Improved Worst-Case Regret Bounds for Randomized Least-Squares Value Iteration

Priyank Agrawal*, **Jinglin Chen***, **Nan Jiang**
University of Illinois at Urbana-Champaign, Urbana, IL, 61801
priyank4@illinois.edu, jinglinc@illinois.edu, nanjiang@illinois.edu

Abstract

This paper 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 $\tilde{\mathrm{O}}(H^2S\sqrt{AT})$ 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.

1 Introduction

We study systematic exploration in reinforcement learning (RL) and the exploration-exploitation trade-off therein. Exploration in RL (Sutton and Barto 2018) has predominantly focused on *Optimism in the face of Uncertainty* (OFU) based algorithms. Since the seminal work of Jaksch, Ortner, and Auer (2010), many provably efficient methods have been proposed but most of them are restricted to either tabular or linear setting (Azar, Osband, and Munos 2017; Jin et al. 2020). A few paper study a more general framework but subjected to computational intractability (Jiang et al. 2017; Sun et al. 2019; Henaff 2019). Another broad category is Thompson Sampling (TS)-based methods (Osband, Russo, and Van Roy 2013; Agrawal and Jia 2017). They are believed to have more appealing empirical results (Chapelle and Li 2011; Osband and Van Roy 2017).

In this work, we investigate a TS-like algorithm, RLSVI (Osband, Van Roy, and Wen 2016; Osband et al. 2019; Russo 2019; Zanette et al. 2020). 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 Lagoudakis and Parr (2003). Thanks to the model-free nature, RLSVI is flexible enough to be extended to general function approximation setting, as shown by Osband et al. (2016); Osband, Aslanides, and Cassirer (2018); Osband et al. (2019), 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 $\tilde{\mathrm{O}}(H^2S\sqrt{AT})$ 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 Russo (2019).
- Our high-probability regret bound matches the sharpest $\tilde{\mathrm{O}}(H^2S\sqrt{AT})$ worst-case regret bound among all TS-based methods (Agrawal and Jia 2017)[2].

Related Works Taking inspirations from Azar, Osband, and Munos (2017); Dann, Lattimore, and Brunskill (2017); Zanette and Brunskill (2019); Yang and Wang (2020), 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 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 (Osband et al. 2016; Fortunato et al. 2018; Pacchiano et al. 2020; Xu and Tewari 2019; Kveton et al. 2019) 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 (Osband et al. 2019; Osband, Van Roy, and Wen 2016). Related to this theme, the

*These two authors contributed equally.

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

[2] Agrawal and Jia (2017) studies weakly communicating MDPs with diameter D. Bounds comparable to our setting (time inhomogeneous) are obtained by augmenting their state space as $S' \to SH$ and noticing $D \geq H$.

worst-case analysis of Agrawal and Jia (2017) 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 Jaksch, Ortner, and Auer (2010) [3]. The gap of $\sqrt{SH}$ is sometimes attributed to the additional cost of exploration in TS-like approaches (Abeille, Lazaric et al. 2017). 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.

2 Preliminaries

Markov Decision Processes We consider the episodic Markov Decision Process (MDP) $M = (H, \mathcal{S}, \mathcal{A}, P, R, s_1)$ described by Puterman (2014), 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

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

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

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:

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

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 (Seldin et al. 2013; Lattimore and Szepesvári 2020) and can be converted to the same order of expected regret bound

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

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 3. 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.

3 C-RLSVI Algorithm

The major goal of this paper 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 (Agrawal and Goyal 2013; Abeille, Lazaric et al. 2017; Russo 2019; Zanette et al. 2020) 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 (Russo 2019). To address this issue, we apply a common clipping technique in RL literature (Azar, Osband, and Munos 2017; Zanette et al. 2020; Yang and Wang 2020).

