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 \( \phi(s, a) \). Under this assumption, Jin et al. [2020] proposed the LSVI-UCB algorithm achieving \( \tilde{O}(d^3H^3T) \) problem independent regret bound and \( \tilde{O}(d^3H^3\text{gap}_{\min}^{-1}\log(T)) \) 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 \( \tilde{O}(dH^2\sqrt{T}) \) 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 \( \tilde{O}(H^2d^2\sqrt{T}) \) 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 \( \phi(s, a, s') \) of the ‘state-action-next state’ triplet. Under this setting, Jia et al. [2020], Ayoub et al. [2020] proposed UCRL-VTR achieving \( \tilde{O}(d^2H^3T) \) problem independent bound for episodic MDP, while He et al. [2021] showed an \( \tilde{O}(d^2H^6\text{gap}_{\min}^{-1}\log^3(T)) \) 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 \( \tilde{O}(d^2\sqrt{T}(1 - \gamma)^2) \) 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 branch 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
We will next provide two other representations. \textbf{Assumption 4.2} To emphasize the difficulties of learning the transition kernel $P$ without modification. Then we define $\pi$ $P$ representation satisfies Assumption 4.2 but the combination of these two will satisfy that assumption. Sampled from the half-normal distribution. Therefore, Assumption 4.2 is satisfied. It is easy to verify that under the current representation $\Lambda$ could always be done since we can multiply different scalars to the generated matrix $M$. Furthermore, we select the matrix $M$ such that for all state-action pair $(s, a) \in S \times A$, $\|\phi_h(s, a)\|_2 \leq 1$. This procedure could always be done since we can multiply different scalars to the generated matrix $M$. To control the norm of $\phi_h$. Therefore, we can verify that for any $(s, a, s', h) \in S \times A \times S \times [H]$, $P_h(s'|s, a) = \phi_h^T(s, a)M_h\psi(s')$ thus it satisfies Assumption 4.2. To emphasize the difficulties of learning the transition kernel $P$.

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

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

Since the transition kernel $P_h(\cdot|\cdot, \cdot)$ 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 $s \in S$. Since $|A| = 3$ and $\pi_h^*(s) \in A$, we can compose the sub-optimal set by $A \setminus \{\pi_h^*(s)\} := \{a_h(s), a_h'(s)\}$. Then we define the two representation...
functions as \(\phi^{(1)}, \phi^{(2)} \in \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^{2d}\) using the following rule:

\[
\begin{align*}
\phi^{(1)}_h(s, \pi^*_h(s)) &= (\phi^T_h(s, \pi^*_h(s)), 0_d)^T, \\
\phi^{(2)}_h(s, \pi^*_h(s)) &= (\phi^T_h(s, \pi^*_h(s)), 0_d)^T, \\
\phi^{(1)}_h(s, a_h(s)) &= (\phi^T_h(s, a_h(s)), 0_d)^T, \\
\phi^{(2)}_h(s, a_h(s)) &= (\phi^T_h(s, a_h(s)), 0_d)^T.
\end{align*}
\]

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

\[
F_h(s'|s, a) = \phi^{(1)\top}_h \widetilde{M}_h \psi(s') = \phi^{(2)\top}_h \widetilde{M}_h \psi(s').
\]

From intuition, these two representation functions \(\phi^{(1)}\) and \(\phi^{(2)}\) may come from two different sensors measuring the same environment. It is obvious that since for both \(\phi^{(1)}\) and \(\phi^{(2)}\), there are at least 1/3 of the whole state-action space is not covered by \(\Lambda_h\), i.e. \(\phi^{(1)}_h(s, a'_h(s)) \notin \text{Im} \Lambda_h \phi^{(1)}\) and \(\phi^{(2)}_h(s, a_h(s)) \notin \text{Im} \Lambda_h \phi^{(2)}\). However, since \(a'_h(s') \neq a_h(s)\) by definition, we can verify that the representation set \(\Phi = \{\phi^{(1)}, \phi^{(2)}\}\) 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 \(\phi^{(1)}\) 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 \(\tilde{\pi}\) 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 \((s, \pi_h(s)), (s, a_h(s))\) and \((s, a'_h(s))\). Considering the representation set \(\phi^{(1)}(\cdot, \cdot) \in \mathbb{R}^{2d}\) and \(\phi^{(2)}(\cdot, \cdot) \in \mathbb{R}^{2d}\) which is defined as

\[
\begin{align*}
\phi^{(1)}_h(s, \pi_h(s)) &= (\phi^T_h(s, \pi_h(s)), 0_d)^T, \\
\phi^{(2)}_h(s, \pi_h(s)) &= (\phi^T_h(s, \pi_h(s)), 0_d)^T, \\
\phi^{(1)}_h(s, a_h(s)) &= (\phi^T_h(s, a_h(s)), 0_d)^T, \\
\phi^{(2)}_h(s, a_h(s)) &= (\phi^T_h(s, a_h(s)), 0_d)^T.
\end{align*}
\]

It is obvious that by using behavior policy \(\tilde{\pi}\), both \(\mathbb{E}_{d_h^\pi} (\phi^{(1)\top}_h \phi^{(1)}_h)\) and \(\mathbb{E}_{d_h^\pi} (\phi^{(2)\top}_h \phi^{(2)}_h)\) would enjoy the format of \(\begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix}\).

