BooVI: Provably Efficient Bootstrapped Value Iteration

Boyi Liu[*] **Qi Cai**[†] **Zhuoran Yang**[‡] **Zhaoran Wang**[§]

Abstract

Despite the tremendous success of reinforcement learning (RL) with function approximation, efficient exploration remains a significant challenge, both practically and theoretically. In particular, existing theoretically grounded RL algorithms based on upper confidence bounds (UCBs), such as optimistic least-squares value iteration (LSVI), are often incompatible with practically powerful function approximators, such as neural networks. In this paper, we develop a variant of bootstrapped LSVI, namely BooVI, which bridges such a gap between practice and theory. Practically, BooVI drives exploration through (re)sampling, making it compatible with general function approximators. Theoretically, BooVI inherits the worst-case $\widetilde{O}(\sqrt{d^3 H^3 T})$-regret of optimistic LSVI in the episodic linear setting. Here d is the feature dimension, H is the episode horizon, and T is the total number of steps.

1 Introduction

Reinforcement learning (RL) with function approximation demonstrates significant empirical success in a broad range of applications [e.g., 16, 38, 40, 45]. However, computationally and statistically efficient exploration of large and intricate state spaces remains a major barrier. Practically, the lack of temporally-extended exploration [31, 28, 30] in existing RL algorithms, e.g., deterministic policy gradient [39] and soft actor-critic [19], hinders them from solving more challenging tasks, e.g., Montezuma's Revenge in Atari Games [42, 17], in a more sample-efficient manner. Theoretically, it remains unclear how to design provably efficient RL algorithms with finite sample complexities or regrets that allow for practically powerful function approximators, e.g., neural networks.

There exist two principled approaches to efficient exploration in RL, namely optimism in the face of uncertainty [e.g., 4, 41, 21, 11, 12, 5, 23, 24] and posterior sampling [e.g., 29, 34, 33, 30, 27, 36].

- The optimism-based approach is often instantiated by incorporating upper confidence bounds (UCBs) into the estimated (action-)value functions as bonuses, which are used to direct exploration. In the tabular setting, the resulting algorithms, e.g., variants of upper confidence bound reinforcement learning (UCRL) [4, 21, 11, 5], are known to attain the optimal worst-case regret [32]. Beyond the tabular setting, it remains unclear how to construct closed-form UCBs in a principled manner for general function approximators, e.g., neural networks. The only exception is the linear setting [51, 52, 24, 8], where optimistic least-squares value iteration (LSVI) [24] is known to attain a near-optimal (with respect to $T = KH$) worst-case regret. However, the closed-form UCB therein is tailored to linear models [1, 9] rather than fully general-purpose.

[*]Northwestern University; boyiliu2018@u.northwestern.edu
[†]Northwestern University; qicai2022@u.northwestern.edu
[‡]Princeton University; zy6@princeton.edu
[§]Northwestern University; zhaoranwang@gmail.com

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

- On the other hand, the posterior-based approach, which originates from Thompson sampling [43, 37], can be instantiated using randomized (action-)value functions [34, 30, 27, 36]. Unlike optimistic LSVI, the resulting algorithm, namely randomized LSVI, straightforwardly allows for general function approximators, as it only requires injecting random noise into the training data of LSVI. Ideally, such injected random noise does not depend on particular function approximators. In the tabular setting, randomized LSVI is known to attain near-optimal worst-case and Bayesian regrets [30, 36]. Meanwhile, in the linear setting, randomized LSVI is very recently shown to attain a near-optimal worst-case regret [54], which is only worse than that of optimistic LSVI by a factor of $\sqrt{dH}$. Here d is the feature dimension and H is the episode horizon. However, achieving such a regret requires a nontrivial modification of the injected random noise, which is tailored to linear models. Such a specialized modification diminishes the supposed practical advantage of randomized LSVI.

To bridge such a gap between practice and theory, we aim to answer the following question: Can we design an RL algorithm that simultaneously achieves the practical advantage of randomized LSVI and the theoretical guarantee of optimistic LSVI?

In this paper, we propose a variant of bootstrapped LSVI, namely BooVI, which combines the advantages of the optimism-based and posterior-based approaches. The key idea of BooVI is to use posterior sampling to implicitly construct an "optimistic version" of the estimated (action-)value functions in a data-driven manner. Unlike randomized LSVI, which samples from the posterior only once, e.g., by injecting random noise into the training data of LSVI, BooVI samples from the posterior multiple times, e.g., via the Langevin dynamics [48, 35]. Upon evaluating an action at a state, BooVI ranks the values of the randomized (action-)value functions sampled from the posterior in descending order and returns a top-ranked value, which can be shown to be approximately optimistic. Generally speaking, BooVI corresponds to bootstrapping the noise in the least-squares regression problem of LSVI. As a result, it can be viewed as a parametric bootstrap counterpart of the nonparametric bootstrap technique used in bootstrapped deep Q-networks (DQNs) [31, 30], which demonstrates significant empirical success in terms of exploration.

Compared with existing RL algorithms with function approximation, the advantage of BooVI is twofold:

- Practically, BooVI bypasses the explicit construction of the closed-form UCB in optimistic LSVI. Like randomized LSVI, BooVI straightforwardly allows for general function approximators, as it only requires injecting random noise into the training process of LSVI, e.g., via the Langevin dynamics.
- Theoretically, BooVI inherits the worst-case $\widetilde{O}(\sqrt{d^3H^3T})$-regret of optimistic LSVI in the linear setting. Here d is the feature dimension, H is the episode horizon, and T is the total number of steps. Such a regret is better than the best known worst-case regret of randomized LSVI by a factor of $\sqrt{dH}$. More importantly, unlike the specialized variant of randomized LSVI studied in [54], BooVI can be applied to the linear setting "as is", without tailoring the injected random noise to linear models.

More Related Work: The idea of using posterior sampling to achieve optimism in a data-driven manner is previously studied in the linear bandit setting [26] and the tabular setting of RL [3]. In fact, our linear setting of RL covers the linear bandit setting as a special case, where the episode horizon H is set to one, and the tabular setting of RL as another special case, where the feature mapping is the canonical basis of the state and action spaces. Although BooVI is practically applicable to general function approximators, our theoretical guarantee on its worst-case regret is only applicable to the linear setting. Despite the recent progress [49, 22, 13, 15], simultaneously achieving provable computational and statistical efficiency in exploration with general function approximators remains challenging. See, e.g., [2, 18] for the recent progress in the contextual bandit setting. Finally, we refer readers to [46, 53, 57, 55] for theories on the generalized models of linear MDPs, which all share similar (approximately) linear structure with the class of linear MDPs considered in the analysis of this paper. We expect our regret bound stays the same for those settings with linear structures, or admits an extra $O(\epsilon T)$ dependency over T for those settings with approximately linear structure, where $\epsilon > 0$ describes the level of nonlinearity in the MDP. We also refer readers to [50] for sharper regret bounds in linear MDPs, which are out of the scope of this paper.

2 Background

In this section, we introduce the general problem setting.

Episodic Markov Decision Process. We consider the episodic Markov decision process represented by a tuple $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r)$, where $\mathcal{S}$ is a compact state space, $\mathcal{A}$ is a finite action space with cardinality A, H is the number of timesteps in each episode, $\mathcal{P} = \{\mathcal{P}_h\}_{h\in[H]}$ with $\mathcal{P}_h : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ for all $h \in [H]$ is the set of transition kernels, and $r = \{r_h\}_{h\in[H]}$ with $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ for all $h \in [H]$ is the set of reward functions.

At each timestep $h \in [H]$ of episode k, an agent at state $s_h^k \in \mathcal{S}$ with policy $\pi^k = \{\pi_h^k\}_{h\in[H]}$, where $\pi_h^k : \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ for all $h \in [H]$, interacts with the environment by first taking an action a_h^k with probability $\pi_h^k(a_h^k \,|\, s_h^k)$ and then receiving the corresponding reward $r_h^k = r_h(s_h^k, a_h^k)$.

We evaluate the performance of policy $\pi = \{\pi_h\}_{h\in[H]}$ starting from timestep h, and state-action pair (s, a) using its action-value function $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, which is defined as

$$Q_h^\pi(s, a) = \mathbb{E}_\pi\biggl[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \,\bigg|\, s_h = s,\ a_h = a\biggr],$$

for all $h \in [H]$. Correspondingly, the value function $V_h^\pi : \mathcal{S} \to \mathbb{R}$ of a policy π is defined as

$$V_h^\pi(s) = \mathbb{E}_\pi\biggl[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \,\bigg|\, s_h = s\biggr] = \mathbb{E}_\pi\bigl[Q_h^\pi(s_h, a) \,\big|\, s_h = s\bigr],$$

for all $h \in [H]$. Here the expectation $\mathbb{E}_\pi[\,\cdot\,]$ is taken over the trajectory generated by π. Also, we let $Q_{H+1}^\pi \equiv 0$ and thus $V_{H+1}^\pi \equiv 0$. Furthermore, we denote by $V_h^*(s)$ the value function corresponding to the optimal policy π^*. Finally, for notational simplicity, we define $[\mathcal{P}_h V](s, a) = \mathbb{E}_{s'\sim\mathcal{P}_h(\,\cdot\,|\, s,a)}[V(s')]$, where $V : \mathcal{S} \to \mathbb{R}$ can be any function.

For any algorithm that generates a sequence of policies $\{\pi^k\}_{k\in[K]}$, we track its performance via the cumulative regret defined by

$$\text{Regret}(K) = \sum_{k=1}^{K}\bigl[V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)\bigr],$$

where K is the number of episodes. Here s_1^k is the initial state of episode k, which is arbitrarily chosen at the start of the episode.

3 Bootstrapped Value Iteration

In this section, we introduce Bootstrapped Value Iteration (BooVI in Algorithm 1). For notational simplicity, we write $r_h(s_h^k, a_h^k)$ as r_h^k throughout the rest of this paper. Also, in this paper, we write $\max\{\min\{\cdot,\cdot\}, 0\}$ as $\min\{\cdot,\cdot\}^+$, and denote by $\|\cdot\|$ the 2-norm for vectors.

Least-Squares Value Iteration. At timestep $h \in [H]$ of episode k, given the estimated action-value function $\widehat{Q}_{h+1}^k(s_{h+1}^\tau, a)$ for all $\tau \in [k-1]$ and $a \in \mathcal{A}$, let $\widehat{V}_{h+1}^k(s_{h+1}^\tau) = \max_{a\in\mathcal{A}} \widehat{Q}_{h+1}^k(s_{h+1}^\tau, a)$. Then, Least-Squares Value Iteration (LSVI) updates the parameter ω of the action-value function via

$$\widehat{\omega}_h^k \leftarrow \operatorname*{argmin}_{\omega\in\mathbb{R}^d}\biggl\{\lambda\cdot\|\omega\|^2 + \sum_{\tau=1}^{k-1}\bigl(r_h^\tau + \widehat{V}_{h+1}^k(s_{h+1}^\tau) - Q(s_h^\tau, a_h^\tau; \omega)\bigr)^2\biggr\}, \tag{3.1}$$

where $Q(\cdot,\cdot\,;\,\cdot) : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$ is the parameterization of action-value function with $\omega \in \mathbb{R}^d$ being its parameter, and $\lambda \geq 0$ is the regularization parameter. Although the deterministic parameter update by LSVI could exploit the historical data well, it has limited ability to address the exploration need in more challenging tasks.

Bootstrapped Value Iteration. To achieve guided exploration, we introduce BooVI (Algorithm 1), which uses bootstrapping to enforce the optimism of the estimated action-value function.

At the timestep h of episode k, given the current data buffer $\mathcal{D}_h^k = \{(s_h^\tau, a_h^\tau, r_h^\tau, s_{h+1}^\tau)\}_{\tau\in[k-1]}$, which contains the data collected from the previous $k-1$ episodes, and the current bootstrapped state values $\{\widetilde{V}_{h+1}^k(s_{h+1}^\tau)\}_{\tau\in[k-1]}$, we define

$$y_h^\tau = r_h^\tau + \widetilde{V}_{h+1}^k(s_{h+1}^\tau), \quad \text{for all } \tau \in [k-1]. \tag{3.2}$$

With $p_0(\omega)$ as prior of ω and $p(y_h^\tau \,|\, (s_h^\tau, a_h^\tau), \omega)$ as the likelihood of y_h^τ, the posterior of ω is given by

$$p\Big(\omega \,\Big|\, \big\{\widetilde{V}_{h+1}^k(s_{h+1}^\tau)\big\}_{\tau\in[k-1]}, \mathcal{D}_h^k\Big) \propto p_0(\omega) \cdot \prod_{\tau=1}^{k-1} p\big(y_h^\tau \,\big|\, (s_h^\tau, a_h^\tau), \omega\big). \tag{3.3}$$

In BooVI (Algorithm 1), we propose to sample repeatedly from such posterior for N_k times, which is equivalent to sampling repeatedly from randomized action-value function since each sample $\omega_h^{k,i}$ corresponds to one randomized action-value function $Q(\cdot,\cdot;\omega_h^{k,i})$. Such a collection of posterior weights is later used to construct the bootstrapped action-value function in Algorithm 2.

To independently sample from the posterior in (3.3), we can consider, e.g., using the Langevin dynamics $\omega(t+1) \leftarrow \omega(t) + \Delta\omega(t)$, where

$$\Delta\omega(t) = \frac{\epsilon_t}{2} \cdot \bigg(\nabla_\omega \log p_0\big(\omega(t)\big) + \sum_{\tau=1}^{k-1} \nabla_\omega \log p\big(y_h^\tau \,\big|\, (s_h^\tau, a_h^\tau), \omega(t)\big)\bigg) + \eta_t, \tag{3.4}$$

where $\epsilon_t > 0$ is the stepsize, and $\eta_t \sim \mathcal{N}(0, \epsilon_t)$. We note here that, after running sufficient many iterations with suitable choices of stepsizes $\epsilon_t > 0$, the Langevin dynamics gives effectively independent parameter samples from the posterior in (3.3).