We now formally introduce our algorithm C-RLSVI as shown in Algorithm 1. C-RLSVI follows a similar approach

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

as RLSVI in Russo (2019). 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. 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 1 C-RLSVI

```
 1: input: variance $\beta_k$ and clipping threshold $\alpha_k$;
 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:         $\hat{Q}_h^k \leftarrow \text{argmin}_{Q \in \mathbb{R}^{SA}} \Big[\mathcal{L}(Q \mid \hat{Q}_{h+1}^k, \dot{\mathcal{D}}_h)$
                    $+\|Q - Q_h^{\text{pri}}\|_2^2\Big]$;
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 $(\pi^k)$ 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 (Osband et al. 2016; Zanette et al. 2020). However, it also has an equivalent model-based interpretation (Russo 2019). In Algorithm 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 (Geurts, Ernst, and Wehenkel 2006; Chen and Jiang 2019) 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 Russo (2019), 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 $\alpha_k = 4H^3 S \log(2HSAk) \log(40SAT/\delta)$ 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 Russo (2019) and therefore obtain a high-probability regret bound. Similar to as discussed in Zanette et al. (2020), 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 (Russo 2019; Osband et al. 2019), it is easy to implement empirically and we will show it does not introduce a major term on regret bound.

4 Main Result

In this section, we present our main result: high-probability regret bound in Theorem 1.

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

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

Theorem 1 shows that C-RLSVI matches the state-of-the-art TS-based method (Agrawal and Jia 2017). Compared to the lower bound (Jaksch, Ortner, and Auer 2010), 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 (Abeille, Lazaric et al. 2017).

It is useful to compare our work with the following contemporaries in related directions.

Comparison with Russo (2019) 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 Russo (2019). 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 1 improves his worst-case expected regret bound $\tilde{\mathrm{O}}(H^{5/2}S^{3/2}\sqrt{AT})$ by $\sqrt{HS}$.

Comparison with Zanette et al. (2020) Very recently, Zanette et al. (2020) proposed frequentist regret analysis for a variant of RLSVI with linear function approximation and obtained high-probability regret bound of $\tilde{\mathrm{O}}\left(H^2d^2\sqrt{T}\right)$, 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 5), directly converting their bound to tabular setting ($d \to SA$) gives us quite loose bound $\tilde{\mathrm{O}}\left(H^2S^2A^2\sqrt{T}\right)$.

Comparison with Azar, Osband, and Munos (2017); Jin et al. (2018) 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 Azar, Osband, and Munos (2017); Dann, Lattimore, and Brunskill (2017); Jin et al. (2018), the estimation error is decomposed as a recurrence. Since RLSVI is only optimistic with a constant probability (see Section 5 for details), their techniques need to be substantially modified to be used in our analysis.

5 Proof Outline

In this section, we outline the proof of our main results, and the details are deferred to the appendix. The major technical flow is presented from Section 5.1 onward. Before that, 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[4].

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

Definition 1 (Confidence set)**.**

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

where we set

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

