}) - V^{\pi^k}_{\tau_k}(s^k_{\tau_k}) \\ &\ge V^\star_{\tau_k}(s^k_{\tau_k}) - V^{\pi^k}_{\tau_k}(s^k_{\tau_k}) \\ &\ge \Delta_{\tau_k}(s^k_{\tau_k}, a^k_{\tau_k}) \\ &\ge \Delta_{\min}. \end{aligned} \tag{34}$$

Here, the second inequality follows from

$$V^{\pi^k}_{\tau_k}(s^k_{\tau_k}) = Q^{\pi^k}_{\tau_k}(s^k_{\tau_k}, a^k_{\tau_k}) \le Q^\star_{\tau_k}(s^k_{\tau_k}, a^k_{\tau_k}),$$

and hence

$$V^\star_{\tau_k}(s^k_{\tau_k}) - V^{\pi^k}_{\tau_k}(s^k_{\tau_k}) \ge V^\star_{\tau_k}(s^k_{\tau_k}) - Q^\star_{\tau_k}(s^k_{\tau_k}, a^k_{\tau_k}) = \Delta_{\tau_k}(s^k_{\tau_k}, a^k_{\tau_k}).$$

The last inequality follows from the definition of τ_k and $\Delta_{\min}$.

Summing Equation (34) over all $k \in \mathcal{K}_\tau$ gives

$$\Delta_{\min} L \le \sum_{k \in \mathcal{K}_\tau} \left(Q_{k,\tau_k}(s^k_{\tau_k}, a^k_{\tau_k}) - Q^{\pi^k}_{\tau_k}(s^k_{\tau_k}, a^k_{\tau_k})\right). \tag{35}$$

For notational convenience, fix an arbitrary dummy action $a^\dagger \in \mathcal{A}$, set $a^k_{H+1} := a^\dagger$ for every $k \in [K]$, and define

$$Q_{k,H+1}(s,a) := Q^{\pi^k}_{H+1}(s,a) := 0$$

for every $(s,a) \in \mathcal{S} \times \mathcal{A}$. Under these conventions, together with $V_{k,H+1} = V_{H+1}^{\pi^k} = 0$, we have

$$V_{k,h+1}(s_{h+1}^k) = Q_{k,h+1}(s_{h+1}^k, a_{h+1}^k), \qquad V_{h+1}^{\pi^k}(s_{h+1}^k) = Q_{h+1}^{\pi^k}(s_{h+1}^k, a_{h+1}^k)$$

for every $k \in [K]$ and $h \in [H]$. Therefore, by Equation (14), for every $k \in [K]$ and $h \in [H]$, we have

$$\begin{aligned}
&Q_{k,h}(s_h^k, a_h^k) - Q_h^{\pi^k}(s_h^k, a_h^k) \\
&\le Q_{k,h+1}(s_{h+1}^k, a_{h+1}^k) - Q_{h+1}^{\pi^k}(s_{h+1}^k, a_{h+1}^k) + \left(\mathbb{P}_{s_h^k, a_h^k, h} - \mathbb{1}_{s_{h+1}^k}\right)\left(V_{k,h+1} - V_{h+1}^{\pi^k}\right) \\
&\quad + 4\min\left\{\beta\sqrt{\boldsymbol{\phi}(s_h^k, a_h^k)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^k, a_h^k)}, H\right\}. \qquad (36)
\end{aligned}$$

For every $k \in \mathcal{K}_\tau$, summing Equation (36) over $h = \tau_k, \tau_k + 1, \ldots, H$ gives

$$\begin{aligned}
&Q_{k,\tau_k}(s_{\tau_k}^k, a_{\tau_k}^k) - Q_{\tau_k}^{\pi^k}(s_{\tau_k}^k, a_{\tau_k}^k) \\
&\le \sum_{h=\tau_k}^{H} \left(\mathbb{P}_{s_h^k, a_h^k, h} - \mathbb{1}_{s_{h+1}^k}\right)\left(V_{k,h+1} - V_{h+1}^{\pi^k}\right) + \sum_{h=\tau_k}^{H} 4\min\left\{\beta\sqrt{\boldsymbol{\phi}(s_h^k, a_h^k)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^k, a_h^k)}, H\right\}.
\end{aligned}$$

Summing the preceding inequality over all $k \in \mathcal{K}_\tau$ yields

$$\begin{aligned}
&\sum_{k \in \mathcal{K}_\tau} \left(Q_{k,\tau_k}(s_{\tau_k}^k, a_{\tau_k}^k) - Q_{\tau_k}^{\pi^k}(s_{\tau_k}^k, a_{\tau_k}^k)\right) \\
&\le \sum_{(k,h) \in \mathcal{J}_\tau} \left(\mathbb{P}_{s_h^k, a_h^k, h} - \mathbb{1}_{s_{h+1}^k}\right)\left(V_{k,h+1} - V_{h+1}^{\pi^k}\right) \\
&\quad + \sum_{h=1}^{H} \sum_{k \in \mathcal{K}_h} 4\min\left\{\beta\sqrt{\boldsymbol{\phi}(s_h^k, a_h^k)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^k, a_h^k)}, H\right\}. \qquad (37)
\end{aligned}$$

For the first term on the right-hand side of Equation (37), on the event $\bigcap_{i=1}^{6} \mathcal{E}_i$, Equation (9) gives $V_{k,h}^{\mathrm{clip}} = V_{k,h}$. Therefore, the event $\mathcal{E}_3$ in Lemma E.1 yields

$$\sum_{(k,h) \in \mathcal{J}_\tau} \left(\mathbb{P}_{s_h^k, a_h^k, h} - \mathbb{1}_{s_{h+1}^k}\right)\left(V_{k,h+1} - V_{h+1}^{\pi^k}\right) \le 2\sqrt{2H^3 L \log(KH/\delta)}. \qquad (38)$$

For the second term on the right-hand side of Equation (37), applying Lemma F.3 to $\mathcal{K}_h$ for every $h \in [H]$ gives

$$\begin{aligned}
&\sum_{h=1}^{H} \sum_{k \in \mathcal{K}_h} 4\min\left\{\beta\sqrt{\boldsymbol{\phi}(s_h^k, a_h^k)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^k, a_h^k)}, H\right\} \\
&\le 40d^4H^5\iota + 40\beta d^4H^3\iota + 8\beta \sum_{h=1}^{H} \sqrt{d\iota \sum_{k \in \mathcal{K}_h} \left(\sigma_{k,h}^2 + H\right)} \\
&\le 40d^4H^5\iota + 40\beta d^4H^3\iota + 8\beta \sqrt{dH\iota \sum_{(k,h) \in \mathcal{J}_\tau} \left(\sigma_{k,h}^2 + H\right)}, \qquad (39)
\end{aligned}$$

where the second inequality follows from the Cauchy–Schwarz inequality.

By Lemma G.2,

$$\sum_{(k,h) \in \mathcal{J}_\tau} \left(\sigma_{k,h}^2 + H\right) = \sum_{(k,h) \in \mathcal{J}_\tau} \sigma_{k,h}^2 + H|\mathcal{J}_\tau| \le O\left(H^2 L + d^{10}H^9\iota_1^3\right).$$

Since $\lambda = 1/H^2$, we have $\beta = O(\sqrt{d}\,\iota_1)$, $\iota \le O(\iota_1)$, and $\log(KH/\delta) \le O(\iota_1)$. Substituting the preceding bound into Equation (38) and Equation (39), and then substituting these two inequalities into Equation (37), gives

$$\begin{aligned}
\sum_{k \in \mathcal{K}_\tau} \left(Q_{k,\tau_k}(s_{\tau_k}^k, a_{\tau_k}^k) - Q_{\tau_k}^{\pi^k}(s_{\tau_k}^k, a_{\tau_k}^k)\right) &\le O\left(d^6H^5\iota_1^3 + \beta\sqrt{dH^3\iota_1 L}\right) \\
&\le O\left(d^6H^5\iota_1^3 + d\sqrt{H^3\iota_1^3 L}\right),
\end{aligned}$$

which proves Equation (32).

Combining Equation (35) and the equation above, we have

$$\sum_{k\in\mathcal{K}_\tau}\left(Q_{k,\tau_k}(s^k_{\tau_k},a^k_{\tau_k})-Q^{\pi^k}_{\tau_k}(s^k_{\tau_k},a^k_{\tau_k})\right)$$
$$\le O\left(d^6H^5\iota_1^3+\beta\sqrt{\frac{dH^3\iota_1}{\Delta_{\min}}\sum_{k\in\mathcal{K}_\tau}\left(Q_{k,\tau_k}(s^k_{\tau_k},a^k_{\tau_k})-Q^{\pi^k}_{\tau_k}(s^k_{\tau_k},a^k_{\tau_k})\right)}\right).$$

Using the fact that $x\le a\sqrt{x}+b$ implies $x\le a^2+2b$, we obtain

$$\sum_{k\in\mathcal{K}_\tau}\left(Q_{k,\tau_k}(s^k_{\tau_k},a^k_{\tau_k})-Q^{\pi^k}_{\tau_k}(s^k_{\tau_k},a^k_{\tau_k})\right)\le O\left(\frac{d\beta^2H^3\iota_1}{\Delta_{\min}}+d^6H^5\iota_1^3\right).$$

This proves Equation (33) and completes the proof. □

On the event $\bigcap_{i=1}^6\mathcal{E}_i$, Lemma F.1 gives

$$\begin{aligned}\sum_{k\in\mathcal{K}_\tau}\left(V^\star_{\tau_k}-V^{\pi^k}_{\tau_k}\right)(s^k_{\tau_k})&\le\sum_{k\in\mathcal{K}_\tau}\left(V_{k,\tau_k}-V^{\pi^k}_{\tau_k}\right)(s^k_{\tau_k})\\&=\sum_{k\in\mathcal{K}_\tau}\left(Q_{k,\tau_k}(s^k_{\tau_k},a^k_{\tau_k})-Q^{\pi^k}_{\tau_k}(s^k_{\tau_k},a^k_{\tau_k})\right)\\&\le O\left(\frac{d\beta^2H^3\iota_1}{\Delta_{\min}}+d^6H^5\iota_1^3\right),\end{aligned}$$

where the last inequality follows from Lemma G.3.

