Frequentist Regret Bounds for Randomized Least-Squares Value Iteration

Andrea Zanette*
Stanford University

David Brandfonbrener*
New York University

Emma Brunskill
Stanford University

Matteo Pirotta
Facebook AI Research

Alessandro Lazaric
Facebook AI Research

Abstract

We consider the exploration-exploitation dilemma in finite-horizon reinforcement learning (RL). When the state space is large or continuous, traditional tabular approaches are unfeasible and some form of function approximation is mandatory. In this paper, we introduce an optimistically-initialized variant of the popular randomized least-squares value iteration (RLSVI), a model-free algorithm where exploration is induced by perturbing the least-squares approximation of the action-value function. Under the assumption that the Markov decision process has low-rank transition dynamics, we prove that the frequentist regret of RLSVI is upper-bounded by $\widetilde{O}(d^2H^2\sqrt{T})$ where d is the feature dimension, H is the horizon, and T is the total number of steps. To the best of our knowledge, this is the first frequentist regret analysis for randomized exploration with function approximation.

1 Introduction

A key challenge in reinforcement learning (RL) is how to balance exploration and exploitation in order to efficiently learn to make good sequences of decisions in a way that is both computationally tractable and statistically efficient. In the tabular case, the exploration-exploitation problem is well-understood for a number of settings (e.g., finite-horizon, average reward, infinite horizon with discount), exploration objectives (e.g., regret minimization and probably approximately correct), and for different algorithmic approaches, where optimism-under-uncertainty (Jaksch et al., 2010; Fruit et al., 2018) and Thompson sampling (TS) (Osband et al., 2016a; Russo, 2019) are the most popular principles. For instance, in the finite-horizon setting, Azar et al. (2017) and Zanette and Brunskill (2019) recently derived minimax optimal and structure adaptive regret bounds for optimistic exploration algorithms. TS-based algorithms have mainly been analyzed in tabular MDPs in terms of Bayesian regret (Osband et al., 2016a; Osband and Roy, 2017; Ouyang et al., 2017), which assumes that the MDP is sampled from a known prior distribution. These bounds do not hold against a fixed MDP and algorithms with small Bayesian regret may still suffer high regret in some hard-to-learn MDPs within the chosen prior. In the tabular setting, frequentist (or worst-case) regret analysis has been developed for TS-based algorithms both in the average reward (Gopalan and Mannor, 2015; Agrawal and Jia, 2017) and finite-horizon case (Russo, 2019). Despite the fact that TS-based approaches have slightly worse regret bounds compared to optimism-based algorithms, their empirical performance is often superior (Chapelle and Li, 2011; Osband and Roy, 2017).

Unfortunately, the performance of tabular exploration methods rapidly degrades with the number of states and actions, thus making them infeasible in large or continuous MDPs. So, one of the most important challenges to improve sample efficiency in large-scale RL is how to combine exploration mechanisms with generalization methods to obtain algorithms with provable regret guarantees. The simplest approach to deal with continuous state is discretization. It has been used in Ortner and Ryabko (2012); Lakshmanan et al. (2015) to derive $\widetilde{O}(T^{3/4})$ and $\widetilde{O}(T^{2/3})$ frequentistic regret bounds for average reward MDPs. Recent work

*Equal Contribution

Proceedings of the 23$^{\text{rd}}$International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

on contextual MDPs (Jiang et al., 2017; Dann et al., 2018) yielded promising sample efficiency guarantees, but such algorithms are computationally intractable, and their bounds are not tight in the tabular settings.

One of the most simple and popular forms of function approximation is to use a linear representation for the action-value functions. When the transition model also has low-rank structure, very recent work has shown that a variant of Q-learning can achieve polynomial sample complexity as a function of the state space dimension when given access to a generative model (Yang and Wang, 2019b). Nonetheless, the generative model assumption removes most of the exploration challenge, as the state space can be arbitrarily sampled. Concurrently to our work, optimism-based exploration has been successfully integrated with linear function approximation both in model-based and model-free algorithms (Yang and Wang, 2019a; Jin et al., 2019). In MDPs with low-rank dynamics, these algorithms are proved to have regret bounds scaling with the dimensionality d of the linear space (i.e., the number of features) instead of the number of states.

On the algorithmic side, TS-based exploration can be easily integrated with linear function approximation as suggested in the Randomized Least-Squares Value Iteration (RLSVI) algorithm (Osband et al., 2016b). Despite promising empirical results, RLSVI has been analyzed only in the tabular case (i.e., when the features are indicators for each state) and for Bayesian regret. While RLSVI is a model-free algorithm, recent work (Russo, 2019) leverages an equivalence between model-free and model-based algorithms in the tabular case to derive frequentist regret bounds. The analysis carefully chooses the variance of the perturbations applied to the estimated solution to ensure that the value estimates are optimistic with constant probability.

In this paper we provide the first frequentist regret analysis for a variant of RLSVI when linear function approximation is used in the finite-horizon setting. Similar to optimistic PSRL for the tabular setting (Agrawal and Jia, 2017), we modify RLSVI to ensure that the perturbed estimates used in the value iteration process are optimistic with constant probability. Following the results in the linear bandit literature (Abeille et al., 2017), we show that the perturbation applied to the the least-squares estimates should be larger than their estimation error. However, in contrast to bandit, perturbed estimates are propagated back through iterations and we need to carefully adjust the perturbation scheme so that the probability of being optimistic does not decay too fast with the horizon and, at the same time, we can control how the perturbations accumulate over iterations. Under the assumption that the system dynamics are low-rank, we show that the frequentist regret of our algorithm is $\widetilde{O}(H^2 d^2\sqrt{T} + H^5 d^4 + \epsilon d H(1+\epsilon d H^2)T)$ where ϵ is the misspecification level, H is the fixed horizon, d is the number of features, and T is the number of samples. Similar to linear bandits, this is worse by a factor of $\sqrt{Hd}$ (i.e., the square root of the dimension of the estimated parameters) than the optimistic algorithm of Jin et al. (2019). Whether this gap can be closed is an open question both in bandits and RL.

