Provably Efficient Representation Selection in Low-rank Markov Decision Processes: From Online to Offline RL (Supplementary Material)

Weitong Zhang[1] **Jiafan He**[1] **Dongruo Zhou**[1] **Amy Zhang**[2,3] **Quanquan Gu**[1]

[1]Department of Computer Science, University of California, Los Angeles, California, USA
[2]Department of Electrical and Computer Engineering, University of Texas at Austin, Texas, USA
[3]Facebook AI Research

1 ADDITIONAL RELATED WORK

Reinforcement Learning with Linear Function Approximation. A large body of literature regarding learning MDP with linear function approximation has emerged recently. Those works can be roughly divided by their assumptions on MDPs: The first one is called Linear MDP [Yang and Wang, 2019, Jin et al., 2020], where the representation function is built on the state action pair $\boldsymbol{\phi}(s, a)$. Under this assumption, Jin et al. [2020] proposed the LSVI-UCB algorithm achieving $\widetilde{\mathcal{O}}(\sqrt{d^3H^3T})$ problem independent regret bound and $\widetilde{\mathcal{O}}(d^3H^5\text{gap}_{\min}^{-1}\log(T))$ problem dependent regret bound due to He et al. [2021]. Here $\text{gap}_{\min}$ is the minimal sub-optimality gap, d is the dimension and H is the time-horizon. Several similar MDP assumptions are studied in the literature: for instance, Jiang et al. [2017] studied a larger class of MDPs with low Bellman rank and proposed an algorithm with polynomial sample complexity. Low inherent Bellman error assumption is proposed by Zanette et al. [2020] and allows a better $\mathcal{O}(dH\sqrt{T})$ regret by considering a global planning oracle. Yang and Wang [2020] considered the bilinear structure of the MDP kernel as a special case of the Linear MDP, and achieved an $\widetilde{\mathcal{O}}(H^2d\sqrt{T})$ problem-independent regret bound. The second linear function approximation assumption is called Linear Mixture MDP [Modi et al., 2020, Ayoub et al., 2020, Zhou et al., 2021b] where the transition kernel of MDP is a linear function $\boldsymbol{\phi}(s, a, s')$ of the 'state-action-next state' triplet. Under this setting, Jia et al. [2020], Ayoub et al. [2020] proposed UCRL-VTR achieving $\mathcal{O}(d\sqrt{H^3T})$ problem independent bound for episodic MDP, while He et al. [2021] showed an $\widetilde{\mathcal{O}}(d^2H^5\text{gap}_{\min}^{-1}\log^3(T))$ problem dependent regret bound for the same algorithm. Zhou et al. [2021b] studied infinite horizon MDP with discounted reward setting and proposed UCLK algorithm to achieve $\widetilde{\mathcal{O}}(d\sqrt{T}(1-\gamma)^2)$ regret. Most recently, Zhou et al. [2021a] proposed nearly minimax optimal algorithms for learning Linear Mixture MDPs in both finite and infinite horizon settings.

However, these works all assume a single representation and do not depend on the quality of the representation as long as it can well approximate the value function. Thus, what a good representation is and what improvement this good representation can bring is still an open question.

Offline Reinforcement Learning with Function Approximation There is a series of works focusing on the offline reinforcement learning with linear function approximation. Jin et al. [2021] introduce the pessimism to offline reinforcement learning and establish a data-dependent upper bound on the sub-optimality for general MDP. They also provide a close-formed data-dependent bound for linear MDPs. Following that, Xie et al. [2021] introduces the notion of Bellman's consistent pessimism for general function approximation. There is also a brunch of work leveraging the variance information in offline RL [Min et al., 2021, Yin et al., 2021, 2022]. Other follow-up works include the partial coverage [Uehara and Sun, 2021] in general function approximation and the statistical barriers for offline RL [Foster et al., 2021].

Model Selection and Representation Learning in Contextual Bandits. Since contextual bandits can be viewed as a special case of MDPs, our work is also related to some previous works on model selection in contextual bandits. The first line of work runs a multi-armed bandit at a high level while each arm corresponds to a low level contextual bandit algorithm. Following this line, Odalric and Munos [2011] used a variant of EXP4 [Auer et al., 2002] as the master algorithm while the EXP3 or UCB algorithm [Auer et al., 2002] serves as the base algorithm. This result is improved by CORRAL [Agarwal et al., 2017], which uses the online mirror descent framework and modifies the base algorithm to be compatible with the

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

master. Pacchiano et al. [2020b] introduced a generic smoothing wrapper that can be directly applied to the base algorithms without modification.

Abbasi-Yadkori et al. [2020] proposed a regret balancing strategy and showed that given the regret bound for the optimal base algorithm as an input, their algorithm can achieve a regret that is close to the regret of the optimal base algorithm. Following that, Pacchiano et al. [2020a] relaxed the requirement in Abbasi-Yadkori et al. [2020] by knowing each base algorithm comes with a candidate regret bound that may or may not hold during all rounds. Despite this progress, how to get the optimal regret guarantee for the general contextual learning problem remains an open question [Foster et al., 2020]. Besides those general model selection algorithms, recent works are focusing on representation learning under several different structures, thus different representations can be used at different rounds in the algorithm. Foster et al. [2019] studied model selection by considering a sequence of feature maps with increasing dimensions where the losses are linear in one of these feature maps. They proposed an algorithm that adaptively learns the optimal feature map, whose regret is independent of the maximum dimension. Chatterji et al. [2020] studied the hidden simple multi-armed bandit structure where the rewards are independent of the contextual information. Ghosh et al. [2021] considered a nested linear contextual bandit problem where the algorithm treats the norm bound or dimension of the weight vector in the linear model as the complexity of the problem and adaptively finds the true complexity for the given dataset.

2 EXPERIMENT DETAILS

2.1 ONLINE REINFORCEMENT LEARNING

Here we describe how to generate the representation functions and the MDP. We denote the d-dimensional half normal distribution by $|\mathbf{x}| \sim \mathcal{H}(\mathbf{I}_d)$ if $\mathbf{x} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$. Considering the following representation sample from the half-normal distribution for all $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$:

$$\widetilde{\boldsymbol{\phi}}(s, a) \sim \mathcal{H}(\mathbf{I}_d), \widetilde{\boldsymbol{\psi}}(s') \sim \mathcal{H}(\mathbf{I}_d).$$

Then we define $\boldsymbol{\psi}$ by $\boldsymbol{\psi}(s') = \widetilde{\boldsymbol{\psi}}(s') / \max_{s \in \mathcal{S}} \|\widetilde{\boldsymbol{\psi}}(s)\|_2$. It is obvious that the Euclidean norm of $\boldsymbol{\psi}(s')$ is bounded by 1 for all $s' \in \mathcal{S}$.

It is easy to tell that each element in $\widetilde{\boldsymbol{\phi}}$ and $\boldsymbol{\psi}$ is non-negative thus we can build the transition kernel as

$$\mathbb{P}_h(s'|s, a) = \frac{\widetilde{\boldsymbol{\phi}}(s, a)^\top \boldsymbol{\psi}(s')}{\sum_{s' \in \mathcal{S}} \widetilde{\boldsymbol{\phi}}(s, a)^\top \boldsymbol{\psi}(s')}.$$

Next, for any step $h \in [H]$, given any non-singular matrix $\mathbf{M}_h \in \mathbb{R}^{d \times d}$, we define representation function $\boldsymbol{\phi}_h(\cdot, \cdot)$ as

$$\boldsymbol{\phi}_h(s, a) = \frac{\mathbf{M}_h^{-1} \widetilde{\boldsymbol{\phi}}(s, a)}{\sum_{s' \in \mathcal{S}} \widetilde{\boldsymbol{\phi}}(s, a)^\top \boldsymbol{\psi}(s')}. \tag{2.1}$$

Furthermore, we select the matrix $\mathbf{M}_h$ such that for all state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, $\|\boldsymbol{\phi}_h(s, a)\|_2 \leq 1$. This procedure could always be done since we can multiply different scalars to the generated matrix $\mathbf{M}_h$ to control the norm of $\boldsymbol{\phi}_h$.

Therefore, we can verify that for any $(s, a, s', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$, $\mathbb{P}_h(s'|s, a) = \boldsymbol{\phi}_h^\top(s, a) \mathbf{M}_h \boldsymbol{\psi}(s')$ thus it satisfies Assumption 4.2. To emphasize the difficulties of learning the transition kernel $\mathbb{P}$.

It is easy to verify that under the current representation $\boldsymbol{\phi}$, with high probability, $\boldsymbol{\Lambda}_{h,\boldsymbol{\phi}} \succeq \mathbf{0}$ since the representation $\boldsymbol{\phi}$ is sampled from the half-normal distribution. Therefore, Assumption 4.2 is satisfied.

We will next provide two other representations $\{\boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(2)}\}$ for the same transition kernel $\mathbb{P}_h(\cdot|\cdot, \cdot)$. Neither of these single representation satisfies Assumption 4.2 but the combination of these two will satisfy that assumption.

Since the transition kernel $\mathbb{P}_h(\cdot|\cdot, \cdot)$ and reward function $r(\cdot, \cdot)$ have already been determined, by Bellman optimality equation (3.1), we can get the optimal action action $\pi_h^*(s)$ for all step $h \in [H]$ and state $s \in \mathcal{S}$. Since $|\mathcal{A}| = 3$ and $\pi_h^*(s) \in \mathcal{A}$, we can compose the sub-optimal set by $\mathcal{A} \setminus \{\pi_h^*(s)\} := \{a_h(s), a_h'(s)\}$. Then we define the two representation

functions as $\boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(2)} \in \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^{2d}$ using the following rule:

$$\begin{cases} \boldsymbol{\phi}_h^{(1)}(s, \pi_h^*(s)) = \left(\boldsymbol{\phi}_h^\top(s, \pi_h^*(s)), \mathbf{0}_d^\top\right)^\top \\ \boldsymbol{\phi}_h^{(1)}(s, a_h(s)) = \left(\boldsymbol{\phi}_h^\top(s, a_h(s)), \mathbf{0}_d^\top\right)^\top \\ \boldsymbol{\phi}_h^{(1)}(s, a_h'(s)) = \left(\mathbf{0}_d^\top, \boldsymbol{\phi}_h^\top(s, a_h'(s))\right)^\top \end{cases}, \begin{cases} \boldsymbol{\phi}_h^{(2)}(s, \pi_h^*(s)) = \left(\boldsymbol{\phi}_h^\top(s, \pi_h^*(s)), \mathbf{0}_d^\top\right)^\top \\ \boldsymbol{\phi}_h^{(2)}(s, a_h(s)) = \left(\mathbf{0}_d^\top, \boldsymbol{\phi}_h^\top(s, a_h(s))\right)^\top \\ \boldsymbol{\phi}_h^{(2)}(s, a_h'(s)) = \left(\boldsymbol{\phi}_h^\top(s, a_h'(s)), \mathbf{0}_d^\top\right)^\top \end{cases}$$

By constructing the new kernel matrix $\widetilde{\mathbf{M}}_h = (\mathbf{M}_h^\top, \mathbf{M}_h^\top)^\top \in \mathbb{R}^{2d \times d}$, we can verify that both $\boldsymbol{\phi}^{(1)}$ and $\boldsymbol{\phi}^{(2)}$ satisfy Assumption 3.2 with dimension $2d$. i.e. for all $(s, a, s', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$

$$\mathbb{P}_h(s'|s, a) = \boldsymbol{\phi}_h^{(1)\top} \widetilde{\mathbf{M}}_h \boldsymbol{\psi}(s') = \boldsymbol{\phi}_h^{(2)\top} \widetilde{\mathbf{M}}_h \boldsymbol{\psi}(s'). \tag{2.2}$$

From intuition, these two representation functions $\boldsymbol{\phi}^{(1)}$ and $\boldsymbol{\phi}^{(2)}$ may come from two different sensors measuring the same environment. It is obvious that since for both $\boldsymbol{\phi}^{(1)}$ and $\boldsymbol{\phi}^{(2)}$, there are at least 1/3 of the whole state-action space is not covered by $\boldsymbol{\Lambda}_h$, i.e. $\boldsymbol{\phi}_h^{(1)}(s, a_h'(s)) \notin \text{Im}\boldsymbol{\Lambda}_{h,\boldsymbol{\phi}^{(1)}}$ and $\boldsymbol{\phi}_h^{(2)}(s, a_h(s)) \notin \text{Im}\boldsymbol{\Lambda}_{h,\boldsymbol{\phi}^{(2)}}$. However, since $a_h'(s') \neq a_h(s)$ by definition, we can verify that the representation set $\Phi = \{\boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(2)}\}$ satisfies Assumption 4.2.

2.2 OFFLINE REINFORCEMENT LEARNING

Here we provide a design for representation functions such that each single representation does not satisfy Assumption 5.1 but the whole representation function set satisfies.

First, the oracle representation which satisfies Assumption 5.1 is generated same as (2.1) with $d = 5$. The underlying MDP is generated same with the online settings with $|\mathcal{S}| = 20, |\mathcal{A}| = 3$. Then we consider an arbitrary behavior policy $\widehat{\pi}$ which is used to generate the offline training data. Since $|\mathcal{A}| = 3$, for any $s \in \mathcal{S}, h \in [H]$, there exists three state-action pairs as $(s, \widehat{\pi}_h(s)), (s, a_h(s))$ and $(s, a_h'(s))$. Considering the representation set $\boldsymbol{\phi}^{(1)}(\cdot, \cdot) \in \mathbb{R}^{2d}$ and $\boldsymbol{\phi}^{(2)}(\cdot, \cdot) \in \mathbb{R}^{2d}$ which is defined as

$$\begin{cases} \boldsymbol{\phi}_h^{(1)}(s, \widehat{\pi}_h(s)) &= \left(\boldsymbol{\phi}_h^\top(s, \widehat{\pi}_h(s))^\top, \mathbf{0}_d^\top\right)^\top \\ \boldsymbol{\phi}_h^{(1)}(s, a_h(s)) &= \left(\boldsymbol{\phi}_h^\top(s, a_h(s))^\top, \mathbf{0}_d^\top\right)^\top \\ \boldsymbol{\phi}_h^{(1)}(s, a_h'(s)) &= \left(\mathbf{0}_d^\top, \boldsymbol{\phi}_h^\top(s, a_h'(s))^\top\right)^\top \end{cases}, \begin{cases} \boldsymbol{\phi}_h^{(2)}(s, \widehat{\pi}_h(s)) &= \left(\boldsymbol{\phi}_h^\top(s, \widehat{\pi}_h(s))^\top, \mathbf{0}_d^\top\right)^\top \\ \boldsymbol{\phi}_h^{(2)}(s, a_h(s)) &= \left(\mathbf{0}_d^\top, \boldsymbol{\phi}_h^\top(s, a_h(s))^\top\right)^\top \\ \boldsymbol{\phi}_h^{(2)}(s, a_h'(s)) &= \left(\boldsymbol{\phi}_h^\top(s, a_h'(s))^\top, \mathbf{0}_d^\top\right)^\top \end{cases}.$$

It is obvious that by using behavior policy $\widehat{\pi}$, both $\mathbb{E}_{d_h^{\widehat{\pi}}}(\boldsymbol{\phi}^{(1)}\boldsymbol{\phi}^{(1)\top})$ and $\mathbb{E}_{d_h^{\widehat{\pi}}}(\boldsymbol{\phi}^{(2)}\boldsymbol{\phi}^{(2)\top})$ would enjoy the format of $\begin{pmatrix} \boldsymbol{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$. Therefore, for $\boldsymbol{\phi}^{(1)}$, it would be easy to verify that $\boldsymbol{\phi}_h^{(1)}$, $\boldsymbol{\phi}^{(1)}(s, a_h'(s))$ is not in $\text{Im}(\mathbb{E}_{d_h^{\widehat{\pi}}}(\boldsymbol{\phi}^{(1)}\boldsymbol{\phi}^{(1)\top}))$ and $(\boldsymbol{\phi}^{(2)}(s, a_h(s))$ is not in $\text{Im}(\mathbb{E}_{d_h^{\widehat{\pi}}}(\boldsymbol{\phi}^{(2)}\boldsymbol{\phi}^{(2)\top}))$. However, it is also easy to show that the union of $\{\boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(2)}\}$ satisfy assumption 5.1.

2.3 ADDITIONAL CONFIGURATION

Parameter Tuning. For both of the offline and online algorithm, we aggregate the parameter $C_{\boldsymbol{\psi}} H \sqrt{\beta_{k,\boldsymbol{\phi}}}$ as a single hyper-parameter C for tuning. We do a grid search for $C = \{1, 3, 10, 30, 100\}$ report the best performance over these values.