Let $\delta=1/(19T)$. By Lemma E.1, the event $\mathcal{E}:=\bigcap_{i=1}^6\mathcal{E}_i$ holds with probability at least $1-19\delta=1-1/T$. Therefore, by Lemma G.1 and the fact that

$$\sum_{k\in\mathcal{K}_\tau}\left(V^\star_{\tau_k}-V^{\pi^k}_{\tau_k}\right)(s^k_{\tau_k})\le HL\le HK=T,$$

we have

$$\begin{aligned}\mathbb{E}\left[\mathrm{Regret}(T)\right]&=\mathbb{E}\left[\sum_{k\in\mathcal{K}_\tau}\left(V^\star_{\tau_k}-V^{\pi^k}_{\tau_k}\right)(s^k_{\tau_k})\right]\\&\le O\left(\frac{d\beta^2H^3\iota_1}{\Delta_{\min}}+d^6H^5\iota_1^3\right)+T\mathbb{P}(\mathcal{E}^c)\\&\le O\left(\frac{d\beta^2H^3\iota_1}{\Delta_{\min}}+d^6H^5\iota_1^3\right)\\&\le O\left(\left(\frac{d^2H^3}{\Delta_{\min}}+d^6H^5\right)\log^3\left(1+dHK\right)\right).\end{aligned}$$

Therefore,

$$\mathbb{E}\left[\mathrm{Regret}(T)\right]=\widetilde{O}\left(\frac{d^2H^3}{\Delta_{\min}}+d^6H^5\right).$$

This completes the proof.

H Proof of Theorem 3.2

In this section, we prove the lower bound for persistent ellipsoidal optimistic algorithms. For a fixed episode budget K, the structural conditions are parameterized by the regularization, confidence radius, regression-weight bounds, covariance staleness, and calibration-failure probability.

For the fixed-initial-state instances used below, the standard performance-difference telescoping argument—the same argument used at the beginning of the proof of Lemma G.1—gives

$$\mathbb{E}[\mathrm{Regret}(T)] = \mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\Delta_h(s_h^k, a_h^k)\right]. \tag{40}$$

We only use this identity after taking expectation; realized action counts are not identified pathwise with pseudo-regret.

H.1 Persistent Ellipsoidal Optimism

Definition H.1 (Persistent ellipsoidal optimistic algorithm)**.** Fix an episode budget K, a regularization parameter $\lambda > 0$, confidence radius $\beta > 0$, deterministic constants $0 < \alpha_- \le \alpha_+ < \infty$, covariance-staleness constant $\kappa \ge 1$, and calibration-failure level $\delta_{\mathrm{cal}} \in (0, 1/16]$. An algorithm satisfies persistent ellipsoidal optimism with parameters $(\lambda, \beta, \alpha_-, \alpha_+, \kappa, \delta_{\mathrm{cal}})$ if the following conditions hold over the first K episodes.

(i) Predictable, immutable regression weights. For every episode k and step h,

$$\Lambda_{k,h}^{\mathrm{all}} = \lambda I_d + \sum_{\ell<k}\alpha_{\ell,h}\phi(s_h^\ell, a_h^\ell)\phi(s_h^\ell, a_h^\ell)^\top,$$

where $\alpha_{\ell,h}$ is chosen before s_{h+1}^ℓ is observed, remains unchanged thereafter, and satisfies

$$\alpha_- \le \alpha_{\ell,h} \le \alpha_+ \qquad \text{on every sample path.}$$

The weighting rule may otherwise depend arbitrarily on the past history and on the algorithm's internal randomization.

(ii) Complete episode-level updates and bounded covariance staleness. At each genuine update index u, corresponding to an episode in which the algorithm recomputes its value estimates, the algorithm uses the current cumulative covariance

$$\Lambda_{u,h} = \Lambda_{u,h}^{\mathrm{all}}$$

and performs the backward update over all steps. Every historical regression target at step h is evaluated with the continuation value $V_{u,h+1}$ from this same update. Values used for decision making in an episode are fixed before the random transitions of that episode occur and remain fixed throughout the episode. Let $\tau(k) \le k$ be the latest genuine update whose values are used in episode k, with $\tau(k) = 0$ if no genuine update has yet occurred, and set $\Lambda_{0,h} = \lambda I_d$. We require

$$\Lambda_{\tau(k),h} \preceq \Lambda_{k,h}^{\mathrm{all}} \preceq \kappa\Lambda_{\tau(k),h} \qquad \forall k, h. \tag{41}$$

(iii) Bellman ridge predictor. At each genuine update u, the algorithm follows one of two reward conventions throughout the run:

$$\widehat{\mathcal{B}}_{u,h}(s,a) = r_h(s,a) + \phi(s,a)^\top\Lambda_{u,h}^{-1}\sum_{\ell<u}\alpha_{\ell,h}\phi(s_h^\ell, a_h^\ell)V_{u,h+1}(s_{h+1}^\ell), \tag{42}$$

$$\widehat{\mathcal{B}}_{u,h}(s,a) = \phi(s,a)^\top\Lambda_{u,h}^{-1}\sum_{\ell<u}\alpha_{\ell,h}\phi(s_h^\ell, a_h^\ell)\big(r_h(s_h^\ell, a_h^\ell) + V_{u,h+1}(s_{h+1}^\ell)\big). \tag{43}$$

Under the first convention, $r_h(s,a)$ is added outside the regression term and satisfies the assumptions in Section 2. Under the second convention, the reward is included in the regression target, as in prior work such as Jin et al. (2020), where the reward is assumed to be linear in the feature map: for every $h \in [H]$, there exists $\eta_h \in \mathbb{R}^d$ with $\|\eta_h\|_2 \le \sqrt{d}$ such that $r_h(s,a) = \phi(s,a)^\top\eta_h$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$.

(iv) Persistent ellipsoidal bonus and one fixed maintenance rule. At each genuine update,

$$\widetilde{Q}_{u,h}(s,a) = \min\Big\{\widehat{\mathcal{B}}_{u,h}(s,a) + \beta\|\phi(s,a)\|_{\Lambda_{u,h}^{-1}},\ H\Big\}.$$

Before interaction, the algorithm fixes one of the two rules

$$\begin{aligned} Q_{u,h} &= \widetilde{Q}_{u,h} && \text{(newest-candidate rule)},\\ Q_{u,h}(s,a) &= \min_{v\le u:\, v \text{ genuine}} \widetilde{Q}_{v,h}(s,a) && \text{(historical-minimum rule)}. \end{aligned} \tag{44}$$

Initially $Q_{0,h}(s,a) = H$, and

$$V_{u,h}(s) = \max_a Q_{u,h}(s,a), \qquad V_{u,H+1}(s) = 0.$$

Under the historical-minimum rule, the maintained Q- and V-values are therefore pointwise nonincreasing across genuine updates.

(v) Fixed deterministic greedy tie-breaking. The finite action set is indexed. At every step,

$$a_h^k = \min \arg\max_a Q_{\tau(k),h}(s_h^k, a). \tag{45}$$

Thus the smallest indexed maximizing action is always selected. The algorithm may still randomize its predictable weights or update decisions.

(vi) Model-wise full-domain calibration and optimism. For each fixed linear MDP M, when the algorithm is run on M for K episodes, there exists an event $\mathcal{E}_{\mathrm{cal}}^M$ satisfying

$$\mathbb{P}_M(\mathcal{E}_{\mathrm{cal}}^M) \ge 1 - \delta_{\mathrm{cal}}. \tag{46}$$

On $\mathcal{E}_{\mathrm{cal}}^M$, simultaneously at every genuine update u within the first K episodes, every step h, and every state-action pair (s,a), including unreachable states,

$$\begin{aligned} &\left|\widehat{\mathcal{B}}_{u,h}(s,a) - \left(r_h(s,a) + \mathbb{P}_{s,a,h}V_{u,h+1}\right)\right| \le \beta\|\phi(s,a)\|_{\Lambda_{u,h}^{-1}},\\ &Q_{u,h}(s,a) \ge Q_h^\star(s,a), \qquad 0 \le Q_{u,h}(s,a), V_{u,h}(s) \le H. \end{aligned} \tag{47}$$

The event may depend on M; no simultaneous event over different MDPs is required. The probability includes all algorithm randomization. The algorithm has only ordinary online interaction and receives no additional observations from unreachable states.

When the same algorithm with the same parameters $(\lambda, \beta, \alpha_-, \alpha_+, \kappa, \delta_{\mathrm{cal}})$ is applied to every member of the hard-instance family constructed in Section H.2, the state and action spaces, feature map, rewards, horizon, initial state, and episode budget are also held fixed across the family; only the unknown transition kernel changes. Consequently, on the same realized history, the algorithm uses the same conditional action, update, weighting, and internal-randomization kernels under every family member.

Remark H.2 (Full-domain calibration). The proof of Theorem H.3 below applies the calibration condition at unreachable states (see Lemma H.5), but does not rely on any observations from those states. This is compatible with the usual ellipsoidal confidence analysis, because the confidence event controls the estimated regression parameter and therefore its prediction at any fixed feature vector.

For any bounded function V, define

$$w_h(V) := \int_{\mathcal{S}} V(s')\, d\boldsymbol{\theta}_h(s') \in \mathbb{R}^d,$$

so that

$$\mathbb{P}_{s,a,h}V = \phi(s,a)^\top w_h(V) \qquad \text{for all } (s,a).$$

Accordingly, the population regression parameter is $w_h(V)$ under the outside-reward convention in Equation (42), and $\eta_h + w_h(V)$ under the inside-reward convention in Equation (43). More generally, if $\widehat{b}$ is the corresponding ridge estimator, b is its population target, and

$$\|\widehat{b} - b\|_\Lambda \le \beta,$$

then, for any feature vector ϕ,

$$|\phi^\top(\widehat{b} - b)| \le \|\widehat{b} - b\|_\Lambda \|\phi\|_{\Lambda^{-1}} \le \beta\|\phi\|_{\Lambda^{-1}}.$$

Thus, the same parameter-space confidence event controls any fixed state-action feature vector, regardless of whether the corresponding state is reachable. For the outside-reward convention, the resulting full-domain Bellman confidence bound with radius β is established in part (b) of Theorem E.1. For the inside-reward convention, under the linear-reward assumption, the analogous full-domain Bellman confidence bound is standard for LSVI-style algorithms; see, e.g., (Jin et al., 2020, Lemma B.4). Hence, the calibration condition in Equation (47) does not depend on observations from unreachable states, which are fixed before interaction and never queried by the algorithm.