Therefore, for \(\phi^{(1)}\), it would be easy to verify that \(\phi^{(1)}_h, \phi^{(1)}(s, a'_h(s))\) is not in \(\text{Im} (\mathbb{E}_{d_h^\pi} (\phi^{(1)\top}_h \phi^{(1)}_h))\) and \(\phi^{(2)}(s, a_h(s))\) is not in \(\text{Im} (\mathbb{E}_{d_h^\pi} (\phi^{(2)\top}_h \phi^{(2)}_h))\). However, it is also easy to show that the union of \(\{\phi^{(1)}, \phi^{(2)}\}\) satisfy assumption 5.1.

2.3 ADDITIONAL CONFIGURATION

Parameter Tuning. For both of the offline and online algorithm, we aggregate the parameter \(C_\psi H \sqrt{\beta_h \phi}\) 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 \(\phi\) and ReLEX-UCB grew very slowly after the first one million episodes. As a comparison, UC-MatrixRL using \(\phi^{(1)}\) or \(\phi^{(2)}\) had a sub-linear regret growth instead of near-constant regret. As for the \(\epsilon\)-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.
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 (∼2.5×) 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 $\epsilon$-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.
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Last averaged reward ↑</th>
<th>Cumulative regret ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>UC-Matrix RL $\phi$ (oracle)</td>
<td>0.7782</td>
<td>2457.71</td>
</tr>
<tr>
<td>ReLEX ${\phi_1,\phi_2}$</td>
<td><strong>0.7780</strong></td>
<td><strong>6827.18</strong></td>
</tr>
<tr>
<td>UC-Matrix RL $\phi_1$</td>
<td><strong>0.7780</strong></td>
<td>11458.83</td>
</tr>
<tr>
<td>UC-Matrix RL $\phi_2$</td>
<td>0.7770</td>
<td>13385.61</td>
</tr>
<tr>
<td>$\epsilon$-greedy $\phi$</td>
<td>0.7754</td>
<td>14906.31</td>
</tr>
<tr>
<td>$\epsilon$-greedy $\phi_1$</td>
<td>0.7756</td>
<td>15042.24</td>
</tr>
<tr>
<td>$\epsilon$-greedy $\phi_2$</td>
<td>0.7751</td>
<td>16470.94</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Last averaged reward ↑</th>
<th>Cumulative regret ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>UC-Matrix RL $\phi$ (oracle)</td>
<td>0.8736</td>
<td>2880.71</td>
</tr>
<tr>
<td>ReLEX ${\phi_1,\phi_2}$</td>
<td><strong>0.8733</strong></td>
<td><strong>7745.49</strong></td>
</tr>
<tr>
<td>UC-Matrix RL $\phi_1$</td>
<td>0.8729</td>
<td>12759.89</td>
</tr>
<tr>
<td>UC-Matrix RL $\phi_2$</td>
<td>0.8730</td>
<td>10411.89</td>
</tr>
<tr>
<td>$\epsilon$-greedy $\phi$</td>
<td>0.8703</td>
<td>18786.77</td>
</tr>
<tr>
<td>$\epsilon$-greedy $\phi_1$</td>
<td>0.8707</td>
<td>19002.81</td>
</tr>
<tr>
<td>$\epsilon$-greedy $\phi_2$</td>
<td>0.8702</td>
<td>20018.97</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Last averaged reward ↑</th>
<th>Cumulative regret ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>UC-Matrix RL $\phi$ (oracle)</td>
<td>0.9749</td>
<td>3085.21</td>
</tr>
<tr>
<td>ReLEX ${\phi_1,\phi_2}$</td>
<td><strong>0.9748</strong></td>
<td><strong>8160.39</strong></td>
</tr>
<tr>
<td>UC-Matrix RL $\phi_1$</td>
<td>0.9743</td>
<td>12946.34</td>
</tr>
<tr>
<td>UC-Matrix RL $\phi_2$</td>
<td>0.9745</td>
<td>14373.86</td>
</tr>
<tr>
<td>$\epsilon$-greedy $\phi$</td>
<td>0.9690</td>
<td>28423.82</td>
</tr>
<tr>
<td>$\epsilon$-greedy $\phi_1$</td>
<td>0.9701</td>
<td>27479.92</td>
</tr>
<tr>
<td>$\epsilon$-greedy $\phi_2$</td>
<td>0.9708</td>
<td>27778.66</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Last averaged reward ↑</th>
<th>Cumulative regret ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>UC-Matrix RL $\phi$ (oracle)</td>
<td>0.9800</td>
<td>3403.65</td>
</tr>
<tr>
<td>ReLEX ${\phi_1,\phi_2}$</td>
<td><strong>0.9792</strong></td>
<td><strong>9733.84</strong></td>
</tr>
<tr>
<td>UC-Matrix RL $\phi_1$</td>
<td>0.9787</td>
<td>10000.15</td>
</tr>
<tr>
<td>UC-Matrix RL $\phi_2$</td>
<td>0.9791</td>
<td><strong>9553.96</strong></td>
</tr>
<tr>
<td>$\epsilon$-greedy $\phi$</td>
<td>0.9763</td>
<td>23301.53</td>
</tr>
<tr>
<td>$\epsilon$-greedy $\phi_1$</td>
<td>0.9759</td>
<td>23392.78</td>
</tr>
<tr>
<td>$\epsilon$-greedy $\phi_2$</td>
<td>0.9758</td>
<td>23218.40</td>
</tr>
</tbody>
</table>