Through an application of Hoeffding's inequality (Jaksch, Ortner, and Auer 2010; Osband, Russo, and Van Roy 2013), it is shown via Lemma 2 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 2. $\sum_{k=1}^{\infty} \mathbb{P}\left(\hat{M}^k \notin \mathcal{M}^k\right) \le 2006HSA.$

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 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 3, where we claim the estimated Q-value function is bounded for all (h,s,a).

Lemma 3 ((Informal) Bound on the estimated Q-value function)**.** *Define $\overline{Q}_k$ as the Q-value function of π^k (as in Algorithm 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 $|(\overline{Q}_{h,k} - Q_h^*)(s,a)| \le H-h+1$.*

See Appendix D for the precise definition of good event and a full proof. Lemma 3 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.

[4] 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.

Optimism The event when none of the rounds in episode k need to be clipped is denoted by $\mathcal{E}^{\text{th}}_k := \{\cap_{h\in[H]}(n^k(h) \geq \alpha_k)\}$. 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, we show that C-RLSVI samples an optimistic value function estimate with at least a constant probability for any (h, k). We emphasize that such difference is fundamental.

Lemma 4. *If $\hat{M}^k \in \mathcal{M}^k$ and the event $\mathcal{E}^{th}_k$ holds, then for any $h \in [H]$*

$$\mathbb{P}\left(\overline{V}_{h,k}(s^k_h) \geq V^*_h(s^k_h) \,|\, \mathcal{H}^{k-1}_H\right) \geq \Phi(-\sqrt{2}),$$

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 (1), and $\mathcal{H}^{k-1}_H$ is all the history of the past observations made by C-RLSVI *till the end of the episode $k-1$.*

Refer to Appendix C for a complete proof and discussion. This result is an extension to a similar one proved in Russo (2019). While we prove Lemma 4 is true for any $h \in [H]$, we only need it for $h = 1$ in our technical analysis.

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

$$\begin{aligned}\text{Reg}(K) \leq\ & \sum_{k=1}^{K} \mathbf{1}\{\mathcal{E}^{\text{th}}_k\}\left(V^*_1 - V^{\pi^k}_1\right)(s^k_1) \qquad (2)\\ +\ & \underbrace{\sum_{k=1}^{K} \mathbf{1}\{\mathcal{E}^{\text{th}\,\complement}_k\}\left(V^*_1 - V^{\pi^k}_1\right)(s^k_1)}_{\text{Warm-up term}} + H\underbrace{\mathbb{P}(\hat{M}^k \notin \mathcal{M}^k)}_{\text{Lemma 2}},\end{aligned}$$

where we use Lemma 2 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 Appendix E.1 for details).

Armed with the necessary tools, over the next few subsections we sketch the proof outline of our main results. Informally, let $\overline{\mathcal{G}}_k$ be the good event under which none of rounds in episode k are clipped (i.e. $\mathcal{E}^{\text{th}}_k$ holds), $\hat{M}^k \in \mathcal{M}^k$ and the results of Lemma 3 hold (i.e. the effect of noise is bounded). We also define an event $\tilde{\mathcal{G}}_k$ (in a similar way as $\overline{\mathcal{G}}_k$, refer to Appendix B for the detail) for the MDP $\tilde{M}^k = (H, \mathcal{S}, \mathcal{A}, \hat{P}^k, \hat{R}^k + \tilde{w}^k, s^k_1)$, where $\tilde{w}^k(h, s, a)$ is sampled independently from the same distribution $N(0, \sigma^2_k(h, s, a))$ conditioned on the history of the algorithm till the last episode. We define $\mathcal{G}_k = \overline{\mathcal{G}}_k \cup \tilde{\mathcal{G}}_k$ and all subsequent discussions are under this good event $\mathcal{G}_k$.

5.1 Regret as Sum of Estimation and Pessimism

Now the regret over K episodes of the algorithm decomposes as

$$\sum_{k=1}^{K}\Big(\underbrace{(V^*_1 - \overline{V}_{1,k})(s^k_1)}_{\text{Pessimism}} + \underbrace{\overline{V}_{1,k}(s^k_1) - V^{\pi^k}_1(s^k_1)}_{\text{Estimation}}\Big). \qquad (3)$$