To instantiate the connection between the posterior (3.3) with the LSVI, we consider Gaussian prior and likelihood as an example. For the Gaussian prior $\omega \sim \mathcal{N}(0, I_d)$ and the Gaussian likelihood $p(y_h^\tau \,|\, (s_h^\tau, a_h^\tau), \omega) \propto \exp\{-(y_h^\tau - Q(s_h^\tau, a_h^\tau; \omega))^2/(2\sigma^2)\}$, the posterior of ω is given by

$$p\Big(\omega \,\Big|\, \big\{\widetilde{V}_{h+1}^k(s_{h+1}^\tau)\big\}_{\tau\in[k-1]}, \mathcal{D}_h^k\Big) \propto \exp\bigg\{-\frac{1}{2}\cdot\|\omega\|^2 - \frac{1}{2\sigma^2}\sum_{\tau=1}^{k-1}\big(y_h^\tau - Q(s_h^\tau, a_h^\tau; \omega)\big)^2\bigg\}, \tag{3.5}$$

where $\sigma > 0$ is an absolute constant. In this case, the maximum a posteriori (MAP) estimate of ω coincides with the weight estimate $\widehat{\omega}_h^k$ obtained by LSVI. Such observation suggests that sampling from the posterior distribution $p(\omega \,|\, \{\widetilde{V}_{h+1}^k(s_{h+1}^\tau)\}_{\tau\in[k-1]}, \mathcal{D}_h^k)$ would allows us to both achieve exploitation based on the available data, and generate noise for randomized exploration. More specifically, the $\Delta\omega(t)$ in (3.4) takes the form of

$$\Delta\omega(t) = \frac{\epsilon_t}{2} \cdot \bigg[-\omega(t) + \frac{1}{\sigma^2}\sum_{\tau=1}^{k-1}\Big(y_h^\tau - Q\big(s_h^\tau, a_h^\tau; \omega(t)\big)\Big) \cdot \nabla_\omega Q\big(s_h^\tau, a_h^\tau; \omega(t)\big)\bigg] + \eta_t,$$

which can be viewed as using stochastic gradient descent to solve for the LSVI update (3.1) with $\lambda = \sigma^2$ and $\widehat{V}_{h+1}^k$ replaced by $\widetilde{V}_{h+1}^k$. See also in Lemma E.1 for the closed form of the posterior in a linear MDP setting with Gaussian prior and likelihood.

We would like to highlight that, although the theoretical guarantee in this paper is built upon the Gaussian prior and likelihood, the choices of prior and likelihood are flexible. For example, we can use uninformative prior for generalized linear model, and Gaussian process prior for kernel of overparameterized neural networks [6]. Moreover, as suggested by theoretical results on Langevin Dynamics [56] and illustrative experiments in Appendix F, the computational overhead caused by replacing the minimization of (3.1) with posterior sampling is mild.

Algorithm 1 Bootstrapped Value Iteration (BooVI)

1: **Require:** MDP $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r)$, action-value function parameterization $Q(\cdot, \cdot; \cdot) : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$, number of episodes K, number of posterior weights $\{N_k\}_{k \in [K]}$, lower and upper bootstrapping ratios $\alpha, \beta \in (0, 1)$
2: Initialize the data buffer $\mathcal{D}_h^1 \leftarrow \{\}$ for $h \in [H]$
3: **For** episode $k = 1, \ldots, K$ **do**
4: Set $N_{k,\alpha} \leftarrow \lceil \alpha \cdot N_k \rceil$,$N_{k,\beta} \leftarrow \lfloor \beta \cdot N_k \rfloor$, and $\omega_{H+1}^{k,i} \leftarrow 0$ for all $i \in [N_k]$
5: Sample n_k uniformly from $\{N_{k,\alpha}, N_{k,\alpha} + 1, \ldots, N_{k,\beta}\}$
6: **For** timestep $h = H, \ldots, 1$ **do**
7:   Generate $\widetilde{Q}_{h+1}^k(s_{h+1}^\tau, a)$ using Algorithm 2 with weights $\{\omega_{h+1}^{k,i}\}_{i \in [N_k]}$ and parameter n_k for all $a \in \mathcal{A}$ and $\tau \in [k-1]$
8:   $\widetilde{V}_{h+1}^k(s_{h+1}^\tau) \leftarrow \max_{a \in \mathcal{A}} \widetilde{Q}_{h+1}^k(s_{h+1}^\tau, a)$ for all $\tau \in [k-1]$
9:   Independently sample $\{\omega_h^{k,i}\}_{i \in [N_k]}$ from the posterior $p(\omega \mid \{\widetilde{V}_{h+1}^k(s_{h+1}^\tau)\}_{\tau \in [k-1]}, \mathcal{D}_h^k)$ defined in (3.3), e.g., using Langevin dynamics in (3.4)
10: **end**
11: **For** timestep $h = 1, \ldots, H$ **do**
12:   Generate $\widetilde{Q}_h^k(s_h^k, a)$ using Algorithm 2 with weights $\{\omega_h^{k,i}\}_{i \in [N_k]}$ and parameter n_k for all $a \in \mathcal{A}$
13:   Take action $a_h^k \leftarrow \mathrm{argmax}_{a \in \mathcal{A}} \widetilde{Q}_h^k(s_h^k, a)$, and observe r_h^k and s_{h+1}^k
14:   Update the data buffer $\mathcal{D}_h^{k+1} \leftarrow \mathcal{D}_h^k \cup \{(s_h^k, a_h^k, r_h^k, s_{h+1}^k)\}$
15: **End**
16: **End**

The following Algorithm 2 serves as the critical building block of BooVI for enforcing the optimism of the estimated action-value. At episode k in BooVI, for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, Algorithm 2 ranks the estimated action-value functions corresponding to the posterior sample obtained in Line 1 of Algorithm 1 in ascending order. Then, in order to enforce the optimism of the estimated action-value function, Algorithm 2 resamples the n_k-th top-ranked value from the ordered estimated action-value functions in the manner of bootstrapping. Finally, to ensure sufficient optimism of the obtained bootstrapped action-value function, we extrapolate with a tunable parameter $\nu > 1$.

Algorithm 2 Bootstrapping Action-Value Function

1: **Require:** Action-value function parameterization $Q(\cdot, \cdot; \cdot) : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$, posterior sample $\{\omega_h^{k,i}\}_{i \in [N_k]}$, integer $n_k \in [N_k]$, extrapolation parameter $\nu > 1$, and state-action pair (s, a)
2: Compute $Q_h^{k,i}(s, a) \leftarrow Q(s, a; \omega_h^{k,i})$ for all $i \in [N_k]$
3: Set $\widehat{Q}_h^k(s, a) \leftarrow (1/N_k) \sum_{i=1}^{N_k} Q_h^{k,i}(s, a)$
4: Rank $\{Q_h^{k,i}(s, a)\}_{i \in [N_k]}$ in ascending order to obtain $\{Q_h^{k,(i)}(s, a)\}_{i \in [N_k]}$
5: **Output:** $\widetilde{Q}_h^k(s, a) \leftarrow \min\{(1 - \nu) \cdot \widehat{Q}_h^k(s, a) + \nu \cdot Q_h^{k,(n_k)}(s, a), H - h + 1\}^+$

Remark 3.1 (Sample Efficiency and Computational Efficiency)**.** Here we clarify the differences between *sample efficiency* and *computational efficiency*. Throughout this paper, we refer *sample efficiency* to the learning efficiency with respect to the total number $T = KH$ of interactions with the environment. Such interactions can often time be very resource demanding. Thus, this paper seeks to provide an algorithm that aims to achieve the *sample efficiency* with respect to the number of such interactions. Although posterior weights are sampled in Algorithm 1, the efficiency of such posterior sampling process is categorized as *computational efficiency* since it is purely based on the current data buffer $\mathcal{D}_h^k$ and does not require any extra interaction with the environment.

Remark 3.2 (Additional Computational Cost)**.** Here we discuss the computational cost of BooVI compared to LSVI. (i) Posterior sampling vs. least-square minimization: Consider the Langevin dynamics in (3.4). Since posterior sampling replaces and admits similar steps as least-square minimization, the computational cost should not differ significantly between LSVI and BooVI. (ii) Computing bootstrapped Q-function: To compute bootstrapped Q-function, at each state-action pair, the extra sorting takes $O(N_k \log N_k)$ time complexity, which makes the action selection procedure $O(N_k \log N_k)$ times slower. In addition, keeping N_k posterior weights takes $O(d \cdot N_k)$ memory. We note that, although the requirement on N_k in our subsequent analysis is large, we should be able to use much smaller N_k in practice due to the pessimistic nature of the analysis. See Appendix F for illustrative experiment results.

It is worth mentioning that BooVI is applicable to any specific parameterization $Q(\cdot, \cdot\,; \cdot) : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$ of the action-value function and is fully data-driven. We also note here that the design of BooVI does not rely on the Gaussian prior/posterior or the Langevin dynamics sampling technique.

4 Main Results

While BooVI (Algorithm 1) is generally applicable to different parameterizations of the action-value function, we establish its regret for a class of linear MDPs.

Linear Markov Decision Process. We consider a class of MDPs, where the transition kernels and the reward functions are linear in the same feature mapping. Specifically, we have the following definition.

Definition 4.1 (Linear MDP)**.** An MDP $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r)$ is called a linear MDP with feature mapping $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, if for any $h \in [H]$, there exists d unknown signed measures $\mu_h = (\mu_h^{(1)}, \ldots, \mu_h^{(d)})$ over $\mathcal{S}$ and an unknown vector $\theta_h \in \mathbb{R}^d$, such that for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\mathcal{P}_h(\cdot \,|\, s, a) = \big\langle \phi(s, a), \mu_h(\cdot) \big\rangle, \quad r_h(s, a) = \big\langle \phi(s, a), \theta_h \big\rangle. \tag{4.1}$$

See [51, 52, 24] for examples of such a class of linear MDPs. Specifically, examples include tabular MDPs where $d = SA$ and the feature mapping is the canonical basis $\phi(s, a) = e_{(s,a)}$, and the simplex feature space where the feature space $\{\phi(s, a) \,|\, (s, a) \in \mathcal{S} \times \mathcal{A}\}$ is a subset of the d-dimensional simplex. See also [14, 44, 25] for related discussions on such a linear representation. For notational simplicity, we write the feature $\phi(s_h^k, a_h^k)$ as ϕ_h^k throughout the rest of this paper.

Based on Definition 4.1, without loss of generality, we make the following assumption.

Assumption 4.2. The MDP $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r)$ is a linear MDP with $\|\phi(s, a)\| \le 1$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $\max\{\|\mu_h(\mathcal{S})\|, \|\theta_h\|\} \le \sqrt{d}$ for all $h \in [H]$.

To motivate the linear parameterization of the action-value function in the subsequent section, we have the following proposition.

Proposition 4.3 (Linear Action-Value Function, Proposition 2.3 in [24])**.** For a linear MDP, for any policy π and $h \in [H]$, there exists $\omega_h^\pi \in \mathbb{R}^d$ such that for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have $Q_h^\pi(s, a) = \langle \phi(s, a), \omega_h^\pi \rangle$.

Regret of BooVI for Linear MDPs. For a linear MDP, by Proposition 4.3, the parameterization $Q(\cdot, \cdot\,; \cdot) : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$ in BooVI should take the form of $Q(s, a; \omega) = \langle \phi(s, a), \omega \rangle$, where $\omega \in \mathbb{R}^d$. Furthermore, when the prior and the likelihood are Gaussian, the posterior of ω is Gaussian distribution with mean being is the weight update by LSVI with chioce of $\lambda = \sigma^2$.

For any bootstrapping ratio $q \in (0, 1)$, we write $C_q = \Phi^{-1}(q)$ throughout the rest of this paper, where $\Phi(\cdot)$ is the cumulative distribution function of the standard Gaussian. Then, we have the following upper bound of regret of BooVI.

Theorem 4.4. Suppose that Assumption 4.2 holds, and that $d \ge 2$. We set $\sigma = 1$, $\nu = dH$, and fix a failure probability $p \in (0, 1]$. Then there exists a sequence of posterior sample sizes $\{N_k\}_{k \in [K]}$, and there exist absolute constants $c_\beta > c_\alpha > 0$, such that if $N_k = O(d^6 T^4 k / p^4)$ for all $k \in [K]$, $C_\alpha = c_\alpha \cdot \sqrt{\iota}$, and $C_\beta = c_\beta \cdot \sqrt{\iota}$, BooVI (Algorithm 1) with Gaussian posterior in (3.5) satisfies

$$\mathrm{Regret}(K) = O\big(\sqrt{d^3 H^3 T \iota^2}\big)$$

with probability at least $1 - p$, where $T = KH$ and $\iota = \log(3dT/p)$.