Table 4: The performance result of Configuration (D) ($S = 40$, $A = 5$)
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Last averaged reward ↑</th>
<th>Cumulative regret ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>UC-Matrix RL ( \phi ) (oracle)</td>
<td>0.9081</td>
<td>1141.63</td>
</tr>
<tr>
<td>UC-Matrix RL ( \phi_1 )</td>
<td>0.9034</td>
<td>4512.82</td>
</tr>
<tr>
<td>UC-Matrix RL ( \phi_2 )</td>
<td>0.9008</td>
<td>4965.15</td>
</tr>
<tr>
<td>UC-Matrix RL ( \phi_3 )</td>
<td>0.9022</td>
<td>4507.18</td>
</tr>
<tr>
<td>ReLEX ( {\phi_1, \phi_2} )</td>
<td>0.9051</td>
<td>2865.86</td>
</tr>
<tr>
<td>ReLEX ( {\phi_1, \phi_3} )</td>
<td>0.9057</td>
<td>2606.99</td>
</tr>
<tr>
<td>ReLEX ( {\phi_2, \phi_3} )</td>
<td>0.90648</td>
<td>2702.94</td>
</tr>
<tr>
<td>ReLEX ( {\phi_1, \phi_2, \phi_3} )</td>
<td><strong>0.9080</strong></td>
<td><strong>2093.32</strong></td>
</tr>
</tbody>
</table>

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 \( \hat{M}_h^k \) is close to the target \( M_h^k \). This definition was originally introduced in (Yang and Wang [2020]).

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

\[
\{ \text{tr} \left( [M_{h,\phi} - M_{h,\phi}^\ast ]^\top U_{h,\phi}^j (M_{h,\phi} - M_{h,\phi}^\ast ) \right) \leq \beta_{j,\phi}, \forall j, k, h \in [H] \} =: \mathcal{E}_k^{\phi}.
\]

With \( \beta_{k,\phi} = c(C_M + C_H^2)d_\phi \log(kHC^2/\delta) \) for some absolute constant \( c > 0 \), we have \( \Pr(\mathcal{E}_k^{\phi}) \geq 1 - \delta \) for all \( \phi \in \Phi \).

**Remark 3.2.** The proof of Lemma 3.1 remains the same since the regression does not depend on the policy \( \pi \). We also make the dependency of \( \delta \) explicit in \( \beta_{k,\phi} \), 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 \( \beta_{k,\phi} \) as in Theorem 4.5. Then suppose \( \mathcal{E}_k^{\phi} \) holds for all \( \phi \in \Phi \). Then with probability at least \( 1 - 3\delta \), the regret for the very first \( k \in [K] \) episodes is controlled by

\[
\text{Regret}(k) \leq \min_{\phi \in \Phi} \left\{ \frac{128C_\phi^2 H^3 d_\phi \beta_{k,\phi} \log(1 + C_\phi kd_\phi)}{\text{gap}_{\min}} + \frac{16}{3} H^2 \log((1 + \log(Hk))k^2|\Phi|/\delta) + 2 + \frac{96H^4 \log(2k(1 + \log(H/\text{gap}_{\min})))|\Phi|/\delta}{\text{gap}_{\min}} \right\}.
\]

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

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

It is easy to verify that when there is only one representation function in \( \Phi \) (let \( d = d_\phi \) for simplicity), Lemma 3.3 yields an \( O(H^5d^2 \log(k/\delta)\text{gap}_{\min}^{-1}) \) problem-dependent bound. Comparing our result with He et al. [2021], ours matches the problem-dependent bound for Linear Mixture MDP \( O(H^5d^2 \log(k/\delta)\text{gap}_{\min}^{-1}) \) and is better than the problem-dependent bound for Linear MDP \( O(H^5d^3 \log(k/\delta)\text{gap}_{\min}^{-1}) \) 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 \( O(H^4d^2 \log(k/\delta)\text{gap}_{\min}^{-1}) \) bound for the summation of the sub-optimality gap. Based on this, the next lemma shows that the “covariance matrix” \( U_{h,\phi}^k \) 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 \( k \in [K], h \in [H], \phi \in \Phi \):

\[
U_{h,\phi}^k \geq (k - 1)\Lambda_{h,\phi} - \lambda I_{h,\phi},
\]

where

\[
\lambda = \frac{C_\phi d_\phi}{\text{gap}_{\min}} \sum_{i=1}^{h} \sum_{j=1}^{k-1} \text{gap}_{\psi}(s_i^j, a_i^j) - 1 + C_\phi d_\phi \sqrt{32H(k - 1)\log(d_\phi|\Phi|Hk(k + 1)/\delta)}.
\]
where the last equation is due to the fact that $k > k^*$. Thus there exists a threshold $\tau$ such that for all $h \in [H]$, $s \in S$ we have

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

Lemma 3.5 suggests that when the episode number $k$ exceeds $\tau$, policy $\pi^k$ will contribute a sub-optimality up to $O(\sqrt{1/k})$. Thus there exists a threshold $k^*$ such that when $k \geq k^*$, the policy $\pi^k$ 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 $\beta_{k, \phi} = c(C_M + C_{\phi}^2) d_\phi \log(kHC_\phi |\Phi|/\delta)$ to make sure with probability at least $1 - \delta$, event $\mathcal{E}_\phi^K$ holds for all $\phi \in \Phi$. 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