In OFU-style analysis, the pessimism term is non-positive and insignificant (Azar, Osband, and Munos 2017; Jin et al. 2018). In TS-based analysis, the pessimism term usually has zero expectation or can be upper bounded by the estimation term (Osband, Russo, and Van Roy 2013; Osband and Van Roy 2017; Agrawal and Jia 2017; Russo 2019). Therefore, the pessimism term is usually relaxed to zero or reduced to the estimation term, and the estimation term can be bounded separately. Our analysis proceeds quite differently. In Section 5.2, we show how the pessimism term is decomposed to terms that are related to the algorithm's trajectory (estimation term and pessimism correction term). In Section 5.3 and Section 5.4, we show how to bound these two terms through two independent recurrences. Finally, in Section 5.5, we reorganize the regret expression whose individual terms can be bounded easily by known concentration results.

5.2 Pessimism in Terms of Estimation

In this section we present Lemma 5, where the pessimism term is bounded in terms of estimation and a correction term $(V^{\pi^k}_h - \underline{V}_{h,k})(s^k_h)$ that will be defined later. This correction term is further handled in Section 5.4. While the essence of Lemma 5 is similar to that given by Zanette et al. (2020), there are key differences: we need to additionally bound the correction term; the nature of the recurrence relations for the pessimism and estimation terms necessitates a distinct solution, hence leading to different order dependence in regret bound. In all, this allows us to obtain stronger regret bounds as compared to Zanette et al. (2020).

The following lemma shows how the pessimism term is decomposed into estimation and a correction term. While we use the lemma only for $h = 1$, here we prove a stronger relation.

Lemma 5. *Under the event $\mathcal{G}_k$,*

$$\begin{aligned}&(V^*_h - \overline{V}_{h,k})(s^k_h) \qquad (4)\\ &\lesssim (\overline{V}_{h,k} - V^{\pi^k}_h)(s^k_h) + (V^{\pi^k}_h - \underline{V}_{h,k})(s^k_h) + \mathcal{M}^w_{h,k},\end{aligned}$$

where $\mathcal{M}^w_{h,k}$ is a martingale difference sequence (MDS).

The detailed proof can be found in Appendix E.3, while we present an informal proof sketch here. The general strategy in bounding $V^*_h(s^k_h) - \overline{V}_{h,k}(s^k_h)$ is that we find an upper bounding estimate of $V^*_h(s^k_h)$ and a lower bounding estimate of $\overline{V}_{h,k}(s^k_h)$, and show that the difference of these two estimates converge. We define $\tilde{V}_{h,k}$ as the optimal value function of $\tilde{M}^k$ and we directly have that $\tilde{V}_{h,k}$ is i.i.d. to $\overline{V}_{h,k}$ under $\mathcal{G}_k$. Since the pseudo-noise is bounded under the event $\mathcal{G}_k$, consider another counterfactual case with MDP $\underline{M}^k = (H, \mathcal{S}, \mathcal{A}, \hat{P}^k, \hat{R}^k + \underline{w}^k, s^k_1)$, where $\underline{w}^k = -\gamma_k$. Let

$\underline{V}_{h,k}$ denote the optimal value function for $\underline{M}^k$. $\underline{V}_{h,k}$ is analogous to $\overline{V}_{h,k}$ but with fixed lower bounds as pseudo-noise terms. $\underline{w}^k$ can also be thought as a solution of following optimization:

$$\min_{w^k_{\text{ptb}} \in \mathbb{R}^{HSA}} V^{w^k_{\text{ptb}}}_{h,k}(s^k_1)$$

$$s.t. \quad |w^k_{\text{ptb}}(h,s,a)| \leq \gamma_k(h,s,a) \quad \forall\, h,s,a.$$

The detailed definition is deferred to Appendix A.2. This ensures $\underline{V}_{h,k} \leq \overline{V}_{h,k}$ and $\underline{V}_{h,k} \leq \tilde{V}_{h,k}$. Thus the pessimism term is now given by

$$(V^*_h - \overline{V}_{h,k})(s^k_h) \leq (V^*_h - \underline{V}_{h,k})(s^k_h). \tag{5}$$