For the remainder of the proof, let

$$A_0 := 9, \qquad m := d - 6, \qquad H_0 := \left\lfloor \frac{H}{4} \right\rfloor, \qquad \mathcal{T} := \{3, \dots, H_0\}, \qquad R_t := H - t.$$

Define

$$N_b := \left\lceil \frac{(\kappa - 1)\lambda + 4096\kappa\beta^2}{\alpha_-} \right\rceil, \qquad n_s := \left\lceil \frac{1024\kappa\beta^2}{\alpha_-\Delta_{\min}^2} \right\rceil, \tag{48}$$

$$K_{\mathrm{burn}} := \lceil \max\{8N_b,\ 8Hn_s,\ 32H\log 128\} \rceil, \tag{49}$$

$$\epsilon_t := \frac{A_0\Delta_{\min}}{R_t\sqrt{m}}, \qquad n_t^{\mathrm{info}} := \frac{m}{2^{19}\epsilon_t^2} = \frac{m^2R_t^2}{2^{19}A_0^2\Delta_{\min}^2}. \tag{50}$$

Finally, define

$$\chi_0 := 2\left(\frac{256}{567}\right)^2 = \frac{131072}{321489}, \qquad c_{\mathrm{cal}} := \frac{127}{256} - \chi_0 - \frac{1}{64} - \delta_{\mathrm{cal}}.$$

Theorem H.3 (Formal statement of Theorem 3.2)**.** *For any feature dimension $d \geq 15$, horizon $H \geq 16$, minimum gap $\Delta_{\min} \in (0, H/16]$, episode number K, and any algorithm satisfying Definition H.1 with parameters $(\lambda, \beta, \alpha_-, \alpha_+, \kappa, \delta_{\mathrm{cal}})$ such that $c_{\mathrm{cal}} > 0$, if*

$$K \geq K_{\mathrm{burn}} + \left\lceil \frac{8H}{c_{\mathrm{cal}}} \max_{t\in\mathcal{T}} n_t^{\mathrm{info}} \right\rceil, \tag{51}$$

then there exists a member $M_{\boldsymbol{\sigma}}$ of the hard-instance family constructed in Section H.2 that is a legal d-dimensional linear MDP with fixed initial state s_{init} and minimum positive gap exactly $\Delta_{\min}$, and satisfies

$$\mathbb{E}_{M_{\boldsymbol{\sigma}}}[\mathrm{Regret}(T)] \geq 2^{-33}\frac{d^2H^3}{\Delta_{\min}}. \tag{52}$$

Moreover, the sufficient budget in Equation (51) *scales as*

$$O\left(1 + \frac{\kappa\lambda}{\alpha_-} + \frac{\kappa\beta^2}{\alpha_-}\left(1 + \frac{H}{\Delta_{\min}^2}\right) + \frac{d^2H^3}{c_{\mathrm{cal}}\Delta_{\min}^2} + H\right). \tag{53}$$

H.2 A Hard-Instance Family

Let

$$\mathfrak{S} := \{-1, +1\}^{m\times\mathcal{T}}.$$

For every sign array $\boldsymbol{\sigma} = (\sigma_{i,t}) \in \mathfrak{S}$, we define an MDP $M_{\boldsymbol{\sigma}}$. All family members share the same known state space, action space, feature map, rewards, horizon, and initial state; only the unknown transition kernel depends on $\boldsymbol{\sigma}$.

Choose orthonormal coordinates

$$e_0, e_1, \dots, e_m, e_\star, e_{\mathrm{init}}, e_w, e_g, e_b \in \mathbb{R}^d.$$

There are exactly $m + 6 = d$ coordinates. Define the countable rational simplex

$$\mathcal{V}_{\mathbb{Q}} := \left\{ v \in \mathbb{Q}^m :\ v_i \geq 0,\ \sum_{i=1}^m v_i = 1 \right\}.$$

The state space, equipped with the discrete sigma-field, is

$$\mathcal{S} = \{s_{\text{init}}, w, g, b, z_s, x_\star, x_1, \dots, x_m\} \cup \{z_v : v \in \mathcal{V}_{\mathbb{Q}}\}, \qquad \mathcal{A} = \{1, 2\}.$$

The feature map is

$$\begin{aligned}
&\phi(x_i, 1) = e_0, && \phi(x_i, 2) = e_i, && i \in [m],\\
&\phi(x_\star, 1) = e_0, && \phi(x_\star, 2) = e_\star,\\
&\phi(z_s, a) = e_0, && \phi(z_v, a) = \sum_{i=1}^{m} v_i e_i,\\
&\phi(s_{\text{init}}, a) = e_{\text{init}}, && \phi(w, a) = e_w, \qquad \phi(g, a) = e_g, \qquad \phi(b, a) = e_b.
\end{aligned}$$

At every step, the reward equals one in g and zero elsewhere. Equivalently,

$$r_h(s, a) = \phi(s, a)^\top e_g.$$

Every episode starts at s_{init}. At step 1, independently of the action,

$$P_1(b \mid s_{\text{init}}, a) = \frac{1}{4}, \qquad P_1(w \mid s_{\text{init}}, a) = \frac{3}{4}.$$

At every other step, s_{init} moves to b. For $2 \le h \le H_0 - 1$, set

$$\begin{aligned}
&P_h(z_s \mid w, a) = \frac{1}{2H},\\
&P_h(x_i \mid w, a) = P_h(x_\star \mid w, a) = \frac{1}{2(m+1)H}, \qquad i \in [m],\\
&P_h(w \mid w, a) = 1 - \frac{1}{H}.
\end{aligned}$$

Outside these steps, w moves to b. The states g and b are absorbing.

For every $t \in \mathcal{T}$, define

$$p_{0,t} := \frac{1}{2} + \epsilon_t + \frac{\Delta_{\min}}{R_t}, \qquad p_{i,t} := \frac{1}{2} + \sigma_{i,t}\epsilon_t, \qquad p_{\star,t} := \frac{1}{2} + \epsilon_t.$$

At step t, action 1 at every x_i and $x_\star$, and both actions at z_s, move to g with probability $p_{0,t}$. Action 2 at x_i moves to g with probability $p_{i,t}$, while action 2 at $x_\star$ moves to g with probability $p_{\star,t}$. Both actions at z_v move to g with probability

$$\sum_{i=1}^{m} v_i p_{i,t}.$$

In every case the remaining probability goes to b. At steps outside $\mathcal{T}$, both actions at $x_i, x_\star, z_s, z_v$ move to b.

Every z_v is unreachable from s_{init}: no transition distribution of a reachable state assigns mass to such a probe. The countable probe family, including all features and transition kernels, is fixed before interaction.

Lemma H.4 (Validity, exact gap, and exogenous reach probabilities)**.** *Every $M_{\boldsymbol{\sigma}}$ is a valid d-dimensional linear MDP satisfying the normalization in Section 2. Its minimum positive gap is exactly $\Delta_{\min}$. Moreover, for every $t \in \mathcal{T}$ and $i \in [m]$, the reach probabilities are independent of the algorithm and of $\boldsymbol{\sigma}$ and satisfy*

$$q_{s,t} := \mathbb{P}(s_t = z_s) \ge \frac{1}{4H}, \qquad q_{i,t} := \mathbb{P}(s_t = x_i) \ge \frac{1}{8mH}. \tag{54}$$

All features observed along an actual trajectory are one-hot in the displayed basis; hence every sampled covariance matrix is diagonal in that basis.

Proof. For $t \in \mathcal{T}$, define the coordinate transition measures corresponding to $e_0, e_i, e_\star$ as the Bernoulli mixtures of δ_g and δ_b with success probabilities $p_{0,t}, p_{i,t}, p_{\star,t}$, respectively. Outside $\mathcal{T}$,

set these coordinate measures to δ_b. The coordinates $e_{\text{init}}, e_w, e_g, e_b$ are assigned the transition distributions specified above. Then the feature $\sum_i v_i e_i$ at z_v produces exactly the stated convex-combination kernel. Every coordinate transition measure is a probability measure, so for every measurable set C,

$$\|\boldsymbol{\theta}_h(C)\|_2^2 \le d.$$

Also, $\|\phi(z_v, a)\|_2 = \|v\|_2 \le \|v\|_1 = 1$, every other feature has norm one, and the reward parameter e_g has norm one. Thus every family member is a legal linear MDP.

Since $m \ge 9$, $A_0 = 9$, and $R_t \ge 3H/4$,

$$R_t \epsilon_t = \frac{A_0 \Delta_{\min}}{\sqrt{m}} \le 3\Delta_{\min}, \qquad \epsilon_t \le \frac{3\Delta_{\min}}{R_t} \le \frac{1}{4}.$$

Therefore $p_{i,t} \in [1/4, 3/4]$, $p_{\star,t} \le 3/4$, and

$$p_{0,t} \le \frac{1}{2} + \frac{4\Delta_{\min}}{R_t} \le \frac{5}{6},$$

so all displayed probabilities are valid. Since

$$V_{t+1}^\star(g) = R_t, \qquad V_{t+1}^\star(b) = 0,$$

the gap of action 2 at x_i is

$$R_t(p_{0,t} - p_{i,t}) = \Delta_{\min} + (1 - \sigma_{i,t}) R_t \epsilon_t \in [\Delta_{\min}, 7\Delta_{\min}], \tag{55}$$

whereas the gap of action 2 at the reachable state $x_\star$ is exactly

$$R_t(p_{0,t} - p_{\star,t}) = \Delta_{\min}.$$

All remaining state-action-step triples have zero gap. Hence the minimum positive gap is exactly $\Delta_{\min}$ for every member of the family.

Finally, for $t \in \mathcal{T}$,

$$q_{s,t} = \frac{3}{4}\left(1 - \frac{1}{H}\right)^{t-3} \frac{1}{2H}, \qquad q_{i,t} = \frac{3}{4}\left(1 - \frac{1}{H}\right)^{t-3} \frac{1}{2(m+1)H}.$$

Since $t \le H_0 \le H/4$, Bernoulli's inequality gives

$$\left(1 - \frac{1}{H}\right)^{t-3} \ge \frac{3}{4}.$$

Thus $q_{s,t} \ge 9/(32H) \ge 1/(4H)$. Also, $m \ge 9$ implies $9m \ge 4(m+1)$, so $q_{i,t} \ge 1/(8mH)$. These transitions are action-independent and use fresh environment randomness, proving the claimed exogeneity. The only non-one-hot features belong to the unreachable probes z_v. □

H.3 Global Calibration Gives a Joint Error Budget

Fix a family member $M_{\boldsymbol{\sigma}}$, a genuine update u, and an active step $t \in \mathcal{T}$. Write

$$D_i := e_i^\top \Lambda_{u,t} e_i, \qquad G := V_{u,t+1}(g), \qquad C := V_{u,t+1}(b), \qquad \Gamma := G - C,$$

and define the actual Bellman-prediction error in hard direction i by

$$\xi_i := \widehat{\mathcal{B}}_{u,t}(x_i, 2) - \big(C + \Gamma p_{i,t}\big).$$

Lemma H.5 (Joint positive and negative error-energy bounds)**.** *On $\mathcal{E}_{\text{cal}}^{M_{\boldsymbol{\sigma}}}$, at every genuine update u and every $t \in \mathcal{T}$,*

$$\sum_{i=1}^m D_i (\xi_i)_+^2 \le \beta^2, \qquad \sum_{i=1}^m D_i (\xi_i)_-^2 \le \beta^2. \tag{56}$$