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

Then we have for all $\phi \in \Phi$,

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

Thus, by Lemma 3.5 we have

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

and it implies $\text{gap}_h(s, \pi^k_h(s)) > \text{gap}_{\min}$ thus $\text{gap}_h(s, \pi^k_h(s)) = 0$. Since from the parameter setting, $\beta_{k, \phi} = O(\log(k))$, it is easy to verify that there exists a threshold $k^*$ satisfying (3.3). Thus we conclude that $\text{gap}_h(s, \pi^k_h(s)) = 0$ for all $k \geq k^*$. Since the optimal policy $\pi^*$ is unique, it follows that $\pi^k = \pi^*$. Thus when $k \geq k^*$, the regret could be decomposed by

$$\text{Regret}(k) = \sum_{j=1}^{k} V^*_1(s^j_1) - V^\pi_1(s^j_1)$$

$$= \sum_{j=1}^{k^*} V^*_1(s^j_1) - V^\pi_1(s^j_1) + \sum_{j=1+k^*}^{k} V^*_1(s^j_1) - V^\pi_1(s^j_1)$$

$$= \text{Regret}(k^*) + 0,$$

where the last equation is due to the fact that $\pi^j = \pi^*$ thus $V^*_1(s) = V^\pi_1(s)$ for all $s \in S$ 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 $\bar{k} = \min\{k, k^*\}$, by Lemma 3.3
we have the regret is bounded by

\[
\text{Regret}(k) \leq \min_{\phi \in \Phi} \left\{ \frac{128C_\phi^2 H^5 d_\phi^2 (C_M + C'_\phi)^2}{\text{gap}_\text{min}} \log \left( 1 + C_\phi \text{gap}_\phi \right) \log \left( \frac{1}{\delta} \right) \right\} + 2 \\
+ \frac{96H^4 \log \left( 2k \left( 1 + \log \left( \frac{H}{\text{gap}_\text{min}} \right) \right) \right)}{\text{gap}_\text{min}} \\
+ \frac{16}{3} H^2 \log \left( \left( 1 + \log \left( \frac{H}{k} \right) \right)^2 \frac{1}{\delta} \right),
\]

with probability at least \(1 - 5\delta\) by taking the union bound of Lemma 3.3, Lemma 3.4 and \(E^K_\phi\) holds.

\[\square\]

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

\[
\mathcal{F}^k_h = \left\{ \{s_i^h, a_i^h\}_{i=1}^{H-k}, \{s_i^k, a_i^k\}_{i=1}^{H} \right\}.
\]

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

Arranging the filtrations as

\[
\mathcal{F} = \{ \mathcal{F}_1^1, \cdots, \mathcal{F}_H^1, \cdots, \mathcal{F}_1^k, \cdots, \mathcal{F}_1^1, \cdots, \mathcal{F}_1^2, \cdots, \mathcal{F}_1^K, \cdots, \mathcal{F}_H^K \},
\]

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 \(Q^k_{h, \phi}\) is always optimistic.

Lemma 4.1. Suppose the event \(E^K_\phi\) holds for all \(\phi \in \Phi\), then for any \((s, a) \in S \times A\), \(Q^*_{h}(s, a) \leq Q^k_{h, \phi}(s, a)\).

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 \(E^K_\phi\) holds, then for any \((s, a) \in S \times A\), \(k \in [K]\) and any policy \(\pi\),

\[
Q^k_h(s, a) - Q^*_{h}(s, a) \leq 2C_\phi H \sqrt{\beta_{k, \phi}(s, a)(U_{h, \phi}^k)^{-1} \phi(s, a) + [\mathbb{P}_h(V_{h+1}^k + V^\tau_{h+1})](s, a)}.
\]

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 \(E^K_\phi\) holds for all \(\phi \in \Phi\), then with probability at least \(1 - \delta\), for any \(k \in [K]\),

\[
\sum_{j=1}^{k} 1[V^*_h(s^j_h) - Q^*_h(s^j_h, a^j_h) \geq \Delta] \leq \frac{16C_\phi H^4 d_\phi \beta_{k, \phi} \log(1 + C_\phi k d_\phi) + 12H^3 \log(2k/\delta)}{\Delta^2}.
\]

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

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

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 $1 - \lceil \log(Hk) \rceil \exp(-\tau)$, we have

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

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

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

By relaxing $[\log(Hk)]$ to $\log(HK)$ and replacing $\delta/k^2$ 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 $\text{gap}_{\text{min}}$, for each $h \in [H], k \in [K]$, we have $V^*_h(s^k_h) - Q^*_h(s^k_h, a^k_h) = 0$ or $\text{gap}_{\text{min}} \leq V^*_h(s^k_h) - Q^*_h(s^k_h, a^k_h) \leq H$. Dividing the interval $[\text{gap}_{\text{min}}, \log(HK)]$ into $N$ intervals $[2^{n-1}\text{gap}_{\text{min}}, 2^n\text{gap}_{\text{min}}]$ where $n \in [N], N = \lceil \log(H/\text{gap}_{\text{min}}) \rceil$, then with probability at least $1 - \lceil \log(H/\text{gap}_{\text{min}}) \rceil \delta$, it holds that

$$\sum_{j=1}^{k} \left(V^*_h(s^j_h) - Q^*_h(s^j_h, a^j_h)\right) \leq \sum_{n=1}^{N} \sum_{j=1}^{k} 2^n \text{gap}_{\text{min}} 1[2^{n-1}\text{gap}_{\text{min}} \leq V^*_h(s^j_h) - Q^*_h(s^j_h, a^j_h) \leq 2^n\text{gap}_{\text{min}}]$$

$$\leq \sum_{n=1}^{N} 2^n \text{gap}_{\text{min}} 1[2^{n-1}\text{gap}_{\text{min}} \leq V^*_h(s^j_h) - Q^*_h(s^j_h, a^j_h) \leq 2^n\text{gap}_{\text{min}}]$$

$$\leq \sum_{n=1}^{N} 2^{n+1} \text{gap}_{\text{min}} 1[2^{n-1}\text{gap}_{\text{min}} \leq V^*_h(s^j_h) - Q^*_h(s^j_h, a^j_h) \leq 2^n\text{gap}_{\text{min}}]$$