2.4 ADDITIONAL RESULTS

2.4.1 Online RL

Figure 1 plots the cumulative regret with respect to the episode number, with the standard deviation indicated by the shadows. We observed that the cumulative regrets for both UC-MatrixRL using $\boldsymbol{\phi}$ and ReLEX-UCB grew very slowly after the first one million episodes. As a comparison, UC-MatrixRL using $\boldsymbol{\phi}^{(1)}$ or $\boldsymbol{\phi}^{(2)}$ had a sub-linear regret growth instead of near-constant regret. As for the ϵ-greedy algorithm, although the greedy policy can learn very fast at the beginning, it eventually had a much higher cumulative regret since it could not explore the environment well.

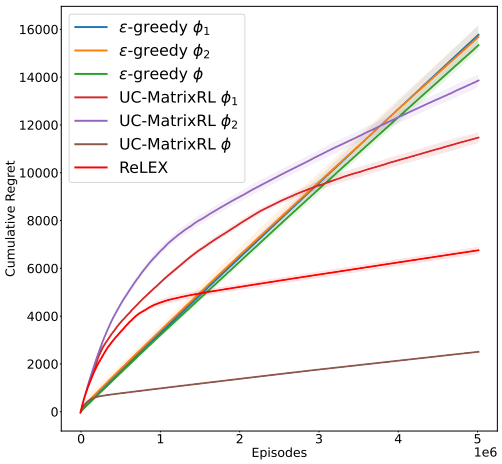

Figure 1: Cumulative regret over 5M episodes for ReLEX-UCB v.s. UC-MatrixRL and ϵ-greedy using a single representation.

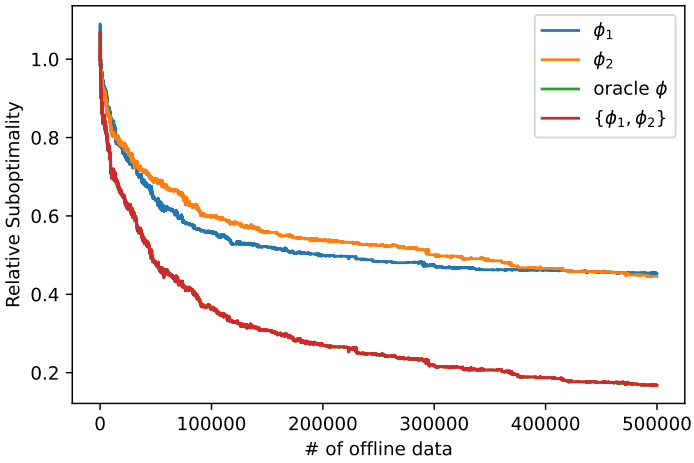

Figure 2: Relative sub-optimality of ReLEX-LCB after 500K offline episodes

2.4.2 Offline RL

From Figure 2 we observe that by selecting over two imperfect representations, ReLEX-LCB can match the performance of the oracle algorithm using a single perfect representation, even if using the two representations separately leads to a larger ($\sim 2.5\times$) sub-optimality on the same offline data.

2.4.3 Ablation studies

We conduct additional experiments on the following different settings in the online setting as the ablation study of our algorithm.

(A) The original setting with same data generation, $S = 20, A = 3, d = d' = 5, H = 10$

(B) The original setting with larger state space $S = 40$, other parameters are not changed

(C) The original setting with larger action $A = 5$, other parameters are not changed

(D) The original setting with larger action $S = 40, A = 5$, other parameters are not changed

(E) The original setting with $|\Phi| = 3, A = 4$, other parameters are not changed

Besides the cumulative regret, we also report the average reward achieved in the last 1000 episodes, which can be considered a "more intuitive performance metric" suggested by Reviewer 4Uhq.

Regarding the data generation, configuration C to E enjoys the same method of generating the data, i.e., arranging the context of sub-optimal actions into other dimensions. We will add the details of generating these data during the revision.

Due to the time limit of the authors' response, we do not repeat the experiments multiple times and we cut experiments E and F with episode $K = 500,000$ instead of the original $K = 5,000,000$ in the paper. We also skipped the ϵ-greedy version for configurations E for the sake of time.

The performance table are presented From Table 1 to Table 5

Computing Resources For both offline and online algorithm, we conduct our experiments on an AWS c5-12xlarge CPU instance with a 48-core Intel® Xeon® Scalable Processors (Cascade Lake).

3 PROOF OF THEOREM 4.5

In this section we will give the key technical lemmas and the proof sketch for Theorem 4.5.

Algorithm	Last averaged reward ↑	Cumulative regret ↓
UC-Matrix RL ϕ (oracle)	0.7782	2457.71
ReLEX $\{\phi_1, \phi_2\}$	**0.7780**	**6827.18**
UC-Matrix RL ϕ_1	**0.7780**	11458.83
UC-Matrix RL ϕ_2	0.7770	13385.61
ϵ-greedy ϕ	0.7754	14906.31
ϵ-greedy ϕ_1	0.7756	15042.24
ϵ-greedy ϕ_2	0.7751	16470.94

Table 1: The performance result of Configuration (A) (The same configuration in the paper)

Algorithm	Last averaged reward ↑	Cumulative regret ↓
UC-Matrix RL ϕ (oracle)	0.8736	2880.71
ReLEX $\{\phi_1, \phi_2\}$	**0.8733**	**7745.49**
UC-Matrix RL ϕ_1	0.8729	12759.89
UC-Matrix RL ϕ_2	0.8730	10411.89
ϵ-greedy ϕ	0.8703	18786.77
ϵ-greedy ϕ_1	0.8707	19002.81
ϵ-greedy ϕ_2	0.8702	20018.97

Table 2: The performance result of Configuration (B) ($S = 40$)

Algorithm	Last averaged reward ↑	Cumulative regret ↓
UC-Matrix RL ϕ (oracle)	0.9749	3085.21
ReLEX $\{\phi_1, \phi_2\}$	**0.9748**	**8160.39**
UC-Matrix RL ϕ_1	0.9743	12946.34
UC-Matrix RL ϕ_2	0.9745	14373.86
ϵ-greedy ϕ	0.9690	28423.82
ϵ-greedy ϕ_1	0.9701	27479.92
ϵ-greedy ϕ_2	0.9708	27778.66

Table 3: The performance result of Configuration (C) ($A = 5$)

Algorithm	Last averaged reward ↑	Cumulative regret ↓
UC-Matrix RL ϕ (oracle)	0.9800	3403.65
ReLEX $\{\phi_1, \phi_2\}$	**0.9792**	9733.84
UC-Matrix RL ϕ_1	0.9787	10000.15
UC-Matrix RL ϕ_2	0.9791	**9553.96**
ϵ-greedy ϕ	0.9763	23301.53
ϵ-greedy ϕ_1	0.9759	23392.78
ϵ-greedy ϕ_2	0.9758	23218.40

Table 4: The performance result of Configuration (D) ($S = 40, A = 5$)

Algorithm	Last averaged reward ↑	Cumulative regret ↓
UC-Matrix RL ϕ (oracle)	0.9081	1141.63
UC-Matrix RL ϕ_1	0.9034	4512.82
UC-Matrix RL ϕ_2	0.9008	4965.15
UC-Matrix RL ϕ_3	0.9022	4507.18
ReLEX $\{\phi_1, \phi_2\}$	0.9051	2865.86
ReLEX $\{\phi_1, \phi_3\}$	0.9057	2606.99
ReLEX $\{\phi_2, \phi_3\}$	0.90648	2702.94
ReLEX $\{\phi_1, \phi_2, \phi_3\}$	**0.9080**	**2093.32**

Table 5: The performance result of Configuration (E) ($S = 20, A = 4, |\Phi| = 3$)

First, we need to define a "good event" which happens with high probability, that the estimation $\mathbf{M}_h^k$ is close to the target $\mathbf{M}_h^*$. This definition was originally introduced in Yang and Wang [2020].

Lemma 3.1 (Lemma 15, Yang and Wang [2020])**.** Define the following event as $\mathcal{E}_{\boldsymbol{\phi}}^k$,

$$\left\{ \operatorname{tr}\left[(\mathbf{M}_{h,\boldsymbol{\phi}}^j - \mathbf{M}_{h,\boldsymbol{\phi}}^*)^\top \mathbf{U}_{h,\boldsymbol{\phi}}^j (\mathbf{M}_{h,\boldsymbol{\phi}}^j - \mathbf{M}_{h,\boldsymbol{\phi}}^*) \right] \le \beta_{j,\boldsymbol{\phi}}, \forall j \le k, \forall h \in [H] \right\} =: \mathcal{E}_{\boldsymbol{\phi}}^k,$$

With $\beta_{k,\boldsymbol{\phi}} = c(C_{\mathbf{M}} + {C'_{\boldsymbol{\psi}}}^2) d_{\boldsymbol{\phi}} \log(kHC_{\boldsymbol{\phi}}/\delta)$ for some absolute constant $c > 0$, we have $\Pr(\mathcal{E}_{\boldsymbol{\phi}}^K) \ge 1 - \delta$ for all $\boldsymbol{\phi} \in \Phi$.

Remark 3.2. The proof of Lemma 3.1 remains the same since the regression does not depend on the policy π. We also make the dependency of δ explicit in $\beta_{k,\boldsymbol{\phi}}$, which can be inferred from the proof of Lemma 15 in Yang and Wang [2020].

The next lemma shows a problem-dependent regret bound for the bilinear MDP in Definition 3.2.

Lemma 3.3. Under Assumption 3.1, setting parameter $\beta_{k,\boldsymbol{\phi}}$ as in Theorem 4.5. Then suppose $\mathcal{E}_{\boldsymbol{\phi}}^K$ holds for all $\boldsymbol{\phi} \in \Phi$. Then with probability at least $1 - 3\delta$, the regret for the very first $k \in [K]$ episodes is controlled by

$$\begin{aligned} \text{Regret}(k) \le \min_{\boldsymbol{\phi} \in \Phi} \left\{ \frac{128 C_{\boldsymbol{\psi}}^2 H^5 d_{\boldsymbol{\phi}} \beta_{k,\boldsymbol{\phi}} \log(1 + C_{\boldsymbol{\phi}} k d_{\boldsymbol{\phi}})}{\text{gap}_{\min}} \right\} + \frac{16}{3} H^2 \log(((1 + \log(Hk)) k^2 |\Phi|/\delta) \\ + 2 + \frac{96 H^4 \log(2k(1 + \log(H/\text{gap}_{\min}))|\Phi|/\delta)}{\text{gap}_{\min}} \end{aligned} \tag{3.1}$$

while the sub-optimality gap for each h is controlled by

$$\sum_{i=1}^{k} (V_h^*(s_h^i) - Q_h^*(s_h^i, a_h^i)) \le \min_{\boldsymbol{\phi} \in \Phi} \left\{ \frac{64 C_{\boldsymbol{\psi}}^2 H^4 d_{\boldsymbol{\phi}} \beta_{k,\boldsymbol{\phi}} \log(1 + C_{\boldsymbol{\phi}} k d_{\boldsymbol{\phi}}) + 48 H^3 \log(2k|\Phi|(1 + \log(H/\text{gap}_{\min})/\delta)}{\text{gap}_{\min}} \right\}. \tag{3.2}$$

It is easy to verify that when there is only one representation function in Φ (let $d = d_{\boldsymbol{\phi}}$ for simplicity), Lemma 3.3 yields an $\mathcal{O}(H^5 d^2 \log(k/\delta) \text{gap}_{\min}^{-1})$ problem-dependent bound. Comparing our result with He et al. [2021], ours matches the problem-dependent bound for Linear Mixture MDP $\mathcal{O}(H^5 d^2 \log(k/\delta) \text{gap}_{\min}^{-1})$ and is better than the problem-dependent bound for Linear MDP $\mathcal{O}(H^5 d^3 \log(k/\delta) \text{gap}_{\min}^{-1})$ by a factor d. This improvement is due to the bilinear MDP structure in Definition 3.2. Moreover, it is obvious that when $|\Phi| > 1$, Algorithm 1 can achieve a regret no worse than any possible regret achieved by a single representation, up to an additive $\log(|\Phi|)$ term. Lemma 3.3 also suggests an $\mathcal{O}(H^4 d^2 \log(k/\delta) \text{gap}_{\min}^{-1})$ bound for the summation of the sub-optimality gap. Based on this, the next lemma shows that the "covariance matrix" $\mathbf{U}_{h,\boldsymbol{\phi}}^k$ is almost linearly growing with respect to k under Assumption 4.2.

Lemma 3.4. Under Assumptions 3.1 and 4.2, with probability at least $1 - \delta$, we have for all $k \in [K], h \in [H], \boldsymbol{\phi} \in \Phi$,

$$\begin{aligned} &\mathbf{U}_{h,\boldsymbol{\phi}}^k \succeq (k-1)\boldsymbol{\Lambda}_{h,\boldsymbol{\phi}} - \iota \mathbf{I}_{d_{\boldsymbol{\phi}}}, \\ &\iota = \frac{C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}}}{\text{gap}_{\min}} \sum_{i=1}^{h} \sum_{j=1}^{k-1} \text{gap}_i(s_i^j, a_i^j) - 1 + C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}} \sqrt{32 H (k-1) \log(d_{\boldsymbol{\phi}} |\Phi| H k (k+1)/\delta)}. \end{aligned}$$

Compared with Lemma 9 in Papini et al. [2021] which shows a similar result for linear contextual bandits, the proof of Lemma 3.4 is more challenging: The distribution of s_h is induced by the optimal policy π^* in Assumption 4.2 but we can only use the estimated policy π^k to sample s_h. As a result, the sub-optimality and the randomness for the steps before h ($i < h$) will all contribute to this distribution mismatch. Therefore, our result contains an additional summation over h to account for this effect.

Finally, equipped with Lemma 3.4, we can provide a constant threshold τ such that if the episode number k goes beyond τ, the sub-optimality gap is bounded by $\mathcal{O}(\sqrt{1/k})$.

Lemma 3.5. Under Assumptions 3.1 and 4.2, assuming the conditions in Lemmas 3.3 and 3.4 hold and $\mathcal{E}_{\boldsymbol{\phi}}^K$ holds for all $\boldsymbol{\phi} \in \Phi$, then there exists a threshold

$$\tau = \text{poly}(d_{\boldsymbol{\phi}}, \sigma_{\boldsymbol{\phi}}^{-1}, H, \log(|\Phi|/\delta), \text{gap}_{\min}^{-1}, C_{\boldsymbol{\phi}}, C_{\boldsymbol{\psi}}, C_{\mathbf{M}}, C'_{\boldsymbol{\psi}})$$

such that for all $\tau \leq k \leq K$, for all $h \in [H], s \in \mathcal{S}$ we have

$$\text{gap}_h(s, \pi_h^k(s)) \leq 2C_{\boldsymbol{\psi}} H^2 \max_{\boldsymbol{\phi} \in \Phi} \left\{ d_{\boldsymbol{\phi}} \sqrt{2C_{\boldsymbol{\phi}} \beta_{k,\boldsymbol{\phi}} / (\sigma_{\boldsymbol{\phi}} k)} \right\}.$$

Lemma 3.5 suggests that when the episode number k exceeds τ, policy π^k will contribute a sub-optimality up to $\mathcal{O}(\sqrt{1/k})$. Thus there exists a threshold k^* such that when $k \geq k^*$, the policy π^k will not contribute any sub-optimality at any step h given the minimal sub-optimality gap $\text{gap}_{\min}$. With that, it suffices to provide the proof for Theorem 4.5.