2 Preliminaries

We consider an undiscounted finite-horizon MDP (Puterman, 1994) $M = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, H)$ with state space $\mathcal{S}$, action space $\mathcal{A}$ and horizon length $H \in \mathbb{N}^+$. For every $t \in [H] \overset{def}{=} \{1, \ldots, H\}$, every state-action pair is characterized by a reward $r_t(s,a) \in [0,1]$ and a transition kernel $\mathbb{P}_t(\cdot|s,a)$ over next state. We assume $\mathcal{S}$ to be a measurable, possibly infinite, space and $\mathcal{A}$ can be any (compact) time and state dependent set (we omit this dependency for brevity). For any $t \in [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, the state-action value function of a non-stationary policy $\pi = (\pi_1, \ldots, \pi_H)$ is defined as $Q_t^\pi(s,a) = r_t(s,a) + \mathbb{E}\left[\sum_{l=t+1}^H r_l(s_l, \pi_l(s_l)) \mid s,a\right]$ and the value function is $V_t^\pi(s) = Q_t^\pi(s, \pi_t(s))$. Since the horizon is finite, under some regularity conditions, (Shreve and Bertsekas, 1978), there always exists an optimal policy $\pi^\star$ whose value and action-value functions are defined as $V_t^\star(s) \overset{def}{=} V_t^{\pi^\star}(s) = \sup_\pi V_t^\pi(s)$ and $Q_t^\star(s,a) \overset{def}{=} Q_t^{\pi^\star}(s,a) = \sup_\pi Q_t^\pi(s,a)$. Both Q^π and $Q^\star$ can be conveniently written as the result of the Bellman equations

$$Q_t^\pi(s,a) = r_t(s,a) + \mathbb{E}_{s' \sim \mathbb{P}_t(\cdot|s,a)}[V_{t+1}^\pi(s')] \qquad (1)$$

$$Q_t^\star(s,a) = r_t(s,a) + \mathbb{E}_{s' \sim \mathbb{P}_t(\cdot|s,a)}[V_{t+1}^\star(s')] \qquad (2)$$

where $V_{H+1}^\pi(s) = V_{H+1}^\star(s) = 0$ and $V_t^\star(s) = \max_{a \in \mathcal{A}} Q_t^\star(s,a)$, for all $s \in \mathcal{S}$. Notice that by boundedness of the reward, for any t and (s,a), all functions Q_t^π, V_t^π, $Q_t^\star$, $V_t^\star$ are bounded in $[0, H-t+1]$.

The learning problem The learning agent interacts with the MDP in a sequence of episodes $k \in [K]$ of fixed length H by playing a nonstationary policy $\pi_k = (\pi_{1k}, \ldots, \pi_{Hk})$ where $\pi_{tk} : \mathcal{S} \to \mathcal{A}$. In each episode, the initial state s_{1k} is chosen arbitrarily and revealed to the agent. The learning agent does not know the transition or reward functions, and it relies on the samples (i.e., states and rewards) observed over episodes to improve its performance over time. Finally, we evaluate the performance of an agent by its regret after K episodes: $\textsc{Regret}(\mathrm{K}) \overset{def}{=} \sum_{k=1}^K V_1^\star(s_{1k}) - V_1^{\pi_k}(s_{1k})$.

Linear function approximation and low-rank MDPs. Whenever the state space $\mathcal{S}$ is too large or continuous, functions above cannot be represented by enumerating their values at each state or state-action pair. A common approach is to define a feature map $\phi_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, possibly different at any $t \in [H]$, embedding each state-action pair (s, a) into a d-dimensional vector $\phi_t(s, a)$. The action-value functions are then represented as a linear combination between the features ϕ_t and a vector parameter $\theta_t \in \mathbb{R}^d$, such that $Q_t(s, a) = \phi_t(s, a)^\top \theta_t$. This representation effectively reduces the complexity of the problem from $\mathcal{S} \times \mathcal{A}$ down to d. Nonetheless, $Q^\star_t$ may not fit into the space spanned by ϕ_t, and approximate value iteration may propagate and accumulate errors over iterations (Munos, 2005; Munos and Szepesvári, 2008), and an exploration algorithm may suffer linear regret. Thus, similar to (Yang and Wang, 2019a,b; Jin et al., 2019), we consider MPDs that are "coherent" with the feature map ϕ_t used to represent action-value functions. In particular, we assume that M has (approximately) low-rank transition dynamics and linear reward in ϕ_t.

Assumption 1 (Approximately Low-Rank MDPs)**.** *We assume that for each $t \in [H]$ there exist a feature map $\psi_t : \mathcal{S} \to \mathbb{R}^d$, $s \mapsto \psi_t(s)$ and a parameter $\theta^r_t \in \mathbb{R}^d$ such that the reward can be decomposed as a linear response and a non-linear term:*

$$r_t(s, a) = \phi_t(s, a)^\top \theta^r_t + \Delta^r_t(s, a) \tag{3}$$

and the dynamics are approximately low-rank:

$$\mathbb{P}_t(s' \mid s, a) = \phi_t(s, a)^\top \psi_t(s') + \Delta^P_t(s' \mid s, a). \tag{4}$$

We denote by ϵ an upper bound on the non-linear terms, as follows:

$$|\Delta^r_t(s, a)| \leqslant \epsilon, \qquad \|\Delta^P_t(\cdot \mid s, a)\|_1 \leqslant \epsilon. \tag{5}$$

We further make the following regularity assumptions:

$$\|\phi_t(s, a)\|_2 \leqslant L_\phi,\ \|\theta^r_t\|_2 \leqslant L_r,\ \int_s \|\psi_t(s)\| \leqslant L_\psi. \tag{6}$$

An important consequence of Asm. 1 in the absence of misspecification ($\epsilon = 0$) is that the Q-function of any policy is linear in the features ϕ.

Proposition 1. *If $\epsilon = 0$, for every policy π and timestep $t \in [H]$ there exists $\theta^\pi_t \in \mathbb{R}^d$ such that*

$$Q^\pi_t(s, a) = \phi_t(s, a)^\top \theta^\pi_t, \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}. \tag{7}$$

Proof. The definition of low-rank MDP from Asm. 1 together with the Bellman equation gives:

$$\begin{aligned} Q^\pi_t(s, a) &= r_t(s, a) + \mathbb{E}_{s'|s,a}[V^\pi_{t+1}(s')] \\ &= \phi_t(s, a)^\top \theta^r_t + \int_{s'} \phi_t(s, a)^\top \psi_t(s') V^\pi_{t+1}(s') \\ &= \phi_t(s, a)^\top \left(\theta^r_t + \int_{s'} \psi_t(s') V^\pi_{t+1}(s') \right) \end{aligned} \tag{8}$$

We define θ^π_t to be the term inside the parentheses. □

To give further intuition about the assumption, consider the case of finite state and action spaces (again with $\epsilon = 0$). Then we can write:

$$\mathbb{P}_t(s, a) = \phi_t(s, a)^\top \Psi_t \tag{9}$$

for a certain $\Psi_t \in \mathbb{R}^{d \times \mathcal{S}}$. Then for any policy π there exists a matrix Φ^π such that the transition matrix of the Markov chain P^π can be expressed by a low-rank factorization:

$$P^\pi_t = \Phi^\pi_t \Psi_t, \quad \Phi^\pi_t \in \mathbb{R}^{S \times d}, \Psi_t \in \mathbb{R}^{d \times S} \tag{10}$$

where in particular Φ^π_t depends on the policy π:

$$\Phi^\pi_t[s, :] = \phi_t(s, \pi(s))^\top, \quad \Psi_t[:, s'] = \psi_t(s'). \tag{11}$$

Since $\textsc{Rank}(\Phi^\pi_t) \leqslant d, \textsc{Rank}(\Psi_t) \leqslant d$ we get $\textsc{Rank}(P^\pi_t) \leqslant d$ (see Golub and Van Loan (2012)).

