IMPROVED WORST-CASE REGRET BOUNDS FOR RANDOMIZED LEAST-SQUARES VALUE ITERATION

BY

PRIYANK AGRAWAL

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois Urbana-Champaign, 2021

Urbana, Illinois

Adviser:

Assistant Professor Nan Jiang

ABSTRACT

This work studies regret minimization with randomized value functions in reinforcement learning. In tabular finite-horizon Markov Decision Processes, we introduce a clipping variant of one classical Thompson Sampling (TS)-like algorithm, randomized least-squares value iteration (RLSVI). Our $\tilde{\mathrm{O}}(H^2 S\sqrt{AT})$ high-probability worst-case regret bound improves the previous sharpest worst-case regret bounds for RLSVI and matches the existing state-of-the-art worst-case TS-based regret bounds.

To all, due to whom I have the privilege of contributing to academic research.

ACKNOWLEDGMENTS

I want to thank my advisor, Professor Nan Jiang, who helped me provide a smooth transition to the domain of reinforcement learning theory. Nan has always ensured a friendly environment, where I could be a part of the research in reinforcement learning theory and could also learn about the best academic research practices.

I am also grateful to my collaborator and friend, Jinglin, with whom I co-authored a paper in the Association for the Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence, the results of which are presented in this thesis.

I would cherish my time at the University of Illinois Urbana-Champaign (UIUC). I am also greatly thankful to the graduate advising office of the Computer Science department, who provided a homely environment right from the start allowing me to focus solely on scholarly pursuits.

I have come to a belief that it is a privilege to be a part of academic research. There are many people whose time, support and care have allowed me this privilege. I will forever be indebted to them.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

We study systematic exploration in reinforcement learning (RL) and the exploration-exploitation trade-off therein. Exploration in RL [1] has predominantly focused on *Optimism in the face of Uncertainty* (OFU) based algorithms. Since the seminal work of [2], many provably efficient methods have been proposed but most of them are restricted to either tabular or linear setting [3, 4]. A few papers study a more general framework but subjected to computational intractability [5, 6, 7]. Another broad category is Thompson Sampling (TS)-based methods [8, 9]. They are believed to have more appealing empirical results [10, 11].

In this work, we investigate a TS-like algorithm, RLSVI [12, 13, 14, 15]. In RLSVI, the exploration is induced by injecting randomness into the value function. The algorithm generates a randomized value function by carefully selecting the variance of Gaussian noise, which is used in perturbations of the history data (the trajectory of the algorithm till the current episode) and then applies the least square policy iteration algorithm of [16]. Thanks to the model-free nature, RLSVI is flexible enough to be extended to general function approximation setting, as shown by [13, 17, 18], and at the same time has less burden on the computational side.

We propose C-RLSVI algorithm, which additionally considers an initial burn-in or warm-up phase on top of the core structure of RLSVI. Theoretically, we prove that C-RLSVI achieves $\tilde{\mathrm{O}}(H^2 S\sqrt{AT})$ high-probability regret bound[1].

Significance of Our Results

- Our high-probability bound improves upon previous $\tilde{\mathrm{O}}(H^{5/2} S^{3/2}\sqrt{AT})$ worst-case expected regret bound of RLSVI in [14].
- Our high-probability regret bound matches the sharpest $\tilde{\mathrm{O}}(H^2 S\sqrt{AT})$ worst-case regret bound among all TS-based methods [9][2].

Related Works Taking inspirations from [3, 19, 20, 21], we introduce clipping to avoid propagation of unreasonable estimates of the value function. Clipping creates a warm-up effect that only affects the regret bound with constant factors (i.e. independent of the total

[1] $\tilde{\mathrm{O}}(\cdot)$ hides dependence on logarithmic factors.

[2] [9] studies weakly communicating MDPs with diameter D. Bounds comparable to our setting (time in-homogeneous) are obtained by augmenting their state space as $S' \to SH$ and noticing $D \geq H$.

number of steps T). With the help of clipping, we prove that the randomized value functions are bounded with high probability.

In the context of using perturbation or random noise methods to obtain provable exploration guarantees, there have been recent works [17, 22, 23, 24, 25] in both theoretical RL and bandit literature. A common theme has been to develop a TS-like algorithm that is suitable for complex models where exact posterior sampling is impossible. RLSVI also enjoys such conceptual connections with Thompson sampling [12, 13]. Related to this theme, the worst-case analysis of [9] should be highlighted, where the authors do not solve for a pure TS algorithm but have proposed an algorithm that samples many times from posterior distribution to obtain an optimistic model. In comparison, C-RLSVI does not require such strong optimistic guarantee.

Our results are not optimal as compared with $\Omega(H\sqrt{SAT})$ lower bounds in [2] [3]. The gap of $\sqrt{SH}$ is sometimes attributed to the additional cost of exploration in TS-like approaches [26]. Whether this gap can be closed, at least for RLSVI, is still an interesting open question. We hope our analysis serves as a building block towards a deeper understanding of TS-based methods.

[3] The lower bound is translated to time-inhomogeneous setting.

CHAPTER 2: LEARNING SET-UP

2.1 INTERACTION MODEL AND LEARNING GOAL

Markov Decision Processes We consider the episodic Markov Decision Process (MDP) $M = (H, \mathcal{S}, \mathcal{A}, P, R, s_1)$ described by [27], where H is the length of the episode, $\mathcal{S} = \{1, 2, \ldots, S\}$ is the finite state space, $\mathcal{A} = \{1, 2, \ldots, A\}$ is the finite action space, $P = [P_1, \ldots, P_H]$ with $P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition function, $R = [R_1, \ldots, R_H]$ with $R_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, and s_1 is the deterministic initial state.

A deterministic (and non-stationary) policy $\pi = (\pi_1, \ldots, \pi_H)$ is a sequence of functions, where each $\pi_h : \mathcal{S} \to \mathcal{A}$ defines the action to take at each state. The RL agent interacts with the environment across K episodes giving us $T = KH$ steps in total. In episode k, the agent start with initial state $s_1^k = s_1$ and then follows policy π^k, thus inducing trajectory $s_1^k, a_1^k, r_1^k, s_2^k, a_2^k, r_2^k, \ldots, s_H^k, a_h^k, r_H^k$.

For any timestep h and state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, the Q-value function of policy π is defined as $Q_h^\pi(s, a) = R_h(s, a) + \mathbb{E}_\pi[\sum_{l=h}^H R_l(s_l, \pi_l(s_l)|s, a)]$ and the state-value function is defined as $V_h^\pi(s) = Q_h^\pi(s, \pi_h(s))$. We use π^* to denote the optimal policy. The optimal state-value function is defined as $V_h^*(s) := V_h^{\pi^*}(s) = \max_\pi V_h^\pi(s)$ and the optimal Q-value function is defined as $Q_h^*(s, a) := Q_h^{\pi^*}(s, a) = \max_\pi Q_h^\pi(s, a)$. Both Q^π and Q^* satisfy Bellman equations

$$Q_h^\pi(s, a) = R_h(s, a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[V_{h+1}^\pi(s')], \tag{2.1}$$

$$Q_h^*(s, a) = R_h(s, a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[V_{h+1}^*(s')], \tag{2.2}$$

where $V_{H+1}^\pi(s) = V_{H+1}^*(s) = 0 \ \forall s$. Notice that by the bounded nature of the reward function, for any (h, s, a), all functions $Q_h^*, V_h^*, Q_h^\pi, V_h^\pi$ are within the range $[0, H - h + 1]$. Since we consider the time-inhomogeneous setting (reward and transition change with timestep h), we have subscript h on policy and value functions, and later traverse over (h, s, a) instead of (s, a).

Regret An RL algorithm is a random mapping from the history until the end of episode $k - 1$ to policy π^k at episode k. We use regret to evaluate the performance of the algorithm:

$$\text{Reg}(K) = \sum_{k=1}^K V_1^*(s_1) - V_1^{\pi^k}(s_1). \tag{2.3}$$

Regret $\text{Reg}(K)$ is a random variable, and we bound it with high probability $1 - \delta$. We emphasize that high-probability regret bound provides a stronger guarantee on each roll-out [28, 29] and can be converted to the same order of expected regret bound:

$$\text{E-Reg}(K) = \mathbb{E}\left[\sum_{k=1}^{K} V_1^*(s_1) - V_1^{\pi^k}(s_1)\right], \tag{2.4}$$

by setting $\delta = 1/T$. However, expected regret bound does not imply small variance for each run. Therefore it can violate the same order of high-probability regret bound. We also point out that both bounds hold for all MDP instances M that have S states, A actions, horizon H, and bounded reward $R \in [0, 1]$. In other words, we consider worst-case (frequentist) regret bound.

Empirical MDP We define the number of visitation of (s, a) pair at timestep h until the end of episode $k-1$ as $n^k(h, s, a) = \sum_{l=1}^{k-1} \mathbf{1}\{(s_h^l, a_h^l) = (s, a)\}$. We also construct empirical reward and empirical transition function as $\hat{R}_{h,s,a}^k = \frac{1}{n^k(h,s,a)+1} \sum_{l=1}^{k-1} \mathbf{1}\{(s_h^l, a_h^l) = (s, a)\} r_h^l$ and $\hat{P}_{h,s,a}^k(s') = \frac{1}{n^k(h,s,a)+1} \sum_{l=1}^{k-1} \mathbf{1}\{(s_h^l, a_h^l, s_{h+1}^l) = (s, a, s')\}$. Finally, we use $\hat{M}^k = (H, \mathcal{S}, \mathcal{A}, \hat{P}^k, \hat{R}^k, s_1^k)$ to denote the empirical MDP. Notice that we have $n^k(h, s, a) + 1$ in the denominator, and it is not standard. The reason we have that is due to the analysis between model-free view and model-based view in Section 2.2. In the current form, $\hat{P}_{h,s,a}^k$ is no longer a valid probability function, and it is for ease of presentation. More formally, we can slightly augment the state space by adding one absorbing state for each level h and let all (h, s, a) transit to the absorbing states with remaining probability.

2.2 C-RLSVI ALGORITHM

The major goal of this work is to improve the regret bound of TS-based algorithms in the tabular setting. Different from using fixed bonus term in the optimism-in-face-of-uncertainty (OFU) approach, TS methods [14, 15, 26, 30] facilitate exploration by making large enough random perturbation so that optimism is obtained with at least a constant probability. However, the range of induced value function can easily grow unbounded and this forms a key obstacle in previous analysis [14]. To address this issue, we apply a common clipping technique in RL literature [3, 15, 21].

We now formally introduce our algorithm C-RLSVI as shown in Algorithm 2.1. C-RLSVI follows a similar approach as RLSVI in [14]. The algorithm proceeds in episodes. In episode k, the agent first samples Q_h^{pri} from prior $N(0, \frac{\beta_k}{2} I)$ and adds random perturbation on the data **(lines 3-10)**, where $\mathcal{D}_h = \{(s_h^l, a_h^l, r_h^l, s_{h+1}^l) : l < k\}$ for $h < H$ and $\mathcal{D}_H = \{(s_H^l, a_H^l, r_H^l, \emptyset) : l < k\}$.

The injection of Gaussian perturbation (noise) is essential for the purpose of exploration and we set $\beta_k = H^3 S \log(2HSAk)$. Later we will see the magnitude of β_k plays a crucial role in the regret bound and it is tuned to satisfy the optimism with a constant probability in Lemma 4.3. Given history data, the agent further conducts regularized least square regression **(lines 11-14)**, where $\mathcal{L}(Q \mid Q', \mathcal{D}) = \sum_{(s,a,r,s') \in \mathcal{D}} (Q(s,a) - r - \max_{a' \in \mathcal{A}} Q'(s', a'))^2$. After clipping on the Q-value function, we obtain $\dot{Q}^k$ **(lines 15-20)**. Finally, clipped Q-value function $\dot{Q}^k$ will be used to extract the greedy policy π^k and the agent rolls out a trajectory with π^k **(lines 21-22)**.

Algorithm 2.1: C-RLSVI

1: **input:** variance β_k and clipping threshold α_k;
2: **for** episode $k = 1, 2, \ldots, K$ **do**
3:   **for** timestep $h = 1, 2, \ldots, H$ **do**
4:     Sample prior $Q_h^{\text{pri}} \sim \mathcal{N}(0, \frac{\beta_k}{2} I)$;
5:     $\dot{D}_h \leftarrow \{\}$;
6:     **for** $(s, a, r, s') \in \mathcal{D}_h$ **do**
7:       Sample $w \sim \mathcal{N}(0, \beta_k/2)$;
8:       $\dot{\mathcal{D}}_h \leftarrow \dot{\mathcal{D}}_h \cup \{(s, a, r + w, s')\}$;
9:     **end for**
10:   **end for**
11:   Define terminal value $Q_{H+1}^k(s, a) \leftarrow 0 \quad \forall s, a$;
12:   **for** timestep $h = H, H-1, \ldots, 1$ **do**
13:     $\hat{Q}_h^k \leftarrow \operatorname{argmin}_{Q \in \mathbb{R}^{SA}} \Big[\mathcal{L}(Q \mid \hat{Q}_{h+1}^k, \dot{\mathcal{D}}_h) + \|Q - Q_h^{\text{pri}}\|_2^2 \Big]$;
14:   **end for**
15:   *(Clipping)* $\forall (h, s, a)$
16:   **if** $n^k(h, s, a) > \alpha_k$ **then**
17:     $\dot{Q}_h^k(s, a) = \hat{Q}_h^k(s, a)$;
18:   **else**
19:     $\dot{Q}_h^k(s, a) = H - h + 1$;
20:   **end if**
21:   Apply greedy policy (π^k) with respect to $(\dot{Q}_1^k, \ldots \dot{Q}_H^k)$ throughout episode;
22:   Obtain trajectory $s_1^k, a_1^k, r_1^k, \ldots s_H^k, a_H^k, r_H^k$;
23: **end for**

C-RLSVI as presented is a model-free algorithm, which can be easily extended to more general setting and achieve computational efficiency [15, 17]. However, it also has an equivalent model-based interpretation [14]. In Algorithm 2.1, the model-free view gives us unclipped Q-value function $\hat{Q}_h^k$ that satisfies, $\hat{Q}_h^k(s, a) | \hat{Q}_{h+1}^k \sim \mathcal{N}(p, q)$, where $p = \hat{R}_{h,s,a}^k + \sum_{s' \in S} \hat{P}_{h,s,a}^k(s') \max_{a' \in \mathcal{A}} \hat{Q}_{h+1}^k(s', a')$ and $q = \frac{\beta_k}{2(n^k(h,s,a)+1)}$. In model-based view, we first define noise term $w^k \in \mathbb{R}^{HSA}$, where $w^k(h,s,a) \sim \mathcal{N}(0, \sigma_k^2(h,s,a))$ and $\sigma_k(h,s,a) = \frac{\beta_k}{2(n^k(h,s,a)+1)}$. Then we construct a perturbed version of empirical MDP $\overline{M}^k = (H, \mathcal{S}, \mathcal{A}, \hat{P}^k, \hat{R}^k + w^k, s_1^k)$. Notice that the (Gaussian) noise term has the same variance in model-free view ($\hat{Q}^k$) and model-based view ($\overline{M}^k$), we can think of each time they sample the same noise w^k. Thus, from the equivalence between running Fitted Q-Iteration [31, 32] with data and using data to first build empirical MDP and then doing planing, one can easily show $\hat{Q}^k$ is the optimal policy of $\overline{M}^k$. If clipping does not happen at episode k, we know that π^k is the greedy policy of $\hat{Q}^k$, so further we know π^k is exactly the optimal policy of $\overline{M}^k$. In the analysis, we will mostly leverage such model-based interpretation.