Proof of Theorem 4.5. We pick $\beta_{k,\boldsymbol{\phi}} = c(C_{\mathbf{M}} + {C'_{\boldsymbol{\psi}}}^2) d_{\boldsymbol{\phi}} \log(kHC_{\boldsymbol{\phi}}|\Phi|/\delta)$ to make sure with probability at least $1-\delta$, event $\mathcal{E}_{\boldsymbol{\phi}}^K$ holds for all $\boldsymbol{\phi} \in \Phi$. By the definition of the sub-optimality gap, we have $\text{gap}_h(s,a) \geq \text{gap}_{\min}(s,a)$ as long as $\text{gap}_h(s,a) \neq 0$. If

$$k > \max \left\{ \frac{8C_{\boldsymbol{\psi}}^2 H^4}{\text{gap}_{\min}^2} \max_{\boldsymbol{\phi} \in \Phi} \left\{ \frac{C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}}^2 \beta_{k,\boldsymbol{\phi}}}{\sigma_{\boldsymbol{\phi}}} \right\}, \tau \right\}. \tag{3.3}$$

Then we have for all $\boldsymbol{\phi} \in \Phi$,

$$2C_{\boldsymbol{\psi}} H^2 d_{\boldsymbol{\phi}} \sqrt{2C_{\boldsymbol{\phi}} \beta_{k,\boldsymbol{\phi}} / (\sigma_{\boldsymbol{\phi}} k)} < \text{gap}_{\min}.$$

Thus, by Lemma 3.5, we have

$$\text{gap}_h(s, \pi_h^k(s)) \leq 2C_{\boldsymbol{\psi}} H^2 \max_{\boldsymbol{\phi} \in \Phi} \left\{ d_{\boldsymbol{\phi}} \sqrt{2C_{\boldsymbol{\phi}} \beta_{k,\boldsymbol{\phi}} / (\sigma_{\boldsymbol{\phi}} k)} \right\},$$

and it implies $\text{gap}_h(s, \pi_h^k(s)) > \text{gap}_{\min}$ thus $\text{gap}_h(s, \pi_h^k(s)) = 0$. Since from the parameter setting, $\beta_{k,\boldsymbol{\phi}} = \mathcal{O}(\log(k))$, it is easy to verify that there exists a threshold $k^* = \text{poly}(C_{\boldsymbol{\psi}}, C'_{\boldsymbol{\psi}}, C_{\mathbf{M}}, C_{\boldsymbol{\phi}}, d_{\boldsymbol{\phi}}, H, \sigma_{\boldsymbol{\phi}}^{-1}, \text{gap}_{\min}^{-1}, \tau)$ such that all $k \geq k^*$ satisfy (3.3). Thus we conclude that $\text{gap}_h(s, \pi_h^k(s)) = 0$ for all $k \geq k^*$. Since the optimal policy π^* is unique, it follows that $\pi^k = \pi^*$.

Thus when $k \geq k^*$, the regret could be decomposed by

$$\begin{aligned}
\text{Regret}(k) &= \sum_{j=1}^{k} V_1^*(s_1^j) - V_1^{\pi^j}(s_1^j) \\
&= \sum_{j=1}^{k^*} V_1^*(s_1^j) - V_1^{\pi^j}(s_1^j) + \sum_{j=1+k^*}^{k} V_1^*(s_1^j) - V_1^{\pi^j}(s_1^j) \\
&= \text{Regret}(k^*) + 0,
\end{aligned}$$

where the last equation is due to the fact that $\pi^j = \pi^*$ thus $V_1^*(s) = V_1^{\pi^j}(s)$ for all $s \in \mathcal{S}$ when $j \geq k^*$. Combining this case with the case $k \leq k^*$, we can conclude that $\text{Regret}(k) \leq \text{Regret}(\min\{k, k^*\})$. Let $\widetilde{k} = \min\{k, k^*\}$, by Lemma 3.3,

we have the regret is bounded by

$$\begin{aligned}\text{Regret}(k) \leq \min_{\boldsymbol{\phi}\in\Phi}\Bigg\{ & \frac{128C_{\boldsymbol{\psi}}^2 H^5 d_{\boldsymbol{\phi}}^2 c(C_{\mathbf{M}} + {C'_{\boldsymbol{\psi}}}^2)}{\text{gap}_{\min}} \log\left(1 + C_{\boldsymbol{\phi}}\widetilde{k}d_{\boldsymbol{\phi}}\right)\log\left(\widetilde{k}HC_{\boldsymbol{\phi}}|\Phi|/\delta\right)\Bigg\} + 2 \\ & + \frac{96H^4\log\left(2\widetilde{k}(1+\log(H/\text{gap}_{\min}))|\Phi|/\delta\right)}{\text{gap}_{\min}} \\ & + \frac{16}{3}H^2\log\left(\left(\left(1+\log(H\widetilde{k}\right)\right)\widetilde{k}^2|\Phi|/\delta\right),\end{aligned}$$

with probability at least $1-5\delta$ by taking the union bound of Lemma 3.3, Lemma 3.4 and $\mathcal{E}_{\boldsymbol{\phi}}^K$ holds. □

4 PROOF OF LEMMAS IN APPENDIX 3

In this section, we provide the proof of the technical lemmas in Appendix 3.

4.1 FILTRATION

To facilitate our proof, we define the filtration list as follows

$$\mathcal{F}_h^k = \left\{\left\{s_i^j, a_i^j\right\}_{i=1,j=1}^{H,k-1}, \left\{s_i^k, a_i^k\right\}_{i=1}^{h}\right\}.$$

It is easy to verify that s_h^k, a_h^k are both $\mathcal{F}_h^k$-measurable. Also, for any function f built on $\mathcal{F}_h^k$, $f(s_{h+1}^k) - [\mathbb{P}_h f](s_h^k, a_h^k)$ is $\mathcal{F}_{h+1}^k$-measurable and it is also a zero-mean random variable conditioned on $\mathcal{F}_h^k$.

Arranging the filtrations as

$$\mathcal{F} = \{\mathcal{F}_1^1, \cdots, \mathcal{F}_H^1, \cdots, \mathcal{F}_1^k, \cdots, \mathcal{F}_h^k, \cdots \mathcal{F}_H^k, \cdots, \mathcal{F}_1^K, \cdots, \mathcal{F}_H^K\},$$

we will use $\mathcal{F}$ as the filtration set for the following proof.

4.2 PROOF OF LEMMA 3.3

To prove this lemma, we first need the following lemma showing the estimator $Q_{h,\boldsymbol{\phi}}^k$ is always optimistic.

Lemma 4.1. Suppose the event $\mathcal{E}_{\boldsymbol{\phi}}^K$ holds for all $\boldsymbol{\phi} \in \Phi$, then for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, $Q_h^*(s,a) \leq Q_{h,\boldsymbol{\phi}}^k(s,a)$.

The next lemma suggests the error between the estimated Q-function and the target Q-function at time-step h can be controlled by the error at $(h+1)$-th step and the UCB bonus term.

Lemma 4.2. Suppose the event $\mathcal{E}_{\boldsymbol{\phi}}^K$ holds, then for any $(s,a) \in \mathcal{S}\times\mathcal{A}, k \in [K]$ and any policy π,

$$Q_h^k(s,a) - Q_h^\pi(s,a) \leq 2C_{\boldsymbol{\psi}}H\sqrt{\beta_{k,\boldsymbol{\phi}}\boldsymbol{\phi}^\top(s,a)(\mathbf{U}_{h,\boldsymbol{\phi}}^k)^{-1}\boldsymbol{\phi}(s,a)} + [\mathbb{P}_h(V_{h+1}^k - V_{h+1}^\pi)](s,a).$$

We also need the following lemma, which is similar to Lemma 6.2 in He et al. [2021].

Lemma 4.3. For any $0 < \Delta \leq H$, if the event $\mathcal{E}_{\boldsymbol{\phi}}^K$ holds for all $\boldsymbol{\phi} \in \Phi$, then with probability at least $1-\delta$, for any $k \in [K]$,

$$\sum_{j=1}^k \mathbb{1}[V_h^*(s_h^j) - Q_h^{\pi^j}(s_h^j, a_h^j) \geq \Delta] \leq \frac{16C_{\boldsymbol{\psi}}H^4 d_{\boldsymbol{\phi}}\beta_{k,\boldsymbol{\phi}}\log(1+C_{\boldsymbol{\phi}}kd_{\boldsymbol{\phi}}) + 12H^3\log(2k/\delta)}{\Delta^2}.$$

Then we need the following lemma from He et al. [2021] to upper-bound the regret by the summation of the sub-optimailty.

Lemma 4.4 (Lemma 6.1, revised, He et al. [2021])**.** For each MDP $\mathcal{M}$, with probability at least $1-2\delta$, for all $k \in [K]$, we have

$$\text{Regret}(k) \leq 2\sum_{j=1}^{k}\sum_{h=1}^{H} \text{gap}_h(s_h^j, a_h^j) + \frac{16H^2}{3}\log(((1+\log(Hk))k^2/\delta) + 2.$$

Remark 4.5. Lemma 4.4 can be easily obtained from Lemma 6.1 in He et al. [2021]. In the original lemma, with probability at least $1 - \lceil \log HK \rceil \exp(-\tau)$,

$$\text{Regret}(k) \leq 2\sum_{j=1}^{k}\sum_{h=1}^{H} \text{gap}_h(s_h^j, a_h^j) + \frac{16H^2\tau}{3} + 2,$$

which implies that with probability at least $1-\delta$,

$$\text{Regret}(k) \leq 2\sum_{j=1}^{k}\sum_{h=1}^{H} \text{gap}_h(s_h^j, a_h^j) + \frac{16H^2\log(\lceil \log(Hk) \rceil/\delta)}{3} + 2.$$

By relaxing $\lceil \log(Hk) \rceil$ to $\log(HK)+1$ and replacing δ with δ/k^2 for different episode number k, the inequality holds with probability at least $1 - \sum_{k=1}^{K} \delta/k^2 \geq 1 - \pi^2\delta/6 \geq 1 - 2\delta$ for all $k \in [K]$ by union bound.

Equipped with these lemmas, we can begin our proof.

Proof of Lemma 3.3. By the definition of $\text{gap}_{\min}$, for each $h \in [H], k \in [K]$, we have $V_h^*(s_h^k) - Q_h^*(s_h^k, a_h^k) = 0$ or $\text{gap}_{\min} \leq V_h^*(s_h^k) - Q_h^*(s_h^k, a_h^k) \leq H$. Dividing the interval $[\text{gap}_{\min}, H]$ into N intervals $[2^{n-1}\text{gap}_{\min}, 2^n\text{gap}_{\min})$ where $n \in [N], N = \lceil \log(H/\text{gap}_{\min}) \rceil$, then with probability at least $1 - \lceil \log(H/\text{gap}_{\min}) \rceil \delta$, it holds that

$$\begin{aligned}
\sum_{j=1}^{k}(V_h^*(s_h^j) - Q_h^*(s_h^j, a_h^j)) &\leq \sum_{n=1}^{N}\sum_{j=1}^{k} 2^n\text{gap}_{\min}\, \mathbb{1}[2^{n-1}\text{gap}_{\min} \leq V_h^*(s_h^j) - Q_h^*(s_h^j, a_h^j) \leq 2^n\text{gap}_{\min}] \\
&\leq \sum_{n=1}^{N}\sum_{j=1}^{k} 2^n\text{gap}_{\min}\, \mathbb{1}[2^{n-1}\text{gap}_{\min} \leq V_h^*(s_h^j) - Q_h^{\pi^j}(s_h^j, a_h^j)] \\
&\leq \sum_{n=1}^{N} \frac{64C_{\psi}^2H^4d_{\phi}\beta_{k,\phi}\log(1+C_{\phi}kd_{\phi}) + 48H^3\log(2k/\delta)}{2^n\text{gap}_{\min}} \\
&\leq \frac{64C_{\psi}^2H^4d_{\phi}\beta_{k,\phi}\log(1+C_{\phi}kd_{\phi}) + 48H^3\log(2k/\delta)}{\text{gap}_{\min}},
\end{aligned}$$

where the first inequality holds by using the "peeling technique", which was used in local Rademacher complexity analysis [Bartlett et al., 2005]. The second inequality in (4.1) is due to $Q_h^*(s,a) \geq Q_h^{\pi^j}(s,a)$ and the third inequality holds due to Lemma 4.3. Finally, the fourth inequality holds due to $\sum_{n=1}^{N} 2^{-n} \leq 1$. Substituting δ with $\delta/(1+\log(H/\text{gap}_{\min}))$, with probability at least $1-\delta$, we have

$$\sum_{j=1}^{k}(V_h^*(s_h^j) - Q_h^*(s_h^j, a_h^j)) \leq \frac{64C_{\psi}^2H^4d_{\phi}\beta_{k,\phi}\log(1+C_{\phi}kd_{\phi}) + 48H^3\log\left(2k(1+\log(H/\text{gap}_{\min})/\delta\right)}{\text{gap}_{\min}}, \tag{4.1}$$

Combining (4.1) with Lemma 4.4, by taking a union bound, with probability at least $1-3\delta$,

$$\begin{aligned}
\text{Regret}(k) &\leq 2\sum_{j=1}^{k}\sum_{h=1}^{H} \text{gap}_h(s_h^k, a_h^k) + \frac{16H^2\log(\lceil HK \rceil/\delta)}{3} + 2 \\
&\leq \frac{128C_{\psi}^2H^5d_{\phi}\beta_{k,\phi}\log(1+C_{\phi}kd_{\phi}) + 96H^4\log(2k(1+\log(H/\text{gap}_{\min}))/\delta)}{\text{gap}_{\min}} \\
&\quad + \frac{16}{3}H^2\log(((1+\log(Hk))k^2/\delta) + 2.
\end{aligned}$$

where the first inequality holds due to Lemma 4.4 , which utilizes the definition of the sub-optimality gap. The second inequality holds due to (4.1). Substituting δ with $\delta/|\Phi|$, the claimed result (3.1) holds for all $\phi \in \Phi$ by taking a union bound. □

4.3 PROOF OF LEMMA 3.4

For brevity, we denote matrix $\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s) = \boldsymbol{\phi}(s, \pi_h^*(s))\boldsymbol{\phi}^\top(s, \pi_h^*(s)) \in \mathbb{R}^{d\times d}$ and fix h, m in the proof. The expectation $\mathbb{E}_{s_h}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)|s_i], i < h$ is taken with respect to the randomness of the states sequence $s_{i+1}, \cdots, s_h$, where $s_{i'+1} \sim \mathbb{P}_{i'}\big(\cdot\,|s_{i'}, \pi_{i'}^*(s_{i'})\big), i \le i' < h$. If the action a_i is given, the expectation $\mathbb{E}_{s_h}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)|s_i, a_i], i < h$ is taken in which $s_{i+1} \sim \mathbb{P}_i(\cdot|s_i, a_i)$ specially. It is worthless to show that $\mathbb{E}_{s_h}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)|s_h^k] = \mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h^k)$. Without specification, we ignore the subscript s_h in the expectation in the proof of this lemma.

To develop the convergence property of the summation $\boldsymbol{\phi}(s,a)\boldsymbol{\phi}^\top(s,a)$, we introduce the following matrix Azuma inequality.

Lemma 4.6. [Matrices Azuma, Tropp [2012]] Let $\{\mathcal{F}_k\}_{k=1}^t$ be a filtration sequence, $\{\mathbf{X}_k\}_{k=1}^t$ be a finite adapted sequence of symmetric matrices where $\mathbf{X}_k \in \mathbb{R}^{d\times d}$ is $\mathcal{F}_{k+1}$-measurable, $\mathbb{E}[\mathbf{X}_k|\mathcal{F}_k] = \mathbf{0}$ and $\mathbf{X}^2 \preceq \mathbf{C}^2$ a.s.. Then with probability at least $1-\delta$,

$$\lambda_{\max}\left(\sum_{k=1}^t \mathbf{X}_k\right) \le \sqrt{8C^2 t\log(d/\delta)},$$

where $C = \|\mathbf{C}\|_2$.

Equipped with this lemma, we can start our proof.