Consequently,

$$\sum_{i=1}^m D_i \xi_i^2 \le 2\beta^2. \tag{57}$$

Proof. The only reachable sample whose feature has an e_i component is action 2 at x_i, and its reward is zero. Hence, under either reward convention in Definition H.1, for every rational-simplex probe z_v,

$$\widehat{\mathcal{B}}_{u,t}(z_v, a) - \mathbb{P}_{z_v,a,t} V_{u,t+1} = \sum_{i=1}^{m} v_i \xi_i.$$

By Lemma H.4, the sampled covariance is diagonal in the displayed basis, so

$$\|\phi(z_v, a)\|^2_{\Lambda^{-1}_{u,t}} = \sum_{i=1}^{m} \frac{v_i^2}{D_i}.$$

Global calibration at the unreachable probe gives

$$\left| \sum_{i=1}^{m} v_i \xi_i \right| \le \beta \sqrt{\sum_{i=1}^{m} \frac{v_i^2}{D_i}} \qquad \forall v \in \mathcal{V}_{\mathbb{Q}}. \tag{58}$$

Both sides are continuous in v, so Equation (58) holds for every point of the real probability simplex. Let

$$S_+ := \sum_i D_i (\xi_i)_+^2, \qquad L_+ := \sum_i D_i (\xi_i)_+.$$

If $L_+ = 0$, the positive-part claim is immediate. Otherwise choose

$$v_i = \frac{D_i (\xi_i)_+}{L_+}.$$

Then Equation (58) gives

$$\frac{S_+}{L_+} \le \beta \frac{\sqrt{S_+}}{L_+},$$

which implies $S_+ \le \beta^2$. Applying the same argument to the negative parts proves the second inequality in Equation (56); summing them gives Equation (57). □

H.4 Burn-in and Maintained-Value Comparisons

For every $t \in \mathcal{T}$, let $\mathcal{F}_t$ be the event that among the first K_{burn} completed episodes, at least N_b episodes take the transition $s_{\mathrm{init}} \to b$ at step 1, and at least n_s episodes visit z_s at step t.

Lemma H.6 (Anchor counts)**.** *For every $M_{\boldsymbol{\sigma}}$ and every $t \in \mathcal{T}$,*

$$\mathbb{P}_{M_{\boldsymbol{\sigma}}}(\mathcal{F}_t^c) \le \frac{1}{64}.$$

Proof. The first count is $\mathrm{Binomial}(K_{\mathrm{burn}}, 1/4)$. The second is binomial with success probability $q_{s,t} \ge 1/(4H)$ by Lemma H.4. For a sum X of independent Bernoulli variables with mean μ, the standard multiplicative Chernoff bound gives

$$\mathbb{P}(X \le \mu/2) \le e^{-\mu/8}.$$

Because $K_{\mathrm{burn}} \ge 8N_b$ and $K_{\mathrm{burn}} \ge 8Hn_s$, the two failure probabilities are at most $e^{-K_{\mathrm{burn}}/32}$ and $e^{-K_{\mathrm{burn}}/(32H)}$, respectively. By Equation (49), each is at most $1/128$. A union bound proves the claim. No independence between the two counts is required, and we do not union-bound over t. □

Lemma H.7 (Cemetery and shared-direction bounds)**.** *Fix $M_{\boldsymbol{\sigma}}$ and $t \in \mathcal{T}$. On $\mathcal{F}_t \cap \mathcal{E}^{M_{\boldsymbol{\sigma}}}_{\mathrm{cal}}$, for every episode $k > K_{\mathrm{burn}}$, the latest update $u = \tau(k)$ is genuine and*

$$V_{u,t+1}(b) \le \frac{R_t}{64}, \qquad V_{u,t+1}(g) \ge R_t, \qquad V_{u,t+1}(g) - V_{u,t+1}(b) \ge \frac{63 R_t}{64}. \tag{59}$$

Moreover, the maintained score of the shared action, which is identical at every x_i, satisfies

$$Q_{u,t}(x_i, 1) \le C + \Gamma p_{0,t} + \frac{\Delta_{\min}}{16} < \frac{7H}{8}. \tag{60}$$

Proof. Every episode whose first transition is $s_{\text{init}} \to b$ supplies an e_b sample at every step $h \geq 2$. Thus, for $k > K_{\text{burn}}$,

$$e_b^\top \Lambda_{k,h}^{\text{all}} e_b \geq \lambda + \alpha_- N_b.$$

If $u = \tau(k) = 0$, Equation (41) would imply

$$\lambda + \alpha_- N_b \leq \kappa\lambda,$$

contradicting the definition of N_b. Hence u is genuine. Again using Equation (41),

$$e_b^\top \Lambda_{u,h} e_b \geq \frac{\lambda + \alpha_- N_b}{\kappa} \geq \lambda + 4096\beta^2. \tag{61}$$

All historical responses in the e_b direction equal the current continuation value at b, and every corresponding reward is zero. Under either reward convention,

$$\widehat{\mathcal{B}}_{u,h}(b,a) = \frac{W_{u,h,b}}{\lambda + W_{u,h,b}} V_{u,h+1}(b) \leq V_{u,h+1}(b),$$

where $W_{u,h,b}$ is the total historical weight in direction e_b. By Equation (61), the bonus is at most $1/64$. Since either maintenance rule gives a maintained score no larger than the current candidate,

$$V_{u,h}(b) \leq V_{u,h+1}(b) + \frac{1}{64}.$$

Backward induction from $V_{u,H+1}(b) = 0$ yields

$$V_{u,t+1}(b) \leq \frac{R_t}{64}.$$

Optimism gives $V_{u,t+1}(g) \geq V_{t+1}^\star(g) = R_t$, proving Equation (59).

At step t, every z_s sample contributes weight at least α_- in direction e_0. Hence

$$e_0^\top \Lambda_{u,t} e_0 \geq \frac{\lambda + \alpha_- n_s}{\kappa} \geq \frac{1024\beta^2}{\Delta_{\min}^2},$$

so the current bonus in direction e_0 is at most $\Delta_{\min}/32$. By calibration, the current candidate for action 1 at x_i is therefore at most

$$C + \Gamma p_{0,t} + \frac{\Delta_{\min}}{16}.$$

The maintained score is no larger than this candidate under either maintenance rule, giving the first inequality in Equation (60). Finally, using $p_{0,t} \leq 5/6$, $G \leq H$, $C \leq H/64$, and $\Delta_{\min}/16 \leq H/256$,

$$C + \Gamma p_{0,t} + \frac{\Delta_{\min}}{16} \leq \left(\frac{5}{6} + \frac{1}{64} + \frac{1}{256}\right) H = \frac{655}{768} H < \frac{7H}{8}.$$

□

The historical-minimum maintenance rule requires one additional comparison because the currently maintained value of an action can have been created at an earlier genuine update.

Lemma H.8 (Frozen-direction comparison)**.** *Fix a sample path, episode k, and $t \in \mathcal{T}$, and let $u = \tau(k) \geq 1$. Suppose the shared score*

$$q_s := Q_{u,t}(x_i, 1)$$

is below H. Under either maintenance rule, if the prescribed greedy action at x_i is action 1, equivalently if $Q_{u,t}(x_i, 2) \leq q_s$, then

$$\widehat{\mathcal{B}}_{u,t}(x_i, 2) + \frac{\beta}{\sqrt{D_i}} \leq q_s. \tag{62}$$

Proof. For the newest-candidate rule, a maintained score no larger than $q_s < H$ cannot be upper-clipped at H, so it equals its raw current candidate, and Equation (62) follows immediately.

For the historical-minimum rule, write

$$q_i := Q_{u,t}(x_i, 2) \le q_s < H.$$

Because the initial score equals H, the value q_i is attained by some genuine historical candidate; let $v \le u$ be an update attaining it. From update v until the current episode, the bad maintained score remains exactly q_i: a strictly smaller intervening candidate would contradict its current value. During the same period the shared maintained score is at least its current value $q_s \ge q_i$, because maintained values are pointwise nonincreasing. Therefore every intervening visit to x_i selects action 1, including at equality by the smallest-index rule Equation (45).

No new sample in direction e_i has therefore been collected since update v. The probes z_v are unreachable, and no other reachable feature has an e_i component. Thus the covariance and immutable historical weights in direction e_i are unchanged. Under historical-minimum maintenance, continuation values are pointwise nonincreasing across genuine updates; all samples in direction e_i have zero reward. Consequently, under either reward convention,

$$\widehat{\mathcal{B}}_{u,t}(x_i, 2) \le \widehat{\mathcal{B}}_{v,t}(x_i, 2).$$

The bonus in direction e_i is unchanged. Since the witness candidate at v equals $q_i < H$, it was not clipped, so its raw value equals q_i. Hence the current raw score is at most $q_i \le q_s$, proving Equation (62). □

H.5 Calibration Forces Exploration or Fine Estimation

Fix $k > K_{\mathrm{burn}}$ and $t \in \mathcal{T}$, and let $u = \tau(k)$. At the information set after the algorithm's pre-episode update and internal randomization but before the current episode's environment transitions, define

$$\mathcal{I}_{k,t} := \{i \in [m] : Q_{u,t}(x_i, 2) > Q_{u,t}(x_i, 1)\}, \qquad r_t(k) := |\mathcal{I}_{k,t}|.$$

A visit to an index in $\mathcal{I}_{k,t}$ selects action 2 with probability one, while outside this set the smallest-index rule selects action 1.

Define an auxiliary sign estimator $\widehat{\boldsymbol{\sigma}}_t(k) \in \{-1, +1\}^m$ from the same pre-episode information. If $u = 0$ or

$$\Gamma := V_{u,t+1}(g) - V_{u,t+1}(b) \le 0,$$

set every estimated sign to $+1$. Otherwise define, for every i,

$$\widehat{p}_i := \frac{\widehat{\mathcal{B}}_{u,t}(x_i, 2) - V_{u,t+1}(b)}{\Gamma}, \qquad \widehat{\sigma}_{i,t}(k) := \begin{cases} +1, & \widehat{p}_i \ge 1/2, \\ -1, & \widehat{p}_i < 1/2. \end{cases} \tag{63}$$

This estimator is defined on the entire history space and is used only in the proof; it does not modify the algorithm.