$$\leq \sum_{n=1}^{N} 2^{n+1} \text{gap}_{\text{min}} 1[2^{n-1}\text{gap}_{\text{min}} \leq V^*_h(s^j_h) - Q^*_h(s^j_h, a^j_h) \leq 2^n\text{gap}_{\text{min}}]$$

$$\leq \sum_{n=1}^{N} 2^{n+1} \text{gap}_{\text{min}} 1[2^{n-1}\text{gap}_{\text{min}} \leq V^*_h(s^j_h) - Q^*_h(s^j_h, a^j_h) \leq 2^n\text{gap}_{\text{min}}]$$

$$\leq \sum_{n=1}^{N} 2^{n+1} \text{gap}_{\text{min}} 1[2^{n-1}\text{gap}_{\text{min}} \leq V^*_h(s^j_h) - Q^*_h(s^j_h, a^j_h) \leq 2^n\text{gap}_{\text{min}}]$$

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 $Q^*_h(s, a) \geq Q^*_h(s, a)$ 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 $\delta$ with $\delta/(1 + \log(H/\text{gap}_{\text{min}}))$, with probability at least $1 - \delta$, we have

$$\sum_{j=1}^{k} \left(V^*_h(s^j_h) - Q^*_h(s^j_h, a^j_h)\right) \leq \frac{64C^2 \psi H^4 d_{\phi, k, \phi} \log(1 + C_{\phi} kd_{\phi}) + 48H^3 \log(2k/\delta)}{\text{gap}_{\text{min}}}$$

(4.1)

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

$$\text{Regret}(k) \leq 2 \sum_{j=1}^{k} \sum_{h=1}^{H} \text{gap}_h(s^k_h, a^k_h) + \frac{16H^2 \log([HK]/\delta)}{3} + 2$$

$$\leq \frac{128C^2 \psi H^5 d_{\phi, k, \phi} \log(1 + C_{\phi} kd_{\phi}) + 96H^4 \log(2k(1 + \log(H/\text{gap}_{\text{min}})) / \delta)}{\text{gap}_{\text{min}}}$$

$$+ \frac{16}{3} H^2 \log((1 + \log(Hk))k^2/\delta) + 2.$$

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 $\delta$ with $\delta/|\Phi|$, the claimed result (3.1) holds for all $\phi \in \Phi$ by taking a union bound.

$\Box$
4.3 PROOF OF LEMMA 3.4

For brevity, we denote matrix $\Lambda_{h,\phi}(s) = \phi(s,\pi^*_i(s))\Phi^T(s,\pi^*_i(s)) \in \mathbb{R}^{d \times d}$ and fix $h$, $m$ in the proof. The expectation $\mathbb{E}_{s_h} \left[ \Lambda_{h,\phi}(s_h) | s_1 \right]$, $i < h$ is taken with respect to the randomness of the states sequence $s_{i+1}, \cdots, s_h$, where $s_{i'+1} \sim \mathbb{P}_{i'} \left[ \cdot | s_i', \pi^*_i(s_i') \right]$, $i \leq i' < h$. If the action $a_i$ is given, the expectation $\mathbb{E}_{s_h} \left[ \Lambda_{h,\phi}(s_h) | s_i \right]$, $i < h$ is taken in which $s_{i+1} \sim \mathbb{P}_i \left[ \cdot | s_i, a_i \right]$ specially. It is worth to show that $\mathbb{E}_{s_h} \left[ \Lambda_{h,\phi}(s_h) | s_h \right] = \Lambda_{h,\phi}(s_h^k)$. 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 $\phi(s, a)\Phi^T(s, a)$, we introduce the following matrix Azuma inequality.

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

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

where $C = \|C\|_2$.

Equipped with this lemma, we can start our proof.

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

$$\phi(s_i^k, a_i^k)\Phi^T(s_i^k, a_i^k) = \Lambda_{h,\phi}(s_i^k) - \mathbb{I} \left[ a_i^k \neq \pi^*_i(s_i^k) \right] \left( \Lambda_{h,\phi}(s_i^k) - \phi(s_i^k, a_i^k)\Phi^T(s_i^k, a_i^k) \right) \geq \Lambda_{h,\phi}(s_i^k) - C_\phi \mathbb{I} \left[ a_i^k \neq \pi^*_i(s_i^k) \right] I_d, \quad (4.2)$$

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

$$E[\Lambda_{h,\phi}(s_h) | s_i^k] = E[\Lambda_{h,\phi}(s_h) | s_{i-1}^k, a_{i-1}^k] - E[\Lambda_{h,\phi}(s_h) | s_{i-1}^k, a_{i-1}^k] - E[\Lambda_{h,\phi}(s_h) | s_i^k]$$

where we denote the first term as $A_i(s_{i-1}^k, a_{i-1}^k)$ while the second term as $e_i^k$ for simplicity. We first consider the term $e_i^k$, it is easy to verify that $e_i^k$ is $F_i^k$-measurable, $d \times d$ symmetric matrix with $E[e_i^k | F_{i-1}^k] = 0$ and

$$\|e_i^k\|_2 \leq \|E[\Lambda_{h,\phi}(s_h) | s_{i-1}^k, a_{i-1}^k]\|_2 + \|E[\Lambda_{h,\phi}(s_h) | s_i^k]\|_2 \leq 2C_\phi d^2 = (4.3)$$

where the inequality holds due to the fact that $\|\Lambda_{h,\phi}(s)\|_2 \leq C_\phi d$. Next, for the term $A_i$, by introducing the indicator showing whether the action $a_{i-1}^k$ is the optimal action $\pi^*_i(s_{i-1}^k)$, we proceed as follows:

$$A_i(s_{i-1}^k, a_{i-1}^k) = \Lambda_i(s_{i-1}^k, \pi^*_i(s_{i-1}^k)) - \mathbb{I} \left[ a_{i-1}^k \neq \pi^*_i(s_{i-1}^k) \right] \left( \Lambda_i(s_{i-1}^k, \pi^*_i(s_{i-1}^k)) - \Lambda_i(s_{i-1}^k, a_{i-1}^k) \right) \geq \Lambda_i(s_{i-1}^k, \pi^*_i(s_{i-1}^k)) - C_\phi \mathbb{I} \left[ a_{i-1}^k \neq \pi^*_i(s_{i-1}^k) \right] I_d$$

$$= E[\Lambda_{h,\phi}(s_h) | s_i^k] - C_\phi \mathbb{I} \left[ a_{i-1}^k \neq \pi^*_i(s_{i-1}^k) \right] I_d, \quad (4.4)$$

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

$$E[\Lambda_{h,\phi}(s_h) | s_i^k, \pi^*_i(s_{i-1}^k)] = E[\Lambda_{h,\phi}(s_h) | s_{i-1}^k].$$

Combining (4.4) and (4.3) together yields

$$E[\Lambda_{h,\phi}(s_h) | s_i^k] \geq E[\Lambda_{h,\phi}(s_h) | s_i^k] - C_\phi \mathbb{I} \left[ a_{i-1}^k \neq \pi^*_i(s_{i-1}^k) \right] I_d - e_i^k.$$
and by telescoping over $i$ we have

$$E[\mathbf{A}_{h,\phi}(s_h)|s_h^k] \geq E[\mathbf{A}_{h,\phi}(s_h)|s_h^1] - \sum_{i=2}^{h} e_i^k - C_{\phi}d_{\phi}\sum_{i=1}^{h-1} \mathbb{I}[a_i^k \neq \pi_i^*(s_i^k)] \mathbf{I}_{d_{\phi}},$$

(4.5)

where $e_i^k$ is $\mathcal{F}_i^k$-measurable and $E[e_i^k|\mathcal{F}_i^h] = 0$, $\|e_i^k\|_2 \leq 2C_{\phi}d_{\phi}$, which is similar to $e_i^k$ above. Plugging (4.5) into (4.2) yields

$$\phi(s_h^k, a_h^k)\mathbf{\phi}^T(s_h^k, a_h^k) \geq E[\mathbf{A}_{h,\phi}(s_h)] - \sum_{i=1}^{h} e_i^k - C_{\phi}d_{\phi}\sum_{i=1}^{h} \mathbb{I}[a_i^k \neq \pi_i^*(s_i^k)] \mathbf{I}_{d_{\phi}},$$

(4.6)

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

By the construction of the “covariance matrix” $\mathbf{U}_{h,\phi}$, (4.6) yields

$$\mathbf{U}_{h,\phi}^k = \mathbf{I}_{d_{\phi}} + \sum_{j=1}^{k-1} \phi(s_h^j, a_h^j)\phi^T(s_h^j, a_h^j) \geq \mathbf{I}_{d_{\phi}} + (k-1)\mathbf{A}_{h,\phi} - \sum_{j=1}^{k-1} \sum_{i=1}^{h} e_i^j - \frac{C_{\phi}d_{\phi}}{\text{gap}_{\min}}\sum_{i=1}^{h} \sum_{j=1}^{k-1} (V_i^*(s_i^j) - Q_i^*(s_i^j, a_i^j))\mathbf{I}_{d_{\phi}}.$$  

(4.7)

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

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

(4.8)

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

\[\square\]

### 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 $\mathcal{E}_{\phi}^K$ holds for all $\phi \in \Phi$. For each $\phi \in \Phi$, there exists a constant threshold

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

such that for any $(s,a) \in S \times A, h \in [H]$, there exists a representation candidate $\phi \in \Phi$ where when $k \geq \tau_{\phi}$, $\mathbf{U}_{h,\phi}^{-1}\phi(s,a) \leq 2C_{\phi}d_{\phi}/(\sigma_{\phi}^k)$. We denote $\tau = \max_{\phi \in \Phi} \tau_{\phi}$ to be the maximum possible threshold over all representations.
Lemma 4.7 suggests that the UCB bonus term is decaying in the rate of $O(1/\sqrt{k})$. 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, $c_{\phi}^k$ holds for all $\phi \in \Phi$. Considering $h = H$, for any state-action pair $(s,a) \in S \times A$, by Lemma 4.7 when $k \geq \tau$, there exists a representation $\phi$ where Lemma 4.2 yields

$$Q^k_H(s,a) - Q^\pi_H(s,a) \leq 2C_{\phi}H \sqrt{\beta_{k,\phi} \Phi(s,a)(U_{H,\phi}^k)^{-1} \phi(s,a) + \{P_H(V^k_{H+1} - V^\pi_{H+1})\}(s,a)}$$

where the second inequality is due to Lemma 4.7 and the fact that $V^k_{H+1}, V^\pi_{H+1}$ are both equal to zero. Thus we have

$$\max_{(s,a) \in S \times A} \{Q^k_H(s,a) - Q^\pi_H(s,a)\} \leq \max_{\phi \in \Phi} \left\{2C_{\phi}H \sqrt{2C_{\phi}d_{k,\phi}/(\sigma_{\phi}k)}\right\}. \quad (4.9)$$