Proof. See Section 5 for a proof sketch. □

5 Proof Sketch

Notation. At episode k, given the number of posterior sample weights N_k and given n_k sampled in Line 1 of Algorithm 1, we define $C_k = \Phi^{-1}(n_k/N_k)$. At timestep h of episode k, further given posterior weights $\{\omega_h^{k,i}\}_{i\in[N_k]}$, corresponding to the bootstrapped action-value function $\widetilde{Q}_h^k$ generated by Algorithm 2, we define the bootstrapped value function as

$$\widetilde{V}_h^k(s) = \max_{a\in\mathcal{A}} \widetilde{Q}_h^k(s,a). \tag{5.1}$$

Next, we define the mean bootstrapped action-value functions as

$$\overline{Q}_h^k(s,a) = \min\Big\{\mathbb{E}_\omega\big[(1-\nu)\cdot \widehat{Q}_h^k(s,a) + \nu\cdot Q_h^{k,(n_k)}(s,a)\big], H-h+1\Big\}^+, \tag{5.2}$$

where $\mathbb{E}_\omega[\,\cdot\,]$ is taken over ω with respect to the posterior $p(\omega\,|\,\{\widetilde{V}_{h+1}^k(s_{h+1}^\tau)\}_{\tau\in[k-1]}, \mathcal{D}_h^k)$ defined in (3.3). Correspondingly, we define the mean bootstrapped value function as

$$\overline{V}_h^k(s) = \max_{a\in\mathcal{A}} \overline{Q}_h^k(s,a). \tag{5.3}$$

Also, we denote by

$$\overline{\omega}_h^k = (\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1} y_h^\tau \cdot \phi_h^\tau \tag{5.4}$$

the mean of the posterior distribution $p(\omega\,|\,\{\widetilde{V}_{h+1}^k(s_{h+1}^\tau)\}_{\tau\in[k-1]}, \mathcal{D}_h^k)$, where y_h^τ is defined in (3.2). For a fixed failure probability $p\in(0,1]$, we write $\iota = \log(3dT/p)$, where $T = KH$. Finally, we define matrix $\Lambda_h^k \in \mathbb{R}^{d\times d}$ by

$$\Lambda_h^k = \sum_{\tau=1}^{k-1} \phi_h^\tau(\phi_h^\tau)^\top + \sigma^2\cdot I_d. \tag{5.5}$$

Concentration Events. Before proving Theorem 4.4, we first present two lemmas, each characterizing an event that is involved throughout the remaining proofs. The first lemma characterizes the concentration behavior of the bootstrapped action-value function $\widetilde{Q}_h^k$.

Lemma 5.1. Let $\sigma = 1$, $\nu = dH$, and $C_\beta = c_\beta\cdot\sqrt{\iota}$ for some $c_\beta > 0$. For a fixed failure probability $p\in(0,1]$, we define $\mathcal{E}$ as the event that the condition

$$\big|\widetilde{Q}_h^k(s,a) - \overline{Q}_h^k(s,a)\big| \le \big(c_\beta/\sqrt{d}+3\big)\cdot H\sqrt{d\iota/k}$$

is satisfied for all $(s,a,k,h)\in\mathcal{S}\times\mathcal{A}\times[K]\times[H]$. Then, there exists a sequence of posterior sample sizes $\{N_k\}_{k\in[K]}$ satisfying $N_k = O(d^6T^4k/p^4)$ for all $k\in[K]$, such that $P(\mathcal{E})\ge 1-p/3$.

Proof. See Appendix C for a detailed proof. □

The second lemma characterizes the concentration behavior of the mean bootstrapped state value function $\overline{V}_h^k$.

Lemma 5.2. Let $\sigma = 1$, $\nu = dH$, and $C_\beta = c_\beta\cdot\sqrt{\iota}$ for some constant $c_\beta > 0$. For a fixed failure $p\in(0,1]$, we define $\mathcal{E}'$ as the event that the condition

$$\Big\|\sum_{\tau=1}^{k-1}\phi_h^\tau\cdot\big(\overline{V}_{h+1}^k(s_{h+1}^\tau) - [\mathcal{P}_h\overline{V}_{h+1}^k](s_h^\tau,a_h^\tau)\big)\Big\|_{(\Lambda_h^k)^{-1}} \le C\cdot dH\sqrt{\chi}$$

is satisfied for all $(k,h)\in[K]\times[H]$, where $\chi = \log[3(1+c_\beta)dT/p)]$. Then we have $\mathbb{P}(\mathcal{E}')\ge 1-p/3$.

Proof. See Appendix D for a detailed proof. □

Model Estimation Error. With the events $\mathcal{E}$ and $\mathcal{E}'$ ready, we can proceed to characterize the model estimation error $\zeta_h^k : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for all $(k, h) \in [K] \times [H]$ defined as

$$\zeta_h^k = \overline{Q}_h^k - \big(r_h + [\mathcal{P}_h \overline{V}_{h+1}^k]\big), \tag{5.6}$$

which can be comprehended as the estimation error of the model at timestep h of episode k induced by the mean bootstrapped value function $\overline{Q}_{h+1}^k$. We have the following lemma charaterizing the model estimation error ζ_h^k defined in (5.6).

Lemma 5.3 (Model Estimation Error)**.** Let $d \geq 2$, $\nu = dH$, $\sigma = 1$, and $p \in (0, 1]$. Then there exists a sequence of posterior sample sizes $\{N_k\}_{k \in [K]}$, and there exist absolute constants $c_\beta > c_\alpha > 0$ such that, if $N_k = O(d^6 T^4 k / p^4)$ for all $k \in [K]$, $C_\alpha = c_\alpha \cdot \sqrt{\iota}$, and $C_\beta = c_\beta \cdot \sqrt{\iota}$, for the model estimation error ζ_h^k defined in (5.6), we under events $\mathcal{E}$ and $\mathcal{E}'$ that

$$0 \leq \zeta_h^k(s, a) \leq (c_\alpha + c_\beta) \cdot dH\sqrt{\iota}\sqrt{\phi(s, a)^\top (\Lambda_h^k)^{-1} \phi(s, a)}$$

for all $(s, a, k, h) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H]$.

Proof. Under Assumption 4.2, using (4.1) we have $[\mathcal{P}_h \overline{V}_{h+1}^k] = \int_{\mathcal{S}} \overline{V}_{h+1}^k(s') \cdot \langle \phi, \mu_h(\mathrm{d}s') \rangle$. Then, with slight abuse of notation, we denote by $\overline{V}_{h+1}^k(\mu_h) = \int_{\mathcal{S}} \overline{V}_{h+1}^k(s) \mu_h(\mathrm{d}s) \in \mathbb{R}^d$ and write

$$\begin{aligned}
\mathcal{P}_h \overline{V}_{h+1}^k &= \big\langle \phi, (\Lambda_h^k)^{-1} (\Lambda_h^k) \overline{V}_{h+1}^k(\mu_h) \big\rangle \\
&= \left\langle \phi, (\Lambda_h^k)^{-1} \left(\sum_{\tau=1}^{k-1} \phi_h^\tau (\phi_h^\tau)^\top \overline{V}_{h+1}^k(\mu_h) \right) \right\rangle + \big\langle \phi, (\Lambda_h^k)^{-1} \overline{V}_{h+1}^k(\mu_h) \big\rangle \\
&= \left\langle \phi, (\Lambda_h^k)^{-1} \left(\sum_{\tau=1}^{k-1} \phi_h^\tau \cdot [\mathcal{P}_h \overline{V}_{h+1}^k](s_h^\tau, a_h^\tau) \right) \right\rangle + \big\langle \phi, (\Lambda_h^k)^{-1} \overline{V}_{h+1}^k(\mu_h) \big\rangle,
\end{aligned}$$

where the second equality follows from (5.5). Thus, we have the following decomposition

$$\big| \langle \phi, \overline{\omega}_h^k \rangle - (r_h + \mathcal{P}_h \overline{V}_{h+1}^k) \big| \leq u_1 + u_2,$$

where $\overline{\omega}_h^k$ is defined in (5.4), and

$$\begin{aligned}
u_1 &= \left| \left\langle \phi, (\Lambda_h^k)^{-1} \left[\sum_{\tau=1}^{k-1} \phi_h^\tau \cdot \big(\widetilde{V}_{h+1}^k(s_{h+1}^\tau) - [\mathcal{P}_h \overline{V}_{h+1}^k](s_h^\tau, a_h^\tau) \big) \right] \right\rangle \right|, \\
u_2 &= \left| \left\langle \phi, (\Lambda_h^k)^{-1} \left(\sum_{\tau=1}^{k-1} r_h^\tau \cdot \phi_h^\tau - \overline{V}_{h+1}^k(\mu_h) \right) \right\rangle - r_h \right|.
\end{aligned}$$

In the sequel, we upper bound u_1 and u_2 separately.

Upper bounding u_1: First, we further decompose u_1 as

$$\left\langle \phi, (\Lambda_h^k)^{-1} \left\{ \sum_{\tau=1}^{k-1} \phi_h^\tau \cdot \Big[\big(\widetilde{V}_{h+1}^k(s_{h+1}^\tau) - \overline{V}_{h+1,}^k(s_{h+1}^\tau) \big) + \big(\overline{V}_{h+1}^k(s_{h+1}^\tau) - [\mathcal{P}_h \overline{V}_{h+1}^k](s_h^\tau, a_h^\tau) \big) \Big] \right\} \right\rangle,$$

applying the Cauchy-Schwartz inequality to which we obtain

$$\begin{aligned}
u_1 \leq{}& \left\langle \phi, (\Lambda_h^k)^{-1} \left[\sum_{\tau=1}^{k-1} \phi_h^\tau \cdot \big(\widetilde{V}_{h+1}^k(s_{h+1}^\tau) - \overline{V}_{h+1}^k(s_{h+1}^\tau) \big) \right] \right\rangle \\
&+ \sqrt{\phi^\top (\Lambda_h^k)^{-1} \phi} \cdot \left\| \sum_{\tau=1}^{k-1} \phi_h^\tau \cdot \big(\overline{V}_{h+1}^k(s_h^\tau, a_h^\tau) - [\mathcal{P}_h \overline{V}_{h+1}^k](s_{h+1}^\tau) \big) \right\|_{(\Lambda_h^k)^{-1}}.
\end{aligned} \tag{5.7}$$

Now we proceed to upper bound the two terms on the right-hand side of (5.7) in the following. For the first term, since by Lemma 5.1 we have under event $\mathcal{E}$ that $|\widetilde{V}_{h+1}^k(s_{h+1}^\tau) - \overline{V}_{h+1}^k(s_{h+1}^\tau)| \leq (c_\beta/\sqrt{d} + 3) \cdot H\sqrt{d\iota/k}$, we have

$$\left| \left\langle \phi, (\Lambda_h^k)^{-1} \left[\sum_{\tau=1}^{k-1} \phi_h^\tau \cdot \big(\widetilde{V}_{h+1}^k(s_{h+1}^\tau) - \overline{V}_{h+1}^k(s_{h+1}^\tau) \big) \right] \right\rangle \right| \leq \big(c_\beta/\sqrt{d} + 3 \big) \cdot H\sqrt{d\iota/k} \sum_{\tau=1}^{k-1} \phi^\top (\Lambda_h^k)^{-1} \phi_h^\tau. \tag{5.8}$$

By the Cauchy-Schwartz inequality and Lemma D.1 in [24] (see Appendix E), we have

$$\sum_{\tau=1}^{k-1} \phi^\top (\Lambda_h^k)^{-1} \phi_h^\tau \le \Bigg[\bigg(\sum_{\tau=1}^{k-1} \phi^\top (\Lambda_h^k)^{-1} \phi_h^\tau\bigg) \cdot \bigg(\sum_{\tau=1}^{k-1} \phi^\top (\Lambda_h^k)^{-1} \phi\bigg)\Bigg]^{1/2} \le \sqrt{d\iota/k} \cdot \sqrt{dk} \sqrt{\phi^\top (\Lambda_h^k)^{-1} \phi},$$

taking which into (5.8) and gives

$$\bigg|\bigg\langle \phi, (\Lambda_h^k)^{-1} \bigg[\sum_{\tau=1}^{k-1} \phi_h^\tau \cdot \big(\widetilde{V}_{h+1}^k(s_{h+1}^\tau) - \overline{V}_{h+1}^k(s_{h+1}^\tau)\big)\bigg]\bigg\rangle\bigg| \le \big(c_\beta/\sqrt{d} + 3\big) \cdot dH\sqrt{\iota}\sqrt{\phi^\top (\Lambda_h^k)^{-1} \phi}. \tag{5.9}$$

Taking (5.9) into (5.7) and applying Lemma 5.2 to the second term on the right-hand side of (5.7), we have under the events $\mathcal{E}$ and $\mathcal{E}'$ that

$$u_1 \le \Big[\big(c_\beta/\sqrt{d} + 3\big) \cdot dH\sqrt{\iota} + C \cdot dH\sqrt{\chi}\Big] \cdot \sqrt{\phi^\top (\Lambda_h^k)^{-1} \phi}. \tag{5.10}$$

Upper bounding u_2: By (4.1), we have

$$u_2 = \Big|\Big\langle \phi, (\Lambda_h^k)^{-1} \big(\theta_h + \overline{V}_{h+1}^k(\mu_h)\big)\Big\rangle\Big|.$$

Then, by the Cauchy-Schwartz inequality, we further obtain

$$\begin{aligned} u_2 &\le \sqrt{\phi^\top (\Lambda_h^k)^{-1} \phi} \cdot \Big(\big\|\overline{V}_{h+1}^k(\mu_h)\big\|_{(\Lambda_h^k)^{-1}} + \big\|\theta_h\big\|_{(\Lambda_h^k)^{-1}}\Big) \\ &\le \sqrt{\phi^\top (\Lambda_h^k)^{-1} \phi} \cdot \Big(\big\|\overline{V}_{h+1}^k(\mu_h)\big\| + \big\|\theta_h\big\|\Big) \le 2H\sqrt{d} \cdot \sqrt{\phi^\top (\Lambda_h^k)^{-1} \phi}, \end{aligned} \tag{5.11}$$

where the second inequality follows from $\Lambda_h^k \succeq I_d$, and the last inequality follows from $\|\theta_h\| \le \sqrt{d}$ in Assumption (4.2) as well as the fact that $\overline{V}_{h+1}^k(s) \le H - h \le H - 1$ for all $s \in \mathcal{S}$.