3 Algorithm

Our primary goal in this work is to provide a Thompson sampling (TS)-based algorithm with linear value function approximation with frequentist regret bounds. A key challenge in frequentist analyses of TS algorithms is to ensure sufficient exploration using randomized (i.e., perturbed) versions of the estimated model or value function. A common way to obtain effective exploration has been to consider perturbations large enough so that the resulting sampled model or value function is optimistic with a fixed probability (Agrawal and Goyal, 2013; Abeille et al., 2017; Russo, 2019). However, such prior work has only considered the bandit or tabular MDP settings. Here we modify RLSVI described by Osband et al. (2016b) to use an optimistic "default" value function during an initial phase and inject carefully-tuned perturbations to enable *frequentist* regret bounds in low-rank MDPs. We refer to the resulting algorithm as OPT-RLSVI and we illustrate it in Alg. 1.

Gaussian noise to encourage exploration. OPT-RLSVI proceeds in episodes. At the beginning of episode k it receives an initial state s_{1k} and runs a value iteration procedure to compute a linear approximation of $Q^\star_t$ at each timestep $t \in [H]$. To encourage exploration, the learned parameter $\widehat{\theta}_{tk}$ is perturbed by adding mean-zero Gaussian noise $\overline{\xi}_{tk} \sim \mathcal{N}(0, \sigma^2 \Sigma^{-1}_{tk})$, obtaining $\overline{\theta}_{tk} = \widehat{\theta}_{tk} + \overline{\xi}_{tk}$. The perturbation (or *pseudonoise*) $\overline{\xi}_{tk}$ has variance proportional to the inverse of the regularized design matrix $\Sigma_{tk} = \sum_{i=1}^{k-1} \phi_{ti}\phi_{ti}^\top + \lambda I$, where the ϕ_{ti}'s are the features encountered in prior episodes; this results in perturbations with higher variance in less explored directions. Finally, we show how to choose the magnitude σ^2 of the variance in Sec. 5.2 to ensure sufficient exploration.

A key contribution of our work is to prove that this strategy can guarantee reliable exploration under Asm. 1. We do this by showing that the algorithm is optimistic with constant probability. Explicitly, we prove that the (random) value function difference $(\overline{V}_{1k} - V_1^{\star})(s_{1k})$ can be expressed as a *one-dimensional biased random walk*, which depends on a high probability bound on the environment noise (the bias of the walk) and on the variance of the injected pseudonoise (the variance of the walk). By setting the pseudonoise to have the appropriate variance we can guarantee that the random walk is "optimistic" enough that the algorithm explores sufficiently. Unfortunately, it is possible to analyze the algorithm as a random walk only if the value function is not perturbed by clipping; otherwise, one cannot write down the walk and the process is difficult to analyze as further bias is introduced by clipping. However, not clipping the value function may give rise to abnormal values.

The issue of abnormal values. A common problem that arises in estimation in RL with function approximation is that as a result of statistical errors combined with the bootstrapping and extrapolation of the next-state value function (Munos, 2005; Munos and Szepesvári, 2008; Farahmand et al., 2010) the value function estimate can take values outside its plausible range. A common solution is to "clip" the bootstrapped value function into the range of plausible values (in this case, between 0 and H). This avoids propagating overly abnormal values to the estimated parameters at prior timesteps which would degrade their estimation accuracy. Clipping the value function is also a solution typically employed in tabular algorithms for exploration (Azar et al., 2017; Dann et al., 2017; Zanette and Brunskill, 2019; Yang and Wang, 2019a; Dann et al., 2019). After adding optimistic bonuses for exploration they "clip" the value function above by H, which is an upper bound on the true optimal value function. Since H is guaranteed to be an optimistic estimate for $V^{\star}$, clipping effectively preserves optimism while keeping the value function bounded for bootstrapping. However, clipping cannot be easily integrated in our setting as it effectively introduces bias in the pseudonoise and it may "pessimistically" affect the value function estimates, reducing the probability of being optimistic.

Default value function. To avoid propagating unreasonable values without using clipping, we define a default value function, similar in the spirit to algorithms such as $R_{\max}$ (Brafman and Tennenholtz, 2002). In particular, we assign the maximum plausible value $\overline{Q}_t(s,a) = H - t + 1$ to an uncertain direction $\phi_t(s,a)$ (as measured by the $\|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}}$ norm). Once a given direction $\phi_t(s,a)$ has been tried a sufficient number of times we can guarantee (under an inductive argument) that the linearity of the representation is accurate enough that with high probability $\phi_t(s,a)^\top \overline{\theta}_{tk} - Q_t^{\star}(s,a) \in [-(H-t+1), 2(H-t+1)]$. In other words, abnormal values are not going to be encountered, and thus clipping becomes unnecessary. Notice that this accuracy requirement is quite minimal because $V_t^{\star}$ has a range of at most $H - t + 1$.

We emphasize that the purpose of the optimistic default function is not to inject further optimism but rather to keep the propagation of the errors under control while ensuring optimism.

Defining the $\overline{Q}$ values. Finally, we also choose our Q function to interpolate between the "default" optimistic value and the linear function of the features as the uncertainty decreases. The main reason is to ensure *continuity* of the function, which facilitates the handling of some of the technical aspects connected to the concentration inequality (in particular in App. E).