Proof of Lemma 3.4. First it is easy to verify that for any $k \in [K]$

$$\begin{aligned}\boldsymbol{\phi}(s_h^k, a_h^k)\boldsymbol{\phi}^\top(s_h^k, a_h^k) &= \mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h^k) - \mathbb{1}\left[a_h^k \ne \pi_h^*(s_h^k)\right]\left(\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h^k) - \boldsymbol{\phi}(s_h^k, a_h^k)\boldsymbol{\phi}^\top(s_h^k, a_h^k)\right)\\ &\succeq \mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h^k) - C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}}\,\mathbb{1}\left[a_h^k \ne \pi_h^*(s_h^k)\right]\mathbf{I}_{d_{\boldsymbol{\phi}}},\end{aligned} \tag{4.2}$$

where the inequality holds due to $\mathbf{0} \preceq \boldsymbol{\phi}(s,a)\boldsymbol{\phi}^\top(s,a) \preceq (C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}})\mathbf{I}_{d_{\boldsymbol{\phi}}}$. By the definition of $\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h^k)$, we have $\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h^k) = \mathbb{E}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)|s_h^k]$ and it suffices to control $\mathbb{E}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)|s_i^k]$ for any $1 < i \le h$. Therefore it follows that

$$\mathbb{E}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)|s_i^k] = \underbrace{\mathbb{E}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)|s_{i-1}^k, a_{i-1}^k]}_{\mathbf{A}_i(s_{i-1}^k, a_{i-1}^k)} - \underbrace{\left(\mathbb{E}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)|s_{i-1}^k, a_{i-1}^k] - \mathbb{E}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)|s_i^k]\right)}_{\boldsymbol{\epsilon}_i^k}$$

where we denote the first term as $\mathbf{A}_i(s_{i-1}^k, a_{i-1}^k)$ while the second term as $\boldsymbol{\epsilon}_i^k$ for simplicity. We first consider the term $\boldsymbol{\epsilon}_i^k$, it is easy to verify that $\boldsymbol{\epsilon}_i^k$ is $\mathcal{F}_i^k$-measurable, $d\times d$ symmetric matrix with $\mathbb{E}[\boldsymbol{\epsilon}_i^k|\mathcal{F}_{i-1}^k] = \mathbf{0}$ and

$$\|\boldsymbol{\epsilon}_i^k\|_2 \le \left\|\mathbb{E}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)|s_{i-1}^k, a_{i-1}^k]\right\|_2 + \left\|\mathbb{E}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)|s_i^k]\right\|_2 \le 2C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}}, \tag{4.3}$$

where the inequality holds due to the fact that $\|\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s)\|_2 \le C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}}$. Next, for the term $\mathbf{A}_i$, by introducing the indicator showing whether the action a_{i-1}^k is the optimal action $\pi_{i-1}^*(s_{i-1}^k)$, we proceed as follows:

$$\begin{aligned}\mathbf{A}_i(s_{i-1}^k, a_{i-1}^k) &= \mathbf{A}_i(s_{i-1}^k, \pi_{i-1}^*(s_{i-1}^k))\\ &\quad - \mathbb{1}\left[a_{i-1}^k \ne \pi_{i-1}^*(s_{i-1}^k)\right]\left(\mathbf{A}_i(s_{i-1}^k, \pi_{i-1}^*(s_{i-1}^k)) - \mathbf{A}_i(s_{i-1}^k, a_{i-1}^k)\right)\\ &\succeq \mathbf{A}_i(s_{i-1}^k, \pi_{i-1}^*(s_{i-1}^k)) - C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}}\,\mathbb{1}\left[a_{i-1}^k \ne \pi_{i-1}^*(s_{i-1}^k)\right]\mathbf{I}_{d_{\boldsymbol{\phi}}}\\ &= \mathbb{E}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)|s_{i-1}^k] - C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}}\,\mathbb{1}\left[a_{i-1}^k \ne \pi_{i-1}^*(s_{i-1}^k)\right]\mathbf{I}_{d_{\boldsymbol{\phi}}},\end{aligned} \tag{4.4}$$

where the inequality holds due to a similar proof of (4.2). The last equality holds due to the definition that

$$\mathbb{E}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)|s_{i-1}^k, \pi_{i-1}^*(s_{i-1}^k)] = \mathbb{E}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)|s_{i-1}].$$

Combining (4.4) and (4.3) together yields

$$\mathbb{E}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)|s_i^k] \succeq \mathbb{E}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)|s_{i-1}^k] - C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}}\,\mathbb{1}\left[a_{i-1}^k \ne \pi_{i-1}^*(s_{i-1}^k)\right]\mathbf{I}_{d_{\boldsymbol{\phi}}} - \boldsymbol{\epsilon}_i^k,$$

and by telescoping over i we have

$$\begin{aligned}\mathbb{E}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)|s_h^k] &\succeq \mathbb{E}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)|s_1^k] - \sum_{i=2}^{h}\boldsymbol{\epsilon}_i^k - C_{\boldsymbol{\phi}}d_{\boldsymbol{\phi}}\sum_{i=1}^{h-1}\mathbb{1}\left[a_i^k \neq \pi_i^*(s_i^k)\right]\mathbf{I}_{d_{\boldsymbol{\phi}}} \\ &= \mathbb{E}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)] - \underbrace{\mathbb{E}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)] - \mathbb{E}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)|s_1^k]}_{\boldsymbol{\epsilon}_1^k} \\ &\quad - \sum_{i=2}^{h}\boldsymbol{\epsilon}_i^k - C_{\boldsymbol{\phi}}d_{\boldsymbol{\phi}}\sum_{i=1}^{h-1}\mathbb{1}\left[a_i^k \neq \pi_i^*(s_i^k)\right]\mathbf{I}_{d_{\boldsymbol{\phi}}}, \end{aligned} \tag{4.5}$$

where $\boldsymbol{\epsilon}_1^k$ is $\mathcal{F}_1^k$-measurable and $\mathbb{E}[\boldsymbol{\epsilon}_1^k|\mathcal{F}_H^{k-1}] = \mathbf{0}$, $\|\boldsymbol{\epsilon}_1^k\|_2 \leq 2C_{\boldsymbol{\phi}}d_{\boldsymbol{\phi}}$, which is similar to ϵ_i^k above. Plugging (4.5) into (4.2) yields

$$\begin{aligned}\boldsymbol{\phi}(s_h^k, a_h^k)\boldsymbol{\phi}^\top(s_h^k, a_h^k) &\succeq \mathbb{E}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)] - \sum_{i=1}^{h}\boldsymbol{\epsilon}_i^k - C_{\boldsymbol{\phi}}d_{\boldsymbol{\phi}}\sum_{i=1}^{h}\mathbb{1}\left[a_i^k \neq \pi_i^*(s_i^k)\right]\mathbf{I}_{d_{\boldsymbol{\phi}}}, \\ &= \mathbb{E}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)] - \sum_{i=1}^{h}\boldsymbol{\epsilon}_i^k - C_{\boldsymbol{\phi}}d_{\boldsymbol{\phi}}\sum_{i=1}^{h}\mathbb{1}\left[Q_i^*(s_i^k, a_i^k) \neq V_i^*(s_i)\right]\mathbf{I}_{d_{\boldsymbol{\phi}}}, \\ &\succeq \mathbb{E}[\mathbf{\Lambda}_{h,\boldsymbol{\phi}}(s_h)] - \sum_{i=1}^{h}\boldsymbol{\epsilon}_i^k - \frac{C_{\boldsymbol{\phi}}d_{\boldsymbol{\phi}}}{\text{gap}_{\min}}\sum_{i=1}^{h}(V_i^*(s_i^j) - Q_i^*(s_i^j, a_i^j))\mathbf{I}_{d_{\boldsymbol{\phi}}} \end{aligned} \tag{4.6}$$

where the equality follows that $a_h^k \neq \pi_i^*(s_h^k)$ is equivalent with $Q_i^*(s_i^k, a_h^k) \neq V_i^*(s_i)$ and the second inequality is from $V_i^*(s_i^k) - Q_i^*(s_i^k, a_h^k) \geq \text{gap}_{\min}\mathbb{1}[Q_i^*(s_i^k, a_h^k) \neq V_i^*(s_i)]$, which is according to the minimal sub-optimality gap condition assumption defined in (3.2).

By the construction of the "covariance matrix" $\mathbf{U}_{h,\boldsymbol{\phi}}^k$, (4.6) yields

$$\begin{aligned}\mathbf{U}_{h,\boldsymbol{\phi}}^k &= \mathbf{I}_{d_{\boldsymbol{\phi}}} + \sum_{j=1}^{k-1}\boldsymbol{\phi}(s_h^j, a_h^j)\boldsymbol{\phi}^\top(s_h^j, a_h^j) \\ &\succeq \mathbf{I}_{d_{\boldsymbol{\phi}}} + (k-1)\mathbf{\Lambda}_{h,\boldsymbol{\phi}} - \sum_{j=1}^{k-1}\sum_{i=1}^{h}\boldsymbol{\epsilon}_i^j - \frac{C_{\boldsymbol{\phi}}d_{\boldsymbol{\phi}}}{\text{gap}_{\min}}\sum_{i=1}^{h}\sum_{j=1}^{k-1}(V_i^*(s_i^j) - Q_i^*(s_i^j, a_i^j))\mathbf{I}_{d_{\boldsymbol{\phi}}}. \end{aligned} \tag{4.7}$$

Recall $\boldsymbol{\epsilon}_i^j$ is a $d_{\boldsymbol{\phi}} \times d_{\boldsymbol{\phi}}$ symmetric matrix, by Lemma 4.6 with $C = 2C_{\boldsymbol{\phi}}d_{\boldsymbol{\phi}}Ib_{d_{\boldsymbol{\phi}}}, t = (k-1)h$, with probability at least $1-\delta$, we have

$$\lambda_{\max}\left(\sum_{j=1}^{k-1}\sum_{i=1}^{h}\boldsymbol{\epsilon}_i^j\right) \leq \sqrt{32C_{\boldsymbol{\phi}}^2d_{\boldsymbol{\phi}}^2h(k-1)\log(d_{\boldsymbol{\phi}}/\delta)}. \tag{4.8}$$

Combining (4.8) with (4.7) and substituting δ with $\delta/Hk(k+1)|\Phi|$, the claim in Lemma 3.4 holds for all $h \in [H], k \in [K], \boldsymbol{\phi} \in \Phi$ by taking a union bound.

□

4.4 PROOF OF LEMMA 3.5

In order to prove Lemma 3.5, we first need the following lemma.

Lemma 4.7. Given the condition in Lemma 3.3 and Lemma 3.4 holds and $\mathcal{E}_{\boldsymbol{\phi}}^K$ holds for all $\boldsymbol{\phi} \in \Phi$. For each $\boldsymbol{\phi} \in \Phi$, there exists a constant threshold

$$\tau_{\boldsymbol{\phi}} = \text{poly}(d_{\boldsymbol{\phi}}, \sigma_{\boldsymbol{\phi}}^{-1}, H, \log(|\Phi|/\delta), \text{gap}_{\min}^{-1}, C_{\boldsymbol{\phi}}, C_{\boldsymbol{\psi}}, C_{\mathbf{M}}, C'_{\boldsymbol{\psi}})$$

such that for any $(s,a) \in \mathcal{S} \times \mathcal{A}, h \in [H]$, there exists a representation candidate $\boldsymbol{\phi} \in \Phi$ where when $k \geq \tau_{\boldsymbol{\phi}}$, $\boldsymbol{\phi}^\top(s,a)(\mathbf{U}_{h,\boldsymbol{\phi}}^k)^{-1}\boldsymbol{\phi}(s,a) \leq 2C_{\boldsymbol{\phi}}d_{\boldsymbol{\phi}}/(\sigma_{\boldsymbol{\phi}}k)$. We denote $\tau = \max_{\boldsymbol{\phi}\in\Phi}\tau_{\boldsymbol{\phi}}$ to be the maximum possible threshold over all representations.

Lemma 4.7 suggests that the UCB bonus term is decaying in the rate of $\mathcal{O}\big(1/\sqrt{k}\big)$. Equipped with this lemma, we can start the proof.

Proof of Lemma 3.5. We will prove this lemma by induction. By the assumption in Lemma 3.5, $\mathcal{E}_{\boldsymbol{\phi}}^K$ holds for all $\boldsymbol{\phi} \in \Phi$. Considering $h = H$, for any state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, by Lemma 4.7, when $k \geq \tau$, there exists a representation $\boldsymbol{\phi}$ where Lemma 4.2 yields

$$\begin{aligned} Q_H^k(s,a) - Q_H^{\pi^k}(s,a) &\leq 2C_{\boldsymbol{\psi}} H \sqrt{\beta_{k,\boldsymbol{\phi}} \boldsymbol{\phi}^\top(s,a)(\mathbf{U}_{H,\boldsymbol{\phi}}^k)^{-1}\boldsymbol{\phi}(s,a)} + [\mathbb{P}_H(V_{H+1}^k - V_{H+1}^\pi)](s,a) \\ &\leq 2C_{\boldsymbol{\psi}} H \sqrt{2C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}} \beta_{k,\boldsymbol{\phi}}/(\sigma_{\boldsymbol{\phi}} k)} + 0, \end{aligned}$$

where the second inequality is due to Lemma 4.7 and the fact that V_{H+1}^k, V_{H+1}^π are both equal to zero. Thus we have

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \{Q_H^k(s,a) - Q_H^{\pi^k}(s,a)\} \leq \max_{\boldsymbol{\phi}\in\Phi} \left\{ 2C_{\boldsymbol{\psi}} H \sqrt{2C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}} \beta_{k,\boldsymbol{\phi}}/(\sigma_{\boldsymbol{\phi}} k)} \right\}.$$

Suppose for step h, we have

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \{Q_h^k(s,a) - Q_h^{\pi^k}(s,a)\} \leq (H-h+1) \max_{\boldsymbol{\phi}\in\Phi} \left\{ 2C_{\boldsymbol{\psi}} H \sqrt{2C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}} \beta_{k,\boldsymbol{\phi}}/(\sigma_{\boldsymbol{\phi}} k)} \right\}, \tag{4.9}$$

then considering time-step $h-1$, by Lemma 4.2 and Lemma 4.7, for each s, a, there exists a $\boldsymbol{\phi} \in \Phi$ such that

$$\begin{aligned} &Q_{h-1}^k(s,a) - Q_{h-1}^{\pi^k}(s,a) \\ &\quad\leq 2C_{\boldsymbol{\psi}} H \sqrt{\beta_{k,\boldsymbol{\phi}} \boldsymbol{\phi}^\top(s,a)(\mathbf{U}_{h-1,\boldsymbol{\phi}}^k)^{-1}\boldsymbol{\phi}(s,a)} + [\mathbb{P}_{h-1}(V_h^k - V_h^{\pi^k})](s,a) \\ &\quad\leq 2C_{\boldsymbol{\phi}} H \sqrt{2C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}} \beta_{k,\boldsymbol{\phi}}/(\sigma_{\boldsymbol{\phi}} k)} + [\mathbb{P}_{h-1}(Q_h^k(\cdot,\pi_h^k(\cdot)) - Q_h^{\pi^k}(\cdot,\pi_h^k(\cdot)))](s,a) \\ &\quad\leq 2C_{\boldsymbol{\phi}} H \sqrt{2C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}} \beta_{k,\boldsymbol{\phi}}/(\sigma_{\boldsymbol{\phi}} k)} + (H-h+1) \max_{\boldsymbol{\phi}\in\Phi} \left\{ 2C_{\boldsymbol{\psi}} H \sqrt{2C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}} \beta_{k,\boldsymbol{\phi}}/(\sigma_{\boldsymbol{\phi}} k)} \right\} \\ &\quad\leq (H-h+2) \max_{\boldsymbol{\phi}\in\Phi} \left\{ 2C_{\boldsymbol{\psi}} H \sqrt{2C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}} \beta_{k,\boldsymbol{\phi}}/(\sigma_{\boldsymbol{\phi}} k)} \right\}. \end{aligned}$$

where the second inequality follows from the definition that $V_h^k(s) = Q_h^k(s, \pi_h^k(s))$ and $V_h^{\pi^k}(s) = Q_h^{\pi^k}(s, \pi_h^k(s))$, the third inequality is due to the induction assumption (4.9) and this result conclude our induction.