Compared with RLSVI in [14], we introduce a clipping technique to handle the abnormal case in the Q-value function. C-RLSVI has simple one-phase clipping and the threshold $\alpha_k = 4H^3 S \log(2HSAk) \log\left(40SAT/\delta\right)$ is designed to guarantee the boundness of the value function. Clipping is the key step that allows us to introduce new analysis as compared to [14] and therefore obtain a high-probability regret bound. Similar to as discussed in [15], we want to emphasize that clipping also hurts the optimism obtained by simply adding Gaussian noise. However, clipping only happens at an early stage of visiting every (h,s,a) tuple. Intuitively, once (h,s,a) is visited for a large number of times, its estimated Q-value will be rather accurate and concentrates around the true value (within $[0, H-h+1]$), which means clipping will not take place. Another effect caused by clipping is we have an optimistic Q-value function at the initial phase of exploration since $Q_h^* \leq H-h+1$. However, this is not the crucial property that we gain from enforcing clipping. Although clipping slightly breaks the Bayesian interpretation of RLSVI [13, 14], it is easy to implement empirically and we will show it does not introduce a major term on regret bound.

CHAPTER 3: MAIN RESULT

In this chapter, we present our main result: high-probability regret bound in Theorem 3.1.

Theorem 3.1. C-RLSVI enjoys the following high-probability regret upper bound, with probability $1-\delta$,

$$\mathrm{Reg}(K) = \tilde{\mathrm{O}}\left(H^2 S\sqrt{AT}\right). \tag{3.1}$$

Theorem 3.1 shows that C-RLSVI matches the state-of-the-art TS-based method [9]. Compared to the lower bound [2], the result is at least off by $\sqrt{HS}$ factor. This additional factor of $\sqrt{HS}$ has eluded all the worst-case analyses of TS-based algorithms known to us in the tabular setting. This is similar to an extra $\sqrt{d}$ factor that appears in the worst-case upper bound analysis of TS for d-dimensional linear bandits [26]. It is useful to compare our work with the following contemporaries in related directions.

Comparison with [14] Other than the notion of clipping (which only contributed to warm-up or burn-in term), the core of C-RLSVI is the same as RLSVI considered by [14]. Their work presents significant insights about randomized value functions but the analysis does not extend to give high-probability regret bounds, and the latter requires a fresh analysis. Theorem 3.1 improves his worst-case expected regret bound $\tilde{\mathrm{O}}(H^{5/2}S^{3/2}\sqrt{AT})$ by $\sqrt{HS}$.

Comparison with [15] Very recently, [15] proposed frequentist regret analysis for a variant of RLSVI with linear function approximation and obtained high-probability regret bound of $\tilde{\mathrm{O}}\left(H^2 d^2\sqrt{T}\right)$, where d is the dimension of the low rank embedding of the MDP. While they present some interesting analytical insights which we use (see Section 4.2), directly converting their bound to tabular setting ($d \to SA$) gives us quite loose bound $\tilde{\mathrm{O}}\left(H^2 S^2 A^2\sqrt{T}\right)$.

Comparison with [3, 33] These OFU works guaranteeing optimism almost surely all the time are fundamentally different from RLSVI. However, they develop key technical ideas which are useful to our analysis, e.g. clipping estimated value functions and estimation error propagation techniques. Specifically, in [3, 19, 33], the estimation error is decomposed as a recurrence. Since RLSVI is only optimistic with a constant probability (see Section 4.2 for details), their techniques need to be substantially modified to be used in our analysis.

CHAPTER 4: TECHNICAL DETAILS

In this chapter we give full proof our main result Theorem 3.1. The Section 4.1 is devoted to developing key notations and definitions required in the analysis.

4.1 PRELIMINARIES

Table 4.1: Notation Table

Symbol	**Explanation**
$\mathcal{S}$	The state space
$\mathcal{A}$	The action space
S	Size of state space
A	Size of action space
H	The length of horizon
K	The total number of episodes
T	The total number of steps across all episodes
π^k	The greedy policy obtained in the Algorithm 2.1 at episode k, $\pi^k = \{\pi_1^k, \cdots, \pi_H^k\}$
π^*	The optimal policy of the true MDP
(s_h^k, a_h^k)	The state-action pair at timestep h in episode k
(s_h^k, a_h^k, r_h^k)	The tuple representing state-action pair and the corresponding reward at timestep h in episode k
$\mathcal{H}_h^k$	$\{(s_l^j, a_l^j, r_l^j) : \text{if } j < k \text{ then} \, h \leq H, \text{ else if } j = k \text{ then } l \leq h\}$ The history (algorithm trajectory) till timestep h of the episode k.
$\overline{\mathcal{H}}_h^k$	$\mathcal{H}_h^k \bigcup \left\{w^k(l, s, a) : l \in [H], s \in \mathcal{S}, a \in \mathcal{A}\right\}$ The union of the history (algorithm trajectory) til timestep h in episode k and the pseudo-noise of all timesteps in episode k
$n^k(h, s, a)$	$\sum_{l=1}^{k-1} \mathbf{1}\{(s_h^l, a_h^l) = (s, a)\}$ The number of visits to state-action pair (s, a) in timestep h upto episode k
P_{h,s_h^k,a_h^k}	The transition distribution for the state action pair (s_h^k, a_h^k)
R_{h,s_h^k,a_h^k}	The reward distribution for the state action pair (s_h^k, a_h^k)
$P_{h,s,a}$	The transition distribution for the state action pair (s, a) at timestep h
$R_{h,s,a}$	The reward distribution for the state action pair (s, a) at

Table 4.1 (Cont.)

	timestep h
$\hat{P}^k_{h,s^k_h,a^k_h}$	The estimated transition distribution for the state action pair (s^k_h, a^k_h)
$\hat{R}^k_{h,s^k_h,a^k_h}$	The estimated reward distribution for the state action pair (s^k_h, a^k_h)
$\mathcal{M}^k$	The confidence set around the true MDP
w^k_{h,s^k_h,a^k_h}	The pseudo-noise used for exploration
$\tilde{w}^k_{h,s^k_h,s^k_h}$	The independently pseudo-noise sample, conditioned on history till episode $k-1$
$\hat{M}^k$	$(H, \mathcal{S}, \mathcal{A}, \hat{P}^k, \hat{R}^k, s^k_1)$ The estimated MDP without perturbation in data in episode k
$\overline{M}^k$	$(H, \mathcal{S}, \mathcal{A}, \hat{P}^k, \hat{R}^k + w^k, s^k_1)$ The estimated MDP with perturbed data in episode k
V^*_h	The optimal value function under true MDP on the sub-episodes consisting of the timesteps $\{h, \cdots, H\}$
$V^{\pi^k}_h$	The state-value function of π^k evaluated on the true MDP on the sub-episodes consisting of the timesteps $\{h, \cdots, H\}$
$\overline{V}_{h,k}$	The estimated state-value function of π^k evaluated on $\overline{M}^k$ on the sub-episodes consisting of the timesteps $\{h, \cdots, H\}$ in episode k
$\overline{Q}_{h,k}$	The estimated Q-value function of π^k evaluated on $\overline{M}^k$ on the sub-episodes consisting of the timesteps $\{h, \cdots, H\}$ in episode k
$\overline{V}^*_{h,k}$	The estimated state-value function of the optimal policy π^* evaluated on $\overline{M}^k$ on the sub-episodes consisting of the timesteps $\{h, \cdots, H\}$ in episode k
$\underline{M}^k$, $\underline{V}_{1,k}$, $\underline{w}^k_{h,s,a}$	Refer to Definition 4.2
$\tilde{M}^k$	$(H, \mathcal{S}, \mathcal{A}, \hat{P}^k, \hat{R}^k + \tilde{w}^k, s^k_1)$ The estimated MDP with i.i.d. perturbations, refer Definition 4.1
$\tilde{V}_{h,k}$	Optimal state-value function of $\tilde{M}^k$
$\overline{\delta}_{h,k}(s^k_h)$	$V^*_h(s^k_h) - \overline{V}_{h,k}(s^k_h)$
$\underline{\delta}_{h,k}(s^k_h)$	$V^*_h(s^k_h) - \underline{V}_{h,k}(s^k_h)$
$\overline{\delta}^{\pi^k}_{h,k}(s^k_h)$	$\overline{V}_{h,k}(s^k_h) - V^{\pi^k}_{h,k}(s^k_h)$
$\underline{\delta}^{\pi^k}_{h,k}(s^k_h)$	$V^{\pi^k}_{h,k}(s^k_h) - \underline{V}_{h,k}(s^k_h)$
$\mathcal{R}^k_{h,s^k_h,a^k_h}$	$\hat{R}^k_{h,s^k_h,a^k_h} - R_{h,s^k_h,a^k_h}$

Table 4.1 (Cont.)

$\mathcal{P}^k_{h,s^k_h,a^k_h}$	$\langle \hat{P}^k_{h,s^k_h,a^k_h} - P_{h,s^k_h,a^k_h}, V^*_{h+1}\rangle$
C	$\frac{1}{\Phi(-\sqrt{2})}$
L	$\log\left(40SAT/\delta\right)$
$\sqrt{\alpha_k}$	$2\sqrt{H^3 S\log\left(2HSAk\right)\log\left(40SAT/\delta\right)}$
$\sigma_k^2(h,s,a)$	$\frac{\beta_k}{2(n^k(h,s,a)+1)} = \frac{H^3 S\log(2HSAk)}{2(n^k(h,s,a)+1)}$
$\gamma_k(h,s,a)$	$\sqrt{\sigma_k^2(h,s,a)L}$
$\sqrt{e_k(h,s,a)}$	$H\sqrt{\frac{\log(2HSAk)}{n^k(h,s,a)+1}}$
β_k	$H^3 S\log(2HSAk)$
$\mathcal{M}_{\overline{\delta}^{\pi^k}_{h,k}(s^k_h)}$	Refer to Appendix 4.1.2
$\mathcal{M}_{\underline{\delta}^{\pi^k}_{h,k}(s^k_h)}$	Refer to Appendix 4.1.2
$\mathcal{M}^w_1$	Refer to Appendix 4.1.2
$\mathcal{C}_k$	$\left\{\hat{M}^k \in \mathcal{M}^k\right\}$
$\mathcal{E}^w_{h,k}$	$\left\{w^k(h,s^k_h,a^k_h) \le \gamma_k(h,s^k_h,a^k_h)\right\}$
$\mathcal{E}^w_k$	$\left\{\bigcap_{h\in[H]}\left(\mathcal{E}^w_{h,k}\right)\right\}$
$\mathcal{E}^{\overline{Q}}_{h,k}$	$\left\{\lvert(\overline{Q}_{h,k} - Q^*_h)(s,a)\rvert \le H-h+1,\ \forall(s,a)\right\}$
$\mathcal{E}^{\overline{Q}}_k$	$\left\{\bigcap_{h\in[H]}\left(\mathcal{E}^{\overline{Q}}_{h,k}\right)\right\}$
$\overline{\mathcal{E}}_k$	$\left\{\mathcal{E}^w_k \cap \mathcal{E}^{\overline{Q}}_k\right\}$
$\mathcal{E}^{\tilde{w}}_{h,k}$	$\left\{\tilde{w}^k(h,s^k_h,a^k_h) \le \gamma_k(h,s^k_h,a^k_h))\right\}$
$\mathcal{E}^{\tilde{w}}_k$	$\left\{\bigcap_{h\in[H]}\left(\mathcal{E}^{\tilde{w}}_{h,k}\right)\right\}$
$\mathcal{E}^{\tilde{Q}}_{h,k}$	$\left\{\lvert(\tilde{Q}_{h,k} - Q^*_h)(s,a)\rvert \le H-h+1,\ \forall(s,a)\right\}$
$\mathcal{E}^{\tilde{Q}}_k$	$\left\{\bigcap_{h\in[H]}\left(\mathcal{E}^{\tilde{Q}}_{h,k}\right)\right\}$
$\tilde{\mathcal{E}}_k$	$\left\{\mathcal{E}^{\tilde{w}}_k \cap \mathcal{E}^{\tilde{Q}}_k\right\}$
$\mathcal{E}^{\text{th}}_{h,k}$	$\left\{n^k(h,s^k_h,a^k_h) \ge \alpha_k\right\}$
$\mathcal{E}^{\text{th}}_k$	$\left\{\bigcap_{h\in[H]}\mathcal{E}^{\text{th}}_{h,k}\right\}$
$\overline{\mathcal{G}}_k$	$\left\{\mathcal{E}^{\text{th}}_k \cap \overline{\mathcal{E}}_k \cap \mathcal{C}_k\right\}$
$\tilde{\mathcal{G}}_k$	$\left\{\mathcal{E}^{\text{th}}_k \cap \tilde{\mathcal{E}}_k \cap \mathcal{C}_k\right\}$
$\mathcal{G}_k$	$\left\{\overline{\mathcal{G}}_k \cap \tilde{\mathcal{G}}_k\right\}$
$\overline{\mathcal{O}}_{h,k}$	$\left\{\overline{V}_{1,k}(s^k_1) \ge V^*_1(s^k_1)\right\}$

Table 4.1 (Cont.)

$\tilde{\mathcal{O}}_{h,k}$	$\left\{\tilde{V}_{1,k}(s_1^k) \geq V_1^*(s_1^k)\right\}$

4.1.1 Definitions of Synthetic Quantities

In this section we define some synthetic quantities required for analysis.

Definition 4.1 ($\tilde{M}^k$ and $\tilde{V}_{h,k}$)**.** Given history $\mathcal{H}_H^{k-1}$, and $\hat{P}^k$ and $\hat{R}^k$ defined in empirical MDP $\overline{M}^k = (H, \mathcal{S}, \mathcal{A}, \hat{P}^k, \hat{R}^k, s_1^k)$, we define independent Gaussian noise term $\tilde{w}^k(h,s,a)|\mathcal{H}_H^{k-1} \sim \mathcal{N}(0, \sigma_k^2(h,s,a))$, perturbed MDP $\tilde{M}^k = (H, \mathcal{S}, \mathcal{A}, \hat{P}^k, \hat{R}^k + \tilde{w}^k, s_1^k)$, and $\tilde{V}_{h,k}$ the optimal value function of $\tilde{M}^k$.

Notice that $\tilde{w}^k$ can be different from the realized noise term w^k sampled in the Algorithm 2.1. They are two independent samples form the same Gaussian distribution. Therefore, conditioned on the history $\mathcal{H}_H^{k-1}$, $\tilde{M}^k$ has the same marginal distribution as $\overline{M}^k$, but is statistically independent of the policy π^k selected by C-RLSVI.

Definition 4.2 ($\underline{V}_{h,k}$)**.** Similar as in Definition 4.1, given history $\mathcal{H}_H^{k-1}$ and any fixed noise $w_{\text{ptb}}^k \in \mathbb{R}^{HSA}$, we define a perturbed MDP $M_{\text{ptb}}^k = (H, \mathcal{S}, \mathcal{A}, \hat{P}^k, \hat{R}^k + w_{\text{ptb}}^k, s_1^k)$ and $V_{h,k}^{w_{\text{ptb}}^k}$ as its optimal value function.