Lemma H.9 (Direct energy charge for a nonexplorable sign error)**.** *On $\mathcal{F}_t \cap \mathcal{E}_{\mathrm{cal}}^{M_\sigma}$, every misclassified index $i \notin \mathcal{I}_{k,t}$ satisfies*

$$D_i \xi_i^2 \ge \beta^2 \left(\frac{\epsilon_t}{\epsilon_t + \Delta_{\min}/R_t + \Delta_{\min}/(16\Gamma)} \right)^2. \tag{64}$$

Consequently, the number of such misclassified nonexplorable indices is at most

$$2 \left(1 + \frac{67\sqrt{m}}{63 A_0} \right)^2 \le \chi_0 m. \tag{65}$$

Proof. On the stated event, Lemma H.7 gives $\Gamma > 0$ and a shared score below H. For $i \notin \mathcal{I}_{k,t}$, Lemma H.8 and Equation (60) imply

$$\xi_i + b_i \le (1 - \sigma_{i,t})\Gamma\epsilon_t + \zeta_t, \qquad b_i := \frac{\beta}{\sqrt{D_i}}, \qquad \zeta_t := \Gamma \frac{\Delta_{\min}}{R_t} + \frac{\Delta_{\min}}{16}. \tag{66}$$

If $\sigma_{i,t} = +1$ is misclassified, then $\widehat{p}_i < 1/2$, so $\xi_i = -x$ for some $x > \Gamma\epsilon_t$. Equation (66) gives $b_i \le x + \zeta_t$, and therefore

$$\frac{|\xi_i|}{b_i} \ge \frac{x}{x + \zeta_t} \ge \frac{\Gamma\epsilon_t}{\Gamma\epsilon_t + \zeta_t}.$$

If $\sigma_{i,t} = -1$ is misclassified, then $\xi_i = x \geq \Gamma\epsilon_t$. Equation (66) gives

$$b_i \leq 2\Gamma\epsilon_t + \zeta_t - x \leq \Gamma\epsilon_t + \zeta_t,$$

and the same ratio bound follows. Since

$$D_i\xi_i^2 = \beta^2 \left(\frac{|\xi_i|}{b_i}\right)^2,$$

we obtain Equation (64).

By Equation (59), $\Gamma \geq 63R_t/64$. Hence

$$\frac{\Delta_{\min}/R_t + \Delta_{\min}/(16\Gamma)}{\epsilon_t} \leq \frac{67\sqrt{m}}{63A_0}.$$

Each nonexplorable mistake therefore consumes at least the reciprocal-square fraction of β^2 displayed in Equation (64). Applying the total energy budget Equation (57) yields the first bound in Equation (65). Finally, $m \geq 9$ and $A_0 = 9$ imply

$$\frac{2}{m}\left(1 + \frac{67\sqrt{m}}{63A_0}\right)^2 = 2\left(\frac{1}{\sqrt{m}} + \frac{67}{567}\right)^2 \leq 2\left(\frac{1}{3} + \frac{67}{567}\right)^2 = \chi_0.$$

□

Lemma H.10 (Exploration–estimation relation)**.** *For every family member $M_{\boldsymbol{\sigma}}$ and every $k > K_{\text{burn}}$,*

$$\mathbb{E}_{M_{\boldsymbol{\sigma}}}[\mathrm{Ham}(\widehat{\boldsymbol{\sigma}}_t(k), \boldsymbol{\sigma}_t)] \leq \mathbb{E}_{M_{\boldsymbol{\sigma}}}[r_t(k)] + m\left(\chi_0 + \frac{1}{64} + \delta_{\text{cal}}\right).$$

Proof. On $\mathcal{F}_t \cap \mathcal{E}_{\text{cal}}^{M_{\boldsymbol{\sigma}}}$, Lemma H.7 ensures that the estimator in Equation (63) uses a positive Γ. There can be at most $r_t(k)$ sign errors on explorable indices, while Lemma H.9 bounds the number of errors outside $\mathcal{I}_{k,t}$ by $\chi_0 m$. Thus

$$\mathrm{Ham}(\widehat{\boldsymbol{\sigma}}_t(k), \boldsymbol{\sigma}_t) \leq r_t(k) + \chi_0 m$$

on this event. The Hamming loss is always at most m. Using Lemma H.6 and Equation (46), without any independence assumption,

$$\mathbb{E}_{M_{\boldsymbol{\sigma}}}[\mathrm{Ham}(\widehat{\boldsymbol{\sigma}}_t(k), \boldsymbol{\sigma}_t)] \leq \mathbb{E}_{M_{\boldsymbol{\sigma}}}[r_t(k)] + m\left(\chi_0 + \frac{1}{64} + \delta_{\text{cal}}\right).$$

□

H.6 Information Lower Bound and Completion of the Proof

For any random quantity X and event E that may depend on the family index $\boldsymbol{\sigma}$, define

$$\overline{\mathbb{E}}[X] := \mathbb{E}_{\boldsymbol{\sigma}\sim\mathrm{Unif}(\mathfrak{S})}[\mathbb{E}_{M_{\boldsymbol{\sigma}}}[X]] = \frac{1}{|\mathfrak{S}|}\sum_{\boldsymbol{\sigma}\in\mathfrak{S}} \mathbb{E}_{M_{\boldsymbol{\sigma}}}[X],$$

$$\overline{\mathbb{P}}(E) := \mathbb{E}_{\boldsymbol{\sigma}\sim\mathrm{Unif}(\mathfrak{S})}[\mathbb{P}_{M_{\boldsymbol{\sigma}}}(E)] = \frac{1}{|\mathfrak{S}|}\sum_{\boldsymbol{\sigma}\in\mathfrak{S}} \mathbb{P}_{M_{\boldsymbol{\sigma}}}(E).$$

For $i \in [m]$, $t \in \mathcal{T}$, and $k \in [K]$, define

$$N_{i,t,2}(k) := \sum_{\ell=1}^{k} \mathbb{I}\{s_t^\ell = x_i,\ a_t^\ell = 2\}, \qquad S_t(k) := \frac{1}{|\mathfrak{S}|}\sum_{\boldsymbol{\sigma}\in\mathfrak{S}}\sum_{i=1}^{m} \mathbb{E}_{M_{\boldsymbol{\sigma}}}[N_{i,t,2}(k)].$$

Lemma H.11 (Information lower bound for the sign estimator)**.** *For any estimator of $\boldsymbol{\sigma}_t$ measurable with respect to the information available before the current episode's environment transitions,*

$$\overline{\mathbb{E}}[\mathrm{Ham}(\widehat{\boldsymbol{\sigma}}_t(k), \boldsymbol{\sigma}_t)] \geq \frac{m}{2} - 2\sqrt{2}\,\epsilon_t\sqrt{mS_t(k-1)}. \tag{67}$$

In particular, if $S_t(k-1) \leq n_t^{\text{info}}$, then

$$\overline{\mathbb{E}}[\mathrm{Ham}(\widehat{\boldsymbol{\sigma}}_t(k), \boldsymbol{\sigma}_t)] \geq \frac{127m}{256}. \tag{68}$$

Proof. Fix an index (i,t) and fix all signs except $\sigma_{i,t}$. Let P_- and P_+ denote the distributions of the pre-episode history under the two models with $\sigma_{i,t}=-1$ and $\sigma_{i,t}=+1$, respectively. The reachable transition kernels of these two models differ only at $(x_i,2,t)$. They also differ at some unreachable probes, but those states have identically zero visit counts.

Conditional on any fixed realized history, the algorithm's action, update, weight, and internal-randomization kernels are the same under the two models. Therefore, the chain rule for relative entropy yields

$$\mathrm{KL}(P_-\|P_+)=\mathbb{E}_-[N_{i,t,2}(k-1)]\,\mathrm{kl}\left(\frac{1}{2}-\epsilon_t,\frac{1}{2}+\epsilon_t\right)\le 32\epsilon_t^2\,\mathbb{E}_-[N_{i,t,2}(k-1)]\,.$$

Indeed, $\epsilon_t\le 1/4$ by Lemma H.4, and $\mathrm{kl}(p,q)\le (p-q)^2/[q(1-q)]\le 32\epsilon_t^2$ for the two displayed Bernoulli parameters. The history may be augmented with the pre-episode update and internal randomization used to form the estimator; these are common conditional kernels and contribute zero conditional relative entropy.

By Pinsker's inequality, any binary test of $\sigma_{i,t}$ satisfies

$$P_-(\widehat{\sigma}_{i,t}=+1)+P_+(\widehat{\sigma}_{i,t}=-1)\ge 1-4\epsilon_t\sqrt{\mathbb{E}_-[N_{i,t,2}(k-1)]}.$$

Average this inequality over all pairs of sign arrays that differ only in coordinate (i,t). The mean count on the negative half-family is at most twice the mean over the full family, so Jensen's inequality gives

$$\overline{\mathbb{P}}(\widehat{\sigma}_{i,t}\ne\sigma_{i,t})\ge\frac{1}{2}-2\sqrt{2}\,\epsilon_t\sqrt{\overline{\mathbb{E}}[N_{i,t,2}(k-1)]}.$$

Summing over i and applying Cauchy–Schwarz proves Equation (67). If $S_t(k-1)\le m/(2^{19}\epsilon_t^2)$, then

$$2\sqrt{2}\,\epsilon_t\sqrt{mS_t(k-1)}\le\frac{m}{256},$$

which yields Equation (68). □

Lemma H.12 (Constant sampling rate below the information threshold)**.** *For every $k>K_{\mathrm{burn}}$, if $S_t(k-1)\le n_t^{\mathrm{info}}$, then*

$$S_t(k)-S_t(k-1)\ge\frac{c_{\mathrm{cal}}}{8H}.$$

Proof. Average Lemma H.10 over the uniform prior on $\mathfrak{S}$ and combine it with Lemma H.11. Under the assumed information threshold,

$$\overline{\mathbb{E}}[r_t(k)]\ge\left(\frac{127}{256}-\chi_0-\frac{1}{64}-\delta_{\mathrm{cal}}\right)m=c_{\mathrm{cal}}m.$$

Condition on the information after the current pre-episode update and before environment transitions. The set $\mathcal{I}_{k,t}$ is then fixed. The root and waiting-state transitions are fresh and policy-independent. For every $i\in\mathcal{I}_{k,t}$, a visit to x_i at step t occurs with probability at least $1/(8mH)$ by Equation (54), and action 2 is then selected with probability one by Equation (45). Therefore

$$S_t(k)-S_t(k-1)\ge\frac{1}{8mH}\overline{\mathbb{E}}[r_t(k)]\ge\frac{c_{\mathrm{cal}}}{8H}.$$

This argument does not condition on the calibration event. □

Proof of Theorem H.3. Lemma H.4 already proves that every member of the family is a legal d-dimensional linear MDP with fixed initial state and minimum positive gap exactly $\Delta_{\min}$.