Combining (5.10) and (5.11), we obtain under event $\mathcal{E}$ that

$$\begin{aligned} &\big|\langle \phi, \overline{\omega}_h^k\rangle - (r_h + \mathcal{P}_h \overline{V}_{h+1}^k)\big| \Big/ \sqrt{\phi^\top (\Lambda_h^k)^{-1} \phi} \\ &\le \Big[\big(c_\beta/\sqrt{d} + 3\big) \cdot dH\sqrt{\iota} + C \cdot dH\sqrt{\chi} + 2H\sqrt{d}\Big] \\ &\le \Big[\big(c_\beta/\sqrt{d} + 5\big) \cdot dH\sqrt{\iota} + C \cdot dH\sqrt{\chi}\Big] = C' \cdot dH\sqrt{\iota}, \end{aligned} \tag{5.12}$$

where $C' > 0$ is an absolute constant. Next, we need to find an absolute constant $c_\beta > 0$ such that $C' \cdot \sqrt{\iota} = (c_\beta/\sqrt{d} + 5) \cdot \sqrt{\iota} + C \cdot \sqrt{\iota + \log(1 + c_\beta)} < c_\beta \cdot \sqrt{\iota}$. Note that $d \ge 2$ and $\iota \ge \log 2$, it suffices to pick a $c_\beta > 0$ such that

$$C \cdot \sqrt{\log 2 + \log(1 + c_\beta)} < \Big[\big(1 - 1/\sqrt{2}\big)c_\beta - 5\Big] \cdot \sqrt{\log 2}, \tag{5.13}$$

which must exist as the left hand side grows logarithmically in c_β and the right-hand side grows linearly in c_β. For $c_\beta > 0$ satisfying (5.13), we pick any $c_\alpha > 0$ such that $C' \le c_\alpha < c_\beta$ and let $C_\alpha = c_\alpha \cdot \sqrt{\iota}$. By (5.12) and $r_h + [\mathcal{P}_h \overline{V}_{h+1}^k] \le H - h + 1$, we obtain under events $\mathcal{E}$ and $\mathcal{E}'$ that

$$\begin{aligned} \zeta_h^k &= \min\Big\{\langle \phi, \overline{\omega}_h^k\rangle + C_k \cdot \nu \cdot \sqrt{\phi^\top (\Lambda_h^k)^{-1} \phi}, H - h + 1\Big\}^+ - (r_h + \mathcal{P}_h \overline{V}_{h+1}^k) \\ &\ge \min\Big\{(c_\alpha - C') \cdot dH\sqrt{\iota}\sqrt{\phi^\top (\Lambda_h^k)^{-1} \phi}, 0\Big\} \ge 0. \end{aligned}$$

On the other hand, by (5.12), we also have under events $\mathcal{E}$ and $\mathcal{E}'$ that

$$\zeta_h^k \le \langle \phi, \overline{\omega}_h^k\rangle - (r_h + \mathcal{P}_h \overline{V}_{h+1}^k) + C_\beta \cdot \nu \cdot \sqrt{\phi^\top (\Lambda_h^k)^{-1} \phi} \le (c_\alpha + c_\beta) \cdot dH\sqrt{\iota}\sqrt{\phi^\top (\Lambda_h^k)^{-1} \phi}.$$

Therefore, we finish the proof of Lemma 5.3. □

As results of Lemma 5.3, we have the following two lemmas characterizing the optimism and the cumulative estimation error of the mean bootstrapped value function $\overline{V}_h^k$, respectively.

Lemma 5.4 (Optimistic Random Value Function)**.** Let $\sigma = 1$ and $\nu = dH$. There there exists a sequence of posterior sample sizes $\{N_k\}_{k\in[K]}$, and there exist absolute constants $c_\beta > c_\alpha > 0$, such that if $N_k = O(d^6T^4k/p^4)$ for all $k \in [K]$, $C_\alpha = c_\alpha \cdot \sqrt{\iota}$, and $C_\beta = c_\beta \cdot \sqrt{\iota}$, we have under the events $\mathcal{E}$ and $\mathcal{E}'$ that

$$\sum_{k=1}^{K}\big[V_1^*(s_1^k) - \overline{V}_1^k(s_1^k)\big] \leq 0.$$

Proof. See Appendix D for a detailed proof. □

Lemma 5.5 (Cumulative Estimation Error)**.** Let $\sigma = 1$, $\nu = dH$, and $p \in (0, 1]$. There exists a sequence of posterior sample sizes $\{N_k\}_{k\in[K]}$, and there exist absolute constants $c_\beta > c_\alpha > 0$, such that if $N_k = O(d^6T^4k/p^4)$ for all $k \in [K]$, $C_\alpha = c_\alpha \cdot \sqrt{\iota}$, and $C_\beta = c_\beta \cdot \sqrt{\iota}$, we have under the events $\mathcal{E}$ and $\mathcal{E}'$ that

$$\begin{aligned}\sum_{k=1}^{K}\big[\overline{V}_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)\big] &\leq \sqrt{18TH^2 \cdot \log(3/p)} + 4(c_\beta/\sqrt{d} + 3) \cdot H^2\sqrt{dK\iota} \\ &\quad + (c_\alpha + c_\beta) \cdot dH^2\sqrt{2dK\iota^2},\end{aligned}$$

with probability at least $1 - p/3$.

Proof. See Appendix D for a detailed proof. □

Finally, the regret bound of BooVI for the class of linear MDPs can be established as a consequence of the optimism (Lemmas 5.4) and the upper bound of the cumulative estimation error (Lemma 5.5) of the bootstrapped value function.

Proof of Theorem 4.4. First, recall that we have $\overline{V}_h^k(s)$ defined in (5.3). We have the regret decomposition

$$\text{Regret}(K) = \sum_{k=1}^{K}\big[V_1^*(s_1^k) - \overline{V}_1^k(s_1^k)\big] + \sum_{k=1}^{K}\big[\overline{V}_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)\big]. \tag{5.14}$$

Applying Lemmas 5.4 and 5.5 to (5.14), we obtain under the events $\mathcal{E}$ and $\mathcal{E}'$ that

$$\begin{aligned}\text{Regret}(K) &\leq 0 + \sqrt{18TH^2 \cdot \log(3/p)} + 4(c_\beta/\sqrt{d} + 3) \cdot H^2\sqrt{dK\iota} \\ &\quad + (c_\alpha + c_\beta) \cdot dH^2\sqrt{2dK\iota^2} \\ &= O\big(\sqrt{d^3H^3T\iota^2}\big) \end{aligned} \tag{5.15}$$

with probability at least $1 - p/3$. Finally, by Lemmas 5.1 and 5.2, we have with at least probability $1 - p/3 - p/3 = 1 - 2p/3$ that the events $\mathcal{E}$ and $\mathcal{E}'$ hold simultaneously. Thus, we have (5.15) hold with probability at least $1 - p$, which concludes the proof. □

6 Conclusion

The cost of collecting data via online experiments is often prohibitive compared to collecting offline data, e.g., via posterior sampling. Thus, designing online learning algorithms that provably explores the environment in a data-driven manner is essential. Moreover, applicability to general environments is critical since reinforcement learning tasks are getting more chanlleging and complex recently. We aspire to motivate the design of novel reinforcement learning algorithms that utilize cheaper data to boost online sample efficiency. Our algorithm and analysis serve as a step towards developing general applicable and provable sample-efficient reinforcement learning. While this paper is mainly on algorithmic and theoretical aspects of the bootstrapping idea in RL, it would be interesting to see the empirical strength of BooVI in more challenging RL environments. We leave this part to our future work.

Acknowledgement

Zhaoran Wang acknowledges National Science Foundation (Awards 2048075, 2008827, 2015568, 1934931), Simons Institute (Theory of Reinforcement Learning), Amazon, J.P. Morgan, and Two Sigma for their supports.

Zhuoran Yang acknowledges Simons Institute (Theory of Reinforcement Learning).

References

[1] Abbasi-Yadkori, Y., Pál, D. and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*.

[2] Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L. and Schapire, R. (2014). Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*.

[3] Agrawal, S. and Jia, R. (2017). Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*.

[4] Auer, P. and Ortner, R. (2006). Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems*.

[5] Azar, M. G., Osband, I. and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*.

[6] Bernardo, J., Berger, J., Dawid, A., Smith, A. et al. (1998). Regression and classification using gaussian process priors. *Bayesian statistics* **6** 475.

[7] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J. and Zaremba, W. (2016). Openai gym. *arXiv preprint arXiv:1606.01540* .

[8] Cai, Q., Yang, Z., Jin, C. and Wang, Z. (2019). Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830* .

[9] Chu, W., Li, L., Reyzin, L. and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*.

[10] Dani, V., Hayes, T. P. and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback. *Conference on Learning Theory* .

[11] Dann, C. and Brunskill, E. (2015). Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*.

[12] 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*.

[13] Dong, K., Peng, J., Wang, Y. and Zhou, Y. (2019). $\sqrt{n}$-regret for learning in Markov decision processes with function approximation and low Bellman rank. *arXiv preprint arXiv:1909.02506* .

[14] Du, S. S., Kakade, S. M., Wang, R. and Yang, L. F. (2019). Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016* .

[15] Du, S. S., Luo, Y., Wang, R. and Zhang, H. (2019). Provably efficient Q-learning with function approximation via distribution shift error checking oracle. *arXiv preprint arXiv:1906.06321* .

[16] Duan, Y., Chen, X., Houthooft, R., Schulman, J. and Abbeel, P. (2016). Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*.

[17] Ecoffet, A., Huizinga, J., Lehman, J., Stanley, K. O. and Clune, J. (2019). Go-explore: A new approach for hard-exploration problems. *arXiv preprint arXiv:1901.10995* .

[18] Foster, D. J., Agarwal, A., Dudík, M., Luo, H. and Schapire, R. E. (2018). Practical contextual bandits with regression oracles. *arXiv preprint arXiv:1803.01088* .

[19] Haarnoja, T., Zhou, A., Abbeel, P. and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290* .

[20] HSU, D., KAKADE, S. and ZHANG, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability* **17**.

[21] JAKSCH, T., ORTNER, R. and AUER, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research* **11** 1563–1600.

[22] 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*.

[23] JIN, C., ALLEN-ZHU, Z., BUBECK, S. and JORDAN, M. I. (2018). Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*.

[24] JIN, C., YANG, Z., WANG, Z. and JORDAN, M. I. (2019). Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388* .

[25] LATTIMORE, T. and SZEPESVARI, C. (2019). Learning with good feature representations in bandits and in RL with a generative model. *arXiv preprint arXiv:1911.07676* .

[26] MAY, B. C., KORDA, N., LEE, A. and LESLIE, D. S. (2012). Optimistic bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research* **13** 2069–2106.

[27] OSBAND, I., ASLANIDES, J. and CASSIRER, A. (2018). Randomized prior functions for deep reinforcement learning. In *Advances in Neural Information Processing Systems*.

[28] OSBAND, I., BLUNDELL, C., PRITZEL, A. and VAN ROY, B. (2016). Deep exploration via bootstrapped DQN. In *Advances in Neural Information Processing Systems*.

[29] OSBAND, I., RUSSO, D. and VAN ROY, B. (2013). (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*.

[30] OSBAND, I., RUSSO, D., WEN, Z. and VAN ROY, B. (2017). Deep exploration via randomized value functions. *Journal of Machine Learning Research* .

[31] OSBAND, I. and VAN ROY, B. (2015). Bootstrapped Thompson sampling and deep exploration. *arXiv preprint arXiv:1507.00300* .

[32] OSBAND, I. and VAN ROY, B. (2016). On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732* .

[33] OSBAND, I. and VAN ROY, B. (2017). Why is posterior sampling better than optimism for reinforcement learning? In *International Conference on Machine Learning*.

[34] OSBAND, I., VAN ROY, B. and WEN, Z. (2014). Generalization and exploration via randomized value functions. *arXiv preprint arXiv:1402.0635* .

[35] RAGINSKY, M., RAKHLIN, A. and TELGARSKY, M. (2017). Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. *arXiv preprint arXiv:1702.03849* .

[36] RUSSO, D. (2019). Worst-case regret bounds for exploration via randomized value functions. In *Advances in Neural Information Processing Systems*.

[37] RUSSO, D. J., VAN ROY, B., KAZEROUNI, A., OSBAND, I. and WEN, Z. (2018). A tutorial on Thompson sampling. *Foundations and Trends® in Machine Learning* **11** 1–96.

[38] SILVER, D., HUANG, A., MADDISON, C. J., GUEZ, A., SIFRE, L., VAN DEN DRIESSCHE, G., SCHRITTWIESER, J., ANTONOGLOU, I., PANNEERSHELVAM, V., LANCTOT, M. ET AL. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature* **529** 484.

[39] SILVER, D., LEVER, G., HEESS, N., DEGRIS, T., WIERSTRA, D. and RIEDMILLER, M. (2014). Deterministic policy gradient algorithms. In *International Conference on Machine Learning*.

[40] SILVER, D., SCHRITTWIESER, J., SIMONYAN, K., ANTONOGLOU, I., HUANG, A., GUEZ, A., HUBERT, T., BAKER, L., LAI, M., BOLTON, A. ET AL. (2017). Mastering the game of Go without human knowledge. *Nature* **550** 354.

[41] STREHL, A. L., LI, L., WIEWIORA, E., LANGFORD, J. and LITTMAN, M. L. (2006). PAC model-free reinforcement learning. In *International Conference on Machine Learning*.

[42] TAÏGA, A. A., FEDUS, W., MACHADO, M. C., COURVILLE, A. and BELLEMARE, M. G. (2019). Benchmarking bonus-based exploration methods on the Arcade Learning Environment. *arXiv preprint arXiv:1908.02388* .

[43] THOMPSON, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25** 285–294.

[44] VAN ROY, B. and DONG, S. (2019). Comments on the Du-Kakade-Wang-Yang lower bounds. *arXiv preprint arXiv:1911.07910* .