Let $\underline{w}^k = -\gamma_k$ and $\underline{V}_{h,k}$ denote the optimal value function of MDP $\underline{M}^k = (H, \mathcal{S}, \mathcal{A}, \hat{P}^k, \hat{R}^k + \underline{w}^k, s_1)$. We immediately know that $\underline{w}^k$ is the solution of following optimization program

$$\begin{aligned} &\min_{w_{\text{ptb}}^k \in \mathbb{R}^{HSA}} V_{h,k}^{w_{\text{ptb}}^k}(s_1^k) \\ &s.t. \quad |w_{\text{ptb}}^k(h,s,a)| \leq \gamma_k(h,s,a) \quad \forall\, h,s,a. \end{aligned} \tag{4.1}$$

The reason is that $\underline{M}^k$ has smallest reward function within the set of MDPs satisfying the constraints, thus obtaining smallest optimal value function. We also know that $\underline{V}_{h,k} \leq V_{h,k}^{w_{\text{ptb}}^k}$ for any $|w_{\text{ptb}}^k| \leq \gamma_k$.

Definition 4.3 (Confidence set, restatement of Definition 4.4)**.**

$$\begin{aligned} \mathcal{M}^k = &\left\{(H, \mathcal{S}, \mathcal{A}, P', R', s_1) : \forall (h,s,a), \left| R'_{h,s,a} - R_{h,s,a} + \langle P'_{h,s,a} - P_{h,s,a}, V_{h+1}^* \rangle \right| \right. \\ &\left. \leq \sqrt{e_k(h,s,a)} \right\}, \end{aligned} \tag{4.2}$$

where we set

$$\sqrt{e_k(h,s,a)} = H\sqrt{\frac{\log(2HSAk)}{n^k(h,s,a)+1}}. \tag{4.3}$$

4.1.2 Martingale Difference Sequences

In this section, we give the filtration sets that consists of the history of the algorithm. Later we enumerate the martingale difference sequences needed for the analysis based on these filtration sets. We use the following to denote the history trajectory:

$$\mathcal{H}_h^k := \{(s_l^j, a_l^j, r_l^j) : \text{if } j < k \text{ then}\, l \in [H], \text{ else if } \; j = k \; \text{ then } \; l \in [h]\}, \tag{4.4}$$

$$\overline{\mathcal{H}}_h^k := \mathcal{H}_h^k \bigcup \left\{ w^k(l, s, a) : l \in [H], s \in \mathcal{S}, a \in \mathcal{A} \right\}. \tag{4.5}$$

With $a_h^k = \pi_h^k(s_h^k)$ as the action taken by C-RLSVI following the policy π_h^k and conditioned on the history of the algorithm, the randomness exists only on the next-step transitions. Specifically, we use the following martingale difference sequences with filtration sets $\{\overline{\mathcal{H}}_h^k\}_{h,k}$ for the following martingale difference sequences:

$$\mathcal{M}_{\overline{\delta}_{h,k}^{\pi^k}(s_h^k)} = \mathbf{1}\{\mathcal{G}_k\} \left[\mathbb{E}\left[\overline{\delta}_{h+1,k}^{\pi^k}(s') \right] - \overline{\delta}_{h+1,k}^{\pi^k}(s_{h+1}^k) \right], \tag{4.6}$$

$$\mathcal{M}_{\underline{\delta}_{h,k}^{\pi^k}(s_h^k)} = \mathbf{1}\{\mathcal{G}_k\} \left[\mathbb{E}\left[\underline{\delta}_{h+1,k}^{\pi^k}(s') \right] - \underline{\delta}_{h+1,k}^{\pi^k}(s_{h+1}^k) \right], \tag{4.7}$$

where the expectation is over next state s' due to the transition distribution: P_{h,s_h^k,a_h^k}. We use with the filtration sets $\{\mathcal{H}_H^{k-1}\}_k$ for the following martingale difference sequence

$$\mathcal{M}_{h,k}^{w} = \mathbf{1}\{\mathcal{G}_k\} \left[\mathbb{E}_{\tilde{w}}\left[\tilde{V}_{h,k}(s_h^k) \right] - \overline{V}_{h,k}(s_h^k) \right]. \tag{4.8}$$

4.2 TECHNICAL PREREQUISITES

The primary technical content is in the Section 4.3. In this section, we present three technical prerequisites: (i) the total probability for the unperturbed estimated $\hat{M}^k$ to fall outside a confidence set is bounded; (ii) the estimated value function $\overline{V}_{h,k}$ (defined as the value function of π^k in MDP $\overline{M}^k$) is an upper bound of the optimal value function V_h^* with at least a constant probability at every episode; (iii) the clipping procedure ensures that $\overline{V}_{h,k}$ is bounded with high probability[1].

Notations To avoid cluttering of mathematical expressions, we abridge our notations to exclude the reference to (s, a) when it is clear from the context. Concise notations are

[1]We drop/hide constants by appropriate use of $\gtrsim, \lesssim, \simeq$ in our mathematical relations. All the detailed analyses can be found in our appendix.

used in the later analysis: $R_{h,s_h^k,a_h^k} \to R_h^k$, $\hat{R}^k_{h,s_h^k,a_h^k} \to \hat{R}_h^k$, $P_{h,s_h^k,a_h^k} \to P_h^k$, $\hat{P}^k_{h,s_h^k,a_h^k} \to \hat{P}_h^k$, $n^k(h, s_h^k, a_h^k) \to n^k(h)$, $w^k_{h,s_h^k,a_h^k} \to w_h^k$.

High probability confidence set In Definition 4.4, $\mathcal{M}^k$ represents a set of MDPs, such that the total estimation error with respect to the true MDP is bounded.

Definition 4.4 (Confidence set)**.**

$$\mathcal{M}^k = \left\{ (H, \mathcal{S}, \mathcal{A}, P', R', s_1) : \forall (h, s, a), \left| R'_{h,s,a} - R_{h,s,a} + \langle P'_{h,s,a} - P_{h,s,a}, V^*_{h+1} \rangle \right| \right. \\ \left. \leq \sqrt{e_k(h, s, a)} \right\}, \tag{4.9}$$

where we set

$$\sqrt{e_k(h, s, a)} = H \sqrt{\frac{\log(2HSAk)}{n^k(h, s, a) + 1}}. \tag{4.10}$$

Through an application of Hoeffding's inequality [2, 8], it is shown via Lemma 4.1 that the empirical MDP does not often fall outside confidence set $\mathcal{M}^k$. This ensures exploitation, i.e., the algorithm's confidence in the estimates for a certain (h, s, a) tuple grows as it visits that tuple many numbers of times.

Lemma 4.1. $\sum_{k=1}^{\infty} \mathbb{P}\left(\hat{M}^k \notin \mathcal{M}^k \right) \leq 2006HSA$.

Bounded Q-function estimates It is important to note the pseudo-noise used by C-RLSVI has both exploratory (optimism) behavior and corrupting effect on the estimated value function. Since the Gaussian noise is unbounded, the clipping procedure (lines 15-20 in Algorithm 2.1) avoids propagation of unreasonable estimates of the value function, especially for the tuples (h, s, a) which have low visit counts. This saves from low rewarding states to be misidentified as high rewarding ones (or vice-versa). Intuitively, the clipping threshold α_k is set such that the noise variance ($\sigma_k(h, s, a) = \frac{\beta_k}{2(n^k(h,s,a)+1)}$) drops below a numerical constant and hence limiting the effect of noise on the estimated value functions. This idea is stated in Lemma 4.2, where we claim the estimated Q-value function is bounded for all (h, s, a).

Lemma 4.2 ((Informal) Bound on the estimated Q-value function)**.** Define $\overline{Q}_k$ as the Q-value function of π^k (as in Algorithm 2.1) in perturbed MDP $\overline{M}^k = (H, \mathcal{S}, \mathcal{A}, \hat{P}^k, \hat{R}^k + w^k, s_1^k)$, where $w^k(h, s, a) \sim \mathcal{N}(0, \sigma_k^2(h, s, a))$. Then under some good event, we have $|(\overline{Q}_{h,k} - Q_h^*)(s, a)| \leq H - h + 1$.

See Appendix A.2 for the precise definition of good event and a full proof. Lemma 4.2 is striking since it suggests that randomized value function needs to be clipped only for constant (i.e. independent of T) number of times to be well-behaved.

Optimism The event when none of the rounds in episode k need to be clipped is denoted by $\mathcal{E}_k^{\text{th}} := \{\cap_{h\in[H]}(n^k(h) \geq \alpha_k)\}$. Due to the randomness in the environment, there is a possibility that a learning algorithm may get stuck on "bad" states, i.e. not visiting the "good" (h, s, a) enough or it grossly underestimates the value function of some states and as result avoid transitioning to those state. Effective exploration is required to avoid these scenarios. To enable correction of faulty estimates, most RL exploration algorithms maintain optimistic estimates almost surely. However, when using randomized value functions, C-RLSVI does not always guarantee optimism. In Lemma 4.3, we show that C-RLSVI samples an optimistic value function estimate with at least a constant probability for any (k). We emphasize that such difference is fundamental.

Lemma 4.3. If $\hat{M}^k \in \mathcal{M}^k$ and the event $\mathcal{E}_k^{\text{th}}$ holds, then

$$\mathbb{P}\left(\overline{V}_{1,k}(s_1^k) \geq V_1^*(s_1^k) \,|\, \mathcal{H}_H^{k-1}\right) \geq \Phi(-\sqrt{2}), \tag{4.11}$$

where $\Phi(\cdot)$ is the CDF of N(0,1) distribution, $\hat{M}^k$ is the estimated MDP, $\mathcal{M}^k$ is the confidence set defined in Eq (4.10), and $\mathcal{H}_H^{k-1}$ is all the history of the past observations made by C-RLSVI till the end of the episode $k-1$.

Lemma 4.3 follows [14], and we reproduce the proof in Appendix A.1 for completeness.

Now, we are in a position to simplify the regret expression as

$$\begin{aligned}\text{Reg}(K) \leq &\sum_{k=1}^{K} \mathbf{1}\{\mathcal{E}_k^{\text{th}}\}\left(V_1^* - V_1^{\pi^k}\right)(s_1^k) + \underbrace{\sum_{k=1}^{K} \mathbf{1}\{\mathcal{E}_k^{\text{th}\,\complement}\}\left(V_1^* - V_1^{\pi^k}\right)(s_1^k)}_{\text{Warm-up term}} \\ &+ H \underbrace{\mathbb{P}(\hat{M}^k \notin \mathcal{M}^k)}_{\text{due to Lemma 4.1}},\end{aligned} \tag{4.12}$$

where we use Lemma 4.1 to show that for any (h, s, a), the edge case that the estimated MDP lies outside the confidence set is a transient term (independent of T) and the warm-up term due to clipping is also independent on T (see Section 4.3.1 for details). Armed with the necessary tools, over the next few subsections we sketch the proof of our main results.

4.3 REGRET DECOMPOSITION

In this section we give full proof of our main result Theorem 4.1 which is a formal version of Theorem 3.1. We will give a high-level sketch proof before jumping into the details of individual parts in Sections 4.3.2 and 4.3.3.

Theorem 4.1. C-RLSVI enjoys the following high probability regret upper bound, with probability at least $1-\delta$,

$$\operatorname{Reg}(K) = \tilde{\mathrm{O}}\left(H^2 S\sqrt{AT} + H^2 S^3 A + H^4 S^2 A\right). \tag{4.13}$$

We first decompose the regret expression into several terms and show bounds for each of the individual terms separately. Consider

$$\operatorname{Reg}(K) \ = \ \sum_{k=1}^{K} \mathbf{1}\{\mathcal{C}_k\}\left(V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)\right) + \sum_{k=1}^{K} \underbrace{\mathbf{1}\{\{\mathcal{C}_k\}^{\complement}\}\left(V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)\right)}_{(1)}. \tag{4.14}$$

This in turn gives:

$$\begin{aligned}\operatorname{Reg}(K) \ =\leq \ & \sum_{k=1}^{K} \mathbf{1}\{\mathcal{E}_k^{\mathrm{th}} \cap \mathcal{C}_k\}\left(\underbrace{V_1^*(s_1^k) - \overline{V}_{1,k}(s_1^k)}_{(2)} + \underbrace{\overline{V}_{1,k}(s_1^k) - V_1^{\pi^k}(s_1^k)}_{(3)}\right) \\ & + \ \sum_{k=1}^{K} \underbrace{\mathbf{1}\{\mathcal{E}_k^{\mathrm{th}\,\complement} \cap \mathcal{C}_k\}\left(V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)\right)}_{(4)} + 2006 H^2 S A. \end{aligned} \tag{4.15}$$

Term (1) is upper bounded due to Lemma A.2 and the fact that $V_h^*(s_h^k) - V_h^{\pi^k}(s_h^k) \leq H, \forall\, k \in [K]$. Term (2), additive inverse of *optimism*, is called *pessimism* [15] and is further decomposed in Lemma 4.8 and Lemma 4.9. Term (3) is a measure of how well estimated MDP tracks the true MDP and is called *estimation error*. It is discussed further by Lemma 4.5, Lemma 4.6 and finally decomposed in Lemma 4.7. Term (4) is the regret due to those episodes in which one or more rounds had clipped estimated Q function and is upper bounded by Lemma 4.4. Since the number of unclipped episodes does not depend on T, the regret can be considered to be due to *burn-in* or *warm-up*. We start with the results that decompose the terms in Eq (4.15) and later aggregate them back to complete the proof of Theorem 4.1.

4.3.1 Bound on the Warm-up Term

Lemma 4.4 (Bound on the warm-up term)**.**

$$\sum_{k=1}^{K} \mathbf{1}\{\mathcal{E}_k^{\mathrm{th}\,\complement} \cap \mathcal{C}_k\} \left(V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)\right) = \tilde{\mathrm{O}}(H^4 S^2 A). \tag{4.16}$$

Proof.

$$\begin{aligned}
&\sum_{k=1}^{K} \mathbf{1}\{\mathcal{E}_k^{\mathrm{th}\,\complement} \cap \mathcal{C}_k\} \left(V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k)\right) \\
&\overset{a}{\leq} \sum_{k=1}^{K} \mathbf{1}\{\mathcal{E}_k^{\mathrm{th}\,\complement}\}\, H \\
&= H \sum_{k=1}^{K} \mathbf{1}\{\underset{h\in[H]}{\cup} n^k(h,s,a) \leq \alpha_k,\ \forall (h,s,a) = (h, s_h^k, a_h^k)\} \\
&\leq H \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbf{1}\{n^k(h,s,a) \leq \alpha_k,\ \forall (h,s,a) = (h, s_h^k, a_h^k)\} \\
&\overset{b}{\leq} H \sum_{a\in\mathcal{A}} \sum_{s\in\mathcal{S}} \sum_{h=1}^{H} \alpha_k \\
&\leq 4H^4 S^2 A \log(2HSAK) \log\left(40SAT/\delta\right) \\
&= \tilde{\mathrm{O}}(H^4 S^2 A).
\end{aligned} \tag{4.17}$$

Step (a) is due to $V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \leq H$. Step (b) is by substituting the value of α_k followed by upper bound for all $4H^3 S \log(2HSAK) \log\left(40SAT/\delta\right)$. QED.

4.3.2 Bound on the Estimation Term

Lemma 4.5 decomposes the deviation term between the Q-value function and its estimate, and the proof relies on Lemma 4.6. This result is extensively used in our analysis. For the purpose of the results in this subsection, we assume the episode index k is fixed and hence dropped from the notation in both the lemma statements and their proofs when it is clear.