Fix $t\in\mathcal{T}$. Suppose, toward a contradiction, that

$$S_t(K)<n_t^{\mathrm{info}}.$$

By monotonicity of $S_t(\cdot)$, the premise of Lemma H.12 then holds for every $k=K_{\mathrm{burn}}+1,\dots,K$. Hence

$$S_t(K)\ge\frac{c_{\mathrm{cal}}}{8H}(K-K_{\mathrm{burn}})\ge n_t^{\mathrm{info}},$$

where the last inequality is exactly Equation (51), a contradiction. Thus

$$S_t(K) \geq n_t^{\text{info}} \qquad \forall t \in \mathcal{T}.$$

Every count in $N_{i,t,2}$ corresponds to a positive gap of at least $\Delta_{\min}$ by Equation (55). Using the expected performance-difference identity Equation (40) and averaging uniformly over the finite family,

$$\overline{\mathbb{E}}[\text{Regret}(T)] \geq \Delta_{\min} \sum_{t \in \mathcal{T}} S_t(K) \geq \frac{m^2}{2^{19} A_0^2 \Delta_{\min}} \sum_{t \in \mathcal{T}} R_t^2.$$

For $d \geq 15$, $m = d - 6 \geq d/2$. For $H \geq 16$,

$$|\mathcal{T}| = \left\lfloor \frac{H}{4} \right\rfloor - 2 \geq \frac{H}{16}, \qquad R_t \geq \frac{3H}{4}.$$

Since $A_0 = 9$,

$$\overline{\mathbb{E}}[\text{Regret}(T)] \geq \frac{9d^2H^3}{2^{29} A_0^2 \Delta_{\min}} = \frac{d^2H^3}{9 \cdot 2^{29} \Delta_{\min}} \geq 2^{-33} \frac{d^2H^3}{\Delta_{\min}}.$$

Therefore at least one fixed family member satisfies Equation (52).

Finally, Equation (53) follows directly from Equation (48)–Equation (50) and the bounds $m \leq d$ and $R_t \leq H$. In particular,

$$\frac{8H}{c_{\text{cal}}} \max_{t \in \mathcal{T}} n_t^{\text{info}} = \frac{m^2 H}{2^{16} A_0^2 c_{\text{cal}} \Delta_{\min}^2} \max_{t \in \mathcal{T}} R_t^2 = O\left(\frac{d^2H^3}{c_{\text{cal}} \Delta_{\min}^2} \right).$$

This completes the proof. □

H.7 Application to LSVI-UCB++

We now verify Definition H.1 and the budget condition in Theorem H.3 for the modified LSVI-UCB++ in Algorithm 1. We use the confidence guarantees already proved in Lemma E.1 and Lemma F.1; no new concentration theorem is needed here. We instantiate the deterministic tie-breaking in Algorithm 1 by choosing the smallest indexed maximizing action. This convention does not affect the upper-bound analysis.

The weighted covariance update of Algorithm 1 is

$$\mathbf{\Sigma}_{k+1,h} = \mathbf{\Sigma}_{k,h} + \bar{\sigma}_{k,h}^{-2} \boldsymbol{\phi}(s_h^k, a_h^k) \boldsymbol{\phi}(s_h^k, a_h^k)^\top,$$

where

$$\bar{\sigma}_{k,h}^2 = \max\left\{ \sigma_{k,h}^2, H, 2d^3H^2 \|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\mathbf{\Sigma}_{k,h}^{-1}} \right\}. \tag{69}$$

Thus

$$\alpha_{k,h} = \bar{\sigma}_{k,h}^{-2}, \qquad \Lambda_{k,h}^{\text{all}} = \mathbf{\Sigma}_{k,h}.$$

The quantity $\bar{\sigma}_{k,h}$ is determined before s_{h+1}^k is observed, so the weights are predictable and remain immutable thereafter.

At a genuine determinant-doubling update, Algorithm 1 recomputes all steps in a single backward pass while keeping the same previous update index fixed throughout that pass. Its optimistic Bellman predictor is

$$\widehat{\mathcal{B}}_{u,h}(s,a) = r_h(s,a) + \hat{\mathbf{w}}_{u,h}^\top \boldsymbol{\phi}(s,a),$$

which is exactly Equation (42), and its current candidate is

$$\min\left\{ r_h(s,a) + \hat{\mathbf{w}}_{u,h}^\top \boldsymbol{\phi}(s,a) + \beta \|\boldsymbol{\phi}(s,a)\|_{\mathbf{\Sigma}_{u,h}^{-1}}, H \right\}.$$

Because each genuine update takes the pointwise minimum with the previously maintained optimistic estimate, the maintained Q-function is precisely the historical-minimum rule in Equation (44).

While an update index u remains in use, $\mathbf{\Sigma}_{k,h} \succeq \mathbf{\Sigma}_{u,h}$ and the determinant-doubling condition has not yet fired, so

$$\det(\mathbf{\Sigma}_{k,h}) < 2\det(\mathbf{\Sigma}_{u,h}).$$

Every eigenvalue of $\mathbf{\Sigma}_{u,h}^{-1/2}\mathbf{\Sigma}_{k,h}\mathbf{\Sigma}_{u,h}^{-1/2}$ is at least one, while their product is smaller than two. Hence every eigenvalue is smaller than two and

$$\mathbf{\Sigma}_{u,h} \preceq \mathbf{\Sigma}_{k,h} \preceq 2\mathbf{\Sigma}_{u,h}. \tag{70}$$

Therefore Equation (41) holds with $\kappa = 2$.

With the standard choice $\lambda = H^{-2}$, we have $\mathbf{\Sigma}_{k,h} \succeq H^{-2}I_d$, and hence

$$\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\mathbf{\Sigma}_{k,h}^{-1}} \leq H.$$

The clipped empirical-variance term in $\sigma_{k,h}^2$ is at most H^2, the two terms defining $E_{k,h}$ are each at most H^2, $D_{k,h} \leq d^3H^3$, and the additive term is H. Thus

$$\sigma_{k,h}^2 \leq H^2 + 2H^2 + d^3H^3 + H \leq 5d^3H^3.$$

The leverage term in Equation (69) is at most $2d^3H^3$, while the outer maximum always gives $\bar{\sigma}_{k,h}^2 \geq H$. Consequently,

$$H \leq \bar{\sigma}_{k,h}^2 \leq 5d^3H^3, \qquad \alpha_- := \frac{1}{5d^3H^3}, \qquad \alpha_+ := \frac{1}{H}. \tag{71}$$

These bounds are pathwise and do not rely on a high-probability event.

It remains to verify modelwise full-domain calibration and optimism. Set

$$\delta_{\mathrm{src}} := \frac{1}{19HK}, \qquad \delta_{\mathrm{cal}} := \frac{1}{16}, \qquad L_K := 1 + \log(1 + 19dH^4K^2),$$

and choose the radii according to Equation (3), with sufficiently large fixed universal constants. Equivalently, for suitable numerical constants $c_\beta, \bar{c}_\beta, \tilde{c}_\beta$,

$$\begin{aligned}
\beta &= c_\beta \sqrt{d}\, L_K, \\
\bar{\beta} &= \bar{c}_\beta \left[\sqrt{d} + d^{3/2}H \log(19dH^4K^2)\right], \\
\tilde{\beta} &= \tilde{c}_\beta \left[H\sqrt{d} + d^{3/2}H^2 \log(19dH^4K^2)\right].
\end{aligned}$$

Lemma E.1 gives

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - 15\delta_{\mathrm{src}} \geq 1 - \delta_{\mathrm{cal}}$$

for $H \geq 16$. On $\mathcal{E}_2$, the Bellman confidence inequality with radius β holds simultaneously for every $(s, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, including the unreachable probe states. Lemma F.1 gives global optimism on $\mathcal{E}_1 \cap \mathcal{E}_2$. Hence, for every fixed linear MDP, we may take

$$\mathcal{E}_{\mathrm{cal}}^M := \mathcal{E}_1 \cap \mathcal{E}_2.$$

Together with Equation (71) and Equation (70), this shows that Algorithm 1 satisfies Definition H.1 with parameters

$$\left(H^{-2}, \beta, \frac{1}{5d^3H^3}, \frac{1}{H}, 2, \frac{1}{16}\right).$$

No simultaneous event over different MDPs is required.

For $\delta_{\mathrm{cal}} = 1/16$, the numerical inequality

$$\chi_0 + \frac{1}{64} + \frac{1}{16} = \frac{131072}{321489} + \frac{5}{64} < \frac{63}{128}$$

implies

$$c_{\mathrm{cal}} > \frac{127}{256} - \frac{63}{128} = \frac{1}{256}.$$

Hence the information term in Equation (51) is at most

$$2048H \max_{t \in \mathcal{T}} n_t^{\text{info}}.$$

Substituting $\lambda = H^{-2}$, $\kappa = 2$, and Equation (71) into Equation (48) gives

$$N_b \le 1 + 5d^3H + 40960d^3H^3\beta^2,$$
$$n_s \le 1 + \frac{10240d^3H^3\beta^2}{\Delta_{\min}^2}.$$

Together with the information term in Equation (50), these bounds imply that there is a universal constant $C_0 > 0$ such that Equation (51) holds whenever

$$K \ge C_0 d^3 H^3 \beta^2 \left(1 + \frac{H}{\Delta_{\min}^2}\right). \tag{72}$$

Since $\beta^2 = O(dL_K^2)$ for the fixed confidence-radius constants, and for $x = dHK \ge 1$ we have $1 + 19dH^4K^2 \le 20(1+x)^4$, it follows that $L_K \le C[1 + \log(1 + dHK)]$ for an absolute constant C. Hence there is another universal constant $C_1 > 0$ such that the sufficient condition

$$K \ge C_1 d^4 H^3 \left(1 + \frac{H}{\Delta_{\min}^2}\right) [1 + \log(1 + dHK)]^2$$

implies Equation (72). The occurrence of K inside the logarithm is harmless and can be removed explicitly as follows.

Lemma H.13 (Removing the implicit logarithmic budget)**.** *For $A, B_1 \ge 1$,*

$$K \ge 64A[\log(eAB_1)]^2 \quad \Longrightarrow \quad K \ge A[\log(B_1K)]^2.$$

Proof. Let $s = \log(eAB_1) \ge 1$ and $K_\star = 64As^2$. Then

$$\log(B_1K_\star) = s - 1 + \log 64 + 2\log s \le 8s,$$

so $K_\star \ge A[\log(B_1K_\star)]^2$. Moreover, $K/[\log(B_1K)]^2$ is increasing whenever $\log(B_1K) > 2$, which holds for every $K \ge K_\star$. □

Let

$$A := C_2 d^4 H^3 \left(1 + \frac{H}{\Delta_{\min}^2}\right), \qquad B_1 := 2edH,$$

where C_2 is a sufficiently large universal constant absorbing the fixed confidence-radius constants and the comparison between L_K and $1 + \log(1 + dHK)$. Then

$$K \ge 64A[\log(eAB_1)]^2 \tag{73}$$

is an explicit polynomial-logarithmic sufficient condition for Equation (51). Therefore Theorem H.3 gives the following direct specialization.