[45] WANG, W. Y., LI, J. and HE, X. (2018). Deep reinforcement learning for NLP. In *Association for Computational Linguistics*.

[46] WANG, Y., WANG, R., DU, S. S. and KRISHNAMURTHY, A. (2019). Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136* .

[47] WATKINS, C. J. C. H. (1989). Learning from delayed rewards .

[48] WELLING, M. and TEH, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*.

[49] WEN, Z. and VAN ROY, B. (2017). Efficient reinforcement learning in deterministic systems with value function generalization. *Mathematics of Operations Research* **42** 762–782.

[50] XIONG, Z., SHEN, R. and DU, S. S. (2021). Randomized exploration is near-optimal for tabular mdp. *arXiv preprint arXiv:2102.09703* .

[51] YANG, L. and WANG, M. (2019). Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*.

[52] YANG, L. F. and WANG, M. (2019). Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. *arXiv preprint arXiv:1905.10389* .

[53] YANG, Z., JIN, C., WANG, Z., WANG, M. and JORDAN, M. (2020). Provably efficient reinforcement learning with kernel and neural function approximations. *Advances in Neural Information Processing Systems* **33**.

[54] ZANETTE, A., BRANDFONBRENER, D., PIROTTA, M. and LAZARIC, A. (2019). Frequentist regret bounds for randomized least-squares value iteration. *arXiv preprint arXiv:1911.00567* .

[55] ZANETTE, A., LAZARIC, A., KOCHENDERFER, M. and BRUNSKILL, E. (2020). Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*. PMLR.

[56] ZHANG, Y., LIANG, P. and CHARIKAR, M. (2017). A hitting time analysis of stochastic gradient langevin dynamics. In *Conference on Learning Theory*. PMLR.

[57] ZHOU, D., GU, Q. and SZEPESVARI, C. (2021). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*. PMLR.

[58] ZHOU, H., CHEN, J., VARSHNEY, L. R. and JAGMOHAN, A. (2020). Nonstationary reinforcement learning with linear function approximation. *arXiv preprint arXiv:2010.04244* .

A Proof of Lemma 5.4

With Lemma 5.3, the proof of Lemma 5.4 mainly follows that of Lemma B.5 in [24]. For self-containedness, we lay out the proof of Lemma 5.4 in the following.

Proof. By the definition of ζ_h^k in (5.6), we have

$$\begin{aligned}Q_h^*(s,a) - \overline{Q}_h^k(s,a) &= Q_h^*(s,a) - \big(r_h + [\mathcal{P}_h \overline{V}_{h+1}^k]\big)(s,a) - \zeta_h^k(s,a)\\ &= \big(r_h + [\mathcal{P}_h V_{h+1}^*]\big)(s,a) - \big(r_h + [\mathcal{P}_h \overline{V}_{h+1}^k]\big)(s,a) - \zeta_h^k(s,a)\\ &= \big[\mathcal{P}_h (V_{h+1}^* - \overline{V}_{h+1}^k)\big](s,a) - \zeta_h^k(s,a),\end{aligned}$$

applying Lemma 5.3 to which we obtain under the events $\mathcal{E}$ and $\mathcal{E}'$ that, for all $(s,a,k,h) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H]$,

$$Q_h^*(s,a) - \overline{Q}_h^k(s,a) \le \big[\mathcal{P}_h (V_{h+1}^* - \overline{V}_{h+1}^k)\big](s,a).$$

Note that by definition we have $Q_{H+1}^* \equiv \overline{Q}_{H+1}^k \equiv 0$, which gives $V_{H+1}^* \equiv \overline{V}_{H+1}^k \equiv 0$. Thus, we have for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ that

$$Q_H^*(s,a) - \overline{Q}_H^k(s,a) \le 0,$$

which gives $V_H^*(s) - \overline{V}_H^k(s) \le 0$ for all $s \in \mathcal{S}$. Thus, we have

$$Q_{H-1}^*(s,a) - \overline{Q}_{H-1}^k(s,a) \le \big[\mathcal{P}_h (V_H^* - \overline{V}_H^k)\big](s,a) \le 0.$$

Finally, by induction, we obtain $V_1^*(s_1) - \overline{V}_1^k(s_1) \le 0$ under events $\mathcal{E}$ and $\mathcal{E}'$, summing up which for all $k \in [K]$ concludes the proof. □

B Proof of Lemma 5.5

First, we present the following lemma that decomposes the term $\overline{V}_{1,\beta}^k(s_1^k) - V_1^{\pi^k}(s_1^k)$.

Lemma B.1. For $(s,a,k,h) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H]$, let $\psi_h^k(s_h^k, a_h^k) = \overline{V}_h^k(s_h^k) - \overline{Q}_h^k(s_h^k, a_h^k)$ and

$$\begin{aligned}\xi_h^k(a_h^k, s_h^k, s_{h+1}^k) = &\; Q_h^{\pi^k}(s_h^k, a_h^k) - V_h^{\pi^k}(s_h^k)\\ &+ \big[\mathcal{P}_h(\overline{V}_{h+1}^k - V_{h+1}^{\pi^k})\big](s_h^k, a_h^k) - (\overline{V}_{h+1}^k - V_{h+1}^{\pi^k})(s_{h+1}^k). \qquad \text{(B.1)}\end{aligned}$$

We have

$$\overline{V}_1^k(s_1^k) - V_1^{\pi^k}(s_1^k) = \sum_{h=1}^H \xi_h^k(s_h^k, a_h^k, s_{h+1}^k) + \sum_{h=1}^H \psi_h^k(s_h^k, a_h^k) + \sum_{h=1}^H \zeta_h^k(s_h^k, a_h^k), \qquad \text{(B.2)}$$

where ζ_h^k is defined in (5.6)

Proof. In this proof, we first consider timesteps within an episode $k \in [K]$. First, by definition of ζ_h^k in (5.6), we have

$$\begin{aligned}\zeta_h^k(s_h^k, a_h^k) &= (\overline{Q}_h^k - Q_h^{\pi^k})(s_h^k, a_h^k) + \Big[Q_h^{\pi^k}(s_h^k, a_h^k) - \big(r_h + [\mathcal{P}_h \overline{V}_{h+1}^k]\big)(s_h^k, a_h^k)\Big]\\ &= (\overline{Q}_h^k - Q_h^{\pi^k})(s_h^k, a_h^k) + \Big[\big(r_h + [\mathcal{P}_h V_{h+1}^{\pi^k}]\big)(s_h^k, a_h^k) - \big(r_h + [\mathcal{P}_h \overline{V}_{h+1}^k]\big)(s_h^k, a_h^k)\Big]\\ &= (\overline{Q}_h^k - Q_h^{\pi^k})(s_h^k, a_h^k) - \big[\mathcal{P}_h(\overline{V}_{h+1}^k - V_{h+1}^{\pi^k})\big](s_h^k, a_h^k),\end{aligned}$$

where the second line follows from Bellman equation. Then we have for all $(h,k) \in [H] \times [K]$ that

$$
\begin{aligned}
&\overline{V}_h^k(s_h^k) - V_h^{\pi^k}(s_h^k) \\
&\quad = \big(\overline{V}_h^k(s_h^k) - V_h^{\pi^k}(s_h^k)\big) - \big(\overline{Q}_h^k(s_h^k, a_h^k) - Q_h^{\pi^k}(s_h^k, a_h^k)\big) + \big[\mathcal{P}_h(\overline{V}_{h+1}^k - V_{h+1}^{\pi^k})\big](s_h^k, a_h^k) + \zeta_h^k(s_h^k, a_h^k) \\
&\quad = \underbrace{Q_h^{\pi^k}(s_h^k, a_h^k) - V_h^{\pi^k}(s_h^k) + \big[\mathcal{P}_h(\overline{V}_{h+1}^k - V_{h+1}^{\pi^k})\big](s_h^k, a_h^k) - (\overline{V}_{h+1}^k - V_{h+1}^{\pi^k})(s_{h+1}^k)}_{=\xi_h^k(s_h^k, a_h^k, s_{h+1}^k)} \\
&\qquad + \underbrace{\overline{V}_h^k(s_h^k) - \overline{Q}_h^k(s_h^k, a_h^k)}_{=\psi_h^k(s_h^k, a_h^k)} + \overline{V}_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k) + \zeta_h^k(s_h^k, a_h^k). \qquad \text{(B.3)}
\end{aligned}
$$

Applying (B.3) recursively for all $h \in [H]$ and recalling that $\overline{Q}_{H+1}^k(s,a) = Q_{H+1}^{\pi^k}(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, we finish the proof. □

Next, we bound the summation of $\xi_h^k(s_h^k, a_h^k, s_{h+1}^k)$ on the right-hand side of (B.2) in the following lemma.

Lemma B.2. Suppose that event $\mathcal{E}$ holds. For $\xi_h^k(s_h^k, a_h^k, s_{h+1}^k)$ defined in (B.1), we have with probability at least $1 - p/3$ that

$$
\sum_{k=1}^{K} \sum_{h=1}^{H} \xi_h^k(s_h^k, a_h^k, s_{h+1}^k) \le \sqrt{18TH^2 \cdot \log(3/p)}.
$$

Proof. First, we define a σ-algebra that starts with $\mathcal{F}_1^1 = \{s_1^1\}$ and is then recursively constructed by

$$
\mathcal{F}_h^k = \mathcal{F}_{h-1}^k \cup \{(a_{h-1}^k, r_{h-1}^k, s_h^k)\}, \quad 2 \le h \le H \qquad \text{(B.4)}
$$

and

$$
\mathcal{F}_1^{k+1} = \mathcal{F}_H^k \cup \{\omega_h^{k+1,i}\}_{h\in[H], i\in[N_k]} \cup \{s_1^{k+1}\} \cup \{n_{k+1}\}. \qquad \text{(B.5)}
$$

Note that given the posterior sample $\{\omega_h^{k,i}\}_{h\in[H], i\in[N_k]}$ and n_k, the action-value function $Q_h^{\pi^k}$ and the bootstrapped action-value function $\widetilde{Q}_h^k$, as functions mapping from $\mathcal{S} \times \mathcal{A}$ to $\mathbb{R}$, are known. Thus, we have

$$
\mathbb{E}\big[\xi_h^k(s_h^k, a_h^k, s_{h+1}^k) \,\big|\, \mathcal{F}_h^k\big] = 0,
$$

which means that the sequence $\{\xi_h^k(s_h^k, a_h^k)\}_{k\in[K], h\in[H]}$ is a martingale sequence adapted to the filtration $\{\mathcal{F}_h^k\}_{k\in[K], h\in[H]}$. Furthermore, by the definition of $\overline{Q}_h^k$ and $\overline{V}_h^k$ in (5.2) and (5.3), we know that the martingale sequence is bounded by $3H$. Therefore, by the Azuma-Hoeffding inequality, we obtain for any $t > 0$ that

$$
\mathbb{P}\bigg(\sum_{k=1}^{K} \sum_{h=1}^{H} \xi_h^k(s_h^k, a_h^k, s_{h+1}^k) > t\bigg) \le \exp\bigg(\frac{-t^2}{2T \cdot 9H^2}\bigg), \qquad \text{(B.6)}
$$

where $T = KH$. Taking $t = \sqrt{18TH^2 \log(3/p)}$ in (B.6), we finish the proof. □

We have the following lemma bounding the summation of $\psi_h^k(s_h^k, a_h^k)$ defined in Lemma B.1.

Lemma B.3. Suppose that event $\mathcal{E}$ holds. For $\psi_h^k(s_h^k, a_h^k)$ defined in Lemma (B.1), we have

$$
\sum_{k=1}^{K} \sum_{h=1}^{H} \psi_h^k(s_h^k, a_h^k) \le 4(c_\beta/\sqrt{d} + 3) \cdot H^2 \sqrt{dK\iota}.
$$

Proof. First, we make the decomposition

$$
\overline{V}_h^k(s_h^k) - \overline{Q}_h^k(s_h^k, a_h^k) = \overline{V}_h^k(s_h^k) - \max_{a\in\mathcal{A}} \widetilde{Q}_h^k(s_h^k, a) + \max_{a\in\mathcal{A}} \widetilde{Q}_h^k(s_h^k, a) - \overline{Q}_h^k(s_h^k, a_h^k). \qquad \text{(B.7)}
$$

By the definition of $\overline{V}_h^k$ in (5.3), we have

$$\begin{aligned}\overline{V}_h^k(s_h^k) - \max_{a\in\mathcal{A}} \widetilde{Q}_h^k(s_h^k, a) &= \max_{a\in\mathcal{A}} \overline{Q}_h^k(s_h^k, a) - \max_{a\in\mathcal{A}} \widetilde{Q}_h^k(s_h^k, a) \\ &\le \max_{a\in\mathcal{A}} \big|\overline{Q}_h^k(s_h^k, a) - \widetilde{Q}_h^k(s_h^k, a)\big|. \end{aligned} \tag{B.8}$$

Meanwhile, since a_h^k is the greedy action with respect to $\widetilde{Q}_h^k(s_h^k, a)$, we have

$$\max_{a\in\mathcal{A}} \widetilde{Q}_h^k(s_h^k, a) - \overline{Q}_h^k(s_h^k, a_h^k) = \widetilde{Q}_h^k(s_h^k, a_h^k) - \overline{Q}_h^k(s_h^k, a_h^k). \tag{B.9}$$

Taking (B.8) and (B.9) into (B.7), we obtain

$$\overline{V}_h^k(s_h^k) - \overline{Q}_h^k(s_h^k, a_h^k) \le \max_{a\in\mathcal{A}} \big|\overline{Q}_h^k(s_h^k, a) - \widetilde{Q}_h^k(s_h^k, a)\big| + \widetilde{Q}_h^k(s_h^k, a_h^k) - \overline{Q}_h^k(s_h^k, a_h^k) \tag{B.10}$$

$$\le 2(c_\beta/\sqrt{d} + 3) \cdot H\sqrt{d\iota/k}, \tag{B.11}$$

where the second inequality holds by the definition of the event $\mathcal{E}$ in Lemma 5.1. Finally, summing up (B.10) for all $h \in [H]$ and $k \in [K]$, we have

$$\begin{aligned}\sum_{k=1}^K \sum_{h=1}^H \overline{V}_h^k(s_h^k) - \overline{Q}_h^k(s_h^k, a_h^k) &\le 2(c_\beta/\sqrt{d} + 3) \cdot H^2\sqrt{d\iota} \sum_{k=1}^K 1/\sqrt{k} \\ &\le 2(c_\beta/\sqrt{d} + 3) \cdot H^2\sqrt{d\iota} \cdot 2\sqrt{K} = 4(c_\beta/\sqrt{d} + 3) \cdot H^2\sqrt{dK\iota}.\end{aligned}$$

Therefore, we conclude the proof of Lemma B.3. □

Now we are ready to prove Lemma 5.5.