Lemma 4.5. With probability at least $1-\delta/4$, for any h, k, s_h, a_h, it follows that

$$\begin{aligned}
&\mathbf{1}\{\mathcal{G}_k\}\left[\overline{Q}_h(s_h,a_h)-Q_h^\pi(s_h,a_h)\right] \\
&\le \mathbf{1}\{\mathcal{G}_k\}\left(\mathcal{P}_{h,s_h,a_h}+\mathcal{R}_{h,s_h,a_h}+w_{h,s_h,a_h}+\overline{\delta}_{h+1}^\pi(s_{h+1})+\mathcal{M}_{\overline{\delta}_h^\pi(s_h)}+4H\sqrt{\frac{SL}{n(h,s_h,a_h)+1}}\right).
\end{aligned} \tag{4.18}$$

Proof. Here the action at any period h is due to the policy of the algorithm, therefore, $a_h=\pi(s_h)$. Applying the Bellman equation, we have the following

$$\begin{aligned}
&\mathbf{1}\{\mathcal{G}_k\}\left[\overline{Q}_h(s_h,a_h)-Q_h^\pi(s_h,a_h)\right] \\
&=\mathbf{1}\{\mathcal{G}_k\}\left[\langle\hat{P}_{h,s_h,a_h},\overline{V}_{h+1}\rangle-\langle P_{h,s_h,a_h},V_{h+1}^\pi\rangle+\hat{R}_{h,s_h,a_h}-R_{h,s_h,a_h}+w_{h,s_h,a_h}\right] \\
&=\mathbf{1}\{\mathcal{G}_k\}\Big[\langle\hat{P}_{h,s_h,a_h},\overline{V}_{h+1}\rangle-\langle P_{h,s_h,a_h},V_{h+1}^\pi\rangle+\hat{R}_{h,s_h,a_h}-R_{h,s_h,a_h}+w_{h,s_h,a_h} \\
&\quad+\langle\hat{P}_{h,s_h,a_h}-P_{h,s_h,a_h},V_{h+1}^*\rangle-\langle\hat{P}_{h,s_h,a_h}-P_{h,s_h,a_h},V_{h+1}^*\rangle\Big] \\
&=\mathbf{1}\{\mathcal{G}_k\}\left[\mathcal{P}_{h,s_h,a_h}+\mathcal{R}_{h,s_h,a_h}+w_{h,s_h,a_h}+\langle P_{h,s_h,a_h},\overline{V}_{h+1}-V_{h+1}^\pi\rangle\right] \\
&\quad+\mathbf{1}\{\mathcal{G}_k\}\langle\hat{P}_{h,s_h,a_h}-P_{h,s_h,a_h},\overline{V}_{h+1}-V_{h+1}^*\rangle \\
&\overset{a}{\le}\mathbf{1}\{\mathcal{G}_k\}\left[\mathcal{P}_{h,s_h,a_h}+\mathcal{R}_{h,s_h,a_h}+w_{h,s_h,a_h}+\overline{\delta}_{h+1}^\pi(s_{h+1})+\mathcal{M}_{\overline{\delta}_h^\pi(s_h)}+4H\sqrt{\frac{SL}{n(h,s_h,a_h)+1}}\right],
\end{aligned} \tag{4.19}$$

where step (a) follows from Lemma 4.6 and by adding and subtracting $\overline{V}_{h+1}^\pi(s_{h+1})-V_{h+1}^\pi(s_{h+1})$ to create $\mathcal{M}_{\overline{\delta}_h^\pi(s_h)}$. QED.

Lemma 4.6. With probability at least $1-\delta/4$, for any h, k, s_h, a_h, it follows that

$$\mathbf{1}\{\mathcal{G}_k\}\langle\hat{P}_{h,s_h,a_h}-P_{h,s_h,a_h},V_{h+1}^*-\overline{V}_{h+1}\rangle\le 4H\sqrt{\frac{SL}{n(h,s_h,a_h)+1}}. \tag{4.20}$$

Proof. Firstly, applying the Cauchy-Schwarz inequality, we get

$$\mathbf{1}\{\mathcal{G}_k\}\langle\hat{P}_{h,s_h,a_h}-P_{h,s_h,a_h},V_{h+1}^*-\overline{V}_{h+1}\rangle\le\|\hat{P}_{h,s_h,a_h}-P_{h,s_h,a_h}\|_1\|\mathbf{1}\{\mathcal{G}_k\}(V_{h+1}^*-\overline{V}_{h+1})\|_\infty. \tag{4.21}$$

Since Lemma A.4 implies $\|\mathbf{1}\{\mathcal{G}_k\}(V_{h+1}^*-\overline{V}_{h+1})\|_\infty\le H$, it suffices to bound $\|\hat{P}_{h,s_h,a_h}-$

$P_{h,s_h,a_h}\|_1$. Note that for any vector $v \in \mathbb{R}^S$, we have

$$\|v\|_1 = \sup_{u \in \{-1,+1\}^S} u^\top v. \tag{4.22}$$

Hence, we will prove the concentration for $u^\top(\hat{P}_{h,s_h,a_h} - P_{h,s_h,a_h})$. For clarity, we will add the superscript k in the following proof.

If the visiting time $n^k(h, s_h^k, a_h^k) = 0$, we know that $\|\hat{P}^k_{h,s_h^k,a_h^k} - P_{h,s_h^k,a_h^k}\|_1 \leq 4H\sqrt{SL}$, which means the final bound holds. Now we consider the case that $n^k(h, s_h^k, a_h^k) \geq 1$. For any fixed k, h, s_h^k, a_h^k and $u \in \{-1,+1\}^S$, applying Hoeffding's inequality, with probability at least $1-\delta'$, we have

$$u^\top \left(\frac{1}{n^k(h, s_h^k, a_h^k)} \sum_{l=1}^{k-1} \mathbf{1}\{(s_h^l, a_h^l, s_{h+1}^l) = (s_h^k, a_h^k, \cdot)\} - P_{h,s_h^k,a_h^k}(\cdot) \right) \leq 2\sqrt{\frac{\log(2/\delta')}{2n^k(h, s_h^k, a_h^k)}} \tag{4.23}$$

This is because $u^\top \left(\frac{1}{n^k(h,s_h^k,a_h^k)} \sum_{l=1}^{k-1} \mathbf{1}\{(s_h^l, a_h^l, s_{h+1}^l) = (s_h^k, a_h^k, \cdot)\} \right)$ is the average of i.i.d. random variables $u^\top \mathbf{e}_{s'}$ with bounded range $[-1, 1]$.

By triangle inequality, we have

$$\begin{aligned}
&\left| \hat{P}^k_{h,s_h^k,a_h^k}(\cdot) - \frac{1}{n^k(h, s_h^k, a_h^k)} \sum_{l=1}^{k-1} \mathbf{1}\{(s_h^l, a_h^l, s_{h+1}^l) = (s_h^k, a_h^k, \cdot)\} \right| \\
&= \frac{1}{n^k(h, s_h^k, a_h^k)(n^k(h, s_h^k, a_h^k) + 1)} \sum_{l=1}^{k-1} \mathbf{1}\{(s_h^l, a_h^l, s_{h+1}^l) = (s_h^k, a_h^k, \cdot)\} \\
&\leq \frac{1}{n^k(h, s_h^k, a_h^k)},
\end{aligned} \tag{4.24}$$

where the last step is by noticing visiting $(s_h^l, a_h^l, s_{h+1}^l) = (s_h^k, a_h^k, \cdot)$ implies visiting (h, s_h^k, a_h^k).

Therefore, we get

$$u^\top \left(\hat{P}^k_{h,s_h^k,a_h^k} - P^k_{h,s_h^k,a_h^k} \right) \leq 3\sqrt{\frac{\log(2/\delta')}{2n^k(h, s_h^k, a_h^k)}}. \tag{4.25}$$

Finally, union bounding over all k, h, s_h^k, a_h^k and $u \in \{-1,+1\}^S$ and set $\delta = \delta'/(2^S SAT)$, we get

$$u^\top \left(\hat{P}^k_{h,s^k_h,a^k_h} - P^k_{h,s^k_h,a^k_h} \right) \le 3\sqrt{\frac{SL}{n^k(h,s^k_h,a^k_h)}} \le 4\sqrt{\frac{SL}{n^k(h,s^k_h,a^k_h)+1}}. \tag{4.26}$$

This implies $\|\hat{P}^k_{h,s^k_h,a^k_h} - P_{h,s^k_h,a^k_h}\|_1 \le 4\sqrt{\frac{SL}{n^k(h,s^k_h,a^k_h)+1}}$, which completes the proof. QED.

The following Lemma 4.7 is the V function version of its Q function version counterpart in Lemma 4.5. It is applied in the proof of final regret decomposition in Theorem 4.1.

Lemma 4.7. With probability at least $1-\delta/4$, for any h, k, s_h, a_h, the following decomposition holds

$$\begin{aligned}&\mathbf{1}\{\mathcal{G}_k\}\left[\overline{V}_h(s_h) - V^\pi_h(s_h)\right]\\ &\le \mathbf{1}\{\mathcal{G}_k\}\left(\mathcal{P}_{h,s_h,a_h} + \mathcal{R}_{h,s_h,a_h} + w_{h,s_h,a_h} + \overline{\delta}^\pi_{h+1}(s_{h+1}) + \mathcal{M}_{\overline{\delta}^\pi_h(s_h)} + 4H\sqrt{\frac{SL}{n(h,s_h,a_h)+1}}\right).\end{aligned} \tag{4.27}$$

Proof. With a_h as the action taken by the algorithm $\pi(s_h)$, it follows that $\overline{V}_h(s_h) = \overline{Q}_h(s_h,a_h)$ and $V^\pi_h(s_h) = Q^\pi_h(s_h,a_h)$. Thus, the proof follows by a direction application of Lemma 4.5. QED.

4.3.3 Bound on Pessimism Term

In this section, we will upper bound the pessimism term with the help of the probability of being optimistic and the bound on the estimation term. The approach generally follows Lemma G.4 of [15]. The difference here is that we also provide a bound for $V^*_1(s^k_1) - \underline{V}_{1,k}(s^k_1)$. This difference enable us to get stronger bounds in the tabular setting as compared to [15]. The pessimism term will be decomposed to the two estimation terms $\overline{V}_{1,k}(s^k_1) - V^{\pi^k}_1(s^k_1)$ and $V^{\pi^k}_1(s^k_1) - \underline{V}_{1,k}(s^k_1)$, and martingale difference term $\mathcal{M}^w_{1,k}$.

Lemma 4.8. For any k, the following decomposition holds,

$$\begin{aligned}\mathbf{1}\{\mathcal{G}_k\}\left(V^*_1(s^k_1) - \overline{V}_{1,k}(s^k_1)\right) &\le \mathbf{1}\{\mathcal{G}_k\}\left(V^*_1(s^k_1) - \underline{V}_{1,k}(s^k_1)\right)\\ &\le C\mathbf{1}\{\mathcal{G}_k\}\left(\overline{V}_{1,k}(s^k_1) - V^{\pi^k}_1(s^k_1) + V^{\pi^k}_1(s^k_1) - \underline{V}_{1,k}(s^k_1) + \mathcal{M}^w_{h,k}\right),\end{aligned} \tag{4.28}$$

where $\mathbf{1}\{\mathcal{G}_k\}\left[V^{\pi^k}_1(s^k_1) - \underline{V}_{1,k}(s^k_1)\right]$ will be further bounded in Lemma 4.9.

Proof. For the purpose of analysis we use two quantities $\tilde{V}_{1,k}(s_1^k)$ and $\underline{V}_{1,k}(s_1^k)$, which are formally stated in the Definitions 4.1 and 4.2 respectively. Thus we can define the event $\tilde{\mathcal{O}}_{1,k} \stackrel{def}{=} \left\{\tilde{V}_{1,k}(s_1^k) \geq V_1^*(s_1^k)\right\}$. For simplicity of exposition, we skip showing dependence on k in the following when it is clear. Also we would use the event $\mathcal{G}_k$ as a concise way of representing the intersection event $\mathcal{G}_k = \overline{\mathcal{G}}_k \cap \tilde{\mathcal{G}}_k = \mathcal{E}_k^{\text{th}} \cap \mathcal{C}_k \cap \overline{\mathcal{E}}_k \cap \tilde{\mathcal{E}}_k$.

By definition, there is no clipping under event $\mathcal{G}_k$, so the equivalence between model-free view and model-based view tells us that $\overline{V}$ would be equal to the optimal value function under $\overline{M}$. Then by Definition 4.1, we know that $\overline{V}_1(s_1)$ and $\tilde{V}_1(s_1)$ are identically distributed conditioned on last round history $\mathcal{H}_H^{k-1}$ and under $\mathcal{G}_k$. From Definition 4.2, under event $\mathcal{G}_k$, it also follows that $\underline{V}_1(s_1) \leq \overline{V}_1(s_1)$ and $\underline{V}_1(s_1) \leq \tilde{V}_1(s_1)$.

Since $\underline{V}_1(s_1) \leq \overline{V}_1(s_1)$ under event $\mathcal{G}_k$, we get

$$\mathbf{1}\{\mathcal{G}_k\}\left[V_1^*(s_1) - \overline{V}_1(s_1)\right] \leq \mathbf{1}\{\mathcal{G}_k\}\left[V_1^*(s_1) - \underline{V}_1(s_1)\right]. \tag{4.29}$$

We also introduce notation $\mathbb{E}_{\tilde{w}}[\cdot]$ to denote the expectation over the pseudo-noise $\tilde{w}$ (recall that $\tilde{w}$ discussed in Definition 4.1). Under event $\tilde{\mathcal{O}}_h$, we have $\tilde{V}_1(s_1) \geq V_1^*(s_1)$. Since $V_1^*(s_1)$ does not depend on $\tilde{w}$, we get $V_1^*(s_1) \leq \mathbb{E}_{\tilde{w}|\tilde{\mathcal{O}}_h}\left[\tilde{V}_1(s_1)\right]$. Using the similar argument for $\underline{V}_1(s_1)$, we know that $\underline{V}_1(s_1) = \mathbb{E}_{\tilde{w}|\tilde{\mathcal{O}}_h}\left[\underline{V}_1(s_1)\right]$. Subtracting this equality from the inequality $V_1^*(s_1) \leq \mathbb{E}_{\tilde{w}|\tilde{\mathcal{O}}_h}\left[\tilde{V}_1(s_1)\right]$, it follows that

$$V_1^*(s_1) - \underline{V}_1(s_1) \leq \mathbb{E}_{\tilde{w}|\tilde{\mathcal{O}}_h}\left[\tilde{V}_1(s_1) - \underline{V}_1(s_1)\right]. \tag{4.30}$$

Therefore, we have

$$\mathbf{1}\{\mathcal{G}_k\}\left[V_1^*(s_1) - \underline{V}_1(s_1)\right] \leq \mathbf{1}\{\mathcal{G}_k\}\mathbb{E}_{\tilde{w}|\tilde{\mathcal{O}}_h}\left[\tilde{V}_1(s_1) - \underline{V}_1(s_1)\right]. \tag{4.31}$$

From the law of total expectation, we can write

$$\mathbb{E}_{\tilde{w}}\left[\tilde{V}_1(s_1) - \underline{V}_1(s_1)\right] = \mathbb{P}(\tilde{\mathcal{O}}_h)\mathbb{E}_{\tilde{w}|\tilde{\mathcal{O}}_h}\left[\tilde{V}_1(s_1) - \underline{V}_1(s_1)\right] + \mathbb{P}(\tilde{\mathcal{O}}_h^{\complement})\mathbb{E}_{\tilde{w}|\tilde{\mathcal{O}}_h^{\complement}}\left[\tilde{V}_1(s_1) - \underline{V}_1(s_1)\right]. \tag{4.32}$$

Since $\tilde{V}_1(s_1) - \underline{V}_1(s_1) \geq 0$ under event $\mathcal{G}_k$, relaxing the second term to 0 and rearranging the above equation yields

$$\mathbf{1}\{\mathcal{G}_k\}\mathbb{E}_{\tilde{w}|\tilde{\mathcal{O}}_h}\left[\tilde{V}_1(s_1) - \underline{V}_1(s_1)\right] \leq \frac{1}{\mathbb{P}(\tilde{\mathcal{O}}_h)}\mathbf{1}\{\mathcal{G}_k\}\mathbb{E}_{\tilde{w}}\left[\tilde{V}_1(s_1) - \underline{V}_1(s_1)\right]. \tag{4.33}$$

Noticing $\tilde{V}$ is an independent sample of $\overline{V}$ we can invoke Lemma A.1 for $\tilde{V}$, it follows that $\mathbb{P}(\tilde{\mathcal{O}}_h) \geq \Phi(-\sqrt{2})$. Set $C = \frac{1}{\Phi(-\sqrt{2})}$ and consider

$$\begin{aligned}\mathbf{1}\{\mathcal{G}_k\}\mathbb{E}_{\tilde{w}}\left[\tilde{V}_1(s_1) - \underline{V}_1(s_1)\right] &\leq C\,\underbrace{\mathbf{1}\{\mathcal{G}_k\}\left(\mathbb{E}_{\tilde{w}}\left[\tilde{V}_1(s_1)\right] - \overline{V}_1(s_1)\right)}_{(1)} \\ &\quad + C\mathbf{1}\{\mathcal{G}_k\}\underbrace{\left(\overline{V}_1(s_1) - \underline{V}_1(s_1)\right)}_{(2)},\end{aligned} \tag{4.34}$$

where the inequality is due to $\tilde{w}$ is independent of $\underline{V}_1(s_1)$.