Definition 1 (Algorithm Q function)**.** *For some constants α_L, α_U and using shorthand for the feature $\phi \overset{def}{=} \phi_t(s,a)$, the default function $B_t \overset{def}{=} H - t + 1$ and the interpolation parameter $\rho \overset{def}{=} \frac{\|\phi\|_{\Sigma_{tk}^{-1}} - \alpha_L}{\alpha_U - \alpha_L}$ define:*

$$\overline{Q}_{tk}(s,a) \overset{def}{=} \begin{cases} \phi^\top \overline{\theta}_{tk}, & \text{if } \|\phi\|_{\Sigma_{tk}^{-1}} \leqslant \alpha_L \\ B_t, & \text{if } \|\phi\|_{\Sigma_{tk}^{-1}} \geqslant \alpha_U \\ \rho\left(\phi^\top \overline{\theta}_{tk}\right) + (1-\rho) B_t, & \text{otherwise.} \end{cases}$$

Algorithm 1 OPT-RLSVI

```
Initialize $\Sigma_{t1} = \lambda I,\ \forall t \in [H]$; Define $\overline{V}_{tk}(s) = \max_a \overline{Q}_{tk}(s,a)$, with $\overline{Q}_{tk}(s,a)$ defined in Def. 1
for $k = 1, 2, \ldots$ do
    Receive starting state $s_{1k}$
    Set $\overline{\theta}_{H+1,k} = 0$
    for $t = H, H-1, \ldots, 1$ do
        $\widehat{\theta}_{tk} = \Sigma_{tk}^{-1}\left(\sum_{i=1}^{k-1} \phi_{ti}[r_{ti} + \overline{V}_{t+1,k}(s_{t+1,i})]\right)$
        Sample $\overline{\xi}_{tk} \sim \mathcal{N}(0, \sigma^2 \Sigma_{tk}^{-1})$
        $\overline{\theta}_{tk} = \widehat{\theta}_{tk} + \overline{\xi}_{tk}$
    end for
    Execute $\overline{\pi}_{tk}(s) = \arg\max_a \overline{Q}_{tk}(s,a)$, see Def. 1
    Collect trajectories of $(s_{tk}, a_{tk}, r_{tk})$ for $t \in [H]$.
    Update $\Sigma_{t,k+1} = \Sigma_{tk} + \phi_{tk}\phi_{tk}^\top$ for $t \in [H]$
end for
```

4 Main Result

We present the first frequentist regret bound for a TS-based algorithm in MDPs with approximate linear reward response and low-rank transition dynamics:

Theorem 1. *Assume Asm. 1 and set the algorithm parameters $\lambda = 1$, $\sigma = \sqrt{H\nu_k(\delta)} = \sqrt{H}(\widetilde{O}(Hd) + L_\phi(3HL_\psi + L_r) + 4\epsilon H\sqrt{dk})$, $\alpha_U = 1/\widetilde{O}(\sigma\sqrt{d})$, and $\alpha_L = \alpha_U/2$ (full definitions with the log terms can be found in App. D). Then for any $0 < \delta < \Phi(-1)/2$, with probability at least $1-\delta$ the regret of* OPT-RLSVI *is bounded jointly for all episodes k up to K by:*

$$\widetilde{O}\left(\sigma dH\sqrt{K} + \frac{H^2 d}{\alpha_L^2} + \epsilon H^2 K\right). \tag{12}$$

If we further assume that $L_\phi = \widetilde{O}(1)$ and $L_r, L_\psi = \widetilde{O}(d)$, then the bound reduces to

$$\widetilde{O}\left(H^2 d^2\sqrt{T} + H^5 d^4 + \epsilon dH(1 + \epsilon dH^2)T\right). \tag{13}$$

For the setting of low-rank MDPs a lower bound is currently missing both in terms of statistical rate and regarding the misspecification. Recently, Du et al. (2019); Lattimore and Szepesvari (2019); Van Roy and Dong (2019) discuss what's possible to achieve regarding the misspecification level.

For finite action spaces OPT-RLSVI can be implemented efficiently in space $O(d^2H + dAHK)$ and time $O(d^2AHK^2)$ where A is the number of actions (Prop. H.1 in appendix).

It is useful to compare our result with Yang and Wang (2019a) and Jin et al. (2019) which study a similar setting but with an approach based on deterministic optimism, and with Russo (2019) which proves worst-case regret bounds of RLSVI for tabular representations.

Comparison with Yang and Wang (2019a). Recently, Yang and Wang (2019a) studied exploration in finite state-spaces and low-rank transitions. They define a model-based algorithm that tries to learn the "core matrix", defined as the middle factor of a three-factor low-rank factorization. While their regularity assumptions on the parameters do not immediately fit in our framework, an important distinction (beyond model-based vs model-free) is that their algorithm potentially needs to compute the value function across all states. This suffers $\Omega(S)$ computational complexity and cannot directly handle continuous state spaces.

Comparison with Jin et al. (2019). A more direct comparison can be done with Jin et al. (2019) which is based on least-square value iteration (like OPT-RLSVI) and uses the same setting as we do when $L_r = L_\psi = \sqrt{d}$ and $L_\phi = 1$. In that case we get the regret in Eqn. (13) which is $\sqrt{Hd}$-times worse in the leading term than Jin et al. (2019).

In terms of feature dimension d, this matches the $\sqrt{d}$ gap in linear bandits between the best bounds for a TS-based algorithm (with regret $\widetilde{O}(d^{3/2}\sqrt{T})$) (Abeille et al., 2017) and the best bounds for an optimistic algorithm (with regret $\widetilde{O}(d\sqrt{T})$) (Abbasi-Yadkori et al., 2011). This happens because the proof techniques for Thompson sampling require the perturbations to have sufficient variance to guarantee optimism (and thus exploration) with some probability. For a geometric interpretation of this, see Abeille et al. (2017). For H-horizon MDPs, the total system dimensionality is dH, and therefore the extra $\sqrt{dH}$ factor is expected.

Comparison with Russo (2019). Recently, Russo (2019) has analyzed RLSVI in tabular finite horizon MDPs. While the core algorithm is similar, function approximation does introduce challenges that required changing RLSVI by, e.g., introducing the default function. While in Russo (2019) the value function can be bounded in high probability thanks to the non-expansiveness of the Bellman operator associated to the estimated model, in our case this has to be handled explicitly. We think that the use of a default optimistic value function could yield better horizon dependence for RLSVI in tabular settings, though this would require changing the algorithm.

5 Proof Outline

In this section we outline the proof of our regret bound for OPT-RLSVI. The four main ingredients are: 1) a one-step expansion of the action-value function difference $\overline{Q}_{tk} - Q_t^{\pi_k}$ in terms of the next-state value function difference; 2) a high probability bound on the noise and pseudonoise; 3) showing that the algorithm is optimistic with constant probability; 4) combining to get the regret bound.For the sake of clarity, we will assume no misspecification ($\epsilon = 0$), no regularization ($\lambda = 0$), and a nonsinigular design matrix $\Sigma_{tk} = \sum_{i=1}^{k-1}\phi_{ti}\phi_{ti}^\top$. The complete proof is reported in the appendix.

5.1 One-Step Analysis of Q functions

In this section we do a "one-step" analysis to decompose the difference in Q functions in the case where $\|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} \leqslant \alpha_L$ so that $\overline{Q}_{tk}$ is linear in the features. The decomposition has three parts: environment noise, pseudonoise, and the difference in value functions at step $t+1$. It reads $(\overline{Q}_{tk} - Q_t^\pi)(s,a) =$

$$\phi_t(s,a)^\top(\overline{\eta}_{tk} + \overline{\xi}_{tk}) + \mathbb{E}_{s'|s,a}(\overline{V}_{t+1,k} - V_{t+1}^\pi)(s') \tag{14}$$

where $\overline{\eta}_{tk}$ is the projected environment noise defined below in Eqn. (18). The complete version of the decomposition is Lem. C.1 in the appendix, while here we give an informal proof sketch of this fact.