Proof of Lemma 5.5. First, summing (B.2) for all $k \in [K]$ we have

$$\sum_{k=1}^K \big[\overline{V}_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)\big] = \sum_{k=1}^K \sum_{h=1}^H \xi_h^k(s_h^k, a_h^k, s_{h+1}^k) + \sum_{k=1}^K \sum_{h=1}^H \psi_h^k(s_h^k, a_h^k) + \sum_{k=1}^K \sum_{h=1}^H \zeta_h^k(s_h^k, a_h^k).$$

By Lemmas B.2 and 5.2, under the event $\mathcal{E}$, we further have with probability at least $1 - p/3$ that

$$\begin{aligned}\sum_{k=1}^K \big[\overline{V}_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)\big] &\le \sqrt{18TH^2 \cdot \log(3/p)} + 4(c_\beta/\sqrt{d} + 3) \cdot H^2\sqrt{dK\iota} \\ &\quad + (c_\alpha + c_\beta) \cdot dH\sqrt{\iota} \sum_{h=1}^H \sum_{k=1}^K \sqrt{(\phi_h^k)^\top (\Lambda_h^k)^{-1} \phi_h^k}. \end{aligned} \tag{B.12}$$

Next, we bound the third term in (B.12). By the Cauchy-Schwartz inequality, we have

$$\begin{aligned}\sum_{k=1}^K \sqrt{(\phi_h^k)^\top (\Lambda_h^k)^{-1} \phi_h^k} &\le \sqrt{K} \cdot \Bigg[\sum_{k=1}^K (\phi_h^k)^\top (\Lambda_h^k)^{-1} \phi_h^k\Bigg]^{1/2} \\ &\le \sqrt{K} \cdot \Bigg[2\log\bigg(\frac{\det(\Lambda_h^K)}{\det(\Lambda_h^0)}\bigg)\Bigg]^{1/2} \le \sqrt{K} \cdot \big[2d \cdot \log(1+K)\big]^{1/2},\end{aligned} \tag{B.13}$$

where the second inequality follows from Lemma E.4, and the last inequality follows from $I_d \preceq \Lambda_h^k \preceq (1+k) \cdot I_d$. Further summing (B.13) up for all $h \in [H]$ and using $\log(1+K) \le \iota$, we have

$$\sum_{k=1}^K \sum_{h=1}^H \sqrt{(\phi_h^k)^\top (\Lambda_h^k)^{-1} \phi_h^k} \le H\sqrt{2dK} \cdot \big[2d \cdot \log(1+K)\big]^{1/2} \le H\sqrt{2dK\iota},$$

taking which back into (B.12) gives

$$\begin{aligned}\sum_{k=1}^K \big[\overline{V}_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)\big] &\le \sqrt{18TH^2 \cdot \log(3/p)} + 4(c_\beta/\sqrt{d} + 3) \cdot H^2\sqrt{dK\iota} \\ &\quad + (c_\alpha + c_\beta) \cdot dH^2\sqrt{2dK\iota^2}.\end{aligned}$$

Thus, we conclude the proof of Lemma 5.5. □

C Proof of Lemma 5.1

Prior to proving the uniform concentration guarantee over all $(s,a) \in \mathcal{S} \times \mathcal{A}$ in Lemma 5.1, we first present the following concentration lemma for a fixed state-action pair (s,a). We drop the truncations at 0 and $H-h+1$ for the bootstrapped action-value function in all the proofs in this section, since for any random variable $z \in \mathbb{R}$ with mean $\overline{z}$,

$$\mathbb{P}\Big(\big|\min\{z, H-h+1\}^{+} - \min\{\overline{z}, H-h+1\}^{+}\big| > \nu\epsilon\Big) \le \mathbb{P}\big(|z-\overline{z}| > \nu\epsilon\big).$$

Lemma C.1. For fixed $(s,a,k,h) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H]$, we have for any $\epsilon > 0$ and any $\nu \ge 2$ that

$$\mathbb{P}\Big(\big|\widetilde{Q}_h^k(s,a) - \overline{Q}_h^k(s,a)\big| > \nu\epsilon\Big) \le 2\exp\Big\{-8\exp\big\{-2(C_\beta+\epsilon)^2\big\} \cdot N_k\epsilon^2/\pi\Big\} + 2\exp\{-2N_k\epsilon^2\}.$$

Proof. For notational simplicity, we omit the state-action pair (s,a), which is fixed throughout this proof. By Lemma E.1, we have

$$\begin{aligned}
&\mathbb{P}\big(\widetilde{Q}_h^k > \overline{Q}_h^k + \nu\epsilon\big) \qquad\qquad (C.1)\\
&\quad = \mathbb{P}\big((1-\nu)\cdot \widehat{Q}_h^k + \nu \cdot Q_h^{k,(n_k)} > \overline{Q}_h^k + \nu\epsilon\big)\\
&\quad = \mathbb{P}\Big((1-\nu)\cdot\big(\widehat{Q}_h^k - \langle \phi, \overline{\omega}_h^k\rangle\big) + \Big[\nu\cdot\big(Q_h^{k,(n_k)} - \langle\phi,\overline{\omega}_h^k\rangle\big) - C_k\cdot\nu\sigma\cdot\sqrt{\phi^\top(\Lambda_h^k)^{-1}\phi}\Big] > \nu\epsilon\Big)\\
&\quad \le \underbrace{\mathbb{P}\Big((1-\nu)\cdot\big(\widehat{Q}_h^k - \langle\phi,\overline{\omega}_h^k\rangle\big) > \nu\epsilon\Big)}_{=p_1} + \underbrace{\mathbb{P}\Big(\big(Q_h^{k,(n_k)} - \langle\phi,\overline{\omega}_h^k\rangle\big) - C_k\sigma\cdot\sqrt{\phi^\top(\Lambda_h^k)^{-1}\phi} > \epsilon\Big)}_{=p_2}.
\end{aligned}$$

Before upper bounding p_1 and p_2, we charatcerize $\langle\phi,\omega_h^{k,i}\rangle$ for $i\in[N_k]$ in the following. By the proof of Lemma E.1, since the state-action pair $(s,a)\in\mathcal{S}\times\mathcal{A}$ is fixed, $\langle\phi,\omega_h^{k,i}\rangle$ is a one-dimensional Gaussian with mean $\langle\phi,\overline{\omega}_h^k\rangle$ and variance $\sigma^2\cdot\phi^\top(\Lambda_h^k)^{-1}\phi$. Also, by definition of Λ_h^k in (5.5), we have $\Lambda_h^k \succeq \sigma^2\cdot I_d$, which, together with the assumption that $\|\phi\|\le 1$, gives

$$\sigma^2\cdot\phi^\top(\Lambda_h^k)^{-1}\phi \le \big\|(\Lambda_h^k)^{-1}\big\|\cdot\|\phi\|^2 \le \sigma^2\cdot(1/\sigma^2)\cdot 1 = 1. \qquad (C.2)$$

Bounding p_1: For $\nu\ge 2$, we have

$$\begin{aligned}
p_1 &= \mathbb{P}\Big(\widehat{Q}_h^k < \langle\phi,\overline{\omega}_h^k\rangle - [\nu/(\nu-1)]\cdot\epsilon\Big)\\
&\le \mathbb{P}\big(\widehat{Q}_h^k < \langle\phi,\overline{\omega}_h^k\rangle - 2\epsilon\big) = \mathbb{P}\Big(\frac{1}{N_k}\sum_{i=1}^{N_k}\langle\phi,\omega_h^{k,i}\rangle < \langle\phi,\overline{\omega}_h^k\rangle - 2\epsilon\Big).
\end{aligned}$$

Since $\langle\phi,\omega_h^{k,i}\rangle$ for $i\in[N_k]$ are one-dimensional Gaussian with mean $\langle\phi,\overline{\omega}_h^k\rangle$ and variance $\sigma^2\cdot\phi^\top(\Lambda_h^k)^{-1}\phi$, by Hoeffding's inequality we further obtain

$$p_1 \le \exp\Big\{-\frac{4\epsilon^2}{2N_k^{-1}\cdot\sigma^2\cdot\phi^\top(\Lambda_h^k)^{-1}\phi}\Big\} \le \exp\{-2N_k\epsilon^2\}, \qquad (C.3)$$

where the second inequality follows from (C.2).

Bounding p_2: Let $\overline{Q}_h^{k,(n_k)} = \min\{\langle\phi,\overline{\omega}_h^k\rangle + C_k\sigma\cdot\sqrt{\phi^\top(\Lambda_h^k)^{-1}\phi}, H-h+1\}$. Note that

$$\Big\{\big(Q_h^{k,(n_k)} - \langle\phi,\overline{\omega}_h^k\rangle\big) - C_k\sigma\cdot\sqrt{\phi^\top(\Lambda_h^k)^{-1}\phi} > \epsilon\Big\} = \Big\{\frac{1}{N_k}\sum_{i=1}^{N_k}\mathbb{1}\big\{Q_h^{k,i} > \overline{Q}_h^{k,(n_k)} + \epsilon\big\} > \frac{n_k}{N_k}\Big\},$$

where the indicators $\mathbb{1}\{Q_h^{k,i} > \overline{Q}_h^{k,(n_k)} + \epsilon\}$ for $i\in[N_k]$ are independently identical distributed random variables with expectation

$$\mathbb{E}\Big[\mathbb{1}\big\{Q_h^{k,i} > \overline{Q}_h^{k,(n_k)} + \epsilon\big\}\Big] = \mathbb{P}\big(Q_h^{k,i} > \overline{Q}_h^k + \epsilon\big) = n_k/N_k - \delta_{h,+\epsilon}^k. \qquad (C.4)$$

Here

$$\delta_{h,+\epsilon}^k = \mathbb{P}\big(Q_h^{k,1} > \overline{Q}_h^{k,(n_k)}\big) - \mathbb{P}\big(Q_h^{k,1} > \overline{Q}_h^{k,(n_k)} + \epsilon\big).$$

Combining (C.1) and (C.4) we have

$$\mathbb{P}\big(\widetilde{Q}_h^k > \overline{Q}_h^{k,(n_k)} + \epsilon\big) = \mathbb{P}\bigg(\frac{1}{N_k}\sum_{i=1}^{N_k} \mathbb{1}\big\{Q_h^{k,i} > \overline{Q}_h^{k,(n_k)} + \epsilon\big\} > \mathbb{E}\Big[\mathbb{1}\big\{Q_h^{k,1} > \overline{Q}_h^{k,(n_k)} + \epsilon\big\}\Big] + \delta_{h,+\epsilon}^k\bigg). \tag{C.5}$$

Since the indicator is bounded within $[0,1]$, by Hoeffding's inequality we obtain

$$\mathbb{P}\big(\widetilde{Q}_h^k > \overline{Q}_h^{k,(n_k)} + \epsilon\big) \le \exp\big(-2N_k \cdot (\delta_{h,\epsilon}^k)^2\big). \tag{C.6}$$

It ramains to characterize $\delta_{h,\epsilon}^k$. Thus, we have

$$\delta_{h,\epsilon}^k = \Phi(C_k) - \Phi\left(C_k + \frac{\epsilon}{\sqrt{\sigma^2 \cdot \phi^\top (\Lambda_h^k)^{-1}\phi}}\right) \ge \frac{2\exp\big\{-(C_\beta+\epsilon)^2\big\}}{\sqrt{\pi}} \cdot \epsilon, \tag{C.7}$$

which is a consequence of $\mathrm{d}(\Phi(z))/\mathrm{d}z = (2/\sqrt{\pi}) \cdot \exp(-z^2)$ and $C_k \le C_\beta$. Taking (C.7) into (C.6), we obtain

$$p_2 \le \exp\Big\{-8\exp\big\{-2(C_\beta+\epsilon)^2\big\} \cdot N_k\epsilon^2/\pi\Big\}. \tag{C.8}$$

Taking (C.3) and (C.8) into (C.1), we have

$$\mathbb{P}\big(\widetilde{Q}_h^k > \overline{Q}_h^k + \nu\epsilon\big) \le \exp\Big\{-8\exp\big\{-2(C_\beta+\epsilon)^2\big\} \cdot N_k\epsilon^2/\pi\Big\} + \exp\{-2N_k\epsilon^2\}. \tag{C.9}$$

Similarly, we also have

$$\mathbb{P}\big(\widetilde{Q}_h^k < \overline{Q}_h^k - \nu\epsilon\big) \le \exp\Big\{-8\exp\big\{-2(C_\beta+\epsilon)^2\big\} \cdot N_k\epsilon^2/\pi\Big\} + \exp\{-2N_k\epsilon^2\},$$

combining which with (C.9) concludes the proof. □

Next, we introduce three useful lemmas to connect the above concentration lemma for a fixed state-action pair to the desired uniform concentration guarantee over all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Lemma C.2 (Proposition 1 in [20]). Let v be a Gaussian random vector in $\mathbb{R}^d$ with mean zero and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. For any $t > 0$, it holds that

$$\mathbb{P}\big(\|v\|^2 > \mathrm{tr}(\Sigma) + 2\sqrt{\mathrm{tr}(\Sigma^2)\cdot t} + 2\|\Sigma\| \cdot t\big) \le \exp(-t).$$