Since $\overline{V}_1(s_1)$ and $\tilde{V}_1(s_1)$ are identically distributed from the definition, we will later show term (1) $\mathbf{1}\{\mathcal{G}_k\}\left[\mathbb{E}_{\tilde{w}}\left[\tilde{V}_1(s_1)\right] - \overline{V}_1(s_1)\right] := \mathcal{M}_h^w$ is a martingale difference sequence in Lemma 4.12. Term (2) can be further decomposed as

$$\overline{V}_1(s_1) - \underline{V}_1(s_1) = \underbrace{\overline{V}_1(s_1) - V_1^\pi(s_1)}_{(3)} + \underbrace{V_1^\pi(s_1) - \underline{V}_1(s_1)}_{(4)}. \tag{4.35}$$

Term (3) in the above equation is same as *estimation* term in Lemma 4.7. For term (4), to make it clearer, we will show a bound separately in Lemma (4.9).

Combining the last three equations, gives us that

$$\begin{aligned}&\mathbf{1}\{\mathcal{G}_k\}\mathbb{E}_{\tilde{w}|\tilde{\mathcal{O}}_h}\left[\tilde{V}_1(s_1) - \underline{V}_1(s_1)\right] \\ &\leq C\mathbf{1}\{\mathcal{G}_k\}\left(V_1^{\pi^k}(s_1^k) - \underline{V}_{1,k}(s_1^k) + \overline{V}_{1,k}(s_1^k) - V_1^{\pi^k}(s_1^k) + \mathcal{M}_{h,k}^w\right).\end{aligned} \tag{4.36}$$

This completes the proof. QED.

In Lemma 4.9, we provide a missing piece in Lemma 4.8. It will be applied when we do the the regret decomposition of major term in Theorem 4.1.

Lemma 4.9. With probability at least $1-\delta/4$, for any s, h, s_h^k, a_h^k, the following decomposition holds with the intersection event $\mathcal{G}_k$

$$\begin{aligned}&\mathbf{1}\{\mathcal{G}_k\}\left[V_h^{\pi^k}(s_h^k) - \underline{V}_{h,k}(s_h^k)\right] \leq \\ &\mathbf{1}\{\mathcal{G}_k\}\left(-\mathcal{P}^k_{h,s_h^k,a_h^k} - \mathcal{R}^k_{h,s_h^k,a_h^k} - \underline{w}^k_{h,s_h^k,a_h^k} + \underline{\delta}^{\pi^k}_{h+1,k}(s^k_{h+1}) + \mathcal{M}_{\underline{\delta}^{\pi^k}_{h,k}(s_h^k)} + 4H\sqrt{\frac{SL}{n^k(h,s_h^k,a_h^k)+1}}\right)\end{aligned} \tag{4.37}$$

Proof. We continue to show how to bound term (4) in Lemma 4.8 and we will also drop the superscript k here.

Noticing that a_h as the action chosen by the algorithm $\pi(s_h)$, we have $V_h^\pi(s_h) = Q_h^\pi(s_h, a_h)$. By definition of value function $\underline{V}_h(s_h) = \max_{a \in \mathcal{A}} \underline{Q}_h(s_h, a)$. This gives $\underline{Q}_h(s_h, a_h) \le \underline{V}_h(s_h)$. Hence,

$$V_h^\pi(s_h) - \underline{V}_h(s_h) = Q_h^\pi(s_h, a_h) - \underline{V}_h(s_h) \le Q_h^\pi(s_h, a_h) - \underline{Q}_h(s_h, a_h). \tag{4.38}$$

Since the decomposition and techniques in Lemma 4.5 and Lemma 4.6 only utilize the property that $\overline{Q}_h$ is a value function (optimal value function under MDP $\bar{M}$ when no clipping happens), we can directly get another version for instance $\underline{Q}_h$ by shifting from $\overline{M}$ to MDP $\underline{M}$. Also noticing that we flip the sign of $V_h^\pi(s_h) - \underline{V}_h(s_h)$, therefore, we obtain the following decomposition for the term (4) in Lemma 4.8

$$\begin{aligned} &\mathbf{1}\{\mathcal{G}_k\} \left[V_h^\pi(s_h) - \underline{V}_h(s_h)\right] \\ \le\ &\mathbf{1}\{\mathcal{G}_k\} \left(-\mathcal{P}_{h,s_h,a_h} - \mathcal{R}_{h,s_h,a_h} - \underline{w}_{h,s_h,a_h} + \underline{\delta}_{h+1}^\pi(s_{h+1}) + \mathcal{M}_{\underline{\delta}_h^\pi(s_h)} + 4H\sqrt{\frac{SL}{n(h, s_h, a_h) + 1}} \right). \end{aligned} \tag{4.39}$$

QED.

4.3.4 Final Bound on Theorem 4.1

Armed with all the supporting lemmas, we present the remaining proof of Theorem 4.1.

Proof. Recall that in the regret decomposition Eq (4.15), it remains to bound

$$\sum_{k=1}^{K} \mathbf{1}\{\mathcal{E}_k^{\text{th}} \cap \mathcal{C}_k\} \left(V_1^*(s_1^k) - \overline{V}_{1,k}(s_1^k) + \overline{V}_{h,k}(s_1^k) - V_1^{\pi^k}(s_1^k)\right). \tag{4.40}$$

From Lemma A.5, we have that with probability at least $1 - \delta/8$, for any k, $\mathcal{C}_k$ holds implies $\overline{\mathcal{E}}_k$ holds. Similarly, we can have a version Lemma A.5 for $\tilde{w}$. This implies with probability at least $1 - \delta/8$, for any k, $\mathcal{C}_k$ holds implies $\tilde{\mathcal{E}}_k$ holds.

Noticing that $\mathcal{G}_k = \mathcal{E}_k^{\text{th}} \cap \overline{\mathcal{E}}_k \cap \tilde{\mathcal{E}}_k \cap \mathcal{C}_k$, with probability at least $1 - \delta/4$, we have

$$
\begin{aligned}
&\sum_{k=1}^{K} \mathbf{1}\{\mathcal{E}_k^{\text{th}} \cap \mathcal{C}_k\} \left(V_1^*(s_1^k) - \overline{V}_{1,k}(s_1^k) + \overline{V}_{1,k}(s_1^k) - V_1^{\pi^k}(s_1^k) \right) \\
= &\sum_{k=1}^{K} \mathbf{1}\{\mathcal{G}_k\} \left(V_1^*(s_1^k) - \overline{V}_{1,k}(s_1^k) + \overline{V}_{1,k}(s_1^k) - V_1^{\pi^k}(s_1^k) \right). \qquad (4.41)
\end{aligned}
$$

Again, we would skip notation dependence on k when it is clear. For each episode k, it suffices to bound

$$
\begin{aligned}
&\mathbf{1}\{\mathcal{G}_k\} \left(V_1^*(s_1) - \overline{V}_1(s_1) + \overline{V}_1(s_1) - V_1^{\pi}(s_1) \right) \\
\leq\ &\mathbf{1}\{\mathcal{G}_k\} \left[V_1^*(s_1) - \overline{V}_1(s_1) \right] + \mathbf{1}\{\mathcal{G}_k\} \left[\overline{V}_1(s_1) - V_1^{\pi}(s_1) \right] \\
=\ &\mathbf{1}\{\mathcal{G}_k\} \overline{\delta}_1(s_1) + \mathbf{1}\{\mathcal{G}_k\} \overline{\delta}_1^{\pi}(s_1). \qquad (4.42)
\end{aligned}
$$

We first use Lemma 4.8 to relax the first term in the above equation. Applying the result of Lemma 4.8 gives us the following

$$
\begin{aligned}
&\mathbf{1}\{\mathcal{G}_k\} \overline{\delta}_1(s_1) \\
=\ &\mathbf{1}\{\mathcal{G}_k\} \left[V_1^*(s_1) - \overline{V}_1(s_1) \right] \\
\leq\ &C \mathbf{1}\{\mathcal{G}_k\} \left(V_1^{\pi}(s_1) - \underline{V}_1(s_1) + \overline{V}_1(s_1) - V_1^{\pi}(s_1) + \mathcal{M}_1^w \right) \\
=\ &C \mathbf{1}\{\mathcal{G}_k\} \left(\overline{\delta}_1^{\pi}(s_1) + \underline{\delta}_1^{\pi}(s_1) + \mathcal{M}_1^w \right). \qquad (4.43)
\end{aligned}
$$

Combining the last two equations, we get

$$
\begin{aligned}
&\mathbf{1}\{\mathcal{G}_k\} \left(V_1^*(s_1) - \overline{V}_1(s_1) + \overline{V}_1(s_1) - V_1^{\pi}(s_1) \right) \\
\leq\ &(C+1) \mathbf{1}\{\mathcal{G}_k\} \overline{\delta}_1^{\pi}(s_1) + C \mathbf{1}\{\mathcal{G}_k\} \left(\mathcal{M}_1^w + \underline{\delta}_1^{\pi}(s_1) \right). \qquad (4.44)
\end{aligned}
$$

We will bound the first and the second terms of the above equation one by one. In the sequence, we always consider the case that Lemma 4.9 and Lemma 4.7 hold. Therefore, the following holds with probability at least $1 - \delta/4 - \delta/4 = 1 - \delta/2$.

For the $\underline{\delta}_1^\pi(s_1)$ term, applying the result of Lemma 4.9 yields

$$\begin{aligned}
&\mathbf{1}\{\mathcal{G}_k\}\underline{\delta}_1^\pi(s_1) \\
&= \mathbf{1}\{\mathcal{G}_k\}\left[V_1^\pi(s_1) - \underline{V}_1(s_1)\right] \\
&\leq \mathbf{1}\{\mathcal{G}_k\}\left(\left|\mathcal{P}_{1,s_1,a_1} + \mathcal{R}_{1,s_1,a_1}\right| + \left|\underline{w}_{1,s_1,a_1}\right| + \underline{\delta}_2^\pi(s_2) + \mathcal{M}_{\underline{\delta}_1^\pi(s_1)} + 4H\sqrt{\frac{SL}{n(1,s_1,a_1)+1}}\right).
\end{aligned} \tag{4.45}$$

For the $\overline{\delta}_1^\pi(s_1)$ term, applying Lemma 4.7 yields

$$\begin{aligned}
&\mathbf{1}\{\mathcal{G}_k\}\overline{\delta}_1^\pi(s_1) \\
&= \mathbf{1}\{\mathcal{G}_k\}\left[\overline{V}_1(s_1) - V_1^\pi(s_1)\right] \\
&\leq \mathbf{1}\{\mathcal{G}_k\}\left(\left|\mathcal{P}_{1,s_1,a_1} + \mathcal{R}_{1,s_1,a_1}\right| + w_{1,s_1,a_1} + \overline{\delta}_2^\pi(s_2) + \mathcal{M}_{\overline{\delta}_1^\pi(s_1)} + 4H\sqrt{\frac{SL}{n(1,s_1,a_1)+1}}\right).
\end{aligned} \tag{4.46}$$

Using the above two relations, gives us, with probability at least $1-\delta/2$,

$$\begin{aligned}
&\mathbf{1}\{\mathcal{G}_k\}\left(V_1^*(s_1) - \overline{V}_1(s_1) + \overline{V}_1(s_1) - V_1^\pi(s_1)\right) \\
&\leq (C+1)\mathbf{1}\{\mathcal{G}_k\}\overline{\delta}_1^\pi(s_1) + C\mathbf{1}\{\mathcal{G}_k\}\left(\mathcal{M}_1^w + \underline{\delta}_1^\pi(s_1)\right) \\
&\leq C\mathbf{1}\{\mathcal{G}_k\}\left(\left|\mathcal{P}_{1,s_1,a_1} + \mathcal{R}_{1,s_1,a_1}\right| + \left|\underline{w}_{1,s_1,a_1}\right| + \underline{\delta}_2^\pi(s_2) + \mathcal{M}_{\underline{\delta}_1^\pi(s_1)} + 4H\sqrt{\frac{SL}{n(1,s_1,a_1)+1}}\right) \\
&\quad + (C+1)\mathbf{1}\{\mathcal{G}_k\}\left(\left|\mathcal{P}_{1,s_1,a_1} + \mathcal{R}_{1,s_1,a_1}\right| + w_{1,s_1,a_1} + \overline{\delta}_2^\pi(s_2) + \mathcal{M}_{\overline{\delta}_1^\pi(s_1)} + 4H\sqrt{\frac{SL}{n(1,s_1,a_1)+1}}\right) \\
&\quad + C\mathbf{1}\{\mathcal{G}_k\}\mathcal{M}_1^w \\
&= C\mathbf{1}\{\mathcal{G}_k\}\underline{\delta}_2^\pi(s_2) + (C+1)\mathbf{1}\{\mathcal{G}_k\}\overline{\delta}_2^\pi(s_2) + C\mathbf{1}\{\mathcal{G}_k\}\mathcal{M}_1^w \\
&\quad + C\mathbf{1}\{\mathcal{G}_k\}\left(\left|\mathcal{P}_{1,s_1,a_1} + \mathcal{R}_{1,s_1,a_1}\right| + \left|\underline{w}_{1,s_1,a_1}\right| + \mathcal{M}_{\underline{\delta}_1^\pi(s_1)} + 4H\sqrt{\frac{SL}{n(1,s_1,a_1)+1}}\right) \\
&\quad + (C+1)\mathbf{1}\{\mathcal{G}_k\}\left(\left|\mathcal{P}_{1,s_1,a_1} + \mathcal{R}_{1,s_1,a_1}\right| + w_{1,s_1,a_1} + \mathcal{M}_{\overline{\delta}_1^\pi(s_1)} + 4H\sqrt{\frac{SL}{n(1,s_1,a_1)+1}}\right).
\end{aligned} \tag{4.47}$$

Keep unrolling the above inequality to timestep H and noticing $\underline{\delta}^{\pi}_{H+1}(s_{H+1}) = \overline{\delta}^{\pi}_{H+1}(s_{H+1}) = 0$ and $\mathcal{M}^{w}_{H+1} = 0$ yields that with probability at least $1 - \delta/2$,

$$\begin{aligned}
&\mathbf{1}\{\mathcal{G}_k\}\left[V_1^*(s_1) - V_1^{\pi}(s_1)\right] \\
&\leq C\sum_{h=1}^{H}\mathbf{1}\{\mathcal{G}_k\}\left(\left|\mathcal{P}_{h,s_h,a_h} + \mathcal{R}_{h,s_h,a_h}\right| + \left|\underline{w}_{h,s_h,a_h}\right| + \mathcal{M}_{\underline{\delta}^{\pi}_h(s_h)} + \mathcal{M}^{w}_h + 4H\sqrt{\frac{SL}{n(h,s_h,a_h)+1}}\right) \\
&\quad + (C+1)\sum_{h=1}^{H}\mathbf{1}\{\mathcal{G}_k\}\left(\left|\mathcal{P}_{h,s_h,a_h} + \mathcal{R}_{h,s_h,a_h}\right| + w_{h,s_h,a_h} + \mathcal{M}_{\overline{\delta}^{\pi}_h(s_h)} + 4H\sqrt{\frac{SL}{n(h,s_h,a_h)+1}}\right).
\end{aligned} \tag{4.48}$$

It suffices to bound each individual term in the above inequality and we will take sum over k outside.