Then, following Lemma 4.1, we have $Q_{h,\boldsymbol{\phi}}^k(s,a) \geq Q_h^*(s,a)$. Thus, $Q_h^k(s,a) = \min_{\boldsymbol{\phi}\in\Phi}\{Q_{h,\boldsymbol{\phi}}^k\} \geq Q_h^*(s,a)$. Then the sub-optimality gap could be bounded by

$$\begin{aligned} \text{gap}_h(s, \pi_h^k(s)) &= Q_h^*(s, \pi_h^*(s)) - Q_h^*(s, \pi_h^k(s)) \\ &\leq Q_h^k(s, \pi_h^*(s)) - Q_h^{\pi^k}(s, \pi_h^k(s)) \\ &\leq Q_h^k(s, \pi_h^k(s)) - Q_h^{\pi^k}(s, \pi_h^k(s)) \\ &\leq (H-h+1) \max_{\boldsymbol{\phi}\in\Phi} \left\{ 2C_{\boldsymbol{\psi}} H \sqrt{2C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}} \beta_{k,\boldsymbol{\phi}}/(\sigma_{\boldsymbol{\phi}} k)} \right\} \\ &\leq 2C_{\boldsymbol{\psi}} H^2 \max_{\boldsymbol{\phi}\in\Phi} \left\{ \sqrt{2C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}} \beta_{k,\boldsymbol{\phi}}/(\sigma_{\boldsymbol{\phi}} k)} \right\}, \end{aligned}$$

where the inequality on the second line holds due to Lemma 4.1, and the inequality on the third line holds due to the greedy policy $\pi_h^k(s) = \text{argmax}_a\, Q_h^k(s,a)$. Finally, the inequality on the forth line is due to the result of induction (4.9) and we finish the proof. □

5 PROOF OF LEMMAS IN APPENDIX 4

5.1 PROOF OF LEMMA 4.1

Lemma 5.1 (Lemma 5 on $B_n^{(2)}$, pp. 23, Yang and Wang [2020])**.** Suppose $\mathcal{E}_{\boldsymbol{\phi}}^K$ holds, then for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\|\boldsymbol{\phi}(s,a)^\top (\mathbf{M}_{h,\boldsymbol{\phi}}^k - \mathbf{M}_{h,\boldsymbol{\phi}}^*)\|_2 \leq \sqrt{\beta_{k,\boldsymbol{\phi}} \boldsymbol{\phi}^\top(s,a)(\mathbf{U}_{h,\boldsymbol{\phi}}^k)^{-1}\boldsymbol{\phi}(s,a)}.$$

Proof of Lemma 4.1. We prove this lemma by induction. First, it is obvious that $Q_{H+1}^k(s,a) = Q_{H+1}^*(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Then assuming for $1 < h \leq H$, we have $Q_{h+1}^k(s,a) \geq Q_{h+1}^*(s,a)$ holds for all (s,a), considering time-step h and representation $\boldsymbol{\phi}$, we have

$$\begin{aligned}
Q_{h,\boldsymbol{\phi}}^k(s,a) &= r(s,a) + \boldsymbol{\phi}^\top(s,a)\mathbf{M}_{h,\boldsymbol{\phi}}^k \boldsymbol{\Psi}^\top \mathbf{v}_{h+1}^k + C_{\boldsymbol{\psi}} H \sqrt{\beta_{k,\boldsymbol{\phi}} \boldsymbol{\phi}^\top(s,a)(\mathbf{U}_{h,\boldsymbol{\phi}}^k)^{-1}\boldsymbol{\phi}(s,a)} \\
&= r(s,a) + \boldsymbol{\phi}^\top(s,a)(\mathbf{M}_{h,\boldsymbol{\phi}}^k - \mathbf{M}_{h,\boldsymbol{\phi}}^*)\boldsymbol{\Psi}^\top \mathbf{v}_{h+1}^k \\
&\quad + C_{\boldsymbol{\psi}} H \sqrt{\beta_{k,\boldsymbol{\phi}} \boldsymbol{\phi}^\top(s,a)(\mathbf{U}_{h,\boldsymbol{\phi}}^k)^{-1}\boldsymbol{\phi}(s,a)} + [\mathbb{P}_h V_{h+1}^k](s,a) \\
&\geq r(s,a) - \|\boldsymbol{\phi}^\top(s,a)(\mathbf{M}_{h,\boldsymbol{\phi}}^k - \mathbf{M}_{h,\boldsymbol{\phi}}^*)\|_2 \|\boldsymbol{\Psi}^\top \mathbf{v}_{h+1}^k\|_2 \\
&\quad + C_{\boldsymbol{\psi}} H \sqrt{\beta_{k,\boldsymbol{\phi}} \boldsymbol{\phi}^\top(s,a)(\mathbf{U}_{h,\boldsymbol{\phi}}^k)^{-1}\boldsymbol{\phi}(s,a)} + [\mathbb{P}_h V_{h+1}^k](s,a) \\
&\geq r(s,a) + [\mathbb{P}_h V_{h+1}^k](s,a), \qquad (5.1)
\end{aligned}$$

where the first inequality comes from the fact that $\langle \mathbf{x}, \mathbf{y} \rangle \geq -\|\mathbf{x}\|_2 \|\mathbf{y}\|_2$, the second inequality holds due to Lemma 5.1 and $\|\boldsymbol{\Psi}^\top \mathbf{v}_{h+1}^k\|_\infty \leq C_{\boldsymbol{\psi}} H$ since $\|\mathbf{v}_{h+1}^k\|_\infty \leq H$. Since $Q_{h+1}^k(s,a) \geq Q_{h+1}^*(s,a)$, then

$$\begin{aligned}
V_{h+1}^k(s) &= \min\{H, Q_{h+1}^k(s, \pi_h^k(s))\} \\
&\geq \min\{H, Q_{h+1}^k(s, \pi_{h+1}^*(s))\} \\
&\geq \min\{H, Q_{h+1}^*(s, \pi_{h+1}^*(s))\} \\
&= V_{h+1}^*(s),
\end{aligned}$$

where the last inequality is due to the fact that $V_{h+1}^*(s) \leq H$. Therefore, (5.1) yields $Q_{h,\boldsymbol{\phi}}^k(s,a) \geq r(s,a) + [\mathbb{P}_h V_{h+1}^*](s,a) = Q_h^*(s,a)$ for all $\boldsymbol{\phi} \in \Phi$. Thus

$$Q_h^k(s,a) = \min_{\boldsymbol{\phi} \in \Phi} \{Q_{h,\boldsymbol{\phi}}^k(s,a)\} \geq Q_h^*(s,a).$$

Then we finish our proof by induction. □

5.2 PROOF OF LEMMA 4.2

Proof of Lemma 4.2. First, the update rule of $Q_{h,\boldsymbol{\phi}}^k$ and Bellman equation yield

$$\begin{aligned}
Q_{h,\boldsymbol{\phi}}^k(s,a) - Q_h^\pi(s,a) &= \underbrace{\boldsymbol{\phi}^\top(s,a)\mathbf{M}_{h,\boldsymbol{\phi}}^k \boldsymbol{\Psi}^\top \mathbf{v}_{h+1}^k}_{I_1} - [\mathbb{P}_h V_{h+1}^\pi](s,a) \\
&\quad + C_{\boldsymbol{\psi}} H \sqrt{\beta_{k,\boldsymbol{\phi}} \boldsymbol{\phi}^\top(s,a)(\mathbf{U}_{h,\boldsymbol{\phi}}^k)^{-1}\boldsymbol{\phi}(s,a)}. \qquad (5.2)
\end{aligned}$$

Since $[\mathbb{P}_h V_{h+1}^k] = \boldsymbol{\phi}^\top(s,a)\mathbf{M}_{h,\boldsymbol{\phi}}^* \boldsymbol{\Psi}^\top \mathbf{v}_{h+1}^k$, I_1 can be decomposed as

$$\begin{aligned}
\boldsymbol{\phi}^\top(s,a)\mathbf{M}_{h,\boldsymbol{\phi}}^k \boldsymbol{\Psi}^\top \mathbf{v}_{h+1}^k &= \boldsymbol{\phi}^\top(s,a)(\mathbf{M}_{h,\boldsymbol{\phi}}^k - \mathbf{M}_{h,\boldsymbol{\phi}}^*)\boldsymbol{\Psi}^\top \mathbf{v}_{h+1}^k + [\mathbb{P}_h V_{h+1}^k](s,a) \\
&\leq \|\boldsymbol{\Psi}^\top \mathbf{v}_{h+1}^k\|_2 \|\boldsymbol{\phi}^\top(s,a)(\mathbf{M}_{h,\boldsymbol{\phi}}^k - \mathbf{M}_{h,\boldsymbol{\phi}}^*)\|_2 + [\mathbb{P}_h V_{h+1}^k](s,a) \\
&\leq C_{\boldsymbol{\psi}} H \sqrt{\beta_{k,\boldsymbol{\phi}} \boldsymbol{\phi}^\top(s,a)(\mathbf{U}_{h,\boldsymbol{\phi}}^k)^{-1}\boldsymbol{\phi}(s,a)} + [\mathbb{P}_h V_{h+1}^k](s,a), \qquad (5.3)
\end{aligned}$$

where the inequality on the second line holds due to $\langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$ and the inequality on the third line comes from Lemma 5.1 with $\|\mathbf{v}_{h+1}^k\|_\infty \leq H$ and Definition 3.2. Plugging (5.3) into (5.2) yields

$$Q_{h,\boldsymbol{\phi}}^k(s,a) - Q_h^\pi(s,a) \leq 2C_{\boldsymbol{\psi}} H \sqrt{\beta_{k,\boldsymbol{\phi}} \boldsymbol{\phi}^\top(s,a)(\mathbf{U}_{h,\boldsymbol{\phi}}^k)^{-1}\boldsymbol{\phi}(s,a)} + [\mathbb{P}_h(V_{h+1}^k - V_{h+1}^\pi)](s,a).$$

Since $Q_h^k(s,a) = \min_{\boldsymbol{\phi}\in\Phi}(s,a)$, we can get the claimed result in Lemma 4.2. □

5.3 PROOF OF LEMMAS 4.3

Lemma 5.2 (Lemma 6.6, He et al. [2021])**.** For any subset $C = \{c_1, \cdots, c_k\} \subseteq [K]$ and any $h \in [H]$,

$$\sum_{i=1}^k \boldsymbol{\phi}_h^\top(s_h^{c_i}, a_h^{c_i})(\mathbf{U}_{h,\boldsymbol{\phi}}^{c_i})^{-1}\boldsymbol{\phi}_h(s_h^{c_i}, a_h^{c_i}) \leq 2d_{\boldsymbol{\phi}} \log(1 + C_{\boldsymbol{\phi}} k d_{\boldsymbol{\phi}})$$

Remark 5.3. Proof of Lemma 5.2 remains the same as He et al. [2021] by changing the norm of $\boldsymbol{\phi}$ from $\|\boldsymbol{\phi}\|_2^2 \leq 1$ to $\|\boldsymbol{\phi}\|_2^2 \leq C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}}$ as Definition 3.2.

Lemma 5.4 (Azuma-Hoeffding's inequality, Azuma 1967)**.** Let $\{x_i\}_{i=1}^n$ be a martingale difference sequence with respect to a filtration $\{\mathcal{F}_i\}_{i=1}^n$ (i.e. $\mathbb{E}[x_i|\mathcal{F}_i] = 0$ a.s. and x_i is $\mathcal{F}_{i+1}$ measurable) such that $|x_i| \leq M$ a.s.. Then for any $0 < \delta < 1$, with probability at least $1-\delta$, $\sum_{i=1}^n x_i \leq M\sqrt{2n\log(1/\delta)}$.

Proof of Lemma 4.3. We fix h and consider the first k episodes in this proof. Let $k_0 = 0$, for any $j \in [k]$, we denote k_j as the minimum index of the episode where the sub-optimality at time-step h is no less than Δ:

$$k_j = \min\left\{\bar{k} : \bar{k} > k_{j-1}, V_h^*(s_h^{\bar{k}}) - Q_h^{\pi^{\bar{k}}}(s_h^{\bar{k}}, a_h^{\bar{k}}) \geq \Delta\right\}.$$

For simplicity, we denote k' to be the number of episodes such that the sub-optimality of this episode at step h is no less than Δ, i.e.

$$k' = \sum_{j=1}^k \mathbb{1}[V_h^*(s_h^j) - Q_h^{\pi^j}(s_h^j, a_h^j) \geq \Delta].$$

Then by the definition of k', it is obvious that

$$\begin{aligned}
\sum_{j=1}^{k'} Q_h^{k_j}(s_h^{k_j}, a_h^{k_j}) - Q_h^{\pi^{k_j}}(s_h^{k_j}, a_h^{k_j}) &\geq \sum_{j=1}^{k'} Q_h^{k_j}(s_h^{k_j}, \pi_h^*(s_h^{k_j})) - Q_h^{\pi^{k_j}}(s_h^{k_j}, a_h^{k_j}) \\
&\geq \sum_{j=1}^{k'} Q_h^*(s_h^{k_j}, \pi_h^*(s_h^{k_j})) - Q_h^{\pi^{k_j}}(s_h^{k_j}, a_h^{k_j}) \\
&= \sum_{j=1}^{k'} V_h^*(s_h^{k_j}) - Q_h^{\pi^{k_j}}(s_h^{k_j}, a_h^{k_j}) \geq \Delta k', \qquad (5.4)
\end{aligned}$$

where the first inequality holds due to $a_h^k = \operatorname{argmax}_a Q_h^k(s_h^k, a)$ and the second inequality follows Lemma 4.1. On the other hand, following Lemma 4.2, when $\mathcal{E}_\phi^k$ holds, for all $i \in [H], j \leq k$

$$\begin{aligned}
Q_i^j(s_i^j, a_i^j) &\leq 2C_{\boldsymbol{\psi}} H \sqrt{\beta_{j,\boldsymbol{\phi}} \boldsymbol{\phi}^\top(s_i^j, a_i^j)(\mathbf{U}_{i,\boldsymbol{\phi}}^j)^{-1}\boldsymbol{\phi}(s_i^j, a_i^j)} + [\mathbb{P}_i(V_{i+1}^j - V_{i+1}^{\pi^j})](s_i^j, a_i^j) \\
&= 2C_{\boldsymbol{\psi}} H \sqrt{\beta_{j,\boldsymbol{\phi}} \boldsymbol{\phi}^\top(s_i^j, a_i^j)(\mathbf{U}_{i,\boldsymbol{\phi}}^j)^{-1}\boldsymbol{\phi}(s_i^j, a_i^j)} + V_{i+1}^j(s_{i+1}^j) - V_{i+1}^{\pi^j}(s_{i+1}^j) + \epsilon_i^j, \qquad (5.5)
\end{aligned}$$

where $\epsilon_i^j = [\mathbb{P}_i(V_{i+1}^j - V_{i+1}^{\pi^j})](s_i^j, a_i^j) - \big(V_{i+1}^j(s_{i+1}^j) - V_{i+1}^{\pi^j}(s_{i+1}^j)\big)$. It is easy to verify that $|\epsilon_i^j| \leq H$, ϵ_i^j is $\mathcal{F}_{i+1}^j$ measurable with $\mathbb{E}[\epsilon_i^j|\mathcal{F}_i^j] = 0$. Taking the telescoping summation on (5.5) over $h \leq i \leq H, j \in \{k_1, \cdots, k_{k'}\}$ using the fact that $V_i^j(s_i^j) = Q_i^j(s_i^j, a_i^j)$ and $V_i^{\pi^j}(s_i^j) = Q_i^{\pi^j}(s_i^j, a_i^j)$ we have

$$\sum_{j=1}^{k'} Q_h^{k_j}(s_h^{k_j}, a_h^{k_j}) - Q_h^{\pi^{k_j}}(s_h^{k_j}, a_h^{k_j}) \leq I_1 + I_2, \qquad (5.6)$$

where

$$I_1 = \sum_{j=1}^{k'} \sum_{i=h}^{H} 2C_{\boldsymbol{\psi}} H \sqrt{\beta_{k_j,\boldsymbol{\phi}} \boldsymbol{\phi}^\top(s_h^{k_j}, a_h^{k_j})(\mathbf{U}_{i,\boldsymbol{\phi}}^{k_j})^{-1} \boldsymbol{\phi}(s_h^{k_j}, a_h^{k_j})}$$
$$I_2 = \sum_{j=1}^{k'} \sum_{i=h}^{H} \epsilon_i^{k_j}.$$

To bound I_1, by Cauchy-Schwarz inequality,

$$\begin{aligned}
I_1 &= \sum_{j=1}^{k'} \sum_{i=h}^{H} 2C_{\boldsymbol{\psi}} H \sqrt{\beta_{k_j,\boldsymbol{\phi}} \boldsymbol{\phi}^\top(s_h^{k_j}, a_h^{k_j})(\mathbf{U}_{i,\boldsymbol{\phi}}^{k_j})^{-1} \boldsymbol{\phi}(s_h^{k_j}, a_h^{k_j})} \\
&\leq 2C_{\boldsymbol{\psi}} H \sqrt{\beta_{k,\boldsymbol{\phi}} k'} \sum_{i=h}^{H} \sqrt{\sum_{j=1}^{k'} \boldsymbol{\phi}^\top(s_h^{k_j}, a_h^{k_j})(\mathbf{U}_{i,\boldsymbol{\phi}}^{k_j})^{-1} \boldsymbol{\phi}(s_h^{k_j}, a_h^{k_j})} \\
&\leq 2C_{\boldsymbol{\psi}} H^2 \sqrt{\beta_{k,\boldsymbol{\phi}} k'} \sqrt{2d_{\boldsymbol{\phi}} \log(1 + C_{\boldsymbol{\phi}} k' d_{\boldsymbol{\phi}})} \\
&\leq 2C_{\boldsymbol{\psi}} H^2 \sqrt{2\beta_{k,\boldsymbol{\phi}} d_{\boldsymbol{\phi}} k' \log(1 + C_{\boldsymbol{\phi}} k d_{\boldsymbol{\phi}})},
\end{aligned}$$

where the second inequity in Line 3 is from Lemma 5.2. To bound I_2, by Lemma 5.4, with probability at least $1 - \delta/k$, we have

$$\sum_{j=1}^{k'} \sum_{i=h}^{H} \epsilon_i^{k_j} \leq \sqrt{2k' H^3 \log(k/\delta)},$$

then taking union bound over all k we can conclude that with probability at least $1 - \delta$,

$$I_2 = \sum_{j=1}^{k'} \sum_{i=h}^{H} \epsilon_i^{k_j} \leq \sqrt{2k' H^3 \log(k/\delta)}$$