First, since we are assuming that $\|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} \leqslant \alpha_L$

and $\epsilon = 0$, we can apply Def. 1 and Prop. 1 to write:

$$(\overline{Q}_{tk} - Q^{\pi}_{t})(s,a) = \phi_t(s,a)^\top(\overline{\theta}_{tk} - \theta^{\pi}_{t}). \tag{15}$$

Decomposing $\overline{\theta}_{tk} = \widehat{\theta}_{tk} + \overline{\xi}_{tk}$ immediately shows how the pseudonoise $\overline{\xi}_{tk}$ appears in Eqn. (14). Now we need to handle the regression term:

$$\widehat{\theta}_{tk} \overset{def}{=} \Sigma^{-1}_{tk} \sum_{i=1}^{k-1} \phi_{ti}(r_{ti} + \overline{V}_{t+1,k}(s_{t+1,i})). \tag{16}$$

To handle this, we need to make an expectation over s' given s_{ti}, a_{ti} (the experienced state and action in timestep t of episode i) appear in each term of the sum so that the value function term will become linear in ϕ_{ti}. To do this, we define the one-step environment noise with respect to $\overline{V}_{t+1,k}$ as

$$\overline{\eta}_{tk}(i) \overset{def}{=} \overline{V}_{t+1,k}(s_{t+1,i}) - \mathbb{E}_{s'|s_{ti},a_{ti}}[\overline{V}_{t+1,k}(s')], \tag{17}$$

Then we define the projected environment noise as:

$$\overline{\eta}_{tk} \overset{def}{=} \Sigma^{-1}_{tk} \sum_{i=1}^{k-1} \phi_{ti}\overline{\eta}_{tk}(i). \tag{18}$$

Putting this into the definition of $\widehat{\theta}_{tk}$ from Eqn. (16),

$$\begin{aligned}\widehat{\theta}_{tk} &= \Sigma^{-1}_{tk} \sum_{i=1}^{k-1} \phi_{ti}(r_{ti} + \mathbb{E}_{s'|s_{ti},a_{ti}}[\overline{V}_{t+1,k}(s')] + \overline{\eta}_{tk}(i)) \\ &= \overline{\eta}_{tk} + \Sigma^{-1}_{tk} \sum_{i=1}^{k-1} \phi_{ti}(r_{ti} + \mathbb{E}_{s'|s_{ti},a_{ti}}[\overline{V}_{t+1,k}(s')]).\end{aligned}$$

But now we note that this reward plus expected value function is linear in the features (thanks to Prop. 1), so we can rewrite the second term as

$$\Sigma^{-1}_{tk} \sum_{i=1}^{k-1} \phi_{ti}\phi_{ti}^\top \left(\theta^r + \int_{s'} \psi_t(s')\overline{V}_{t+1,k}(s')\right) \tag{19}$$

$$= \theta^r + \int_{s'} \psi_t(s')\overline{V}_{t+1,k}(s'). \tag{20}$$

Finally, comparing with the definition of θ^{π}_{t} (Eqn. (8)) we see that the θ^r terms cancel and we get

$$\overline{\theta}_{tk} - \theta^{\pi}_{t} = \overline{\xi}_{tk} + \overline{\eta}_{tk} + \int_{s'} \psi_t(s')(\overline{V}_{t+1,k} - V^{\pi}_{t+1})(s').$$

Premultiplying by $\phi_t(s,a)^\top$ gives Eqn. (14).

5.2 High probability bounds on the noise

To ensure that our estimates concentrate around the true Q functions, we need to ensure that the $\overline{\eta}_{tk}$ and $\overline{\xi}_{tk}$ are not too large. This is achieved with similar ideas of self-normalizing processes as is done for linear bandits (Abbasi-Yadkori et al., 2011), with an additional union bound over possible value functions $\overline{V}_{t+1,k}$ which depend on $\overline{\theta}_{tk}$ and Σ^{-1}_{tk}. In the end, we prove in Lem. E.6 that indeed with high probability for any ϕ:

$$|\phi^\top \overline{\eta}_{tk}| \leqslant \|\phi\|_{\Sigma^{-1}_{tk}} \|\overline{\eta}_{tk}\|_{\Sigma_{tk}} = \sqrt{\nu_k(\delta)}\|\phi\|_{\Sigma^{-1}_{tk}} \tag{21}$$

where $\sqrt{\nu_k(\delta)} = \widetilde{O}(dH)$ is defined fully in App. D. While we defer the computation of the "right" amount of pseudonoise to the next subsection, here we mention that for the choice we make $\overline{\xi}_{tk} \sim \mathcal{N}(0, H\nu_k(\delta)\Sigma^{-1}_{tk})$ we obtain w.h.p.:

$$|\phi^\top \overline{\xi}_{tk}| \leqslant \|\phi\|_{\Sigma^{-1}_{tk}} \|\overline{\xi}_{tk}\|_{\Sigma_{tk}} = \sqrt{\gamma_k(\delta)}\|\phi\|_{\Sigma^{-1}_{tk}} \tag{22}$$

where $\sqrt{\gamma_k(\delta)} = \widetilde{O}((dH)^{3/2})$ is also defined fully in App. D. Note the pseudonoise *worst-case* bound is $\sqrt{Hd}$ worse than the corresponding environment noise.

5.3 Stochastic Optimism and Random Walk

We now want to show that OPT-RLSVI injects enough pseudonoise that the estimated value function $\overline{V}_{1k}(s_{1k})$ at the initial state s_{1k} is optimistic with constant probability (see App. F). We call this event $\mathcal{O}_k$:

$$\mathcal{O}_k \overset{def}{=} \left\{(\overline{V}_{1k} - V^\star_1)(s_{1k}) \geqslant 0\right\}. \tag{23}$$

Note that the optimal policy $\pi^\star$ maximizes $Q^\star$ and not the $\overline{Q}$ computed by the algorithm and thus

$$(\overline{V}_{1k} - V^\star_1)(s_{1k}) \geqslant (\overline{Q}_{1k} - Q^\star_1)(s_{1k}, \pi^\star_1(s_{1k})). \tag{24}$$