Suppose for step $h$, we have

$$\max_{(s,a) \in S \times A} \{Q^k_h(s,a) - Q^\pi_h(s,a)\} \leq (H - h + 1) \max_{\phi \in \Phi} \left\{2C_{\phi}H \sqrt{2C_{\phi}d_{k,\phi}/(\sigma_{\phi}k)}\right\},$$

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

$$Q^k_{h-1}(s,a) - Q^\pi_{h-1}(s,a)$$

$$\leq 2C_{\phi}H \sqrt{\beta_{k,\phi} \Phi(s,a)(U_{H-1,\phi}^k)^{-1} \phi(s,a) + \{P_{H-1}(V^k_{h-1} - V^\pi_{h-1})\}(s,a)}$$

$$\leq 2C_{\phi}H \sqrt{2C_{\phi}d_{k,\phi}/(\sigma_{\phi}k) + \{P_{H-1}(Q^k_{h-1}(\cdot, \pi^k_h(\cdot)) - Q^\pi_{h-1}(\cdot, \pi^\pi_h(\cdot)))\}(s,a)}$$

$$\leq 2C_{\phi}H \sqrt{2C_{\phi}d_{k,\phi}/(\sigma_{\phi}k) + (H - h + 1) \max_{\phi \in \Phi} \left\{2C_{\phi}H \sqrt{2C_{\phi}d_{k,\phi}/(\sigma_{\phi}k)}\right\}}$$

$$\leq (H - h + 2) \max_{\phi \in \Phi} \left\{2C_{\phi}H \sqrt{2C_{\phi}d_{k,\phi}/(\sigma_{\phi}k)}\right\}.$$

where the second inequality follows from the definition that $V^k_{h}(s) = Q^k_{h}(s, \pi^k_h(s))$ and $V^\pi_{h}(s) = Q^\pi_{h}(s, \pi^\pi_h(s))$, the third inequality is due to the induction assumption (4.9) and this result conclude our induction.

Then, following Lemma 4.1 we have $Q^k_{h,\phi}(s,a) \geq Q^\pi_h(s,a)$. Thus, $Q^k_h(s,a) = \min_{\phi \in \Phi} \{Q^k_{h,\phi}\} \geq Q^\pi_h(s,a)$. Then the sub-optimality gap could be bounded by

$$\text{gap}_h(s, \pi^k_h(s)) = Q^\pi_h(s, \pi^\pi_h(s)) - Q^\pi_h(s, \pi^k_h(s))$$

$$\leq Q^k_h(s, \pi^k_h(s)) - Q^\pi_h(s, \pi^k_h(s))$$

$$\leq Q^k_h(s, \pi^k_h(s)) - Q^\pi_h(s, \pi^\pi_h(s))$$

$$\leq (H - h + 1) \max_{\phi \in \Phi} \left\{2C_{\phi}H \sqrt{2C_{\phi}d_{k,\phi}/(\sigma_{\phi}k)}\right\}$$

$$\leq 2C_{\phi}H^2 \max_{\phi \in \Phi} \left\{\sqrt{2C_{\phi}d_{k,\phi}/(\sigma_{\phi}k)}\right\},$$

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 $\pi^k_h(s) = \arg\max_a Q^k_h(s, a)$. Finally, the inequality on the forth line is due to the result of induction (4.9) and we finish the proof. \qed
5 PROOF OF LEMMAS IN APPENDIX 4

5.1 PROOF OF LEMMA 4.1

Lemma 5.1 (Lemma 5 on 2020, Yang and Wang). Suppose $C_{\Phi}^k$ holds, then for any $(s, a) \in S \times A$, we have

$$\|\phi(s, a)^{\top}(M_{h, \phi}^k - M_{h, \phi}^*)\|_2 \leq \sqrt{\beta_{k, \phi} \phi^\top(s, a)(U_{h, \phi}^k)^{-1} \phi(s, a)}.$$  \hspace{1cm} (5.1)

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

$$Q_{h, \phi}^k(s, a) = r(s, a) + \phi^\top(s, a)M_{h, \phi}^k \Psi^\top v_{h+1}^k + C_{\phi}H \sqrt{\beta_{k, \phi} \phi^\top(s, a)(U_{h, \phi}^k)^{-1} \phi(s, a)}$$

$$= r(s, a) + \phi^\top(s, a)(M_{h, \phi}^k - M_{h, \phi}^*) \Psi^\top v_{h+1}^k$$

$$+ C_{\phi}H \sqrt{\beta_{k, \phi} \phi^\top(s, a)(U_{h, \phi}^k)^{-1} \phi(s, a) + [P_h V_{h+1}^k](s, a)}$$

$$\geq r(s, a) - \|\phi^\top(s, a)(M_{h, \phi}^k - M_{h, \phi}^*)\|_2 \|\Psi^\top v_{h+1}^k\|_2$$

$$+ C_{\phi}H \sqrt{\beta_{k, \phi} \phi^\top(s, a)(U_{h, \phi}^k)^{-1} \phi(s, a) + [P_h V_{h+1}^k](s, a)}$$

$$\geq r(s, a) + [P_h V_{h+1}^k](s, a),$$  \hspace{1cm} (5.2)

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

$$V_{h+1}^k(s) = \min\{H, Q_{h+1}^k(s, \pi_h^k(s))\}$$

$$\geq \min\{H, Q_{h+1}^k(s, \pi_{h+1}^* (s))\}$$

$$\geq \min\{H, Q_{h+1}^k(s, \pi_{h+1}^* (s))\}$$

$$= V_{h+1}^k(s),$$

where the last inequality is due to the fact that $V_{h+1}^k(s) \leq H$. Therefore, (5.1) yields $Q_{h, \phi}^k(s, a) \geq r(s, a) + [P_h V_{h+1}^k](s, a) = Q_{h}^k(s, a) \geq Q_{h}^k(s, a).$