Lemma 4.11 gives us the bound on transition function and reward function

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbf{1}\{\mathcal{G}_k\}\left|\mathcal{P}^k_{h,s^k_h,a^k_h} + \mathcal{R}^k_{h,s^k_h,a^k_h}\right| = \tilde{\mathrm{O}}(\sqrt{H^3SAT}). \tag{4.49}$$

Following the steps in Lemma 4.11, we also get the bound

$$\sum_{k=1}^{K}\sum_{h=1}^{H} H\sqrt{\frac{SL}{n^k(h,s^k_h,a^k_h)+1}} = \tilde{\mathrm{O}}\left(H^{3/2}S\sqrt{AT}\right). \tag{4.50}$$

Lemma 4.12 bounds the martingale difference sequences. Replacing δ by δ' in Lemma 4.12 gives us that with probability at least $1 - \delta'$,

$$\left|\sum_{k=1}^{K}\mathbf{1}\{\mathcal{G}_k\}\sum_{h=1}^{H}\mathcal{M}_{\underline{\delta}^{\pi^k}_{h,k}(s^k_h)}\right| = \tilde{\mathrm{O}}(H\sqrt{T}), \tag{4.51}$$

$$\left|\sum_{k=1}^{K}\mathbf{1}\{\mathcal{G}_k\}\sum_{h=1}^{H}\mathcal{M}_{\overline{\delta}^{\pi^k}_{h,k}(s^k_h)}\right| = \tilde{\mathrm{O}}(H\sqrt{T}), \tag{4.52}$$

$$\left|\sum_{k=1}^{K}\mathbf{1}\{\mathcal{G}_k\}\sum_{h=1}^{H}\mathcal{M}^{w}_{h,k}\right| = \tilde{\mathrm{O}}(H\sqrt{T}). \tag{4.53}$$

For the noise term, we first notice that under event $\mathcal{G}_k$, $w^k_{h,s^k_h,\pi^k(s^k_h)}$ can be upper bounded by $\left|\underline{w}^k_{h,s^k_h,\pi^k(s^k_h)}\right|$. Applying Lemma 4.10 and (replacing δ by δ' in Lemma 4.10) gives us, with probability at least $1-2\delta'$

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbf{1}\{\mathcal{G}_k\}w^k_{h,s^k_h,\pi^k(s^k_h)} = \tilde{\mathrm{O}}(H^2S\sqrt{AT}) \tag{4.54}$$

and

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbf{1}\{\mathcal{G}_k\}\left|\underline{w}^k_{h,s^k_h,\pi^k(s^k_h)}\right| = \tilde{\mathrm{O}}(H^2S\sqrt{AT}). \tag{4.55}$$

Putting all these pieces together and setting $\delta' = \delta/12$ yields that with probability at least $1-\delta$,

$$\sum_{k=1}^{K}\mathbf{1}\{\mathcal{E}^{\mathrm{th}}_k \cap \mathcal{C}_k\}\left(V^*_1(s_1) - \overline{V}_{1,k}(s_1) + \overline{V}_{h,k}(s_1) - V^{\pi^k}_1(s_1)\right) = \tilde{\mathrm{O}}\left(H^2S\sqrt{AT} + H^2S^3A\right), \tag{4.56}$$

which together with the *warm-up* regret bounds in Lemma 4.4 completes the proof of Theorem 4.1. QED.

4.4 SUPPORTING LEMMAS

4.4.1 Bound on the Noise Term

Lemma 4.10. With $\underline{w}^k_{h,s^k_h,a^k_h}$ as defined in Definition 4.2 and $a^k_h = \pi^k(s^k_h)$, the following bound holds:

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbf{1}\{\mathcal{G}_k\}\left|\underline{w}^k_{h,s^k_h,\pi^k(s^k_h)}\right| = \tilde{\mathrm{O}}\left(H^2S\sqrt{AT}\right). \tag{4.57}$$

Proof. We have:

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H}\left|\underline{w}^k_{h,s^k_h,\pi^k(s^k_h)}\right| &= \sqrt{\frac{\beta_k L}{2}}\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{1}{n^k(h,s^k_h,a^k_h)+1}} \\ &= \sqrt{\frac{\beta_k L}{2}}\sum_{h,s,a}\sum_{n=1}^{n^K(h,s,a)}\sqrt{\frac{1}{n}}. \end{aligned}\tag{4.58}$$

Upper bounding by integration followed by an application of Cauchy-Schwarz inequality gives:

$$\begin{aligned}\sum_{h,s,a}\sum_{n=1}^{n^K(h,s,a)}\sqrt{\frac{1}{n}} &\leq \sum_{h,s,a}\int_0^{n^K(h,s,a)}\sqrt{\frac{1}{x}}dx \\ &= 2\sum_{h,s,a}\sqrt{n^K(h,s,a)} \\ &\leq 2\sqrt{HSA\sum_{h,s,a}n^K(h,s,a)} \\ &= \mathrm{O}\left(\sqrt{HSAT}\right). \end{aligned} \tag{4.59}$$

This leads to the bound of $\mathrm{O}\left(\sqrt{\beta_k L}\sqrt{HSAT}\right) = \tilde{\mathrm{O}}\left(H^2 S\sqrt{AT}\right)$. QED.

4.4.2 Bound on Estimation Error

Lemma 4.11. For $a_h^k = \pi^k(s_h^k)$, the following bound holds

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbf{1}\{\mathcal{G}_k\}\left|\hat{R}^k_{h,s_h^k,a_h^k} - R_{h,s_h^k,a_h^k} + \langle \hat{P}^k_{h,s_h^k,a_h^k} - P_{h,s_h^k,a_h^k}, V^*_{h+1}\rangle\right| = \tilde{\mathrm{O}}\left(H^{3/2}\sqrt{SAT}\right). \tag{4.60}$$

Proof. Under the event $\mathcal{G}_k$, the estimated MDP $\hat{M}^k$ lies in the confidence set defined in Notation Table of Section 4.1. Hence

$$\left|\hat{R}^k_{h,s_h^k,a_h^k} - R_{h,s_h^k,a_h^k} + \langle \hat{P}^k_{h,s_h^k,a_h^k} - P_{h,s_h^k,a_h^k}, V^*_{h+1}\rangle\right| \leq \sqrt{e_k(h,s_h^k,a_h^k)}, \tag{4.61}$$

where $\sqrt{e_k(h,s_h^k,a_h^k)} = H\sqrt{\frac{\log(2HSAk)}{n^k(h,s_h^k,a_h^k)+1}}$.

We bound the denominator as

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{1}{n^k(h,s_h^k,a_h^k)+1}} \leq \sum_{h,s,a}\sum_{n=1}^{n^K(h,s,a)}\sqrt{\frac{1}{n}} \leq \sum_{h,s,a}\int_0^{n^K(h,s,a)}\sqrt{\frac{1}{x}}dx \\ &\leq 2\sum_{h,s,a}\sqrt{n^K(h,s,a)} \overset{a}{\leq} 2\sqrt{HSA\sum_{h,s,a}n^K(h,s,a)} = \mathrm{O}(\sqrt{HSAT}),\end{aligned} \tag{4.62}$$

where step (a) follows Cauchy-Schwarz inequality.

Therefore we get

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{e_k(h, s_h^k, a_h^k)} = H\tilde{\mathrm{O}}\left(\sqrt{HSAT}\right) = \tilde{\mathrm{O}}\left(H^{3/2}\sqrt{SAT}\right). \tag{4.63}$$

QED.

In the next section, we show concentration bounds on the martingale difference sequences created in the analysis. For each of the martingale difference sequences, we identify the filtration set ($\sigma-$algebra) and the terms of the martingales and apply Hoeffding's inequality.

4.4.3 Bounds on Martingale Difference Sequences

Lemma 4.12. The following martingale difference summations enjoy the specified upper bounds with probability at least $1-\delta$,

$$\left|\sum_{k=1}^{K}\mathbf{1}\{\mathcal{G}_k\}\sum_{h=1}^{H}\mathcal{M}_{\underline{\delta}_{h,k}^{\pi^k}(s_h^k)}\right| = \tilde{\mathrm{O}}(H\sqrt{T}), \tag{4.64}$$

$$\left|\sum_{k=1}^{K}\mathbf{1}\{\mathcal{G}_k\}\sum_{h=1}^{H}\mathcal{M}_{\overline{\delta}_{h,k}^{\pi^k}(s_h^k)}\right| = \tilde{\mathrm{O}}(H\sqrt{T}) \tag{4.65}$$

$$\left|\sum_{k=1}^{K}\mathbf{1}\{\mathcal{G}_k\}\sum_{h=1}^{H}\mathcal{M}_{h,k}^{w}\right| = \tilde{\mathrm{O}}(H\sqrt{T}). \tag{4.66}$$

Here $\mathcal{M}_{\underline{\delta}_{h,k}^{\pi^k}(s_h^k)}, \mathcal{M}_{\overline{\delta}_{h,k}^{\pi^k}(s_h^k)}$ are considered under filtration $\overline{\mathcal{H}}_h^k$, while $\mathcal{M}_{h,k}^{w}$ is considered under filtration $\mathcal{H}_H^{k-1}$. Noticing the definition of martingale difference sequences, we can also drop $\mathbf{1}\{\mathcal{G}_k\}$ in the lemma statement.

Proof. This proof has two parts. We show (i) above are summations of martingale difference sequences and (ii) these summations concentrate under the event $\mathcal{G}_k$ due to Azuma-Hoeffding inequality [34]. We only present the proof for $\left\{\mathcal{M}_{\overline{\delta}_{h,k}^{\pi^k}(s_h^k)}\right\}$ and $\{\mathcal{M}_{h,k}^{w}\}$, and another one follow like-wise.

We first consider $\mathcal{M}_{\overline{\delta}_{h,k}^{\pi^k}(s_h^k)}$ term. Given the filtration set $\overline{\mathcal{H}}_h^k$, we observe that

$$\mathbb{E}\left[\mathbf{1}\{\mathcal{G}_k\}\overline{\delta}_{h+1,k}^{\pi^k}(s_{h+1}^k)\middle|\overline{\mathcal{H}}_h^k\right] = \mathbb{E}\left[\mathbf{1}\{\mathcal{G}_k\}\mathbb{E}_{s'\sim P^k_{h,s_h^k,\pi_h^k(s_h^k)}}\left[\overline{\delta}_{h+1,k}^{\pi^k}(s')\right]\middle|\overline{\mathcal{H}}_h^k\right]. \tag{4.67}$$

This is because the randomness is due to the random transitions of the algorithms when conditioning on $\overline{\mathcal{H}}_h^k$. Thus we have $\mathbb{E}\left[\mathcal{M}_{\overline{\delta}_{h,k}^{\pi^k}(s_h^k)}\middle|\overline{\mathcal{H}}_h^k\right] = 0$ and $\left\{\mathcal{M}_{\overline{\delta}_{h,k}^{\pi^k}(s_h^k)}\right\}$ is indeed a martingale difference on the filtration set $\overline{\mathcal{H}}_h^k$.

Under event $\mathcal{G}_k$, we also have $\overline{\delta}_{h+1,k}^{\pi^k}(s_{h+1}^k) = \overline{V}_{h+1,k}(s_{h+1}^k) - V_{h+1,k}^{\pi^k}(s_{h+1}^k) \leq 2H$. Applying Azuma-Hoeffding inequality (e.g. [3]), for any fixed $K' \in [K]$ and $H' \in [H]$, we have with probability at least $1-\delta'$,

$$\left|\sum_{k=1}^{K'}\sum_{h=1}^{H'}\mathcal{M}_{\overline{\delta}_{h,k}^{\pi^k}(s_h^k)}\right| \leq H\sqrt{4T\log\left(\frac{2T}{\delta'}\right)} = \tilde{\mathrm{O}}\left(H\sqrt{T}\right). \tag{4.68}$$

Union bounding over all K' and H', we know the following holds for any $K' \in [K]$ and $H' \in [H]$ with probability at least $1-\delta'$

$$\left|\sum_{k=1}^{K'}\sum_{h=1}^{H'}\mathcal{M}_{\overline{\delta}_{h,k}^{\pi^k}(s_h^k)}\right| \leq H\sqrt{4T\log\left(\frac{2T}{\delta'}\right)} = \tilde{\mathrm{O}}\left(H\sqrt{T}\right). \tag{4.69}$$

Then we consider $\mathcal{M}_{h,k}^w$ term. Given filtration $\mathcal{H}_H^{k-1}$ and under event $\mathcal{G}_k$, we know that $\tilde{V}_{h,k}$ has identical distribution as $\overline{V}_{h,k}$. Therefore, for any state s, we have

$$\mathbb{E}\left[\mathbf{1}\{\mathcal{G}_k\}\tilde{V}_{h,k}(s)\middle|\mathcal{H}_H^{k-1}\right] = \mathbb{E}\left[\mathbf{1}\{\mathcal{G}_k\}\overline{V}_{h,k}(s)\middle|\mathcal{H}_H^{k-1}\right]. \tag{4.70}$$

Besides, from the definition of $\mathbb{E}_{\tilde{w}}$ and $\tilde{w}$ is the only randomness given $\mathcal{H}_H^{k-1}$, we have that for any state s,

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}\{\mathcal{G}_k\}\mathbb{E}_{\tilde{w}}\left[\tilde{V}_{h,k}(s)\right]\middle|\mathcal{H}_H^{k-1}\right] \\
&= \mathbb{E}\left[\mathbf{1}\{\mathcal{G}_k\}\mathbb{E}_{\tilde{w}}\left[\tilde{V}_{h,k}(s)\middle|\mathcal{H}_H^{k-1}\right]\middle|\mathcal{H}_H^{k-1}\right] \\
&= \mathbb{E}\left[\mathbf{1}\{\mathcal{G}_k\}\mathbb{E}\left[\tilde{V}_{h,k}(s)\middle|\mathcal{H}_H^{k-1}\right]\middle|\mathcal{H}_H^{k-1}\right] \\
&= \mathbb{E}\left[\mathbf{1}\{\mathcal{G}_k\}\tilde{V}_{h,k}(s)\middle|\mathcal{H}_H^{k-1}\right].
\end{aligned} \tag{4.71}$$

Combining these two equations and setting $s = s_h^k$, we have $\mathbb{E}\left[\mathcal{M}_{h,k}^w\middle|\mathcal{H}_H^{k-1}\right]$. Therefore the sequence $\{\mathcal{M}_{h,k}^{w}\}$ is indeed a martingale difference.