Corollary H.14 (Persistent-ellipsoidal lower bound for LSVI-UCB++)**.** *Let $d \ge 15$, $H \ge 16$, and $\Delta_{\min} \in (0, H/16]$. Run Algorithm 1 with $\lambda = H^{-2}$, the confidence radii in Equation* (3) *with $\delta = 1/(19HK)$, sufficiently large fixed universal constants, and smallest-index tie-breaking. If K satisfies Equation* (73)*, then there exists a d-dimensional linear MDP with fixed initial state and minimum positive gap exactly $\Delta_{\min}$ such that*

$$\mathbb{E}[\text{Regret}(T)] \ge 2^{-33} \frac{d^2H^3}{\Delta_{\min}}.$$

The lower bound in Corollary H.14 matches the $d^2H^3/\Delta_{\min}$ dependence of the leading term in Theorem 3.1 up to logarithmic factors.

I PROOF OF THEOREM 3.3

For fixed integers $d, H \geq 1$, let $\mathfrak{M}_{d,H}$ denote the class of horizon-H linear MDPs with ambient feature dimension d, deterministic known rewards in $[0,1]$, and the standard linear-MDP norm bounds.

For probability measures P and Q with $P \ll Q$, define

$$\mathrm{KL}(P\|Q) := \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P.$$

For $p, q \in [0,1]$, define

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

with the usual conventions at the boundary of $[0,1]$.

Proof. Fix any possibly randomized learning algorithm Alg and parameters d, H, $\Delta_{\min}$, and K satisfying the conditions of Theorem 3.3. Let

$$d' := \left\lfloor \frac{d-3}{2} \right\rfloor, \qquad H' := \left\lfloor \frac{H}{2} \right\rfloor.$$

We construct a family of linear MDPs indexed by

$$\Sigma := \{-1,+1\}^{[d']\times\{2,\dots,H'\}}.$$

For each $\sigma \in \Sigma$, let $\mathbb{Q}_\sigma^K$ denote the probability law of the complete interaction history through episode K generated by Alg on the corresponding MDP $\mathcal{M}_\sigma$, and let $\mathbb{E}_\sigma$ denote the corresponding expectation.

Construction of the hard-instance family. Consider the state space

$$\mathcal{S} := \{w, g, b, x_1, \dots, x_{d'}\}$$

and the action space $\{1,2\}$. The initial state is w in every episode. The state w is a waiting state, $x_1, \dots, x_{d'}$ are decision states, and g and b are absorbing states. For every $t \in [H]$ and $a \in \{1,2\}$, define the deterministic known rewards by

$$r_t(g,a) = 1, \qquad r_t(s,a) = 0 \quad \text{for every } s \neq g.$$

Thus, all rewards are deterministic and belong to $[0,1]$.

For every $t \in [H'-1]$ and $a \in \{1,2\}$, let

$$P_t^\sigma(x_i \mid w, a) = \frac{1}{d'H}, \quad i \in [d'], \qquad P_t^\sigma(w \mid w, a) = 1 - \frac{1}{H}.$$

For every $t \in \{H', \dots, H\}$ and $a \in \{1,2\}$, let $P_t^\sigma(b \mid w, a) = 1$. Moreover, for every $t \in [H]$ and $a \in \{1,2\}$, let

$$P_t^\sigma(g \mid g, a) = 1, \qquad P_t^\sigma(b \mid b, a) = 1.$$

For every $i \in [d']$ and $t \in \{2, \dots, H'\}$, define

$$P_t^\sigma(g \mid x_i, 1) = \frac{1}{2}, \qquad P_t^\sigma(b \mid x_i, 1) = \frac{1}{2},$$

and

$$P_t^\sigma(g \mid x_i, 2) = \frac{1}{2} + \sigma_{i,t}\frac{\Delta_{\min}}{H-t}, \qquad P_t^\sigma(b \mid x_i, 2) = \frac{1}{2} - \sigma_{i,t}\frac{\Delta_{\min}}{H-t}.$$

At every other step, both actions at every x_i transition deterministically to b. Since $t \leq H'$ and $\Delta_{\min} \leq H/4$, we have $\Delta_{\min}/(H-t) \leq 1/2$, so all transition probabilities are valid.

If the process enters g after the transition at step t, it receives reward 1 at steps $t+1, \dots, H$. Hence, for every $i \in [d']$ and $t \in \{2, \dots, H'\}$,

$$Q_t^\star(x_i, 1) = \frac{H-t}{2}, \qquad Q_t^\star(x_i, 2) = \frac{H-t}{2} + \sigma_{i,t}\Delta_{\min}.$$

Therefore, when $\sigma_{i,t} = -1$, action 1 is optimal and action 2 has gap $\Delta_{\min}$, while when $\sigma_{i,t} = +1$, action 2 is optimal and action 1 has gap $\Delta_{\min}$. All other state-action-step triples have zero gap. Thus, every $\mathcal{M}_\sigma$ has minimum positive gap $\Delta_{\min}$.

Verification of the linear-MDP structure. Let $\{e_j\}_{j=1}^d$ be the standard basis of $\mathbb{R}^d$ and define

$$\boldsymbol{\phi}(x_i, 1) = e_{2i-1}, \qquad \boldsymbol{\phi}(x_i, 2) = e_{2i}, \qquad i \in [d'],$$

and

$$\boldsymbol{\phi}(w, a) = e_{2d'+1}, \qquad \boldsymbol{\phi}(g, a) = e_{2d'+2}, \qquad \boldsymbol{\phi}(b, a) = e_{2d'+3}, \qquad a \in \{1, 2\}.$$

Since $2d' + 3 \le d$, this defines a feature map in $\mathbb{R}^d$. More explicitly, $2d' + 3 = d$ when d is odd and $2d' + 3 = d - 1$ when d is even; in the latter case, the last coordinate is set to zero in the feature map and all linear parameters. Clearly, $\|\boldsymbol{\phi}(s, a)\|_2 = 1$.

For each step t, define the coordinates of the vector-valued signed measure $\boldsymbol{\theta}_t^\sigma$ by

$$(\theta_t^\sigma)^{(2i-1)}(\cdot) = P_t^\sigma(\cdot \mid x_i, 1), \quad (\theta_t^\sigma)^{(2i)}(\cdot) = P_t^\sigma(\cdot \mid x_i, 2), \qquad i \in [d'],$$

$$(\theta_t^\sigma)^{(2d'+1)}(\cdot) = P_t^\sigma(\cdot \mid w, 1), \quad (\theta_t^\sigma)^{(2d'+2)}(\cdot) = P_t^\sigma(\cdot \mid g, 1), \qquad (\theta_t^\sigma)^{(2d'+3)}(\cdot) = P_t^\sigma(\cdot \mid b, 1),$$

and set every unused coordinate to the zero measure. Since the two actions have identical transitions at w, g, and b,

$$P_t^\sigma(\cdot \mid s, a) = \langle \boldsymbol{\phi}(s, a), \boldsymbol{\theta}_t^\sigma(\cdot) \rangle$$

for every (s, a, t). Moreover, for every measurable $B \subseteq \mathcal{S}$, each nonzero coordinate of $\theta_t^\sigma(B)$ lies in $[0, 1]$, and there are at most $2d' + 3 \le d$ such coordinates. Hence $\|\theta_t^\sigma(B)\|_2^2 \le 2d' + 3 \le d$.

Therefore, $\mathcal{M}_\sigma \in \mathfrak{M}_{d,H}$ for every $\sigma \in \Sigma$.

Reach probabilities and regret decomposition. For every $i \in [d']$ and $t \in \{2, \dots, H'\}$, because both actions have identical transition kernels at w, the probability of reaching x_i at step t is independent of the algorithm and of σ, and equals

$$p_{i,t} := \left(1 - \frac{1}{H}\right)^{t-2} \frac{1}{d'H}.$$

Since $t \le H'$, Bernoulli's inequality gives

$$\left(1 - \frac{1}{H}\right)^{t-2} \ge 1 - \frac{t-2}{H} \ge \frac{1}{2}.$$

Since $d' \le d/2$, it follows that

$$p_{i,t} \ge \frac{1}{2d'H} \ge \frac{1}{dH}.$$

For $a \in \{1, 2\}$, define

$$N_{i,t,a}(K) := \sum_{k=1}^{K} \mathbb{I}\{s_t^k = x_i,\ a_t^k = a\}, \qquad N_{i,t}(K) := N_{i,t,1}(K) + N_{i,t,2}(K).$$

The transition draws in different episodes are independent, and the probability of reaching x_i at step t is always $p_{i,t}$. Therefore,

$$N_{i,t}(K) \sim \mathrm{Binomial}(K, p_{i,t})$$

under every $\mathcal{M}_\sigma$.

By the expected performance-difference identity and the gap structure above,

$$\mathbb{E}_\sigma[\mathrm{Regret}(T)] = \Delta_{\min} \sum_{i=1}^{d'} \sum_{t=2}^{H'} \left(\mathbb{I}\{\sigma_{i,t} = -1\}\mathbb{E}_\sigma[N_{i,t,2}(K)] + \mathbb{I}\{\sigma_{i,t} = +1\}\mathbb{E}_\sigma[N_{i,t,1}(K)]\right).$$

A two-instance lower bound. Fix any $i \in [d']$ and $t \in \{2, \dots, \lfloor H/4 \rfloor\}$, and fix all entries of σ other than $\sigma_{i,t}$. Let σ^- and σ^+ denote the two resulting sign vectors obtained by setting $\sigma_{i,t} = -1$ and $\sigma_{i,t} = +1$, respectively. The MDPs $\mathcal{M}_{\sigma^-}$ and $\mathcal{M}_{\sigma^+}$ agree everywhere except at $(x_i, 2, t)$, where

$$P_t^{\sigma^-}(g \mid x_i, 2) = \frac{1}{2} - \frac{\Delta_{\min}}{H - t}, \qquad P_t^{\sigma^+}(g \mid x_i, 2) = \frac{1}{2} + \frac{\Delta_{\min}}{H - t}.$$

We claim that

$$\max\left\{\mathbb{E}_{\sigma^-}[N_{i,t,2}(K)], \mathbb{E}_{\sigma^+}[N_{i,t,1}(K)]\right\} \geq \frac{(H-t)^2}{256\Delta_{\min}^2}\log K.$$

Suppose, toward a contradiction, that

$$\mathbb{E}_{\sigma^-}[N_{i,t,2}(K)] < \frac{(H-t)^2}{256\Delta_{\min}^2}\log K, \qquad \mathbb{E}_{\sigma^+}[N_{i,t,1}(K)] < \frac{(H-t)^2}{256\Delta_{\min}^2}\log K.$$