Combining (5.4) with (5.6), we can obtain

$$\Delta k' \leq 2C_{\boldsymbol{\psi}} H^2 \sqrt{2\beta_{k,\boldsymbol{\phi}} d_{\boldsymbol{\phi}} k' \log(1 + C_{\boldsymbol{\phi}} k d_{\boldsymbol{\phi}})} + \sqrt{2k' H^3 \log(k/\delta)}. \tag{5.7}$$

By $(a+b)^2 \leq 2a^2 + 2b^2$, (5.7) immediately implies

$$k' \leq \frac{16 C_{\boldsymbol{\psi}}^2 H^4 d_{\boldsymbol{\phi}} \beta_{k,\boldsymbol{\phi}} \log(1 + C_{\boldsymbol{\phi}} k d) + 4H^3 \log(k/\delta)}{\Delta^2} \tag{5.8}$$

Since event $\mathcal{E}_{\boldsymbol{\phi}}^K$ directly implies $\mathcal{E}_{\boldsymbol{\phi}}^k$ for all $k \leq K$, we can get the claimed result (5.8) holds for all $k \leq K$ with probability $1 - \delta$. Replace δ with $\delta/k(k+1)$ for different k, taking union bound for all possible k, we have with probability at least $1 - \delta$, for all possible k,

$$\begin{aligned}
\sum_{j=1}^{k} \mathbb{1}[V_h^*(s_h^j) - Q_h^{\pi^j}(s_h^j, a_h^j)] &\leq \frac{16 C_{\boldsymbol{\psi}}^2 H^4 d_{\boldsymbol{\phi}} \beta_{k,\boldsymbol{\phi}} \log(1 + C_{\boldsymbol{\phi}} k d_{\boldsymbol{\phi}}) + 4H^3 \log(k^2(k+1)/\delta)}{\Delta^2} \\
&\leq \frac{16 C_{\boldsymbol{\psi}}^2 H^4 d_{\boldsymbol{\phi}} \beta_{k,\boldsymbol{\phi}} \log(1 + C_{\boldsymbol{\phi}} k d_{\boldsymbol{\phi}}) + 12H^3 \log(2k/\delta)}{\Delta^2}.
\end{aligned}$$

□

5.4 PROOF OF LEMMA 4.7

Proof of Lemma 4.7. For any state-action pair (s, a) at step h, according to Assumption 4.2, we consider the set $\mathcal{Z}_{h,\boldsymbol{\phi}}$ where $(s, a) \in \mathcal{Z}_{h,\boldsymbol{\phi}}$ and the corresponding representation $\boldsymbol{\phi}$. By Lemma 3.4, we denote $\mathbf{B}$ as

$$\mathbf{B} := (k-1)\boldsymbol{\Lambda}_{h,\boldsymbol{\phi}} - \iota \mathbf{I}_{d_{\boldsymbol{\phi}}} \preceq \mathbf{U}_{h,\boldsymbol{\phi}}^k,$$
$$\iota = \frac{C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}}}{\text{gap}_{\min}} \sum_{i=1}^{h} \sum_{j=1}^{k-1} \text{gap}_i(s_i^j, a_i^j) + C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}} \sqrt{32H(k-1)\log(d_{\boldsymbol{\phi}}|\Phi|Hk(k+1)/\delta)} - 1.$$

Decomposing $\boldsymbol{\Lambda}_{h,\boldsymbol{\phi}} = \mathbf{Q}^\top \mathbf{D} \mathbf{Q}$ where $\mathbf{Q} \in \mathbb{R}^{d_{\boldsymbol{\phi}} \times d_{\boldsymbol{\phi}}}$ is the orthogonal matrix and $\mathbf{D}$ is the diagonal matrix, we have $\mathbf{B} = \mathbf{Q}^\top((k-1)\mathbf{D} - \iota \mathbf{I}_{d_{\boldsymbol{\phi}}})\mathbf{Q}$.

We first prove the non-singular property of $\mathbf{B}$. Considering the zero diagonal element $\mathbf{D}_{[ii]}$, we have

$$((k-1)\mathbf{D} - \iota \mathbf{I}_{d_{\boldsymbol{\phi}}})_{[ii]} \leq -\iota \leq -C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}} \sqrt{32H(k-1)\log(d_{\boldsymbol{\phi}}|\Phi|Hk(k+1)/\delta)} + 1,$$

where the second inequality is due to $\text{gap}_h(s, a) \geq 0$. As a result, it is obvious to verify that there exists a constant K_1 such that once $k \geq K_1$, $((k-1)\mathbf{D} - \iota \mathbf{I}_{d_{\boldsymbol{\phi}}})_{[ii]} < 0$ for all zero diagonal element $\mathbf{D}_{[ii]}$ in $\mathbf{D}$. Next we consider the non-zero diagonal value $\mathbf{D}_{[jj]}$. By Assumption 4.2, $\mathbf{D}_{[jj]} \geq \sigma_{\boldsymbol{\phi}}$. Therefore, the corresponding diagonal value $((k-1)\mathbf{D} - \iota \mathbf{I}_{d_{\boldsymbol{\phi}}})_{[jj]}$ could be bounded by

$$((k-1)\mathbf{D} - \iota \mathbf{I}_{d_{\boldsymbol{\phi}}})_{[jj]} \geq \sigma_{\boldsymbol{\phi}}(k-1) - \iota.$$

Removing the minimum operator in (3.2) in Lemma 3.3, we have

$$\begin{aligned}((k-1)\mathbf{D} - \iota \mathbf{I}_{d_{\boldsymbol{\phi}}})_{[jj]} \geq 1 + \sigma_{\boldsymbol{\phi}}(k-1) - C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}} \sqrt{32H(k-1)\log(d_{\boldsymbol{\phi}}|\Phi|Hk(k+1)/\delta)} \\ - \frac{64 C_{\boldsymbol{\psi}}^2 H^4 d_{\boldsymbol{\phi}}^2 \beta_{k,\boldsymbol{\phi}} \log(1 + C_{\boldsymbol{\phi}} k d_{\boldsymbol{\phi}}) + 48H^3 \log\big(2k(1 + \log(H/\text{gap}_{\min})/\delta\big)}{\text{gap}_{\min}}\end{aligned}$$

It's easy to verify that the increasing term $\sigma_{\boldsymbol{\phi}} k$ is $\mathcal{O}(k)$ while the decreasing term is in the order of $\mathcal{O}(\sqrt{k})$ and $\mathcal{O}(\log(k))$ where $\beta_{k,\boldsymbol{\phi}} = \mathcal{O}(\log(k))$ as shown in Lemma 3.1, thus there exists a constant threshold

$$\tau_\phi = \text{poly}(d_{\boldsymbol{\phi}}, \sigma_{\boldsymbol{\phi}}^{-1}, H, \log(|\Phi|/\delta), \text{gap}_{\min}^{-1}, C_{\boldsymbol{\phi}}, C_{\boldsymbol{\psi}}, C_{\mathbf{M}}, C'_{\boldsymbol{\psi}})$$

such that for any $k \geq \tau_{\boldsymbol{\phi}}$, $((k-1)\mathbf{D} - \iota \mathbf{I}_{d_{\boldsymbol{\phi}}})_{[jj]} \geq \sigma_{\boldsymbol{\phi}}/2$. Since we have shown that all of the diagonal value for $(k-1)\mathbf{D} - \iota$ is either strictly smaller than zero or strictly greater than zero, $\mathbf{B}$ is invertible.

By the definition of $\mathcal{Z}_{h,\boldsymbol{\phi}}$ in Assumption 4.2, there exists a vector $\mathbf{x} \in \mathbb{R}^{d_{\boldsymbol{\phi}}}$ such that $\boldsymbol{\Lambda}\mathbf{x} = \boldsymbol{\phi}(s, a)/\|\boldsymbol{\phi}(s, a)\|_2$. Since $\mathbf{U}_{h,\boldsymbol{\phi}}^k \succeq \mathbf{B}$ and $\mathbf{U}_{h,\boldsymbol{\phi}}^k, \mathbf{B}$ are both invertible, it follows

$$\boldsymbol{\phi}^\top(s, a)(\mathbf{U}_{h,\boldsymbol{\phi}}^k)^{-1}\boldsymbol{\phi}(s, a) \leq \|\boldsymbol{\phi}(s, a)\|_2^2 \underbrace{\frac{\boldsymbol{\phi}^\top(s, a)}{\|\boldsymbol{\phi}(s, a)\|_2} \mathbf{B}^{-1} \frac{\boldsymbol{\phi}(s, a)}{\|\boldsymbol{\phi}(s, a)\|_2}}_{I_1}, \tag{5.9}$$

where I_1 could be rewrote by

$$\begin{aligned} I_1 &= \mathbf{x}^\top \boldsymbol{\Lambda} \mathbf{B}^{-1} \boldsymbol{\Lambda} \mathbf{x} \\ &= \mathbf{x}^\top \mathbf{Q}^\top \mathbf{D} \mathbf{Q} \mathbf{Q}^\top ((k-1)\mathbf{D} - \iota \mathbf{I}_{d_{\boldsymbol{\phi}}})^{-1} \mathbf{Q} \mathbf{Q}^\top \mathbf{D} \mathbf{Q} \mathbf{x} \\ &= \mathbf{x}^\top \mathbf{Q}^\top \mathbf{D} ((k-1)\mathbf{D} - \iota \mathbf{I}_{d_{\boldsymbol{\phi}}})^{-1} \mathbf{D} \mathbf{Q} \mathbf{x}. \end{aligned} \tag{5.10}$$

Since $\|\boldsymbol{\Lambda}\mathbf{x}\|_2 = 1$, it is easy to verify that $\mathbf{x}^\top \mathbf{Q}^\top \mathbf{D} \mathbf{D} \mathbf{Q} \mathbf{x} = \mathbf{x}^\top \boldsymbol{\Lambda} \boldsymbol{\Lambda} \mathbf{x} = \|\boldsymbol{\Lambda}\mathbf{x}\|_2^2 = 1$. We hereby denote $\mathbf{y}$ as $\mathbf{D}\mathbf{Q}\mathbf{x}$ and we have $\|\mathbf{y}\|_2^2 = 1$. Furthermore, it is obvious that $\mathbf{y}_{[i]} = 0$ as long as $\mathbf{D}_{[ii]} = 0$. Therefore, $\sum_{i=1, \mathbf{D}_{[ii]} \neq 0}^{d_{\boldsymbol{\phi}}} \mathbf{y}_{[i]}^2 = 1$. Then plugging the notation of $\mathbf{y}$ into (5.10) yields

$$I_1 = \mathbf{y}^\top ((k-1)\mathbf{D} - \iota \mathbf{I}_{d_{\boldsymbol{\phi}}})^{-1} \mathbf{y} = \sum_{i=1, \mathbf{D}_{[ii]} \neq 0}^{d_{\boldsymbol{\phi}}} \frac{\mathbf{y}_{[i]}^2}{((k-1)\mathbf{D} - \iota \mathbf{I}_{d_{\boldsymbol{\phi}}})_{[ii]}},$$

since we have shown that the $((k-1)\mathbf{D} - \iota \mathbf{I}_{d_{\boldsymbol{\phi}}})_{[ii]} \geq \sigma_{\boldsymbol{\phi}} k/2$ when $\mathbf{D}_{[ii]} \neq 0$ and $k \geq \tau_{\boldsymbol{\phi}}$. Thus we can easily conclude that $I_1 \leq 2/(\sigma_{\boldsymbol{\phi}} k)$, plugging this into (5.9) we can get the claimed result. □

6 PROOF OF THEOREM 5.4

In this section, we provide the proof of Theorem 5.4, which bounds the sample complexity of the offline version algorithm ReLEX-LCB. The offline training process favors a similar "good event" with its online counterpart (Lemma 3.1) which is formalized as the lemma below:

Lemma 6.1 (Lemma 15, Yang and Wang [2020], offline ver.)**.** Define the following event as $\mathcal{E}_{\boldsymbol{\phi}}^k$:

$$\left\{\mathrm{tr}\left[(\mathbf{M}_{h,\boldsymbol{\phi}} - \mathbf{M}_{h,\boldsymbol{\phi}}^*)^\top \mathbf{U}_{h,\boldsymbol{\phi}}(\mathbf{M}_{h,\boldsymbol{\phi}} - \mathbf{M}_{h,\boldsymbol{\phi}}^*)\right] \le \beta_{\boldsymbol{\phi}}, \forall h \in [H]\right\} =: \mathcal{E}_{\boldsymbol{\phi}}.$$

With $\beta_{\boldsymbol{\phi}} = Cd_{\boldsymbol{\phi}}\log(KH/\delta)$ for some absolute constant $C > 0$, we have $\Pr(\mathcal{E}_{\boldsymbol{\phi}}) \ge 1 - \delta$ for all $\boldsymbol{\phi} \in \Phi$.

Proof. The proof is similar with the original proof in Yang and Wang [2020] by changing k to K. The remaining part is unchanged given the offline training data. □

Then the next lemma is esentially the first part of Theorem 5.4, which provides an upper bound of the sub-optimality planned by Algorithm 2.

Lemma 6.2. Let β set as Lemma 6.1. If the event $\mathcal{E}_{\boldsymbol{\phi}}$ in Lemma 6.1 holds for all $\boldsymbol{\phi} \in \Phi$, for all state $s \in \mathcal{S}$ and $h \in [H]$,

$$V_h^*(s) - V_h^\pi(s) \le 2C_{\boldsymbol{\psi}} H \sum_{h'=h}^{H} \mathbb{E}_{\pi^*}\left[\min_{\boldsymbol{\phi}\in\Phi}\left\{\sqrt{\beta_{\boldsymbol{\phi}}}\|\boldsymbol{\phi}(s,a)\|_{\mathbf{U}_{h',\boldsymbol{\phi}}^{-1}}\right\}\middle| s_h = s\right],$$

where the expectation is taken with respect to the trajectory induced by the optimal policy π^* given the fixed covariance matrix $\mathbf{U}_{h,\boldsymbol{\phi}}$.

Comparing with Jin et al. [2021], our results adapts the minimal uncertainty $\|\boldsymbol{\phi}\|_{\mathbf{U}^{-1}}$ over all representation $\boldsymbol{\phi} \in \Phi$. Therefore, even if each single representation $\boldsymbol{\phi}$ cannot satisfy Assumption 5.1, we can still get sample complexity bound which Jin et al. [2021] failed to provide.

Then the next lemma suggests that the uncertainty $\|\boldsymbol{\phi}\|_{\mathbf{U}^{-1}}$ is bounded by $\widetilde{\mathcal{O}}(1/\sqrt{K})$ where the K is the size of the offline dataset.

Lemma 6.3. With probability at least $1-\delta$, for $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, there exists a $\boldsymbol{\phi} \in \Phi$ such that when

$$K > \frac{32C_{\boldsymbol{\phi}}^2 d_{\boldsymbol{\phi}}^2 \log(Hd_{\boldsymbol{\phi}}|\Phi|/\delta)}{\widetilde{\sigma}_{h,\boldsymbol{\phi}}^2}\left(1 + \frac{C_{\boldsymbol{\psi}}^2 H^4 \beta_{\boldsymbol{\phi}} C_{\boldsymbol{\phi}} \widetilde{\sigma}_{h,\boldsymbol{\phi}}}{4\mathrm{gap}_{\min}^2 C_{\boldsymbol{\phi}}^2 d_{\boldsymbol{\phi}} \log(Hd_{\boldsymbol{\phi}}|\Phi|/\delta)}\right), \tag{6.1}$$

we have $\|\boldsymbol{\phi}(s,a)\|_{\mathbf{U}_{h,\boldsymbol{\phi}}^{-1}} < \mathrm{gap}_{\min}/(2H^2 C_{\boldsymbol{\psi}}\sqrt{\beta_{\boldsymbol{\phi}}})$. Here $\widetilde{\sigma}_{h,\boldsymbol{\phi}}$ is the minimum non-zero eigen value of expected offline matrix $\mathbb{E}_{d_h^{\tilde{\pi}}}[\boldsymbol{\phi}\boldsymbol{\phi}^\top]$ and K is the number of trajectories in offline data.

Equipped with these lemmas, we can start our proof.