Let event $\tilde{\mathcal{O}}_{h,k} := \{\tilde{V}_{h,k}(s^k_h) \geq V^*_h(s^k_h)\}$ and $\mathbb{E}_{\tilde{w}}[\cdot]$ denote the expectation over the pseudo-noise $\tilde{w}$. Since $V^*_h(s^k_h)$ does not depend on $\tilde{w}$, we get $V^*_h(s^k_h) \leq \mathbb{E}_{\tilde{w}|\tilde{\mathcal{O}}_{h,k}}\left[\tilde{V}_{h,k}(s^k_h)\right]$. We can further upper bound Eq (5) as

$$(V^*_h - \underline{V}_{h,k})(s^k_h) \leq \mathbb{E}_{\tilde{w}|\tilde{\mathcal{O}}_{h,k}}[(\tilde{V}_{h,k} - \underline{V}_{h,k})(s^k_h)]. \tag{6}$$

Thus, we are able to relate pessimism to quantities which only depend on the algorithm's trajectory. Further we upper bound the expectation over marginal distribution $\mathbb{E}_{\tilde{w}|\tilde{\mathcal{O}}_{h,k}}[\cdot]$ by $\mathbb{E}_{\tilde{w}}[\cdot]$. This is possible because we are taking expectation of non-negative entities. We can show:

$$\begin{aligned}&\mathbb{E}_{\tilde{w}|\tilde{\mathcal{O}}_{h,k}}[(\tilde{V}_{h,k} - \underline{V}_{h,k})(s^k_h)] \\ &\simeq \mathcal{M}^w_{h,k} + \overline{V}_{h,k}(s^k_h) - \underline{V}_{h,k}(s^k_h),\end{aligned} \tag{7}$$

Now consider

$$\begin{aligned}&(\overline{V}_{h,k} - \underline{V}_{h,k})(s^k_h) \\ &= \underbrace{(\overline{V}_{h,k} - V^{\pi^k}_h)(s^k_h)}_{\text{Estimation term}} + \underbrace{(V^{\pi^k}_h - \underline{V}_{h,k})(s^k_h)}_{\text{Correction term}}.\end{aligned} \tag{8}$$

In Eq (8), the estimation term is decomposed further in Section 5.3. The correction term is simplified in Section 5.4

5.3 Bounds on Estimation Term

In this section we show the bound on estimation term. Under the high probability good event $\mathcal{G}_k$, we show decomposition for the estimation term $(\overline{V}_{h,k} - V^{\pi^k}_{h,k})(s^k_h)$ holds with high probability. Applying Bellman equation yields

$$\begin{aligned}(\overline{V}_{h,k} - V^{\pi^k}_h)(s^k_h) = &\underbrace{\langle \hat{P}^k_h - P^k_h, \overline{V}_{h+1,k}\rangle}_{(1)} \\ &+ \underbrace{\langle P^k_h, \overline{V}_{h+1,k} - V^{\pi^k}_{h+1}\rangle}_{(1')} + \hat{R}^k_h - R^k_h + w^k_h.\end{aligned} \tag{9}$$

We first decompose Term (1) as

$$(1) = \underbrace{\langle \hat{P}^k_h - P^k_h, V^*_{h+1}\rangle}_{(2)} + \underbrace{\langle \hat{P}^k_h - P^k_h, \overline{V}_{h+1,k} - V^*_{h+1}\rangle}_{(3)}. \tag{10}$$

Term (2) represents the error in estimating the transition probability for the optimal value function V^*_h, while Term (3) is an offset term. The total estimation error, $\epsilon^{\text{err}}_{h,k} := \left|\text{Term (2)} + \hat{R}^k_h - R^k_h\right|$ is easy to bound since the empirical MDP $\hat{M}^k$ lies in the confidence set (Eq (1)). Then we discuss how to bound Term (3). Unlike OFU-styled analysis, here we do not have optimism almost surely. Therefore we cannot simply relax $\overline{V}_{h+1,k} - V^*_{h+1}$ to $\overline{V}_{h+1,k} - V^{\pi^k}_{h+1}$ and form the recurrence. Instead, we will apply (L_1, L_∞) Cauchy-Schwarz inequality to separate the deviation of transition function estimation and the deviation of value function estimation, and then further bound these two deviation terms. Noticing that $\overline{V}_{h+1,k} - V^*_{h+1}$ might be unbounded, we use Lemma 3 to assert that $\|V^*_{h+1} - \overline{V}_{h+1}\|_\infty \leq H$ under event $\mathcal{G}_k$. With the boundedness of the deviation of value function estimation, it suffices to bound the remaining $\|\hat{P}_{h,s_h,a_h} - P_{h,s_h,a_h}\|_1$ term. Proving an L_1 concentration bound for multinomial distribution with careful application of the Hoeffding's inequality shows

$$\|\hat{P}^k_h - P^k_h\|_1 \leq 4\sqrt{\frac{SL}{n^k(h)+1}},$$

where $L = \log\left(40SAT/\delta\right)$. In Eq (9), we also decomposes Term (1') to a sum of the next-state estimation and a MDS.

Clubbing all the terms starting from Eq (9), with high probability, the upper bound on estimation is given by