Now, the goal is to leverage Eqn. (14) to inductively expand this inequality by unrolling a trajectory under the policy $\pi^\star$. To access the result in Eqn. (14) we need to have $\|\phi_1(s_{1k}, \pi^\star_1(s_{1k}))\|_{\Sigma^{-1}_{1k}} \leqslant \alpha_L$. For now, we just assume that this is the case to motivate the idea. In that case, applying Eqn. (14) gives us

$$\begin{aligned}\left(\overline{V}_{1k} - V^\star_1\right)(s_{1k}) \geqslant{}& \phi_1(s_{1k}, \pi^\star_1(s_{1k}))^\top \left(\overline{\xi}_{1k} + \overline{\eta}_{1k}\right) \\ &+ \mathbb{E}_{s'|s_{1k},\pi^\star_1(s_{1k})}[\left(\overline{V}_{2k} - V^\star_2\right)(s')].\end{aligned} \tag{25}$$

Now we can inductively apply the same reasoning to the term inside of the expectation (assuming that we always get features with small Σ^{-1}-norm). Using x_t to denote the states sampled under $\pi^\star$ to avoid confusion with s_{tk} observed by the algorithm, we get

$$\geqslant \sum_{t=1}^{H} \mathbb{E}_{x_t \sim \pi^\star | s_{1k}} \left[\phi_t(x_t, \pi^\star_t(x_t))^\top(\overline{\xi}_{tk} + \overline{\eta}_{tk})\right] \tag{26}$$

Since these trajectories over x come from $\pi^\star$ and the environment, they do not depend on the algorithm's policy and with respect to the pseudonoise

$\overline{\xi}$, they are non-random. If we let $\phi_t^\star$ denote $\mathbb{E}_{x_t \sim \pi^\star | s_{1k}} \phi_t(x_t, \pi^\star(x_t))$, and apply Eqn. (21) we get with probability at least $1-\delta$ that:

$$\sum_{t=1}^{H} (\phi_t^\star)^\top (\overline{\xi}_{tk} + \overline{\eta}_{tk}) \geqslant \sum_{t=1}^{H} [(\phi_t^\star)^\top \overline{\xi}_{tk} - \sqrt{\nu_k(\delta)} \|\phi_t^\star\|_{\Sigma_{tk}^{-1}}]$$

$$\geqslant \sum_{t=1}^{H} (\phi_t^\star)^\top \overline{\xi}_{tk} - \sqrt{H \nu_k(\delta)} \left(\sum_{t=1}^{H} \|\phi_t^\star\|^2_{\Sigma_{tk}^{-1}} \right)^{1/2} \tag{27}$$

where the second inequality is Cauchy-Schwarz.

Note that the only randomness in this quantity comes from the pseudonoise we inject. We can think of this sum as a one-dimensional normal random walk over H steps with a negative bias. Moreover, if we chose each $\overline{\xi}_{tk} \sim \mathcal{N}(0, H\nu_k(\delta)\Sigma_{tk}^{-1})$, we know that

$$\sum_{t=1}^{H} (\phi_t^\star)^\top \overline{\xi}_{tk} \sim \mathcal{N}\left(0, \sum_{t=1}^{H} H\nu_k(\delta) \|\phi_t^\star\|^2_{\Sigma_{tk}^{-1}}\right). \tag{28}$$

Comparing this with Eqn. (27) we can immediately see that the standard deviation of the sum of pseudonoise terms is exactly the bound on the bias induced by the high probability bound on the sum of the environment noise $\overline{\eta}_{tk}$. Thus we can conclude that

$$\mathbf{P}\left((\overline{V}_{1k} - V_1^\star)(s_{1k}) \geqslant 0\right) \geqslant \Phi(-1) \tag{29}$$

where Φ is the normal CDF. This is just the result that we are looking for. However, this presentation avoided the technicalities of handling the cases where $\|\phi_t(x_t, \pi_t^\star(x_t))\|_{\Sigma_{tk}^{-1}} > \alpha_L$ and $\overline{Q}_{tk}$ takes the default value. At a high level the default value is optimistic and so it cannot reduce the probability of optimism. This is handled carefully in Lem. F.1 and F.2 of the appendix, where we obtain a recursion structurally similar to Eqn. (27) albeit with a less interpretable definition of $\phi_t^\star$. One important detail is that our choice of when to default does not depend on the $\overline{\xi}_{tk}$ and is thus non-random with respect to the pseudonoise.

5.4 High Probability Regret Bound

In this section we provide a high level sketch of the main argument that allows us to obtain a high probability regret bound for OPT-RLSVI under Asm. 1. In particular, we assume that the "good event" holds, which lets us use the bounds in Eqn. (21) and (22).

First, we recall the definition of regret up to episode K from the preliminaries and further add and subtract the randomized value functions $\overline{V}_{1k}$ to get that Regret(K) decomposes as

$$\sum_{k=1}^{K} \Big(\underbrace{V_1^\star - \overline{V}_{1k}}_{\text{Pessimism}} + \underbrace{\overline{V}_{1k} - V_1^{\pi_k}}_{\text{Estimation}} \Big)(s_{1k}) \tag{30}$$

5.4.1 Bound on estimation

We need to distinguish between cases where $\|\phi_{tk}\|_{\Sigma_{tk}^{-1}} \leqslant \alpha_L$, which we will denote by $\mathcal{S}_{tk}$ for small feature, or not, which we will denote by $\mathcal{S}_{tk}^c$ for its complement. Under $\mathcal{S}_{tk}$ linearity of the representation can be used via Eqn. (14) and under $\mathcal{S}_{tk}^c$ we can use the trivial upper bound of H on the difference in values:

$$\left(\overline{V}_{1k} - V_1^{\pi_k}\right)(s_{1k}) \leqslant H \mathbb{1}\left\{\mathcal{S}_{1k}^c\right\} + \tag{31}$$
$$\left(\phi_{1k}^\top \left(\overline{\xi}_{1k} + \overline{\eta}_{1k}\right) + \underbrace{\mathbb{E}_{s'|s_{1k},a_{1k}}[\left(\overline{V}_{2k} - V_2^{\pi_k}\right)(s')]}_{=\dot{\zeta}_{1k} + (\overline{V}_{2k} - V_2^{\pi_k})(s_{2k})}\right) \mathbb{1}\left\{\mathcal{S}_{1k}\right\}$$

where $\dot{\zeta}_{tk} \overset{def}{=} \mathbb{1}\left\{\mathcal{S}_{1k}\right\}\big(\mathbb{E}_{s'|s_{tk},a_{tk}} \left(\overline{V}_{t+1,k} - V_{t+1}^{\pi_k}\right)(s') - (\overline{V}_{t+1,k} - V_{t+1}^{\pi_k})(s_{t+1,k})\big)$ is a bounded martingale difference sequence on the good event. Induction and summing over k eventually yields:

$$\leqslant \sum_{k=1}^{K} \sum_{t=1}^{H} \underbrace{H \mathbb{1}\left\{\mathcal{S}_{tk}^c\right\}}_{\text{Warmup}} + \underbrace{\phi_{tk}^\top \left(\overline{\xi}_{tk} + \overline{\eta}_{tk}\right) \mathbb{1}\left\{\mathcal{S}_{tk}\right\}}_{\text{Linear Regime}} + \underbrace{\dot{\zeta}_{tk} \mathbb{1}\left\{\mathcal{S}_{tk}\right\}}_{\text{Martingale}}.$$

The martingale term can be bounded with high probability by $\widetilde{O}(H\sqrt{T})$ using Azuma-Hoeffding.