Proof of Theorem 5.4. The proof for the first part of the theorem have been shown in Lemma 6.2, where we assume the event $\mathcal{E}_{\boldsymbol{\phi}}$ in Lemma 6.1 holds. Then suppose the event in Lemma 6.3 holds, let K be greater than the threshold (6.1) provided in Lemma 6.3, i.e.

$$K > \max_{\boldsymbol{\phi}\in\Phi, h\in[H]}\left\{\frac{32C_{\boldsymbol{\phi}}^2 d_{\boldsymbol{\phi}}^2 \log(Hd_{\boldsymbol{\phi}}|\Phi|/\delta)}{\widetilde{\sigma}_{h,\boldsymbol{\phi}}^2}\left(1 + \frac{C_{\boldsymbol{\psi}}^2 H^4 \beta_{\boldsymbol{\phi}} C_{\boldsymbol{\phi}} \widetilde{\sigma}_{h,\boldsymbol{\phi}}}{4\mathrm{gap}_{\min}^2 C_{\boldsymbol{\phi}}^2 d_{\boldsymbol{\phi}} \log(Hd_{\boldsymbol{\phi}}|\Phi|/\delta)}\right)\right\},. \tag{6.2}$$

By Lemma 6.3, for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, there exists a $\boldsymbol{\phi} \in \Phi$ such that $\|\boldsymbol{\phi}(s,a)\|_{\mathbf{U}_{h,\boldsymbol{\phi}}^{-1}} < \Delta/(2H^2 C_{\boldsymbol{\psi}}\sqrt{\beta_{\boldsymbol{\phi}}})$. Then by Lemma 6.2, for any state $s \in \mathcal{S}$ at step $h \in [H]$, the sub-optimality is bounded by

$$\begin{aligned} V_h^*(s) - V_h^\pi(s) &\le 2C_{\boldsymbol{\psi}} H \sum_{h'=h}^{H} \mathbb{E}_{\pi^*}\left[\min_{\boldsymbol{\phi}\in\Phi}\left\{\sqrt{\beta_{\boldsymbol{\phi}}}\|\boldsymbol{\phi}(s,a)\|_{\mathbf{U}_{h',\boldsymbol{\phi}}^{-1}}\right\}\middle| s_h = s\right] \\ &< \sum_{h'=h}^{H} \mathbb{E}_{\pi^*}\left[\frac{\mathrm{gap}_{\min}}{H}\middle| s_h = s\right] \\ &< \mathrm{gap}_{\min}. \end{aligned} \tag{6.3}$$

On the other hand, since $Q_h^*(s,\pi(s)) \geq V_h^\pi(s)$, by the definition of sub-optimality gap in Definition 3.2,

$$V_h^*(s) - V_h^\pi(s) \geq V_h^*(s) - Q_h^*(s,\pi(s)) \geq \mathbb{1}[\pi_h(s) \neq \pi_h^*(s)]\text{gap}_{\min}, \tag{6.4}$$

where the last inequality follows the uniqueness of the optimal policy and all other action will lead to sub-optimality. Combining (6.3) and (6.4) together suggests that when K satisfies the condition in (6.2),

$$\mathbb{1}[\pi_h(s) \neq \pi_h^*(s)]\text{gap}_{\min} < \text{gap}_{\min},$$

which yields that $\pi_h(s) = \pi_h^*(s)$. Applying this to all $(s,h) \in \mathcal{S} \times [H]$ and replacing δ with $\delta/(2|\Phi|)$, we can get the claimed result in Theorem 5.4 by union bound. □

7 PROOF OF LEMMAS IN APPENDIX 6

7.1 PROOF OF LEMMA 6.2

First we need to introduce the extended value difference lemma provided in Jin et al. [2021], Cai et al. [2020]

Lemma 7.1 (Extended value difference Cai et al. [2020], Lemma A.1 Jin et al. [2021])**.** Let $\{\pi\}_h, \{\pi'\}_h$ by any two policies and let $\{\widehat{Q}\}_h$ be any estimated Q-function. For any $h \in [H]$, define the estimated value function as $\widehat{V}_h(s) = \widehat{Q}_h(s,\pi_h(s))$. For all $s \in \mathcal{S}$ we have

$$\begin{aligned}\widehat{V}_h(s) - V_h^{\pi'}(s) = &\sum_{h'=h}^{H} \mathbb{E}_{\pi'}\left[\widehat{Q}_{h'}(s_{h'},\pi_{h'}(s_{h'})) - \widehat{Q}_{h'}(s_{h'},\pi'_{h'}(s_{h'}))\Big)\Big|s_h = s\right] \\ &+ \sum_{h'=h}^{H} \mathbb{E}_{\pi'}\left[\widehat{Q}_{h'}(s_{h'},a_{h'}) - r(s_{h'},a_{h'}) - [\mathbb{P}\widehat{V}_{h'+1}](s_{h'},a_{h'})\Big|s_h = s\right],\end{aligned}$$

where $\mathbb{E}_{\pi'}$ is taken with respect to the trajectory generated by π' using underlying MDP and $a_{h'}$ is defined by $a_{h'} = \pi'_{h'}(s_{h'})$.

Proof. The proof of this lemma is same with Section B.1 in Cai et al. [2020] by replacing the initial state from 1 to any arbitrary step h. □

We also provide an error control lemma similar with Lemma 5.1 in online setting

Lemma 7.2 (Lemma 5 on $B_n^{(2)}$, pp. 23, Yang and Wang [2020])**.** Suppose $\mathcal{E}_{\boldsymbol{\phi}}$ holds, then for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\|\boldsymbol{\phi}(s,a)^\top(\mathbf{M}_{h,\boldsymbol{\phi}} - \mathbf{M}^*_{h,\boldsymbol{\phi}})\|_2 \leq \sqrt{\beta_{\boldsymbol{\phi}}\boldsymbol{\phi}^\top(s,a)\mathbf{U}^{-1}_{h,\boldsymbol{\phi}}\boldsymbol{\phi}(s,a)}.$$

Proof. The proof is similar with Yang and Wang [2020] by fixing k to K. □

Then our proof starts by following the idea in Jin et al. [2021].

Proof. First it is obvious that $V_h^*(s) - V_h^\pi(s) = (V_h^*(s) - V_h(s)) - (V_h^\pi(s) - V_h(s))$ where V_h is the estimated value function in Line 10 in Algorithm 2. Note that $V(s) = Q(s,\pi(s))$ where π is the output policy from Algorithm 2, Lemma 7.1 suggests that by setting $\pi' = \pi^*$, $V_h^*(s) - V_h(s)$ can be written by

$$\begin{aligned}V_h(s) - V_h^*(s) = &\sum_{h'=h}^{H} \mathbb{E}_{\pi^*}\left[Q_{h'}(s_{h'},\pi_{h'}(s_{h'})) - Q_{h'}(s_{h'},\pi^*_{h'}(s_{h'}))\right)|s_h = s] \\ &+ \sum_{h'=h}^{H} \mathbb{E}_{\pi^*}\left[Q_{h'}(s_{h'},a_{h'}) - r(s_{h'},a_{h'}) - [\mathbb{P}V_{h'+1}](s_{h'},a_{h'})|s_h = s\right] \\ \geq &\sum_{h'=h}^{H} \mathbb{E}_{\pi^*}\left[Q_{h'}(s_{h'},a_{h'}) - r(s_{h'},a_{h'}) - [\mathbb{P}V_{h'+1}](s_{h'},a_{h'})|s_h = s\right],\end{aligned} \tag{7.1}$$

where the last inequality is due to the fact that we are executing the greedy policy i.e. $\pi_h(s) = \operatorname{argmax} Q_h(s,a)$ thus $Q_h(s,\pi_h(s)) \geq Q_h(s,\pi_h^*(s))$. Meanwhile, letting $\pi = \pi' = \pi$, Lemma 7.1 suggests that

$$V_h(s) - V_h^{\pi}(s) = \sum_{h'=h}^{H} \mathbb{E}_{\pi}\left[Q_{h'}(s_{h'},a_{h'}) - r(s_{h'},a_{h'}) - [\mathbb{P}V_{h'+1}](s_{h'},a_{h'})|s_h = s\right]. \tag{7.2}$$

Noticing that both (7.1) and (7.2) are the summation about the $Q_h(s,a) - r(s,a) - [\mathbb{P}V_{h+1}](s,a)$, which we will bound next. Recall the calculation rule of Q-function in Line 10 suggests that

$$\begin{aligned}
&Q_h(s,a) - r(s,a) - [\mathbb{P}V_{h+1}](s,a) \\
&= \max_{\boldsymbol{\phi}\in\Phi}\left\{r(s,a) + \sum_{s'\in\mathcal{S}} \boldsymbol{\phi}(s,a)^{\top}\mathbf{M}_{h,\boldsymbol{\phi}}\boldsymbol{\psi}(s')V_{h+1}(s') - \Gamma_{h,\boldsymbol{\phi}}(s,a)\right\} - r(s,a) \\
&\quad - \sum_{s'\in\mathcal{S}} \boldsymbol{\phi}(s,a)^{\top}\mathbf{M}^*_{h,\boldsymbol{\phi}}\boldsymbol{\psi}(s')V_{h+1}(s') \\
&= \max_{\boldsymbol{\phi}\in\Phi}\left\{\sum_{s'\in\mathcal{S}} \boldsymbol{\phi}(s,a)^{\top}(\mathbf{M}_{h,\boldsymbol{\phi}} - \mathbf{M}^*_{\boldsymbol{\phi}})\boldsymbol{\psi}(s')V_{h+1}(s') - \Gamma_{h,\boldsymbol{\phi}}(s,a)\right\} \\
&= \max_{\boldsymbol{\phi}\in\Phi}\left\{\boldsymbol{\phi}(s,a)^{\top}(\mathbf{M}_{h,\boldsymbol{\phi}} - \mathbf{M}^*_{h,\boldsymbol{\phi}})\boldsymbol{\Psi}\mathbf{v}_{h+1} - C_{\boldsymbol{\psi}}H\sqrt{\beta_{\boldsymbol{\phi}}}\|\boldsymbol{\phi}(s,a)\|_{\mathbf{U}^{-1}_{h,\phi}}\right\}
\end{aligned} \tag{7.3}$$

where the last inequality utilize the notation $\boldsymbol{\Psi} = \left(\boldsymbol{\psi}(s_1),\boldsymbol{\psi}(s_2),\cdots,\boldsymbol{\psi}(s_{|\mathcal{S}|})\right)^{\top} \in \mathbb{R}^{|\mathcal{S}|\times d'}$ and $\mathbf{v}_{h+1} = \left(V_{h+1}(s_1),V_{h+1}(s_2),\cdots,V_{h+1}(s_{|\mathcal{S}|})\right)^{\top} \in \mathbb{R}^{|\mathcal{S}|}$. For each $\boldsymbol{\phi}\in\Phi$, lemma 7.2 suggests that

$$\begin{aligned}
\left|\boldsymbol{\phi}(s,a)^{\top}(\mathbf{M}_{h,\boldsymbol{\phi}} - \mathbf{M}^*_{h,\boldsymbol{\phi}})\boldsymbol{\Psi}\mathbf{v}_{h+1}\right| &\leq \left\|(\mathbf{M}_{h,\boldsymbol{\phi}} - \mathbf{M}^*_{h,\boldsymbol{\phi}})\boldsymbol{\phi}(s,a)\right\|_2 \|\boldsymbol{\Psi}\mathbf{v}_{h+1}\|_2 \\
&\leq C_{\boldsymbol{\psi}}H\sqrt{\beta_{\boldsymbol{\phi}}}\|\boldsymbol{\phi}(s,a)\|_{\mathbf{U}^{-1}_{h,\boldsymbol{\phi}}},
\end{aligned}$$

where the first inequality follows the C-S inequality and the second inequality utilizes the fact that $\|\boldsymbol{\Psi}\mathbf{v}_{h+1}\|_2 \leq C_{\boldsymbol{\psi}}\|\mathbf{v}_{h+1}\|_{\infty} \leq C_{\boldsymbol{\psi}}H$. Therefore for all $\boldsymbol{\phi}\in\Phi$, for any $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$:

$$-2C_{\boldsymbol{\psi}}H\sqrt{\beta_{\boldsymbol{\phi}}}\|\boldsymbol{\phi}(s,a)\|_{\mathbf{U}^{-1}_{h,\boldsymbol{\phi}}} \leq \boldsymbol{\phi}(s,a)^{\top}(\mathbf{M}_{h,\boldsymbol{\phi}} - \mathbf{M}^*_{h,\boldsymbol{\phi}})\boldsymbol{\Psi}\mathbf{v}_{h+1} - C_{\boldsymbol{\psi}}H\sqrt{\beta_{\boldsymbol{\phi}}}\|\boldsymbol{\phi}(s,a)\|_{\mathbf{U}^{-1}_{h,\boldsymbol{\phi}}} \leq 0.$$

Plugging this into (7.3) yields

$$\begin{aligned}
&Q_h(s,a) - r(s,a) - [\mathbb{P}V_{h+1}](s,a) \\
&\quad = \max_{\boldsymbol{\phi}\in\Phi}\left\{\boldsymbol{\phi}(s,a)^{\top}(\mathbf{M}_{h,\boldsymbol{\phi}} - \mathbf{M}^*_{h,\boldsymbol{\phi}})\boldsymbol{\Psi}\mathbf{v}_{h+1} - C_{\boldsymbol{\psi}}H\sqrt{\beta_{\boldsymbol{\phi}}}\|\boldsymbol{\phi}(s,a)\|_{\mathbf{U}^{-1}_{h,\boldsymbol{\phi}}}\right\} \\
&\quad \geq \max_{\boldsymbol{\phi}\in\Phi}\left\{-2C_{\boldsymbol{\psi}}H\sqrt{\beta_{\boldsymbol{\phi}}}\|\boldsymbol{\phi}(s,a)\|_{\mathbf{U}^{-1}_{h,\boldsymbol{\phi}}}\right\} \\
&\quad = -2C_{\boldsymbol{\psi}}H\min_{\boldsymbol{\phi}\in\Phi}\left\{\sqrt{\beta_{\boldsymbol{\phi}}}\|\boldsymbol{\phi}(s,a)\|_{\mathbf{U}^{-1}_{h,\boldsymbol{\phi}}}\right\}
\end{aligned}$$

and $Q_h(s,a) - r(s,a) - [\mathbb{P}_h V_{h+1}](s,a) \leq 0$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. Plugging the bound back to (7.1) yields

$$\begin{aligned}
V_h(s) - V_h^*(s) &\geq \sum_{h'=h}^{H} \mathbb{E}_{\pi^*}\left[Q_{h'}(s_{h'},a_{h'}) - r(s_{h'},a_{h'}) - [\mathbb{P}V_{h'+1}](s_{h'},a_{h'})|s_h = s\right] \\
&\geq -2C_{\boldsymbol{\psi}}H\sum_{h'=h}^{H}\mathbb{E}_{\pi^*}\left[\min_{\boldsymbol{\phi}\in\Phi}\left\{\sqrt{\beta_{\boldsymbol{\phi}}}\|\boldsymbol{\phi}(s,a)\|_{\mathbf{U}^{-1}_{h',\boldsymbol{\phi}}}\right\}\middle| s_h = s\right]
\end{aligned} \tag{7.4}$$

and back to (7.2) yields

$$V_h(s) - V_h^{\pi}(s) = \sum_{h'=h}^{H} \mathbb{E}_{\pi^*}\left[Q_{h'}(s_{h'},a_{h'}) - r(s_{h'},a_{h'}) - [\mathbb{P}V_{h'+1}](s_{h'},a_{h'})|s_h = s\right] \leq 0. \tag{7.5}$$

substituting (7.4) from (7.5) yields the claimed result:

$$V_h^*(s) - V_h^{\pi}(s) \leq 2C_{\boldsymbol{\psi}}H\sum_{h'=h}^{H}\mathbb{E}_{\pi^*}\left[\min_{\boldsymbol{\phi}\in\Phi}\left\{\sqrt{\beta_{\boldsymbol{\phi}}}\|\boldsymbol{\phi}(s,a)\|_{\mathbf{U}^{-1}_{h',\boldsymbol{\phi}}}\right\}\middle| s_h = s\right]$$

□

7.2 PROOF OF LEMMA 6.3

Proof. First we show that the covariance matrix $\mathbf{U}_{\boldsymbol{\phi},h}$ is almost linearly growth with respect to the expectation $\mathbb{E}_{d_h^{\widehat{\pi}}}[\boldsymbol{\phi}\boldsymbol{\phi}^\top]$ and the size of offline data. Considering the formalization of covariance matrix

$$\begin{aligned}\mathbf{U}_{\boldsymbol{\phi},h} &= \mathbf{I} + \sum_{(s,a,s')\in\mathcal{D}_h} \boldsymbol{\phi}(s,a)\boldsymbol{\phi}^\top(s,a) \\ &= \mathbf{I} - \sum_{(s,a,s')\in\mathcal{D}_h} \underbrace{\mathbb{E}_{d_h^{\widehat{\pi}}}[\boldsymbol{\phi}(s,a)\boldsymbol{\phi}^\top(s,a)] - \boldsymbol{\phi}(s,a)\boldsymbol{\phi}^\top(s,a)}_{\boldsymbol{\epsilon}_h} + |\mathcal{D}_h|\mathbb{E}_{d_h^{\widehat{\pi}}}[\boldsymbol{\phi}(s,a)\boldsymbol{\phi}^\top(s,a)].\end{aligned}$$