$$\begin{aligned}&(\overline{V}_{h,k} - V^{\pi^k}_h)(s^k_h) \\ &\lesssim \underbrace{(\overline{V}_{h+1,k} - V^{\pi^k}_{h+1})(s^k_{h+1})}_{\text{Next-state estimation}} + \epsilon^{\text{err}}_{h,k} + w^k_h + \mathcal{M}_{\overline{\delta}^{\pi^k}_{h,k}(s^k_h)} \\ &\quad + 4H\sqrt{\frac{SL}{n^k(h)+1}},\end{aligned} \tag{11}$$

where $\mathcal{M}_{\overline{\delta}^{\pi^k}_{h,k}(s^k_h)}$ is a Martingale difference sequence (MDS). Thus, via Eq (11) we are able decompose estimation term in terms of total estimation error, next-step estimation, pseudo-noise, a MDS term, and a $\tilde{\mathrm{O}}\left(\sqrt{1/n^k(h)}\right)$ term. From the form of Eq (11), we can see that it forms a recurrence. Due to this style of proof, our Theorem 1 is $\sqrt{HS}$ superior than the previous state-of-art result (Russo 2019), and we are able to provide a high probability regret bound instead of just the expected regret bound.

5.4 Bounds on Pessimism Correction

In this section, we give the decomposition of the pessimism correction term $(V^{\pi^k}_h - \underline{V}_{h,k})(s^k_h)$. Shifting from MDP $\overline{M}_k$ to MDP $\underline{M}_k$ and re-tracing the steps of Section 5.3, with high probability, it follows

$$\begin{aligned}(V^{\pi^k}_h - \underline{V}_{h,k})(s^k_h) \lesssim &\underbrace{(V^{\pi^k}_{h+1} - \underline{V}_{h+1,k})(s^k_{h+1})}_{\text{Next-state pessimism correction}} + \epsilon^{\text{err}}_{h,k} \\ &+ \left|\underline{w}^k_h\right| + \mathcal{M}_{\underline{\delta}^{\pi^k}_{h,k}(s^k_h)} + 4H\sqrt{\frac{SL}{n^k(h)+1}}.\end{aligned} \tag{12}$$

The decomposition Eq (12) also forms a recurrence. The recurrences due to Eq (11) and Eq (12) are later solved in Section 5.5.

5.5 Final High-Probability Regret Bound

To solve the recurrences of Eq (11) and Eq (12), we keep unrolling these two inequalities from $h = 1$ to $h = H$. Then with high probability, we get

$$\mathrm{Reg}(K) \lesssim \sum_{k=1}^{K} \sum_{h=1}^{H} \Bigg(\left|\epsilon_{h,k}^{\mathrm{err}}\right| + \left|\underline{w}_h^k\right| + w_h^k + \mathcal{M}_{h,k}^{w} + \mathcal{M}_{\underline{\delta}_{h,k}^{\pi^k}(s_h^k)} + \mathcal{M}_{\bar{\delta}_{h,k}^{\pi^k}(s_h^k)} + 4H\sqrt{\frac{SL}{n^k(h)+1}} \Bigg).$$

Bounds of individual terms in the above equation are given in Appendix F, and here we only show the order dependence.

The maximum estimation error that can occur at any round is limited by the size of the confidence set Eq (1). Lemma 19 sums up the confidence set sizes across the h and k to obtain $\sum_{k=1}^{K}\sum_{h=1}^{H}\left|\epsilon_{h,k}^{\mathrm{err}}\right| = \tilde{\mathrm{O}}(\sqrt{H^3SAT})$. In Lemma 20, we use Azuma-Hoeffding inequality to bound the summations of the martingale difference sequences with high probability by $\tilde{\mathrm{O}}(H\sqrt{T})$. The pseudo-noise $\sum_{k=1}^{K}\sum_{h=1}^{H} w_h^k$ and the related term $\sum_{k=1}^{K}\sum_{h=1}^{H} \underline{w}_h^k$ are bounded in Lemma 18 with high probability by $\tilde{\mathrm{O}}(H^2S\sqrt{AT})$. Finally we have $\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{SL}{n^k(h)+1}} = \tilde{\mathrm{O}}(H^2S\sqrt{AT})$ again from Lemma 18. Putting all these together yields the high-probability regret bound of Theorem 1.

6 Discussions and 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 (Azar, Osband, and Munos 2017), 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 bound on Term (3) of Eq (10). 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 (Azar, Osband, and Munos 2017). 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.

Acknowledgements

We gratefully thank the constructive comments from Zhihan Xiong and Anonymous Reviewers.

References

Abeille, M.; Lazaric, A.; et al. 2017. Linear thompson sampling revisited. *Electronic Journal of Statistics* 11(2): 5165–5197.