The first term measures regret during "warmup", when the algorithm cannot guarantee that the value function estimates are bounded and needs to use the default function. In Lem. G.5 we bound it and obtain:

$$\widetilde{O}\left(\frac{H^2 d}{\alpha_L^2}\right) = \widetilde{O}\left(H^5 d^4\right) \tag{32}$$

which is $\sqrt{T}$-free and is thus a lower order term.

For the dominant linear regime term we can use the high probability bounds from Eqn. (21) and (22) along with two applications of Cauchy-Schwarz:

$$\leqslant \sum_{k=1}^{K} \sum_{t=1}^{H} \|\phi_{tk}\|_{\Sigma_{tk}^{-1}} \left(\sqrt{\gamma_k(\delta)} + \sqrt{\nu_k(\delta)}\right) \tag{33}$$

$$\leqslant \sqrt{K} \times \sum_{t=1}^{H} \underbrace{\sqrt{\sum_{k=1}^{K} \|\phi_{tk}\|^2_{\Sigma_{tk}^{-1}}}}_{\widetilde{O}(\sqrt{d})} \times \Big(\underbrace{\sqrt{\gamma_K(\delta)}}_{\widetilde{O}(H^{3/2} d^{3/2})} + \underbrace{\sqrt{\nu_K(\delta)}}_{\widetilde{O}(Hd)} \Big)$$

This final bound on the sum of the squared norm of the features is a standard quantity that arises in linear bandit computations (Abbasi-Yadkori et al., 2011). We can see that the estimation term gives the same regret bound reported in the Thm. 1. Now we show that the pessimism term is of the same order.

5.4.2 Bound on Pessimism

For optimistic algorithms the pessimism term of the regret $\sum_{k=1}^{K} (V_1^\star - \overline{V}_{1k}^{\pi_k})(s_{1k})$ is negative by construction; here we need to work a little more. As seen above, the algorithm has at least a *constant* probability of being optimistic. When it is, it makes progress similar to a deterministic optimistic algorithm, and when it is not, it is still choosing a reasonable policy (using shrinking confidence intervals) so that the mistakes it makes become less and less severe. Ultimately, we would like to transform the pessimism term into an estimation argument that we can handle as before. So, we first upper bound $V_1^\star$ and then lower bound $\overline{V}_{1k}$ by randomized value functions with specific choices for the pseudonoise. As more samples are collected, the pseudonoise shrinks and the estimates converge.

Upper Bound on $V_1^\star$. Consider drawing $\widetilde{\xi}_{tk}$'s defined as independent and identically distributed copies of the $\overline{\xi}_{tk}$'s. Let $\widetilde{\mathcal{O}}_k$ be the event that in episode k the algorithm obtains an optimistic value function $\widetilde{V}_{1k}$ using these $\widetilde{\xi}_{tk}$ in place of $\overline{\xi}_{tk}$. Explicitly,

$$\widetilde{\mathcal{O}}_k = \{(\widetilde{V}_{1k} - V_{1k}^\star)(s_{1k}) \geqslant 0\}. \tag{34}$$

Note that since the $\widetilde{\xi}_{tk}$ are iid copies of the $\overline{\xi}_{tk}$ we have that $\mathbf{P}(\widetilde{\mathcal{O}}_k)$ is equal to $\mathbf{P}(\mathcal{O}_k) = \Phi(-1)$ from Sec. 5.3. Taking conditional expectation $\mathbb{E}_{\widetilde{\xi}|\widetilde{\mathcal{O}}_k}$ over the $\widetilde{\xi}_{tk}$ for $t \in [H]$ gives us an upper bound:

$$V_{1k}^\star(s_{1k}) \leqslant \mathbb{E}_{\widetilde{\xi}|\widetilde{\mathcal{O}}_k} \widetilde{V}_{1k}(s_{1k}) \tag{35}$$

by definition of the event $\widetilde{\mathcal{O}}_k$.

Lower Bound on $\overline{V}_{1k}$. Under the high probability bound on the pseudonoise of Eqn. (22) we consider the below optimization program over the *optimization variables* ξ_{tk}'s, which are constrained to satisfy the same bound on the pseudonoise of Eqn. (22):

$$\min_{\{\xi_{tk}\}_{t=1,\dots,H}} V_{1k}^{\xi}(s_{1k}) \tag{36}$$
$$\|\xi_{tk}\|_{\Sigma_{tk}} \leqslant \sqrt{\gamma_k(\delta)}, \quad \forall t \in [H]$$

where V_{1k}^{ξ} is analogous to $\overline{V}_{1k}$ derived from our algorithm, but with the optimization variables ξ_{tk} in place of $\overline{\xi}_{tk}$. Solving the program above would give a value function $\underline{V}_{1k}$ such that:

$$\underline{V}_{1k}(s_{1k}) \leqslant \overline{V}_{1k}(s_{1k}) \tag{37}$$

whenever the $\overline{\xi}_{tk}$'s obey the high probability bound.

Putting it together. Now we chain the upper bound of Eqn. (35) with the lower bound of Eqn. (37):

$$\left(V_{1k}^\star - \overline{V}_{1k}\right)(s_{1k}) \leqslant \mathbb{E}_{\widetilde{\xi}|\widetilde{\mathcal{O}}_k}[(\widetilde{V}_{1k} - \underline{V}_{1k})(s_{1k})]. \tag{38}$$

We can connect this conditional expectation with the probability of optimism to get to a concentration bound by applying the law of total expectation:

$$\begin{aligned}\mathbb{E}_{\widetilde{\xi}}[(\widetilde{V}_{1k}-\underline{V}_{1k})(s_{1k})] &= \mathbb{E}_{\widetilde{\xi}|\widetilde{\mathcal{O}}_k}[(\widetilde{V}_{1k} - \underline{V}_{1k})(s_{1k})]\,\mathbf{P}(\widetilde{\mathcal{O}}_k) \\ &\quad + \underbrace{\mathbb{E}_{\widetilde{\xi}|\widetilde{\mathcal{O}}_k^c}[(\widetilde{V}_{1k} - \underline{V}_{1k})(s_{1k})]\,\mathbf{P}(\widetilde{\mathcal{O}}_k^c)}_{\geqslant 0}.\end{aligned} \tag{39}$$