Define

$$\mathcal{L}_K := \left\{N_{i,t,2}(K) \leq \frac{p_{i,t}K}{2}\right\}.$$

Under $\mathcal{M}_{\sigma^-}$, Markov's inequality gives

$$\mathbb{Q}_{\sigma^-}^K(\mathcal{L}_K^c) \leq \frac{2\mathbb{E}_{\sigma^-}[N_{i,t,2}(K)]}{p_{i,t}K} < \frac{(H-t)^2\log K}{128p_{i,t}K\Delta_{\min}^2} \leq \frac{dH^3\log K}{128K\Delta_{\min}^2}.$$

The assumed regime $K \geq (dH^3/\Delta_{\min}^2)^2$ implies

$$\frac{dH^3}{\Delta_{\min}^2} \leq \sqrt{K}.$$

Moreover, $\Delta_{\min} \leq H/4$, $d \geq 7$, and $H \geq 8$ imply $K \geq (dH^3/\Delta_{\min}^2)^2 \geq 896^2 > 2^{16}$, so the assumed regime in particular ensures that K is large enough that $\log K \leq K^{1/4}$. Hence,

$$\mathbb{Q}_{\sigma^-}^K(\mathcal{L}_K^c) \leq \frac{1}{128K^{1/4}} < \frac{1}{4},$$

and therefore

$$\mathbb{Q}_{\sigma^-}^K(\mathcal{L}_K) \geq \frac{3}{4}.$$

Under $\mathcal{M}_{\sigma^+}$, if $\mathcal{L}_K$ occurs and $N_{i,t}(K) > 3p_{i,t}K/4$, then

$$N_{i,t,1}(K) = N_{i,t}(K) - N_{i,t,2}(K) > \frac{p_{i,t}K}{4}.$$

Thus, by a multiplicative Chernoff bound and Markov's inequality,

$$\begin{aligned}\mathbb{Q}_{\sigma^+}^K(\mathcal{L}_K) &\leq \mathbb{Q}_{\sigma^+}^K\left(N_{i,t}(K) \leq \frac{3p_{i,t}K}{4}\right) + \mathbb{Q}_{\sigma^+}^K\left(N_{i,t,1}(K) > \frac{p_{i,t}K}{4}\right)\\ &\leq \exp\left(-\frac{p_{i,t}K}{32}\right) + \frac{4\mathbb{E}_{\sigma^+}[N_{i,t,1}(K)]}{p_{i,t}K}\\ &< \exp\left(-\frac{K}{32dH}\right) + \frac{dH^3\log K}{64K\Delta_{\min}^2}.\end{aligned}$$

Using again $dH^3/\Delta_{\min}^2 \leq \sqrt{K}$ and $\log K \leq K^{1/4}$,

$$\frac{dH^3\log K}{64K\Delta_{\min}^2} \leq \frac{1}{64K^{1/4}}.$$

Furthermore, since $dH^3/\Delta_{\min}^2 \geq 16dH$ and $\sqrt{K} \geq dH^3/\Delta_{\min}^2$,

$$\frac{K}{32dH} \geq \frac{\sqrt{K}}{2}.$$

For every $K \geq 896^2$, $\frac{\sqrt{K}}{2} \geq \frac{1}{4}\log K + \log 64$, and hence

$$\exp\left(-\frac{K}{32dH}\right) \leq \frac{1}{64K^{1/4}}.$$

Consequently,

$$\mathbb{Q}_{\sigma^+}^K(\mathcal{L}_K) \leq \frac{1}{32K^{1/4}} \leq K^{-1/4}.$$

Although the histories and selected actions may have different distributions under the two MDPs, given any fixed history before an action is selected, the algorithm uses the same conditional action-selection kernel under both MDPs. Hence, the action-selection factors cancel when the likelihood ratio of the same realized history is formed. Since the initial state and reward functions are also identical and the two transition kernels differ only at $(x_i, 2, t)$, the chain rule for relative entropy gives

$$\mathrm{KL}(\mathbb{Q}^K_{\sigma^-} \| \mathbb{Q}^K_{\sigma^+}) = \mathbb{E}_{\sigma^-}[N_{i,t,2}(K)]\,\mathrm{kl}\left(\frac{1}{2} - \frac{\Delta_{\min}}{H-t}, \frac{1}{2} + \frac{\Delta_{\min}}{H-t}\right).$$

Since $\mathcal{L}_K$ is determined by the complete interaction history, the data-processing inequality gives

$$\mathrm{KL}(\mathbb{Q}^K_{\sigma^-} \| \mathbb{Q}^K_{\sigma^+}) \geq \mathrm{kl}\left(\mathbb{Q}^K_{\sigma^-}(\mathcal{L}_K), \mathbb{Q}^K_{\sigma^+}(\mathcal{L}_K)\right).$$

For $p, q \in [0,1]$,

$$\mathrm{kl}(p,q) \geq p \log\left(\frac{1}{q}\right) - \log 2.$$

Using $\mathbb{Q}^K_{\sigma^-}(\mathcal{L}_K) \geq 3/4$ and $\mathbb{Q}^K_{\sigma^+}(\mathcal{L}_K) \leq K^{-1/4}$, we obtain

$$\mathrm{KL}(\mathbb{Q}^K_{\sigma^-} \| \mathbb{Q}^K_{\sigma^+}) \geq \frac{3}{16}\log K - \log 2.$$

On the other hand, using $\mathrm{kl}(p,q) \leq (p-q)^2/[q(1-q)]$ and $\Delta_{\min}/(H-t) \leq 1/3$,

$$\mathrm{kl}\left(\frac{1}{2} - \frac{\Delta_{\min}}{H-t}, \frac{1}{2} + \frac{\Delta_{\min}}{H-t}\right) \leq \frac{32\Delta^2_{\min}}{(H-t)^2}.$$

Together with the assumed upper bound on $\mathbb{E}_{\sigma^-}[N_{i,t,2}(K)]$, this gives

$$\mathrm{KL}(\mathbb{Q}^K_{\sigma^-} \| \mathbb{Q}^K_{\sigma^+}) < \frac{1}{8}\log K.$$

Since the assumed regime implies $K > 2^{16}$, we have

$$\frac{3}{16}\log K - \log 2 > \frac{1}{8}\log K,$$

which is a contradiction. Therefore,

$$\max\{\mathbb{E}_{\sigma^-}[N_{i,t,2}(K)], \mathbb{E}_{\sigma^+}[N_{i,t,1}(K)]\} \geq \frac{(H-t)^2}{256\Delta^2_{\min}}\log K.$$

Averaging over the hard-instance family. Fix $i \in [d']$ and $t \in \{2, \ldots, \lfloor H/4 \rfloor\}$. The set Σ can be partitioned into exactly

$$2^{d'(H'-1)-1}$$

disjoint pairs (σ^-, σ^+) which agree on every coordinate except (i,t) and satisfy $\sigma^-_{i,t} = -1$ and $\sigma^+_{i,t} = +1$. For every such pair, the preceding two-instance bound implies

$$\mathbb{E}_{\sigma^-}[N_{i,t,2}(K)] + \mathbb{E}_{\sigma^+}[N_{i,t,1}(K)] \geq \frac{(H-t)^2}{256\Delta^2_{\min}}\log K.$$

Therefore,

$$\begin{aligned}\frac{1}{|\Sigma|}\sum_{\sigma\in\Sigma}\left(\mathbb{I}\{\sigma_{i,t}=-1\}\mathbb{E}_\sigma[N_{i,t,2}(K)] + \mathbb{I}\{\sigma_{i,t}=+1\}\mathbb{E}_\sigma[N_{i,t,1}(K)]\right)\\ \geq \frac{2^{d'(H'-1)-1}}{2^{d'(H'-1)}}\frac{(H-t)^2}{256\Delta^2_{\min}}\log K = \frac{(H-t)^2}{512\Delta^2_{\min}}\log K.\end{aligned}$$

Averaging the regret decomposition over all $\sigma \in \Sigma$ and then applying the preceding bound to every (i, t) yields

$$\begin{aligned}
&\frac{1}{|\Sigma|}\sum_{\sigma\in\Sigma}\mathbb{E}_\sigma[\mathrm{Regret}(T)] \\
&= \Delta_{\min}\sum_{i=1}^{d'}\sum_{t=2}^{H'}\frac{1}{|\Sigma|}\sum_{\sigma\in\Sigma}\left(\mathbb{I}\{\sigma_{i,t}=-1\}\mathbb{E}_\sigma[N_{i,t,2}(K)]+\mathbb{I}\{\sigma_{i,t}=+1\}\mathbb{E}_\sigma[N_{i,t,1}(K)]\right) \\
&\geq \Delta_{\min}\sum_{i=1}^{d'}\sum_{t=2}^{\lfloor H/4\rfloor}\frac{1}{|\Sigma|}\sum_{\sigma\in\Sigma}\left(\mathbb{I}\{\sigma_{i,t}=-1\}\mathbb{E}_\sigma[N_{i,t,2}(K)]+\mathbb{I}\{\sigma_{i,t}=+1\}\mathbb{E}_\sigma[N_{i,t,1}(K)]\right) \\
&\geq \frac{\log K}{512\Delta_{\min}}\sum_{i=1}^{d'}\sum_{t=2}^{\lfloor H/4\rfloor}(H-t)^2.
\end{aligned}$$

Since $t \leq \lfloor H/4 \rfloor$,

$$(H-t)^2 \geq \frac{9H^2}{16}.$$

Therefore,

$$\frac{1}{|\Sigma|}\sum_{\sigma\in\Sigma}\mathbb{E}_\sigma[\mathrm{Regret}(T)] \geq \frac{\log K}{512\Delta_{\min}}\, d'(\lfloor\frac{H}{4}\rfloor - 1)\frac{9H^2}{16}.$$

For $d \geq 7$ and $H \geq 8$,

$$d' = \left\lfloor\frac{d-3}{2}\right\rfloor \geq \frac{d}{4}, \qquad \left\lfloor\frac{H}{4}\right\rfloor - 1 \geq \frac{H}{11}.$$

Hence,

$$\frac{1}{|\Sigma|}\sum_{\sigma\in\Sigma}\mathbb{E}_\sigma[\mathrm{Regret}(T)] \geq \frac{1}{40050}\frac{dH^3}{\Delta_{\min}}\log K.$$

Since the maximum is at least the average,

$$\max_{\sigma\in\Sigma}\mathbb{E}_\sigma[\mathrm{Regret}(T)] \geq \frac{1}{40050}\frac{dH^3}{\Delta_{\min}}\log K.$$

Thus, there exists $\sigma^\star \in \Sigma$ such that

$$\mathbb{E}_{\sigma^\star}[\mathrm{Regret}(T)] \geq \frac{1}{40050}\frac{dH^3}{\Delta_{\min}}\log K.$$

Since $\mathcal{M}_{\sigma^\star} \in \mathfrak{M}_{d,H}$ and has minimum positive gap $\Delta_{\min}$, this completes the proof of Theorem 3.3 with $c = 1/40050$. □