Agrawal, S.; and Goyal, N. 2013. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, 127–135.

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

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

Chapelle, O.; and Li, L. 2011. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, 2249–2257.

Chen, J.; and Jiang, N. 2019. Information-Theoretic Considerations in Batch Reinforcement Learning. In *International Conference on Machine Learning*, 1042–1051.

Dann, C.; Lattimore, T.; and Brunskill, E. 2017. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, 5713–5723.

Fortunato, M.; Azar, M. G.; Piot, B.; Menick, J.; Hessel, M.; Osband, I.; Graves, A.; Mnih, V.; Munos, R.; Hassabis, D.; et al. 2018. Noisy Networks For Exploration. In *International Conference on Learning Representations*.

Geurts, P.; Ernst, D.; and Wehenkel, L. 2006. Extremely randomized trees. *Machine learning* 63(1): 3–42.

Henaff, M. 2019. Explicit explore-exploit algorithms in continuous state spaces. In *Advances in Neural Information Processing Systems*, 9377–9387.

Jaksch, T.; Ortner, R.; and Auer, P. 2010. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research* 11(Apr): 1563–1600.

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

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

Jin, C.; Yang, Z.; Wang, Z.; and Jordan, M. I. 2020. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, 2137–2143.

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

Lagoudakis, M. G.; and Parr, R. 2003. Least-squares policy iteration. *Journal of machine learning research* 4(Dec): 1107–1149.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit algorithms*. Cambridge University Press.

Osband, I.; Aslanides, J.; and Cassirer, A. 2018. Randomized prior functions for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, 8617–8629.

Osband, I.; Blundell, C.; Pritzel, A.; and Van Roy, B. 2016. Deep exploration via bootstrapped DQN. In *Advances in neural information processing systems*, 4026–4034.

Osband, I.; Russo, D.; and Van Roy, B. 2013. (More) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, 3003–3011.

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

Osband, I.; Van Roy, B.; Russo, D. J.; and Wen, Z. 2019. Deep Exploration via Randomized Value Functions. *Journal of Machine Learning Research* 20(124): 1–62.

Osband, I.; Van Roy, B.; and Wen, Z. 2016. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, 2377–2386. PMLR.

Pacchiano, A.; Ball, P.; Parker-Holder, J.; Choromanski, K.; and Roberts, S. 2020. On Optimism in Model-Based Reinforcement Learning. *arXiv preprint arXiv:2006.11911* .

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

Russo, D. 2019. Worst-case regret bounds for exploration via randomized value functions. In *Advances in Neural Information Processing Systems*, 14410–14420.

Seldin, Y.; Szepesvári, C.; Auer, P.; and Abbasi-Yadkori, Y. 2013. Evaluation and analysis of the performance of the EXP3 algorithm in stochastic environments. In *European Workshop on Reinforcement Learning*, 103–116.

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

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

Xu, Z.; and Tewari, A. 2019. Worst-Case Regret Bound for Perturbation Based Exploration in Reinforcement Learning. *Ann Arbor* 1001: 48109.

Yang, L. F.; and Wang, M. 2020. Reinforcement Learning in Feature Space: Matrix Bandit, Kernels, and Regret Bound. In *International Conference on Machine Learning*, 4943–4953.

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

Zanette, A.; and Brunskill, E. 2019. Tighter Problem-Dependent Regret Bounds in Reinforcement Learning without Domain Knowledge using Value Function Bounds. In *International Conference on Machine Learning*, 7304–7312.