One can verify that $\mathbb{E}[\boldsymbol{\epsilon}_h] = \mathbf{0}$ where the expectation is taken with respect to the randomness in the generation of the offline data. Since we have $\|\boldsymbol{\phi}(s,a)\|_2^2 \le C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}}$, it is obvious that $\|\boldsymbol{\epsilon}_h\|_2 \le \|\boldsymbol{\phi}\boldsymbol{\phi}^\top\|_2 + \|\mathbb{E}[\boldsymbol{\phi}\boldsymbol{\phi}^\top]\|_2 \le 2\|\boldsymbol{\phi}\|_2^2 \le 2C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}}$ by triangle's inequality. Then by Lemma 4.6, with probability at least $1-\delta$, for $|\mathcal{D}_h| = K$ data,

$$\lambda_{\max}\left(\sum_{(s,a,s')\in\mathcal{D}_h} \boldsymbol{\epsilon}_h\right) \le 4C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}} \sqrt{2K\log(d_{\boldsymbol{\phi}}/\delta)}.$$

Therefore, it's suffice to show that

$$\mathbf{U}_{\boldsymbol{\phi},h} \succeq K\mathbb{E}_{d_h^{\widehat{\pi}}}[\boldsymbol{\phi}(s,a)\boldsymbol{\phi}^\top(s,a)] + \left(1 - 4C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}} \sqrt{2K\log(d_{\boldsymbol{\phi}}/\delta)}\right)\mathbf{I} \tag{7.6}$$

Furthermore, noticing that $\mathbb{E}_{d_h^{\widehat{\pi}}}[\boldsymbol{\phi}(s,a)\boldsymbol{\phi}^\top(s,a)]$ can be always written as

$$\mathbb{E}_{d_h^{\widehat{\pi}}}[\boldsymbol{\phi}(s,a)\boldsymbol{\phi}^\top(s,a)] = \mathbf{Q}^\top \mathrm{diag}(\mathbf{d}_r, \mathbf{0}_{d_{\boldsymbol{\phi}}-r})\mathbf{Q}, \tag{7.7}$$

where $\mathbf{Q}$ is an orthogonal matrix, $\mathbf{d}_r \in \mathbb{R}^r$ is the non-zero eigen values of $\mathbb{E}_{d_h^{\widehat{\pi}}}[\boldsymbol{\phi}(s,a)\boldsymbol{\phi}^\top(s,a)]$ its minimal element as $\widetilde{\sigma}_{h,\boldsymbol{\phi}}$. $r = \mathrm{rank}(\mathbb{E}_{d_h^{\widehat{\pi}}}[\boldsymbol{\phi}(s,a)\boldsymbol{\phi}^\top(s,a)])$. Then (7.6) can be formalized as

$$\mathbf{Q}\mathbf{U}_{\boldsymbol{\phi},h}\mathbf{Q}^\top \succeq \mathrm{diag}\left(K\mathbf{d}_r - a\mathbf{1}_r + \mathbf{1}_r, -a\mathbf{1}_{d_{\boldsymbol{\phi}}-r} + \mathbf{1}_{d_{\boldsymbol{\phi}}-r}\right), \tag{7.8}$$

where $a = 4C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}} \sqrt{2K\log(d_{\boldsymbol{\phi}}/\delta)}$ is in the order of $\sqrt{K}$. On the other hand, since $\mathbf{U}_{\boldsymbol{\phi},h} \succeq \mathbf{I}$, then $\mathbf{Q}\mathbf{U}_{\boldsymbol{\phi},h}\mathbf{Q}^\top \succeq \mathbf{Q}\mathbf{Q}^\top = \mathbf{I}$. Combining this with (7.8) we can conclude that

$$\mathbf{Q}\mathbf{U}_{\boldsymbol{\phi},h}\mathbf{Q}^\top \succeq \mathrm{diag}\left(K\mathbf{d}_r - a\mathbf{1}_r + \mathbf{1}_r, \mathbf{1}_{d_{\boldsymbol{\phi}}-r}\right). \tag{7.9}$$

Noticing the minimal element of $\mathbf{d}_r$ is $\widetilde{\sigma}_{h,\boldsymbol{\phi}}$, then when $K\widetilde{\sigma}_{h,\boldsymbol{\phi}} \ge a$, which we will verify later, the RHS of (7.9) is positive definite, which implies that

$$\mathbf{Q}\mathbf{U}_{\boldsymbol{\phi},h}^{-1}\mathbf{Q}^\top \preceq \mathrm{diag}\left(K\mathbf{d}_r - a\mathbf{1}_r + \mathbf{1}_r, \mathbf{1}_{d_{\boldsymbol{\phi}}-r}\right)^{-1}. \tag{7.10}$$

By union bound (7.10) holds for all $\boldsymbol{\phi} \in \Phi$ and $h \in [H]$ with probability at least $1 - H|\Phi|\delta$. Then for any $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, according to Assumption 5.1, there exists $\boldsymbol{\phi} \in \Phi$ and $\mathbf{y} \in \mathbb{R}^d$ such that $\boldsymbol{\phi}(s,a) = \mathbb{E}_{d_h^{\widehat{\pi}}}[\boldsymbol{\phi}(s,a)\boldsymbol{\phi}^\top(s,a)]\mathbf{y}$, combining this with (7.7) and (7.10) yields that

$$\begin{aligned}&\boldsymbol{\phi}^\top(s,a)\mathbf{U}_{h,\boldsymbol{\phi}}^{-1}\boldsymbol{\phi}(s,a) \\ &= \mathbf{y}^\top\mathbb{E}_{d_h^{\widehat{\pi}}}[\boldsymbol{\phi}(s,a)\boldsymbol{\phi}^\top(s,a)]^\top\mathbf{U}_{h,\boldsymbol{\phi}}^{-1}\mathbb{E}_{d_h^{\widehat{\pi}}}[\boldsymbol{\phi}(s,a)\boldsymbol{\phi}^\top(s,a)]\mathbf{y} \\ &= \mathbf{y}^\top\mathbf{Q}^\top\mathrm{diag}(\mathbf{d}_r, \mathbf{0}_{d_{\boldsymbol{\phi}}-r})\mathbf{Q}\mathbf{U}_{h,\boldsymbol{\phi}}^{-1}\mathbf{Q}^\top\mathrm{diag}(\mathbf{d}_r, \mathbf{0}_{d_{\boldsymbol{\phi}}-r})\mathbf{Q}\mathbf{y} \\ &\le \mathbf{y}^\top\mathbf{Q}^\top\mathrm{diag}(\mathbf{d}_r, \mathbf{0}_{d_{\boldsymbol{\phi}}-r})\mathrm{diag}(K\mathbf{d}_r - a\mathbf{1}_r + \mathbf{1}_r, \mathbf{1}_{d_{\boldsymbol{\phi}}-r})^{-1}\mathrm{diag}(\mathbf{d}_r, \mathbf{0}_{d_{\boldsymbol{\phi}}-r})\mathbf{Q}\mathbf{y},\end{aligned} \tag{7.11}$$

where the last inequality follows (7.10). Noticing that $\mathrm{diag}(\mathbf{d}_r, \mathbf{0}_{d_{\boldsymbol{\phi}}-r})\mathbf{Q}\mathbf{y}$ can be written as $\mathrm{diag}(\mathbf{d}_r, \mathbf{0}_{d_{\boldsymbol{\phi}}-r})\mathbf{Q}\mathbf{y} = \left(\mathbf{z}_r^\top, \mathbf{0}_{d_{\boldsymbol{\phi}}-r}^\top\right)^\top$. Therefore (7.11) becomes

$$\begin{aligned}\boldsymbol{\phi}^\top(s,a)\mathbf{U}_{h,\boldsymbol{\phi}}^{-1}\boldsymbol{\phi}(s,a) &\le \left(\mathbf{z}_r^\top, \mathbf{0}_{d_{\boldsymbol{\phi}}-r}^\top\right)^\top \mathrm{diag}(K\mathbf{d}_r - a\mathbf{1}_r + \mathbf{1}_r, \mathbf{1}_{d_{\boldsymbol{\phi}}-r})^{-1}\left(\mathbf{z}_r^\top, \mathbf{0}_{d_{\boldsymbol{\phi}}-r}^\top\right)^\top \\ &= \mathbf{z}_r^\top\mathrm{diag}(K\mathbf{d}_r - a\mathbf{1}_r + \mathbf{1}_r)^{-1}\mathbf{z}_r.\end{aligned} \tag{7.12}$$

Noticing that

$$\|\mathbf{z}_r\|_2 = \left\|\mathrm{diag}(\mathbf{d}_r, \mathbf{0}_{d_{\boldsymbol{\phi}}-r})\mathbf{Q}\mathbf{y}\right\| = \|\mathbf{Q}\boldsymbol{\phi}(s,a)\|_2 = \|\boldsymbol{\phi}(s,a)\|_2 \le \sqrt{C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}}},$$

where the last equality comes form Definition 3.2, (7.12) finally becomes

$$\boldsymbol{\phi}^\top(s,a)\mathbf{U}_{h,\boldsymbol{\phi}}^{-1}\boldsymbol{\phi}(s,a) \le \mathbf{z}_r^\top \mathrm{diag}(K\mathbf{d}_r - a\mathbf{1}_r + \mathbf{1}_r)^{-1}\mathbf{z}_r \le \frac{C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}}}{K\widetilde{\sigma}_{h,\boldsymbol{\phi}} - a + 1} \le \frac{C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}}}{K\widetilde{\sigma}_{h,\boldsymbol{\phi}} - a}, \tag{7.13}$$

where the second last inequality is due to $\lambda_{\max}\left(\mathrm{diag}(K\mathbf{d}_r - a\mathbf{1}_r + \mathbf{1}_r)^{-1}\right) = (K\widetilde{\sigma}_{h,\boldsymbol{\phi}} - a + 1)^{-1}$ and the last inequality utilizes the assumption that $K\widetilde{\sigma}_{h,\boldsymbol{\phi}} \ge a0$

Next, to control $\|\boldsymbol{\phi}(s,a)\|_{\mathbf{U}_{h,\boldsymbol{\phi}}^{-1}} \le \Delta/(2C_{\boldsymbol{\psi}}H^2\sqrt{\beta_{\boldsymbol{\phi}}}) =: B$, by (7.13), it suffices to control

$$\frac{C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}}}{K\widetilde{\sigma}_{h,\boldsymbol{\phi}} - 4C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}}\sqrt{2K\log(d_{\boldsymbol{\phi}}/\delta)}} \le B^2. \tag{7.14}$$

Denoting $\sqrt{K} = 4C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}}\sqrt{2\log(d_{\boldsymbol{\phi}}/\delta)}\widetilde{\sigma}_{h,\boldsymbol{\phi}}^{-1}x$, then the constrain $K\widetilde{\sigma}_{h,\boldsymbol{\phi}} \ge 4C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}}\sqrt{2K\log(d_{\boldsymbol{\phi}}/\delta)}$ is equivalent with $x \ge 1$, and (7.14) becomes

$$\frac{C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}}}{B^2} \le \frac{32C_{\boldsymbol{\phi}}^2 d_{\boldsymbol{\phi}}^2 \log(d_{\boldsymbol{\phi}}/\delta)}{\widetilde{\sigma}_{h,\boldsymbol{\phi}}}(x^2 - x). \tag{7.15}$$

Since the sufficient condition of inequality $x^2 - x - c > 0$ is $x > (1+\sqrt{1+2c})/2$ which could be implied by $x > \sqrt{1+2c}$ by C-S inequality, the sufficient condition of (7.15) could be written as

$$x > \sqrt{1 + \frac{C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}}\widetilde{\sigma}_{h,\boldsymbol{\phi}}}{16B^2C_{\boldsymbol{\phi}}^2 d_{\boldsymbol{\phi}}^2\log(d_{\boldsymbol{\phi}}/\delta)}}, \tag{7.16}$$

which can imply the constrain that $x \ge 1$. Plugging the notations $B := \Delta/(2C_{\boldsymbol{\psi}}H^2\sqrt{\beta_{\boldsymbol{\phi}}})$ and $\sqrt{K} = 4C_{\boldsymbol{\phi}} d_{\boldsymbol{\phi}}\sqrt{2\log(d_{\boldsymbol{\phi}}/\delta)}\widetilde{\sigma}_{h,\boldsymbol{\phi}}^{-1}x$ back into (7.16) yields for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, with probability at least $1 - H|\Phi|\delta$, there exists $\boldsymbol{\phi} \in \Phi$ such that when

$$K > \frac{32C_{\boldsymbol{\phi}}^2 d_{\boldsymbol{\phi}}^2 \log(d_{\boldsymbol{\phi}}/\delta)}{\widetilde{\sigma}_{h,\boldsymbol{\phi}}^2}\left(1 + \frac{C_{\boldsymbol{\psi}}^2 H^4 \beta_{\boldsymbol{\phi}} C_{\boldsymbol{\phi}}\widetilde{\sigma}_{h,\boldsymbol{\phi}}}{4\Delta^2 C_{\boldsymbol{\phi}}^2 d_{\boldsymbol{\phi}}\log(d_{\boldsymbol{\phi}}/\delta)}\right),$$

$\|\boldsymbol{\phi}(s,a)\|_{\mathbf{U}_{h,\boldsymbol{\phi}}} < \mathrm{gap}_{\min}/(2C_{\boldsymbol{\psi}}H^2\sqrt{\beta_{\boldsymbol{\phi}}})$. Replacing δ by $\delta/(H|\Phi|)$ we can get the claimed result. □

References

Yasin Abbasi-Yadkori, Aldo Pacchiano, and My Phan. Regret balancing for bandit and rl model selection. *arXiv preprint arXiv:2006.05491*, 2020.

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38. PMLR, 2017.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

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

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19 (3):357–367, 1967.

Peter L Bartlett, Olivier Bousquet, Shahar Mendelson, et al. Local rademacher complexities. *The Annals of Statistics*, 33(4): 1497–1537, 2005.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

Niladri Chatterji, Vidya Muthukumar, and Peter Bartlett. Osom: A simultaneously optimal algorithm for multi-armed and linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 1844–1854. PMLR, 2020.

Dylan Foster, Akshay Krishnamurthy, and Haipeng Luo. Model selection for contextual bandits. *Advances in neural information processing systems*, 2019.

Dylan J Foster, Akshay Krishnamurthy, and Haipeng Luo. Open problem: Model selection for contextual bandits. In *Conference on Learning Theory*, pages 3842–3846. PMLR, 2020.

Dylan J Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. *arXiv preprint arXiv:2111.10919*, 2021.

Avishek Ghosh, Abishek Sankararaman, and Ramchandran Kannan. Problem-complexity adaptive model selection for stochastic linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 1396–1404. PMLR, 2021.

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

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

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

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

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

Yifei Min, Tianhao Wang, Dongruo Zhou, and Quanquan Gu. Variance-aware off-policy evaluation with linear function approximation. *Advances in neural information processing systems*, 34, 2021.

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

Maillard Odalric and Rémi Munos. Adaptive bandits: Towards the best history-dependent strategy. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 570–578. JMLR Workshop and Conference Proceedings, 2011.

Aldo Pacchiano, Christoph Dann, Claudio Gentile, and Peter Bartlett. Regret bound balancing and elimination for model selection in bandits and rl. *arXiv preprint arXiv:2012.13045*, 2020a.

Aldo Pacchiano, My Phan, Yasin Abbasi Yadkori, Anup Rao, Julian Zimmert, Tor Lattimore, and Csaba Szepesvari. Model selection in contextual stochastic bandit problems. *Advances in Neural Information Processing Systems*, 33, 2020b.

Matteo Papini, Andrea Tirinzoni, Marcello Restelli, Alessandro Lazaric, and Matteo Pirotta. Leveraging good representations in linear contextual bandits. In *International Conference on Machine Learning*. PMLR, 2021.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4): 389–434, 2012.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline rl: Pac bounds and posterior sampling under partial coverage. *arXiv e-prints*, pages arXiv–2107, 2021.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34, 2021.

Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.

Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal offline reinforcement learning via double variance reduction. *Advances in neural information processing systems*, 34, 2021.

Ming Yin, Yaqi Duan, Mengdi Wang, and Yu-Xiang Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. In *International Conference on Learning Representations*, 2022.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.

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

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