This inequality holds by the same reasoning as Eqn. (37) with high probability since the $\widetilde{\xi}_{tk}$ are also in the set over which $\underline{V}_{1k}$ is minimized. Dividing by $\mathbf{P}(\widetilde{\mathcal{O}}_k)$ and chaining with Eqn. (38) gives us:

$$\left(V_{1k}^\star - \overline{V}_{1k}\right)(s_{1k}) \leqslant \mathbb{E}_{\widetilde{\xi}}[(\widetilde{V}_{1k} - \underline{V}_{1k})(s_{1k})]/\,\mathbf{P}(\widetilde{\mathcal{O}}_k).$$

Now, since the $\widetilde{\xi}_{tk}$ are iid copies of the $\overline{\xi}_{tk}$ that the algorithm computes we have that $\mathbb{E}_{\widetilde{\xi}}[\widetilde{V}_{1k}(s_{1k})] = \mathbb{E}_{\overline{\xi}}[\overline{V}_{1k}(s_{1k})]$ and $\mathbf{P}(\mathcal{O}_k) = \mathbf{P}(\widetilde{\mathcal{O}}_k)$. So we can define a martingale difference sequence $\ddot{\zeta}_k \overset{def}{=} \mathbb{E}_{\widetilde{\xi}}[\widetilde{V}_{1k}(s_{1k})] - \overline{V}_{1k}(s_{1k})$ and get our final bound on the pessimism as:

$$\left(V_{1k}^\star - \overline{V}_{1k}\right)(s_{1k}) \leqslant \frac{(\overline{V}_{1k} - \underline{V}_{1k})(s_{1k}) + \ddot{\zeta}_k}{\mathbf{P}(\mathcal{O}_k)}. \tag{40}$$

When summing over the episodes $k \in [K]$, the martingale can be bounded with high probability by Azuma-Hoeffding as $\sum_{k=1}^{K} \ddot{\zeta}_k = \widetilde{O}(H\sqrt{K})$. To bound the remaining term we add and subtract $V_1^{\pi_k}$ to get:

$$\left(\sum_{k=1}^{K}[(\overline{V}_{1k} - V_1^{\pi_k})(s_{1k}) + (V_1^{\pi_k} - \underline{V}_{1k})(s_{1k})]\right)/\,\mathbf{P}(\mathcal{O}_k).$$

Each of these is bounded by arguments similar to those in Sec. 5.4.1. We discuss this in detail in Lem. G.4.

It is instructive to re-examine Eqn. (40), ignoring the martingale term. While the left hand side is negative for optimistic algorithms, for OPT-RLSVI it is upper bounded by a difference in estimated value functions (which shrinks with more data) times the inverse probability of being optimistic $1/\mathbf{P}(\mathcal{O}_k)$. In other words, roughly once every $1/\mathbf{P}(\mathcal{O}_k)$ episodes the algorithm is optimistic and exploration progress is made.

6 Concluding Remarks

This work proposes the first high probability regret bounds for (a modified version of) RLSVI with function approximation, confirming its sound exploration principles. Perhaps unsurprisingly, we inherit an extra $\sqrt{dH}$ regret factor compared to an optimistic approach which can be explained by analogy to the bandit literature. Whether Thompson sampling-based algorithms need to suffer this extra factor compared to their optimistic counterparts remains a fundamental research question in exploration. Our work enriches the literature on provably efficient exploration algorithms with function approximation with a new algorithmic design as well as a new set of analytical techniques.

Acknowledgments

Andrea Zanette is partially supported by the Total Innovation Fellowship program. David Brandfonbrener is supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

References

Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

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

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In *AISTATS*, volume 31 of *JMLR Workshop and Conference Proceedings*, pages 99–107. JMLR.org, 2013.

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 1184–1194. Curran Associates, Inc., 2017.

Mohammad Gheshlaghi Azar, Ian Osband, and Remi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2017.

Ronen I. Brafman and Moshe Tennenholtz. R-MAX - A general polynomial time algorithm for near-optimal reinforcement learning. *J. Mach. Learn. Res.*, 3:213–231, 2002.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *NIPS*, pages 2249–2257, 2011.

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

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient pac rl with rich observations. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1429–1439, 2018.

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

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. 2018.

Gene H Golub and Charles F Van Loan. *Matrix Computations*. JHU Press, 2012.

Aditya Gopalan and Shie Mannor. Thompson sampling for learning parameterized markov decision processes. In *COLT*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 861–898. JMLR.org, 2015.

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

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In Doina Precup and Yee Whye Teh, editors, *International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL `http://proceedings.mlr.press/v70/jiang17c.html`.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*, 2019.

K. Lakshmanan, Ronald Ortner, and Daniil Ryabko. Improved regret bounds for undiscounted continuous reinforcement learning. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 524–532. JMLR.org, 2015.

Tor Lattimore and Csaba Szepesvari. Learning with good feature representations in bandits and in rl with a generative model. *arXiv preprint arXiv:1911.07676*, 2019.

Rémi Munos. Error bounds for approximate value iteration. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2005.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(May):815–857, 2008.

Ronald Ortner and Daniil Ryabko. Online regret bounds for undiscounted continuous reinforcement learning. In *NIPS*, pages 1772–1780, 2012.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International Conference on Machine Learning (ICML)*, 2017.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped DQN. In *Advances in Neural Information Processing Systems (NIPS)*, 2016a.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning (ICML)*, 2016b.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown markov decision processes: A thompson sampling approach. In *NIPS*, pages 1333–1342, 2017.

David Pollard. Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–86. JSTOR, 1990.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons, Inc., New York, NY, USA, 1994. ISBN 0471619779.

Daniel Russo. Worst-case regret bounds for exploration via randomized value functions. *arXiv preprint arXiv:1906.02870*, 2019.

Steven E Shreve and Dimitri P Bertsekas. Alternative theoretical frameworks for finite horizon discrete-time stochastic optimal control. *SIAM Journal on control and optimization*, 16(6):953–978, 1978.

Benjamin Van Roy and Shi Dong. Comments on the du-kakade-wang-yang lower bounds. *arXiv preprint arXiv:1911.07910*, 2019.

Lin F Yang and Mengdi Wang. Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. *arXiv preprint arXiv:1905.10389*, 2019a.

Lin F Yang and Mengdi Wang. Sample-optimal parametric q-learning with linear transition models. *arXiv preprint arXiv:1902.04779*, 2019b.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning (ICML)*, 2019. URL `http://proceedings.mlr.press/v97/zanette19a.html`.