Under event $\mathcal{G}_k$, we also have $\left|\mathbb{E}_w\left[\overline{V}_{h,k}(s_h^k)\right] - \overline{V}_{h,k}(s_h^k)\right| \leq 2H$ from Lemma A.4. Applying from Azum-Hoeffding inequality (e.g. [3]) and similar union bounding argument above, for any $K' \in [K]$ and $H' \in [H]$, with probability at least $1-\delta'$, we have

$$\left|\sum_{k=1}^{K'}\sum_{h=1}^{H'}\mathcal{M}_{h,k}^{w}\right| \leq H\sqrt{4T\log\left(\frac{2T}{\delta'}\right)} = \tilde{\mathrm{O}}\left(H\sqrt{T}\right). \tag{4.72}$$

The remaining results as in the lemma statement is proved like-wise. Finally let $\delta' = \delta/3$ and uniform bounding over these 3 martingale difference sequences completes the proof. QED.

CHAPTER 5: CONCLUSIONS

In this work, we provide a sharper regret analysis for a variant of RLSVI and advance our understanding of TS-based algorithms. Compared with the lower bound, the looseness mainly comes from the magnitude of the noise term in random perturbation, which is delicately tuned for obtaining optimism with constant probability. Specifically, the magnitude of β_k is $\tilde{\mathrm{O}}(\sqrt{HS})$ larger than sharpest bonus term [3], which leads to an additional $\tilde{\mathrm{O}}(\sqrt{HS})$ dependence. Naively using a smaller noise term will affect optimism, thus breaking the analysis. Another obstacle to obtaining $\tilde{\mathrm{O}}(\sqrt{S})$ results is attributed to the proof of Lemma 4.6. Regarding the dependence on the horizon, one $\mathrm{O}(\sqrt{H})$ improvement may be achieved by applying the law of total variance type of analysis in [3]. The future direction of this work includes bridging the gap in the regret bounds and the extension of our results to the time-homogeneous setting.

APPENDIX A: OPTIMISM AND EVENT CONCENTRATION

In this Appendix, we give the missing proofs from the Section 4.2.

A.1 PROOF OF OPTIMISM

Optimism is required since it is used for bounding the pessimism term in the regret bound calculation. We only care about the probability of a timestep 1 in an episode k being optimistic when the event $\mathcal{E}_k^{\text{th}}$ holds. The following proof of Lemma A.1 follows Lemma 4 in [14].

Lemma A.1 (Optimism with a constant probability, restatement of Lemma 4.3)**.** Conditioned on history $\mathcal{H}_H^{k-1}$, if the estimated MDP $\hat{M}^k \in \mathcal{M}^k$ and the event $\mathcal{E}_k^{\text{th}}$ holds, then

$$\mathbb{P}\left(\overline{V}_{1,k}(s_1^k) \geq V_1^*(s_1^k) \,|\, \mathcal{H}_H^{k-1}\right) \geq \Phi(-\sqrt{2}), \tag{A.1}$$

where $\mathcal{M}^k$ is the confidence set defined in Eq (4.3).

Proof. The analysis is valid for any episode k and hence we skip k from the subsequent notations in this proof. When the event $\mathcal{E}_k^{\text{th}}$ holds then none of the rounds in the episode k are clipped. By the equivalence between two views of RLSVI, we know that π is the optimal policy under $\overline{M}$. Let $\overline{V}_1^{\pi^*}(s_1)$ represents the value function of optimal policy π^* (of the true MDP) in MDP $\overline{M}$. Notice that $\overline{V}_1^{\pi^*}(s_1)$ is not the optimal value function under MDP $\overline{M}$. As the first step, we show $\overline{V}_1^{\pi^*}(s_1) \geq V_1^*(s_1)$ with probability at least $\Phi(-\sqrt{2})$.

Let $(s_h, \cdots, s_H)$ be the random sequence of states drawn by the policy π^* (optimal policy under true MDP M in the estimated MDP $\overline{M}$ and $a_h = \pi^*(s_h)$, then applying Bellman equation, we have

$$\begin{aligned}
&\overline{V}_h^{\pi^*}(s_h) - V_h^*(s_h) \\
&= \hat{R}_{h,s_h,a_h} + w_{h,s_h,a_h} + \langle \hat{P}_{h,s_h,a_h}\,,\, \overline{V}_{h+1}^{\pi^*} \rangle - R_{h,s_h,a_h} - \langle P_{h,s_h,a_h}\,,\, V_{h+1}^* \rangle \\
&= \hat{R}_{h,s_h,a_h} - R_{h,s_h,a_h} + \langle \hat{P}_{h,s_h,a_h}\,,\, \overline{V}_{h+1}^{\pi^*} - V_{h+1}^* \rangle + \langle \hat{P}_{h,s_h,a_h} - P_{h,s_h,a_h}\,,\, V_{h+1}^* \rangle + w_{h,s_h,a_h} \\
&= \hat{R}_{h,s_h,a_h} - R_{h,s_h,a_h} + \langle \hat{P}_{h,s_h,a_h} - P_{h,s_h,a_h}\,,\, V_{h+1}^* \rangle + \mathbb{E}_{\pi^*,\overline{M}}\left[\overline{V}_{h+1}^{\pi^*}(s_{h+1}) - V_{h+1}^*(s_{h+1})\right] + w_{h,s_h,a_h}
\end{aligned} \tag{A.2}$$

Further:

$$
\begin{aligned}
&\overline{V}_h^{\pi^*}(s_h) - V_h^*(s_h) \\
&\overset{a}{=} \mathbb{E}_{\pi^*,\overline{M}}\left[\sum_{i=h}^{H} w_{i,s_i,a_i} + \mathcal{R}_{i,s_i,a_i} + \mathcal{P}_{i,s_i,a_i}\right] \\
&\overset{b}{\geq} (H-h+1)\mathbb{E}_{\pi^*,\overline{M}}\left[\frac{1}{H-h+1}\sum_{i=h}^{H}\left(w_{i,s_i,a_i} - \sqrt{e(i,s_i,a_i)}\right)\right].
\end{aligned}
\tag{A.3}
$$

Step (a) follows by expanding the recursion. Step (b) is from the definition of the confidence sets (Definition 4.3). Now substitute $h = 1$ in the inequality above and define $d(i,s) = \frac{1}{H}\mathbb{P}(s_i = s)$ for any $s \in \mathcal{S}$. The inequality can be written as

$$
\begin{aligned}
&\frac{1}{H}\left(\overline{V}_1^{\pi^*}(s_1) - V_1^*(s_1)\right) \\
&\geq \sum_{s\in\mathcal{S},1\leq i\leq H} d(i,s)\left(w_{i,s_i,a_i} - \sqrt{e(i,s_i,a_i)}\right) \\
&\geq \left(\sum_{s\in\mathcal{S},1\leq i\leq H} d(i,s)w_{i,s_i,a_i}\right) - \sqrt{HS}\sqrt{\sum_{s\in\mathcal{S},1\leq i\leq H} d(i,s)^2 e(i,s_i,a_i)} \\
&:= X(w).
\end{aligned}
\tag{A.4}
$$

The second inequality is due to Cauchy-Shwartz. Since $d(i,s)w_{i,s_i,a_i} \sim N\left(0, d(i,s)^2\sigma_k^2\right)$ with $\sigma_k^2 = \frac{HSe(i,s,a)}{2}$, we get

$$
X(w) \sim N\left(-\sqrt{HS\sum_{s\in\mathcal{S},1\leq i\leq H} d(i,s)^2 e(i,s_i,a_i)},\ HS\sum_{s\in\mathcal{S},1\leq i\leq H} d(i,s)^2 e(i,s_i,a_i)/2\right). \tag{A.5}
$$

Upon converting to standard Gaussian distribution it follows that

$$
\mathbb{P}\left(X(W) \geq 0\right) = \mathbb{P}\left(N(0,1) \geq \sqrt{2}\right) = \Phi(-\sqrt{2}). \tag{A.6}
$$

Therefore, $\mathbb{P}\left(\overline{V}_1^{\pi^*}(s_1) - V_1^*(s_1) \geq 0\right) \geq \Phi(-\sqrt{2})$. Since the policy π is optimal under the MDP $\overline{M}$ when $\mathcal{E}_k^{\text{th}}$ holds and $\overline{V}$ is defined as the value function of π under MDP $\overline{M}$, we know that $\overline{V}_1(s_1) \geq \overline{V}_1^{\pi^*}(s_1)$. This implies $\mathbb{P}\left(\overline{V}_1(s_1) \geq V_1^*(s_1)\right) \geq \Phi(-\sqrt{2})$ and thus completes the proof. QED.

A.2 CONCENTRATION OF EVENTS

Lemma A.2 (Bound on the confident set, restatement of Lemma 4.1)**.** $\sum_{k=1}^{\infty} \mathbb{P}(\mathcal{C}_k) = \sum_{k=1}^{\infty} \mathbb{P}(\hat{M}^k \notin \mathcal{M}^k) \leq 2006HSA$.

Proof. Similar as [14], we construct "stack of rewards" as in [29]. For every tuple $z = (h, s, a)$, we generate two i.i.d sequences of random variables $r_{z,n} \sim R_{h,s,a}$ and $s_{z,n} \sim P_{h,s,a}(\cdot)$. Here $r_{(h,s,a),n}$ and $s_{(h,s,a),n}$ denote the reward and state transition generated from the nth time action a is played in state s, timestep h. Set

$$Y_{z,n} = r_{z,n} + V^*_{h+1}(s_{z,n}) \qquad n \in \mathbb{N}. \tag{A.7}$$

They are i.i.d, with $Y_{z,n} \in [0, H]$ since $\|V^*_{h+1}\|_\infty \leq H - 1$, and satisfies

$$\mathbb{E}[Y_{z,n}] = R_{h,s,a} + \langle P_{h,s,a}\,,\, V^*_{h+1}\rangle. \tag{A.8}$$

Now let $n = n^k(h, s, a)$. First consider the case $n \geq 0$. From the definition of empirical MDP, we have

$$\hat{R}^k_{h,s,a} + \langle \hat{P}^k_{h,s,a}\,,\, V^*_{h+1}\rangle = \frac{1}{n+1}\sum_{i=1}^{n} Y_{(h,s,a),i} = \frac{1}{n}\sum_{i=1}^{n} Y_{(h,s,a),i} - \frac{1}{n(n+1)}\sum_{i=1}^{n} Y_{(h,s,a),i}. \tag{A.9}$$

Applying triangle inequality gives us

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{R}^k_{h,s,a} - R_{h,s,a} + \langle \hat{P}^k_{h,s,a} - P_{h,s,a}\,,\, V^*_{h+1}\rangle\right| \geq H\sqrt{\frac{\log(2/\delta_n)}{2n}}\right) \\
=\, &\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_{(h,s,a),i} - R_{h,s,a} - \langle P_{h,s,a}\,,\, V^*_{h+1}\rangle - \frac{1}{n(n+1)}\sum_{i=1}^{n} Y_{(h,s,a),i}\right| \geq H\sqrt{\frac{\log(2/\delta_n)}{2n}}\right) \\
\leq\, &\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_{(h,s,a),i} - R_{h,s,a} - \langle P_{h,s,a}\,,\, V^*_{h+1}\rangle\right| + \left|\frac{1}{n(n+1)}\sum_{i=1}^{n} Y_{(h,s,a),i}\right| \geq H\sqrt{\frac{\log(2/\delta_n)}{2n}}\right) \\
\leq\, &\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_{(h,s,a),i} - R_{h,s,a} - \langle P_{h,s,a}\,,\, V^*_{h+1}\rangle\right| + \frac{1}{n+1}H \geq H\sqrt{\frac{\log(2/\delta_n)}{2n}}\right) \\
=\, &\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_{(h,s,a),i} - R_{h,s,a} - \langle P_{h,s,a}\,,\, V^*_{h+1}\rangle\right| \geq H\sqrt{\frac{\log(2/\delta_n)}{2n}} - \frac{1}{n+1}H\right).
\end{aligned} \tag{A.10}$$

When $n \geq 126$, we have

$$
\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_{(h,s,a),i} - R_{h,s,a} - \langle P_{h,s,a}\,,\, V_{h+1}^*\rangle\right| \geq H\sqrt{\frac{\log(2/\delta_n)}{2n}} - \frac{1}{n+1}H\right) \\
&\leq \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_{(h,s,a),i} - R_{h,s,a} - \langle P_{h,s,a}\,,\, V_{h+1}^*\rangle\right| \geq H\sqrt{\frac{\log(2/\delta_n)}{2n}} - \frac{H}{8}\sqrt{\frac{\log(2/\delta_n)}{2n}}\right) \\
&= \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_{(h,s,a),i} - R_{h,s,a} - \langle P_{h,s,a}\,,\, V_{h+1}^*\rangle\right| \geq \frac{7H}{8}\sqrt{\frac{\log(2/\delta_n)}{2n}}\right). \qquad \text{(A.11)}
\end{aligned}
$$

By Hoeffding's inequality, for any $\delta_n \in (0,1)$,

$$
\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_{(h,s,a),i} - R_{h,s,a} - \langle P_{h,s,a}\,,\, V_{h+1}^*\rangle\right| \geq \frac{7H}{8}\sqrt{\frac{\log(2/\delta_n)}{2n}}\right) \leq \sqrt[64]{2^{15}\delta_n^{49}}. \qquad \text{(A.12)}
$$

For $\delta_n = \frac{1}{HSAn^2}$, a union bound over HSA values of $z = (h,s,a)$ and all possible $n \geq 127$ yields

$$
\begin{aligned}
&\mathbb{P}\left(\bigcup_{h\in[H],s\in[S],a\in[A],n\geq 126}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} Y_{(h,s,a),i} - R_{h,s,a} - \langle P_{h,s,a}\,,\, V_{h+1}^*\rangle\right| \geq H\sqrt{\frac{\log(2/\delta_n)}{2n}}\right\}\right) \\
&\leq \mathbb{P}\left(\bigcup_{h\in[H],s\in[S],a\in[A],n\geq 126}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} Y_{(h,s,a),i} - R_{h,s,a} - \langle P_{h,s,a}\,,\, V_{h+1}^*\rangle\right| \geq \frac{7H}{8}\sqrt{\frac{\log(2/\delta_n)}{2n}}\right\}\right) \\
&\leq \sum_{s=1}^{S}\sum_{a=1}^{A}\sum_{h=1}^{H}\sum_{n=126}^{\infty}\sqrt[64]{2^{15}\left(\frac{1}{HSAn^2}\right)^{49}} \\
&= (HSA)\sum_{n=126}^{\infty}\sqrt[64]{2^{15}\left(\frac{1}{HSAn^2}\right)^{49}} \\
&\leq 2(HSA)^{15/64}\left(\int_{x=1}^{\infty}\left(\frac{1}{x}\right)^{49/32}dx + 1\right) \\
&\leq 6(HSA)^{15/64}. \qquad \text{(A.13)}
\end{aligned}
$$

For $1 \leq n \leq 125$, we instead have

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_{(h,s,a),i} - R_{h,s,a} - \langle P_{h,s,a}\,,\, V_{h+1}^{*}\rangle\right| \geq H\sqrt{\frac{\log(2/\delta_n)}{2n}} - \frac{1}{n+1}H\right) \\
\leq\ &\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_{(h,s,a),i} - R_{h,s,a} - \langle P_{h,s,a}\,,\, V_{h+1}^{*}\rangle\right| \geq H\sqrt{\frac{\log(2/\delta_n)}{2n}} - \frac{H}{2}\sqrt{\frac{\log(2/\delta_n)}{2n}}\right) \\
=\ &\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_{(h,s,a),i} - R_{h,s,a} - \langle P_{h,s,a}\,,\, V_{h+1}^{*}\rangle\right| \geq \frac{H}{2}\sqrt{\frac{\log(2/\delta_n)}{2n}}\right). \qquad \text{(A.14)}
\end{aligned}$$