Let $v_h^{k,i} = \omega_h^{k,i} - \overline{\omega}_h^k \in \mathbb{R}^d$ for $i \in [N]$. We have $v_h^{k,i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 \cdot (\Lambda_h^k)^{-1})$ for $i \in [N]$. By Lemma C.2, we have with probability at least $1 - \exp(-t)$ that

$$\begin{aligned}\|v_h^{k,i}\|^2 &\le \mathrm{tr}\big((\Lambda_h^k)^{-1}\big) + 2\sqrt{\mathrm{tr}\big((\Lambda_h^k)^{-2}\big)\cdot t} + 2\big\|(\Lambda_h^k)^{-1}\big\| \cdot t \\ &\le d/\sigma^2 + 2\sqrt{dt}/\sigma^2 + 2t/\sigma^2.\end{aligned} \tag{C.10}$$

Combining Lemma E.2 and (C.10), we further have with probability at least $1 - \exp(t)$ that

$$\begin{aligned}\|\omega_h^{k,i}\|^2 = \|\overline{\omega}_h^k + v_h^{k,i}\|^2 \le 2\|\overline{\omega}_h^k\|^2 + 2\|v_h^{k,i}\|^2 &\le \big(8dH^2k + 2d + 4\sqrt{dt} + 4t\big)/\sigma^2 \\ &\le \underbrace{\big(10dH^2k + 4\sqrt{dt} + 4t\big)/\sigma^2}_{W(t)^2}.\end{aligned} \tag{C.11}$$

Lemma C.3. For $\{w^i\}_{i\in[N]} \subset \mathbb{R}^d$ such that $\|w^i\| \le W$ for all $i \in [N]$ and vectors $\phi, \phi' \in \mathbb{R}^d$ such that $\|\phi - \phi'\| \le \varepsilon$, let $\widetilde{Q}$ and $\widetilde{Q}'$ be the m-th ($1 \le m \le N$) order statistics in $\{\langle\phi, w^i\rangle\}_{i\in[N]}$ and $\{\langle\phi', w^i\rangle\}_{i\in[N]}$ respectively. Then it holds that $|\widetilde{Q} - \widetilde{Q}'| \le \varepsilon W$.

Proof. By definition of $\widetilde{Q}$, there exists an index set $I \subset [N]$ such that $|I| = m$ and $\langle \phi, \omega^i \rangle \geq \widetilde{Q}$ for any $i \in I$ and $\langle \phi, \omega^i \rangle \leq \widetilde{Q}$ for any $i \notin I$. Since $\|\phi - \phi'\| \leq \varepsilon$ and $\|w^i\| \leq W$ for all $i \in [N]$, by Hölder's inequality it holds that $\langle \phi', w^i \rangle \geq \widetilde{Q}' - \varepsilon W$ for all $i \in I$, which implies $\widetilde{Q}' \geq \widetilde{Q} - \varepsilon W$. Similarly we also have $\widetilde{Q}' \leq \widetilde{Q} + \varepsilon W$ and thus we finish the proof. □

Lemma C.4. For any $\phi, \phi' \in \mathbb{R}^d$ such that $\|\phi - \phi'\| \leq \varepsilon$, let $\overline{Q}_q, \overline{Q}_q'$ be the q-quantiles of the random variables $\langle \phi, \omega \rangle$ and $\langle \phi', \omega \rangle$ respectively, where the random vector ω follows distribution $N(\overline{\omega}, \Sigma)$ with $\|\overline{\omega}\| \leq 2H\sqrt{dk}/\sigma$ and $\Sigma \preceq I_d$. Then it holds that $|\overline{Q}_q - \overline{Q}_q'| \leq \varepsilon \cdot (2H\sqrt{dk}/\sigma + C_q)$.

Proof. Recall that $C_q = \Phi^{-1}(q)$. It holds that

$$\overline{Q}_q = \langle \phi, \overline{\omega} \rangle + C_q \cdot \sqrt{\phi^\top \Sigma \phi}, \quad \overline{Q}_q' = \langle \phi', \overline{\omega} \rangle + C_q \cdot \sqrt{(\phi')^\top \Sigma \phi'}.$$

Thus we have

$$\begin{aligned} |\overline{Q}_q - \overline{Q}_q'| &= \Big| \langle \phi - \phi', \overline{\omega} \rangle + c_q \cdot \Big(\sqrt{\phi^\top \Sigma \phi} - \sqrt{(\phi')^\top \Sigma \phi'} \Big) \Big| \\ &\leq \|\phi - \phi'\| \cdot \|\overline{\omega}\| + C_q \cdot \big\| A(\phi - \phi') \big\| \\ &\leq \varepsilon \cdot \big(\|\overline{\omega}\| + C_q \cdot \|A\| \big) \leq \varepsilon \cdot \big(2H\sqrt{dk}/\sigma + C_q \big), \end{aligned}$$

where $\Sigma = A^\top A$, and the last inequality follows from $\|A\| = \|\Sigma\|^{1/2} \leq 1$. □

Now we are ready to prove Lemma 5.1.

Proof of Lemma 5.1. Let $N_\varepsilon(\mathcal{B}) = (1 + 2/\varepsilon)^d$ be the ε-covering number of the unit ball $\mathcal{B}$ in $\mathbb{R}^d$. For any $(s, a) \in \mathcal{S} \times \mathcal{A}$, since $\|\phi(s, a)\| \leq 1$, there exists a vector $\phi' \in \mathbb{R}^d$ in the ε-covering such that

$$\big\| \phi(s, a) - \phi' \big\| \leq \varepsilon.$$

We denote by $\widetilde{Q}_h^{k'}$ and $\overline{Q}_h^{k'}$ the empirical and true (n_k/N_k)-quantile of $\langle \phi', \omega_h^{k,1} \rangle$ respectively. Then we have the following decomposition

$$\big| \widetilde{Q}_h^k(s, a) - \overline{Q}_h^k(s, a) \big| \leq \big| \widetilde{Q}_h^k(s, a) - \widetilde{Q}_h^{k'} \big| + \big| \widetilde{Q}_h^{k'} - \overline{Q}_h^{k'} \big| + \big| \overline{Q}_h^{k'} - \overline{Q}_h^k(s, a) \big|. \tag{C.12}$$

Plugging Lemmas C.1, C.3 and C.4 into (C.12), and recalling $W(t)$ defined in (C.11) and using $\sigma = 1$, we have

$$\big| \widetilde{Q}_h^k(s, a) - \overline{Q}_h^k(s, a) \big| \leq \varepsilon \cdot W(t) + \nu \epsilon + \varepsilon \cdot \big(2H\sqrt{dk} + C_k \big), \tag{C.13}$$

for all $(x, a, k, h) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H]$, $p \in (0, 1]$ and $t, \epsilon, \varepsilon > 0$, with probability at least

$$1 - 2N_\varepsilon(\mathcal{B})T \cdot \bigg(\exp\Big\{ -8 \exp\big\{ -2(C_\beta + \epsilon)^2 \big\} \cdot N_k \epsilon^2 / \pi \Big\} + \exp\big\{ -2N_k \epsilon^2 \big\} \bigg) - NT \exp(-t). \tag{C.14}$$

In what remains, we take $\epsilon = \sqrt{\iota/(dk)}$ and $\varepsilon = 1/(5dk)$. First, we have $(C_\beta + \epsilon)^2 \leq 2C_\beta^2 + 2\epsilon^2 = 2c_\beta^2 \cdot \iota + 2\iota/dk \leq 2(c_\beta^2 + 1) \cdot \iota$, taking which into (C.14) and recalling that $\iota = \log(3dT/p)$, we obtain

$$8 \exp\big\{ -2(C_\beta + \epsilon)^2 \big\} / \pi \geq 8 \exp\{ -4(c_\beta^2 + 1) \} \cdot \big(3dT/p \big)^{-4} / \pi = (1/c_\beta') \cdot \big(dT/p \big)^{-4},$$

where $c_\beta' = 81\pi \exp\{4(c_\beta^2 + 1)\}/8$. Thus, if

$$N_k = 2c_\beta' \cdot d^6 T^4 k / p^4 \geq \frac{\log\big(18 N_\varepsilon(\mathcal{B}) \cdot T/p \big)}{8 \exp\big\{ -2(C_\beta + \epsilon)^2 \big\} \cdot \epsilon^2 / \pi}, \tag{C.15}$$

and

$$t = \log\big(9 N_k T/p \big), \tag{C.16}$$

we have the probability in (C.14) being at least $1 - p/3$. Taking (C.15) into (C.16), we further have

$$\sqrt{t} = \sqrt{\iota + \log(3N_k/d)} \leq \sqrt{6\iota + 4(1 + c_\beta^2)} \leq \sqrt{6\iota} + 2 + 2c_\beta. \tag{C.17}$$

Next, taking (C.17) into (C.13), we obtain

$$W(t) = \big(10dH^2k + 4\sqrt{dt} + 4t\big)^{1/2} \leq (7 + 4c_\beta) \cdot H\sqrt{dk\iota}. \tag{C.18}$$

Finally, using $C_k \leq C_\beta = c_\beta \cdot \sqrt{\iota}$ and (C.18), we know that with probability at least $1 - p/3$,

$$\begin{aligned}
\big|\widetilde{Q}_h^k(s,a) - \overline{Q}_h^k(s,a)\big| &\leq \varepsilon \cdot \big(W(t) + 2H\sqrt{dk} + C_\beta\big) + \nu\epsilon \\
&= \frac{1}{5dk} \cdot \big((7 + 4c_\beta) \cdot H\sqrt{dk\iota} + 2H\sqrt{dk}\big) + \frac{c_\beta \cdot \sqrt{\iota}}{5dk} + H\sqrt{d\iota/k} \\
&\leq \big(c_\beta/\sqrt{d} + 3\big) \cdot H\sqrt{d\iota/k},
\end{aligned}$$

which concludes the proof. ☐

D Proof of Lemma 5.2

Lemma D.1 (Lemma D.4 in [24])**.** Let $\{s^\tau\}_{\tau\geq 0}$ be a stochastic process in the state space $\mathcal{S}$ with corresponding filtration $\{\mathcal{F}^\tau\}_{\tau\geq 0}$. Let $\{\phi^\tau\}_{\tau\geq 0} \subset \mathbb{R}^d$ be a stochastic process where $\phi^\tau \in \mathcal{F}^{\tau-1}$ and $\|\phi^\tau\| \leq 1$ for all $\tau \geq 0$. Let $\Lambda^k = \sum_{\tau=1}^k \phi^\tau(\phi^\tau)^\top + \sigma^2 \cdot I_d$. Then for any $\delta > 0$, with probability at least $1 - p$, for all $k \geq 0$ and any $V \in \mathcal{V}$ so that $\sup_s |V(s)| \leq H$, we have

$$\bigg\|\sum_{\tau=1}^k \phi^\tau \cdot \Big(V(s^\tau) - \mathbb{E}\big[V(s^\tau)\,\big|\,\mathcal{F}^{\tau-1}\big]\Big)\bigg\|_{(\Lambda^k)^{-1}}^2 \leq 4H^2 \cdot \bigg[\frac{d}{2} \cdot \log\bigg(\frac{k+\sigma^2}{\sigma^2}\bigg) + \log\frac{N_\varepsilon(\mathcal{V})}{p}\bigg] + \frac{8k^2\varepsilon^2}{\sigma^2},$$

where $N_\varepsilon(\mathcal{V})$ is the ε-covering number of the function class $\mathcal{V}$ with respect to the distance $d(V, V') = \sup_{s\in\mathcal{S}} |V(s) - V'(s)|$.