By Hoeffding's inequality, for any $\delta_n \in (0,1)$, we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_{(h,s,a),i} - R_{h,s,a} - \langle P_{h,s,a}\,,\, V_{h+1}^{*}\rangle\right| \geq \frac{H}{2}\sqrt{\frac{\log(2/\delta_n)}{2n}}\right) \leq \sqrt[4]{8\delta_n}. \qquad \text{(A.15)}$$

For $\delta_n = \frac{1}{HSAn^2}$, a union bound over HSA values of $z = (h,s,a)$ and all possible $1 \leq n \leq 125$ gives

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{h\in[H],s\in[S],a\in[A],1\leq n\leq 125}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} Y_{(h,s,a),i} - R_{h,s,a} - \langle P_{h,s,a}\,,\, V_{h+1}^{*}\rangle\right| \geq H\sqrt{\frac{\log(2/\delta_n)}{2n}}\right\}\right) \\
\leq\ &\mathbb{P}\left(\bigcup_{h\in[H],s\in[S],a\in[A],1\leq n\leq 125}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} Y_{(h,s,a),i} - R_{h,s,a} - \langle P_{h,s,a}\,,\, V_{h+1}^{*}\rangle\right| \geq \frac{H}{2}\sqrt{\frac{\log(2/\delta_n)}{2n}}\right\}\right) \\
\leq\ &\sum_{s=1}^{S}\sum_{a=1}^{A}\sum_{h=1}^{H}\sum_{n=1}^{125}\sqrt[4]{8\frac{1}{HSAn^2}} \\
=\ &(HSA)^{3/4}\sum_{n=1}^{125}\sqrt[4]{8\frac{1}{n^2}} \\
\leq\ &2000(HSA)^{3/4}. \qquad \text{(A.16)}
\end{aligned}$$

Combining the above two cases, we have

$$\mathbb{P}\left(\exists(k,h,s,a): n>0, \left|\hat{R}^k_{h,s,a} - R_{h,s,a} + \langle \hat{P}^k_{h,s,a} - P_{h,s,a}, V^*_{h+1}\rangle\right| \geq H\sqrt{\frac{\log(2HSAn)}{2n}}\right)$$
$$\leq 3(HSA)^{15/64} + 2000(HSA)^{3/4}$$
$$\leq 2006HSA. \tag{A.17}$$

Note that by definition, when $n = n^k(h,s,a) > 0$ we have

$$\sqrt{e_k(h,s,a)} \geq H\sqrt{\frac{\log\left(2HSAn^k(h,s,a)\right)}{2n^k(h,s,a)}} \tag{A.18}$$

and hence this concentration inequality holds with $\sqrt{e_k(h,s,a)}$ on the right hand side.

When $n = n^k(h,s,a) = 0$, we have $\hat{R}^k_{h,s,a} = 0$ and $\hat{P}^k_{h,s,a}(\cdot) = 0$ by definition, so we trivially have

$$\left|\hat{R}^k_{h,s,a} - R_{h,s,a} + \langle \hat{P}^k_{h,s,a} - P_{h,s,a}\ ,\ V^*_{h+1}\rangle\right| = |R_{h,s,a} + \langle P_{h,s,a}\ ,\ V^*_{h+1}\rangle| \leq H \leq e_k(h,s,a). \tag{A.19}$$

QED.

Lemma A.3 (Bound on the noise)**.** For $w^k(h,s,a) \sim \mathcal{N}(0, \sigma^2_k(h,s,a))$, where $\sigma^2_k(h,s,a) = \frac{H^3 S \log(2HSAk)}{2(n^k(h,s,a)+1)}$, we have that for any $k \in [K]$, the event $\mathcal{E}^w_k$ holds with probability at least $1-\delta/8$.

Proof. For any fix s,a,h,k, the random variable $w^k(h,s,a)$ follows Gaussian distribution $\mathcal{N}(0,\sigma^2_k)$. Therefore, Chernoff concentration bounds (see e.g. [34]) suggests

$$\mathbb{P}\left[|w^k(h,s,a)| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2\sigma^2_k}\right). \tag{A.20}$$

Substituting the value of σ^2_k and rearranging, with probability at least $1-\delta'$, we can write

$$\left|w^k(h,s,a)\right| \leq \sqrt{\frac{H^3 S \log(2HSAk)\log(2/\delta')}{n^k(h,s,a)+1}}. \tag{A.21}$$

Union bounding over all s,a,k,h (i.e. over state, action, timestep, and episode) imply that

$\forall s, a, k, h$, the following hold with probability at least $1-\delta'$,

$$\left|w^k(h,s,a)\right| \leq \sqrt{\frac{H^3 S \log(2HSAk)\log(2SAT/\delta')}{n^k(h,s,a)+1}}. \tag{A.22}$$

Setting $\delta' = \delta/8$, for any $s \in [S], a \in [A], h \in [H], k \in [K]$, we have that

$$\left|w^k(h,s,a)\right| \leq \sqrt{\frac{H^3 S \log(2HSAk)\log(16SAT/\delta)}{n^k(h,s,a)+1}} \leq \gamma_k(h,s,a). \tag{A.23}$$

Finally recalling the definition of $\mathcal{E}_k^w$, we complete the proof. QED.

Lemma A.4 (Bounds on the estimated action-value function, restatement of Lemma 4.2)**.** When the events $\mathcal{C}_k$ and $\mathcal{E}_k^w$ hold then for all (h,s,a)

$$\left|\left(\overline{Q}_{h,k} - Q_h^*\right)(s,a)\right| \leq H-h+1. \tag{A.24}$$

Proof. For simplicity, we set $\overline{Q}_{H+1,k}(s,a) = Q_{H+1}^*(s,a) = 0$ and it is a purely virtual value for the purpose of the proof. The proof goes through by backward induction for $h = H+1, H, \ldots, 1$.

Firstly, consider the base case $h = H+1$. The condition $|\overline{Q}_{H+1,k}(s,a) - Q_{H+1}^*(s,a)| = 0 \leq H-(H+1)+1$ directly holds from the definition.

Now we do backward induction. Assume the following inductive hypothesis to be true

$$\left|\left(\overline{Q}_{h+1,k} - Q_{h+1}^*\right)(s,a)\right| \leq H-h. \tag{A.25}$$

We consider two cases:

Case 1: $n^k(h,s,a) \leq \alpha_k$

Then the Q-function is clipped and hence $\overline{Q}_{h,k} = H-h+1$. Also by definition of the optimal Q-function $0 \leq Q_h^* \leq H-h+1$, therefore it is trivially satisfied that

$$\left|\left(\overline{Q}_{h,k} - Q_h^*\right)(s,a)\right| \leq H-h+1. \tag{A.26}$$

Case 2: $n^k(h,s,a) > \alpha_k$

In this case, we don't have clipping. From discussion of the equivalence between model-free view and model-based view in the main text, we know that $\overline{Q}_{h,k}$ is the optimal value function of $\overline{M}^k$. Using Bellman equation, we have the following decomposition

$$\begin{aligned}
&\left|\overline{Q}_{h,k}(s,a) - Q_h^*(s,a)\right| \\
&= \left|\hat{R}^k_{h,s,a} + w^k_{h,s,a} + \langle \hat{P}^k_{h,s,a}\,,\, \overline{V}_{h+1,k}\rangle - R^k_{h,s,a} - \langle P^k_{h,s,a}\,,\, V^*_{h+1}\rangle\right| \\
&\le \underbrace{\left|\langle \hat{P}^k_{h,s,a}\,,\, \overline{V}_{h+1,k} - V^*_{h+1}\rangle\right|}_{(1)} + \underbrace{\left|\hat{R}^k_{h,s,a} - R^k_{h,s,a} + \langle \hat{P}^k_{h,s,a} - P^k_{h,s,a}\,,\, V^*_{h+1}\rangle\right|}_{(2)} + \underbrace{\left|w^k_{h,s,a}\right|}_{(3)}. \qquad \text{(A.27)}
\end{aligned}$$

Term (1) is bounded by $H-h$ due to the inductive hypothesis. Under the event $\mathcal{C}_k$, term (2) is bounded by $\sqrt{e_k(h,s,a)} = H\sqrt{\frac{\log(2HSAk)}{n^k(h,s,a)+1}}$. Finally, term (3) is bounded by $\gamma_k(h,s,a)$ as the event $\mathcal{E}^w_{h,k}$ holds. With the choice of α_k, it follows that the sum of terms (2) and (3) is bounded by 1 as

$$\frac{\sqrt{H^2\log(2HSAk)} + \sqrt{H^3 S\log(2HSAk)L}}{\sqrt{n^k(h,s,a)}} < 1. \qquad \text{(A.28)}$$

Thus the sums of all the three terms is upper bounded by $H-h+1$. This completes the proof. QED.

Lemma A.5 (Intersection event probability)**.** For any episode $k \in [K]$, when the event $\mathcal{C}_k$ holds (i.e. $\hat{M}^k \in \mathcal{M}^k$), the intersection event intersection event $\overline{\mathcal{E}}_k = \mathcal{E}^w_k \cap \mathcal{E}^{\overline{Q}}_k$ holds with probability at least $1-\delta/8$. In other words, whenever the unperturbed estimated MDP lies in the confidence set (Definition 4.3), the each pseudo-noise and the estimated $\overline{Q}$ function are bounded with high probability $1-\delta/8$. Similarly defined, $\tilde{\mathcal{E}}_k$ also holds with probability $1-\delta/8$ when $\mathcal{C}_k$ happens.

Proof. The event, $\mathcal{E}^w_k$ holds with probability at least $1-\delta/8$ from Lemma A.3. Lemma A.4 gives that whenever $\left(\mathcal{C}_k \cap \mathcal{E}^w_k\right)$ holds then almost surely $\mathcal{E}^{\overline{Q}}_k$ holds. Therefore, $\mathcal{E}_k$ holds with probability $1-\delta/8$, whenever $\mathcal{C}_k$ holds. QED.

REFERENCES

[1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction.* MIT press, 2018.

[2] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning," *Journal of Machine Learning Research*, vol. 11, no. Apr, pp. 1563–1600, 2010.

[3] M. G. Azar, I. Osband, and R. Munos, "Minimax regret bounds for reinforcement learning," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70.* JMLR. org, 2017, pp. 263–272.

[4] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan, "Provably efficient reinforcement learning with linear function approximation," in *Conference on Learning Theory*, 2020, pp. 2137–2143.

[5] N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire, "Contextual decision processes with low bellman rank are pac-learnable," in *International Conference on Machine Learning.* PMLR, 2017, pp. 1704–1713.

[6] W. Sun, N. Jiang, A. Krishnamurthy, A. Agarwal, and J. Langford, "Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches," in *Conference on Learning Theory*, 2019, pp. 2898–2933.

[7] M. Henaff, "Explicit explore-exploit algorithms in continuous state spaces," in *Advances in Neural Information Processing Systems*, 2019, pp. 9377–9387.

[8] I. Osband, D. Russo, and B. Van Roy, "(more) efficient reinforcement learning via posterior sampling," in *Advances in Neural Information Processing Systems*, 2013, pp. 3003–3011.

[9] S. Agrawal and R. Jia, "Optimistic posterior sampling for reinforcement learning: worst-case regret bounds," in *Advances in Neural Information Processing Systems*, 2017, pp. 1184–1194.

[10] O. Chapelle and L. Li, "An empirical evaluation of thompson sampling," in *Advances in neural information processing systems*, 2011, pp. 2249–2257.

[11] I. Osband and B. Van Roy, "Why is posterior sampling better than optimism for reinforcement learning?" in *Proceedings of the 34th International Conference on Machine Learning-Volume 70.* JMLR. org, 2017, pp. 2701–2710.

[12] I. Osband, B. Van Roy, and Z. Wen, "Generalization and exploration via randomized value functions," in *International Conference on Machine Learning.* PMLR, 2016, pp. 2377–2386.

[13] I. Osband, B. Van Roy, D. J. Russo, and Z. Wen, "Deep exploration via randomized value functions," *Journal of Machine Learning Research*, vol. 20, no. 124, pp. 1–62, 2019.

[14] D. Russo, "Worst-case regret bounds for exploration via randomized value functions," in *Advances in Neural Information Processing Systems*, 2019, pp. 14 410–14 420.

[15] A. Zanette, D. Brandfonbrener, E. Brunskill, M. Pirotta, and A. Lazaric, "Frequentist regret bounds for randomized least-squares value iteration," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 1954–1964.

[16] M. G. Lagoudakis and R. Parr, "Least-squares policy iteration," *Journal of machine learning research*, vol. 4, no. Dec, pp. 1107–1149, 2003.

[17] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy, "Deep exploration via bootstrapped dqn," in *Advances in neural information processing systems*, 2016, pp. 4026–4034.

[18] I. Osband, J. Aslanides, and A. Cassirer, "Randomized prior functions for deep reinforcement learning," in *Advances in Neural Information Processing Systems*, 2018, pp. 8617–8629.

[19] C. Dann, T. Lattimore, and E. Brunskill, "Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 5713–5723.

[20] A. Zanette and E. Brunskill, "Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds," in *International Conference on Machine Learning*, 2019, pp. 7304–7312.

[21] L. F. Yang and M. Wang, "Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound," in *International Conference on Machine Learning*, 2020, pp. 4943–4953.

[22] M. Fortunato, M. G. Azar, B. Piot, J. Menick, M. Hessel, I. Osband, A. Graves, V. Mnih, R. Munos, D. Hassabis et al., "Noisy networks for exploration," in *International Conference on Learning Representations*, 2018.

[23] A. Pacchiano, P. Ball, J. Parker-Holder, K. Choromanski, and S. Roberts, "On optimism in model-based reinforcement learning," *arXiv preprint arXiv:2006.11911*, 2020.

[24] Z. Xu and A. Tewari, "Worst-case regret bound for perturbation based exploration in reinforcement learning," *Ann Arbor*, vol. 1001, p. 48109, 2019.

[25] B. Kveton, C. Szepesvari, S. Vaswani, Z. Wen, T. Lattimore, and M. Ghavamzadeh, "Garbage in, reward out: Bootstrapping exploration in multi-armed bandits," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3601–3610.

[26] M. Abeille, A. Lazaric et al., "Linear thompson sampling revisited," *Electronic Journal of Statistics*, vol. 11, no. 2, pp. 5165–5197, 2017.

[27] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons, 2014.

[28] Y. Seldin, C. Szepesvári, P. Auer, and Y. Abbasi-Yadkori, "Evaluation and analysis of the performance of the exp3 algorithm in stochastic environments," in *European Workshop on Reinforcement Learning*, 2013, pp. 103–116.

[29] T. Lattimore and C. Szepesvári, *Bandit algorithms.* Cambridge University Press, 2020.

[30] S. Agrawal and N. Goyal, "Thompson sampling for contextual bandits with linear payoffs," in *International Conference on Machine Learning*, 2013, pp. 127–135.

[31] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[32] J. Chen and N. Jiang, "Information-theoretic considerations in batch reinforcement learning," *arXiv preprint arXiv:1905.00360*, 2019.

[33] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan, "Is q-learning provably efficient?" in *Advances in Neural Information Processing Systems*, 2018, pp. 4863–4873.

[34] M. J. Wainwright, *Basic tail and concentration bounds*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019, p. 21–57.