Lemma D.2 (Lemma D.6 in [24])**.** Let $\mathcal{V}_h$ be the class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with the form

$$V_h(\cdot) = \max_{a\in\mathcal{A}}\bigg\{\min\Big\{\langle\phi(\cdot,a),\omega\rangle + C_k \cdot \nu\sigma \cdot \sqrt{\phi(\cdot,a)^\top\Lambda^{-1}\phi(\cdot,a)}, H - h + 1\Big\}^+\bigg\},$$

where the parameters (ω, C_k, Λ) satisfy $\|\omega\| \leq L$, $C_k \in [C_\alpha, C_\beta]$, and $\lambda_{\min}(\Lambda) \geq \sigma^2$. Assume that $\|\phi(s,a)\| \leq 1$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, and let $N_\varepsilon(\mathcal{V}_h)$ be the ε-covering number of $\mathcal{V}$ with respect to the distance $d(V_h, V_h') = \sup_{s\in\mathcal{S}} |V_h(s) - V_h'(s)|$. Then,

$$\log N_\varepsilon(\mathcal{V}_h) \leq d \cdot \log\bigg(1 + \frac{4L}{\varepsilon}\bigg) + d^2 \cdot \log\bigg(1 + \frac{8C_\beta^2 \cdot \nu^2\sqrt{d}}{\varepsilon^2}\bigg).$$

Proof of Lemma 5.2. First, recall that we defined in (B.4) and (B.5) the filtration $\{\mathcal{F}_h^k\}_{k\in[K],h\in[H]}$, based on which we can define another filtration $\{\mathcal{G}_h^k\}_{k\in[K],h\in[H]}$ by

$$\mathcal{G}_h^{k-1} = \mathcal{F}_h^k \cup \{a_h^k\}, \quad (k,h) \in [K]\times[H].$$

Then, by definition of $[\mathcal{P}_h V]$, we have for all $\tau \in [k-1]$ that

$$\begin{aligned}
[\mathcal{P}_h\overline{V}_{h+1}^k](s_h^\tau, a_h^\tau) &= \mathbb{E}_{s_{h+1}\sim\mathcal{P}_h(\cdot\,|\,s_h^\tau,a_h^\tau)}\big[\overline{V}_{h+1}^k(s_{h+1})\,\big|\,s_h^\tau, a_h^\tau\big] \\
&= \mathbb{E}_{s_{h+1}\sim\mathcal{P}_h(\cdot\,|\,s_h^\tau,a_h^\tau)}\big[\overline{V}_{h+1}^k(s_{h+1})\,\big|\,\mathcal{G}_h^{\tau-1}\big].
\end{aligned}$$

Thus, by Lemmas D.1,D.2 and E.2, we have with probability at least $1 - p/3$ that

$$\bigg\|\sum_{\tau=1}^{k-1} \phi_h^\tau \cdot \big(\overline{V}_{h+1}^k(s_{h+1}^\tau) - [\mathcal{P}_h\overline{V}_{h+1}^k](s_h^\tau, a_h^\tau)\big)\bigg\|_{(\Lambda_h^k)^{-1}}^2 \tag{D.1}$$

$$\leq 4H^2 \cdot \bigg[\frac{d}{2} \cdot \log\bigg(\frac{k+\sigma^2}{\sigma^2}\bigg) + \log\bigg(\frac{3}{p}\bigg) + d \cdot \log\bigg(1 + \frac{8H\sqrt{dk}}{\varepsilon\sigma}\bigg) + d^2 \cdot \log\bigg(1 + \frac{8C_\beta^2 \cdot \nu^2\sqrt{d}}{\varepsilon^2}\bigg)\bigg] + \frac{8k^2\varepsilon^2}{\sigma^2}.$$

In (D.1), taking $\sigma = 1$, $\varepsilon = dH/k$, $\nu = dH$ and $C_\beta = c_\beta \cdot \sqrt{\iota}$, we obtain that there exists an absolute constant $C > 0$ such that

$$\bigg\| \sum_{\tau=1}^{k-1} \phi_h^\tau \cdot \big(\overline{V}_{h+1}^k(s_{h+1}^\tau) - [\mathcal{P}_h \overline{V}_{h+1}^k](s_h^\tau, a_h^\tau)\big) \bigg\|_{(\Lambda_h^k)^{-1}}^2 \leq C^2 \cdot d^2 H^2 \chi,$$

for all $(k, h) \in [K] \times [H]$, where $\chi = \log[3(1 + c_\beta) dT/p]$. Therefore, we finish the proof. $\square$

E Supporting Lemmas

Lemma E.1. For $\overline{Q}_h^k(s, a)$ defined in (5.2), we have for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ that

$$\overline{Q}_h^k(s, a) = \min\Big\{ \big\langle \phi(s, a), \overline{\omega}_h^k \big\rangle + C_k \cdot \nu\sigma \cdot \sqrt{\phi(s, a)^\top (\Lambda_h^k)^{-1} \phi(s, a)}, H - h + 1 \Big\},$$

where $\overline{\omega}_h^k = (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} y_h^\tau \cdot \phi_h^\tau$ is defined in (5.4).

Proof. When the prior and the likelihood are Gaussian, the posterior of ω is Gaussian distribution with mean

$$\begin{aligned} \overline{\omega}_h^k &= \mathbb{E}\big[\omega \,\big|\, \{\widetilde{V}_{h+1}^k(s_{h+1}^\tau)\}_{\tau \in [k-1]}, \mathcal{D}_h^k\big] \\ &= \operatorname*{argmax}_{\omega \in \mathbb{R}^d} \Big\{ -\frac{1}{2} \cdot \|\omega\|^2 - \frac{1}{\sigma^2} \sum_{\tau=1}^{k-1} \big(y_h^\tau - Q(s_h^\tau, s_h^\tau; \omega)\big)^2 \Big\} = (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} y_h^\tau \cdot \phi_h^\tau, \end{aligned}$$

where Λ_h^k is defined in (5.5), and y_h^τ is defined in (3.2). Also, we have the covariance of the posterior as

$$\mathrm{Cov}\Big(\omega \,\Big|\, \big\{\widetilde{V}_{h+1}^k(s_{h+1}^\tau)\big\}_{\tau \in [k-1]}, \mathcal{D}_h\Big) = \sigma^2 \cdot (\Lambda_h^k)^{-1},$$

from which we obtain for all $i \in [N_k]$ that

$$\big\langle \phi(s, a), \omega_h^{k,i} \big\rangle \sim \mathcal{N}\Big(\big\langle \phi(s, a), \overline{\omega}_h^k \big\rangle, \sigma^2 \cdot \phi(s, a)^\top (\Lambda_h^k)^{-1} \phi(s, a) \Big).$$

Thus, we have

$$\begin{aligned} &\mathbb{E}_\omega\big[(1 - \nu) \cdot \widehat{Q}_h^k(s, a) + \nu \cdot Q_h^{k,(n_k)}(s, a)\big] \\ &\quad = (1 - \nu) \cdot \big\langle \phi(s, a), \overline{\omega}_h^k \big\rangle + \nu \cdot \Big(\big\langle \phi(s, a), \overline{\omega}_h^k \big\rangle + \Phi^{-1}(n_k/N_k) \cdot \sqrt{\sigma^2 \cdot \phi(s, a)^\top (\Lambda_h^k)^{-1} \phi(s, a)} \Big) \\ &\quad = \big\langle \phi(s, a), \overline{\omega}_h^k \big\rangle + C_k \cdot \nu\sigma \cdot \sqrt{\phi(s, a)^\top (\Lambda_h^k)^{-1} \phi(s, a)}, \end{aligned}$$

plugging which into definition of $\overline{Q}_h^k$ in (5.2) concludes the proof. $\square$

Lemma E.2 (Lemma B.2 in [24])**.** For $\overline{\omega}_h^k$ defined in Lemma (E.1) and all $(k, h) \in [K] \times [H]$, we have

$$\|\overline{\omega}_h^k\| \leq 2H\sqrt{dk}/\sigma.$$

Lemma E.3 (Lemma D.1 in [24])**.** Let $\Lambda^k = \sum_{\tau=1}^k \phi^\tau (\phi^\tau)^\top + \lambda \cdot I_d$, where $\phi^\tau \in \mathbb{R}^d$ and $\lambda > 0$. Then

$$\sum_{\tau=1}^k (\phi^\tau)^\top (\Lambda^k)^{-1} \phi^\tau \leq d.$$

Lemma E.4 (Elliptical Potential Lemma in, e.g., [10, 24])**.** Let $\{\phi^\tau\}_{\tau \geq 0} \subset \mathbb{R}^d$ be a sequence satisfying $\sup_{\tau \geq 0} \|\phi^\tau\| \leq 1$. Let $\Lambda^0 \in \mathbb{R}^{d \times d}$ be a positive definite matrix. For any $k \geq 0$, we define $\Lambda^k = \Lambda^0 + \sum_{\tau=1}^k \phi^\tau (\phi^\tau)^\top$. Then, if $\lambda_{\min}(\Lambda^0) \geq 1$, we have

$$\log\bigg(\frac{\det(\Lambda^k)}{\det(\Lambda^0)}\bigg) \leq \sum_{\tau=1}^k (\phi^\tau)^\top (\Lambda^{\tau-1})^{-1} \phi^\tau \leq 2\log\bigg(\frac{\det(\Lambda^k)}{\det(\Lambda^0)}\bigg).$$

F Illustrative Experiments

F.1 Synthetic Environment

In this section, we perform illustrative experiments on a synthetic dataset. We compare BooVI with three baseline algorithms: Random-Exploration, Epsilon-Greedy [47], and LSVI-UCB [24].

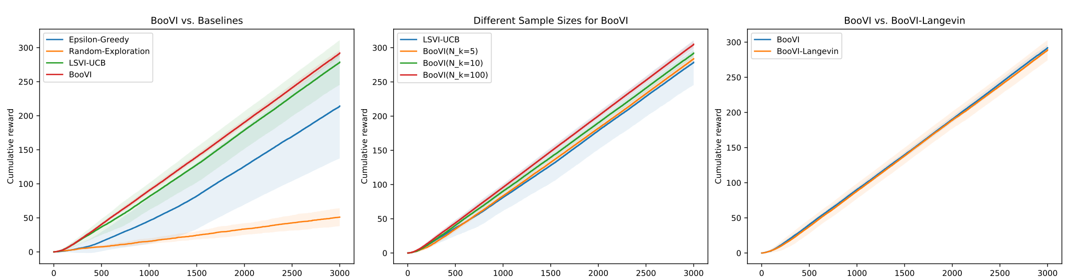
Environment. The construction of the synthetic MDP environment with simplex feature follows that of [58]. The constructed MDP has $|\mathcal{S}| = 15$, $|\mathcal{A}| = 7$, $H = 10$, $d = 10$, and $K = 300$. In the constructed MDP, there is only one good chain that leads to a huge reward at (only) the end of the episode. Otherwise, the agent receives small positive rewards in the suboptimal chains. Such an MDP requires the agent to perform deep exploration [30] in order to reach near-optimal policy rather than being attracted to the small suboptimal rewards.

Baseline. The Random-Exploration baseline takes actions uniformly randomly throughout all the episodes. The Epsilon-Greedy baseline takes the greedy action according to the current estimates of $\{Q_h\}_{h\in[H]}$ with probability of $1-\epsilon$, and takes the uniformly random action with probability ϵ. Here we set $\epsilon = 0.05$. In LSVI-UCB, we tune the parameter β for a best outcome. In the reported results, we use $\beta = 0.7$.

BooVI Setup. Due to the linear structure in the synthetic MDP, we are able to compute the matrix inverse $(\Lambda_h^k)^{-1}$. Thus, we perform posterior weight sampling via directly sampling from the exact posterior distribution $\mathcal{N}(\omega_h^k, (\Lambda_h^k)^{-1})$. In the reported results, we use the extrapolation parameter $\nu = 1.4$, the order parameter $n_k/N_k = 0.6$, and three different of posterior sample sizes $N_k \equiv 5$, $N_k \equiv 10$, and $N_k \equiv 100$.

In addition, to test the case with approximate posterior sampling, we perform BooVI-Langevin, which uses Langevin dynamics in (3.4) to sample posterior weights. In BooVI-Langevin, we use the same ν and n_k as BooVI, and we use $N_k \equiv 10$. For the Langevin dynamics sampling part, we choose the starting weight as ω_h^k and use the stepsize $\eta_t = 0.01/t$. To avoid getting too correlated posterior weights, we use 5 Langevin iterates to warm up, which are thrown away, and keep 1 posterior weight every 3 iterates after the warm-up phase.

Result. In the following, we report the cumulative rewards for BooVI, BooVI-Langevin, and the three baselines. The error bars reflect one standard deviation with 10 trials.

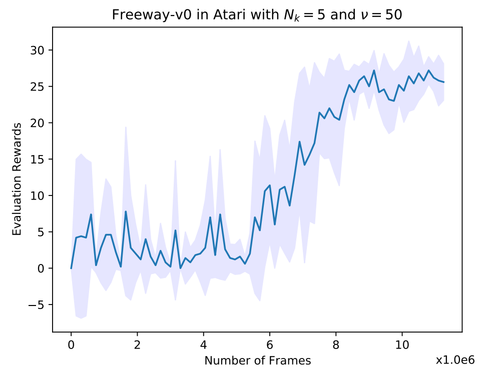


Left: Comparison between BooVI with $N_k \equiv 10$ and the baselines.
Middle: Comparison for different sample sizes $N_k = 5, 10, 100$ with LSVI-UCB as the baseline.
Right: Comparison between BooVI with $N_k = 10$ and BooVI-Langevin with $N_k = 10$.

In the left figure, we can observe that BooVI attains slightly better performance as LSVI-UCB (both outperform Epsilon-Greedy and Random-Exploration) but has much smaller standard deviation on the performance among trials. The middle figure shows that the larger N_k only improves the performance of BooVI marginally. The right figure shows that BooVI and BooVI-Langevin also has similar performance. However, BooVI-Langevin has larger standard deviation on the performance due to the extra stochasticity brought by the Langevin dynamics.

F.2 Freeway Game

We present the following illustrative result on the freeway game in Atari from OpenAI gym [7]. Here we choose $N_k \equiv 5$ posterior sample size and $\nu = 50$ extrapolation parameter throughout the training process. The standard deviation is plotted using 5 trials.



Freeway_v0: Scores achieved by BooVI with $N_k \equiv 5$ and $\nu = 50$.

Without any modification, BooVI achieves final performances ranging from 25 to 30 in freeway. Considering that only $N_k \equiv 5$ is used, the computational overhead is much less than what' suggested by the theory.

F.3 Discussion

As we see in the experimental results in synthetic environment, BooVI and BooVI-Langevin have comparable performance compared to LSVI-UCB at the cost of extra memory allocation for N_k copies of Q-function parameters and extra computational cost for evaluating the sampled Q-functions for N_k times. However, the requirements on the extrapolation parameter ν, and the sample size N_k are much milder than the requirements in the theoretical analysis, which results in milder computational overhead and memory allocation compared to LSVI/LSVI-UCB. We speculate that this is due to the pessimism nature of the analysis for establishing upper bounds. Moreover, in the comparison between BooVI and BooVI-Langevin, we see similar performance without using large number of Langevin iterates (35 total iterates). This suggests that BooVI is robust to approximate posterior sampling.

The aforementioned phenomenons in the experiment results for the synthetic MDP are encouraging. It would be interesting to see the performance of BooVI in more challenging environments, which we leave to our future work.