Then we finish our proof by induction. \hspace{1cm} \square

5.2 PROOF OF LEMMA 4.2

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

$$Q_{h, \phi}^k(s, a) - Q_{h}^k(s, a) = \phi^\top(s, a)M_{h, \phi}^k \Psi^\top v_{h+1}^k - [P_h V_{h+1}^\pi](s, a)$$

$$+ C_{\phi}H \sqrt{\beta_{k, \phi} \phi^\top(s, a)(U_{h, \phi}^k)^{-1} \phi(s, a)}.$$  \hspace{1cm} (5.2)

Since $[P_h V_{h+1}^k] = \phi^\top(s, a)M_{h, \phi}^k \Psi^\top v_{h+1}^k$, $I_1$ can be decomposed as

$$\phi^\top(s, a)M_{h, \phi}^k \Psi^\top v_{h+1}^k = \phi^\top(s, a)(M_{h, \phi}^k - M_{h, \phi}^*) \Psi^\top v_{h+1}^k + [P_h V_{h+1}^k](s, a)$$

$$\leq \|\Psi^\top v_{h+1}^k\|_2 \|\phi^\top(s, a)(M_{h, \phi}^k - M_{h, \phi}^*)\|_2 + [P_h V_{h+1}^k](s, a)$$

$$\leq C_{\phi}H \sqrt{\beta_{k, \phi} \phi^\top(s, a)(U_{h, \phi}^k)^{-1} \phi(s, a) + [P_h V_{h+1}^k](s, a)},$$  \hspace{1cm} (5.3)
where the inequality on the second line holds due to \( \|x, y\| \leq \|x\|_2 \|y\|_2 \) and the inequality on the third line comes from Lemma 5.1 with \( \|V_h^{k+1}\|_\infty \leq H \) and Definition 3.2. Plugging (5.3) into (5.2) yields

\[
Q_h^k(s, a) - Q_h^\pi(s, a) \leq 2C_{\phi}H \sqrt{\beta_{k, \phi}(s,a)}(U_{h,\phi})^{-1}\phi(s,a) + |\mathbb{P}_h(V_h^k - V_h^\pi)(s,a)|.
\]

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

\[\square\]

### 5.3 Proof of Lemmas 4.3

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

\[
\sum_{i=1}^k \phi_h^i(s_n^i, a_n^i)(U_{h,\phi})^{-1}\phi_h(s_n^i, a_n^i) \leq 2d_\phi \log(1 + C_{\phi}kd_\phi).
\]

**Remark 5.3.** Proof of Lemma 5.2 remains the same as He et al. [2021] by changing the norm of \( \phi \) from \( \|\phi\|^2 \leq 1 \) to \( \|\phi\|^2 \leq C_{\phi}d_\phi \), as defined in 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 \( \{\mathcal{F}_t\}_{t=1}^n \) (i.e. \( \mathbb{E}[x_t|\mathcal{F}_t] = 0 \) a.s. and \( x_t \) is \( \mathcal{F}_{t+1} \) measurable) such that \( |x_t| \leq \mathcal{M} \) a.s. Then for any \( 0 < \delta < 1 \), with probability at least \( 1-\delta \),

\[
\sum_{i=1}^n x_i \leq \mathcal{M}\sqrt{2n\log(1/\delta)}.
\]

**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 \( \Delta \):

\[
k_j = \min\left\{ \bar{k} : \bar{k} > k_{j-1}, V_h^\star(s_{\bar{k}}^j) - Q_h^\pi(s_{\bar{k}}^j, a_{\bar{k}}^j) \geq \Delta \right\}.
\]

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 \( \Delta \), i.e.

\[
k' = \sum_{j=1}^k 1[V_h^\star(s_{\bar{k}}^j) - Q_h^\pi(s_{\bar{k}}^j, a_{\bar{k}}^j) \geq \Delta] = \sum_{j=1}^{k'} Q_h^\pi(s_{\bar{k}}^j, a_{\bar{k}}^j) \geq \Delta k'.
\]

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

\[
\sum_{j=1}^{k'} Q_h^\pi(s_{\bar{k}}^j, a_{\bar{k}}^j) \geq \sum_{j=1}^{k'} Q_h^\pi(s_{\bar{k}}^j, a_{\bar{k}}^j) - Q_h^\pi(s_{\bar{k}}^j, a_{\bar{k}}^j)
\]

\[
\geq \sum_{j=1}^{k'} Q_h^\pi(s_{\bar{k}}^j, a_{\bar{k}}^j) - Q_h^\pi(s_{\bar{k}}^j, a_{\bar{k}}^j)
\]

\[
= \sum_{j=1}^{k'} V_h^\star(s_{\bar{k}}^j) - Q_h^\pi(s_{\bar{k}}^j, a_{\bar{k}}^j) \geq \Delta k',
\]

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

\[
Q_i^\star(s_i^j, a_i^j) \leq 2C_{\phi}H \sqrt{\beta_{j,\phi}(s_i^j, a_i^j)(U_{j,\phi})^{-1}\phi(s_i^j, a_i^j)} + |\mathbb{P}_i(V_i^j - V_i^\pi)(s_i^j, a_i^j)|
\]

\[
= 2C_{\phi}H \sqrt{\beta_{j,\phi}(s_i^j, a_i^j)(U_{j,\phi})^{-1}\phi(s_i^j, a_i^j)} + V_i^j - V_i^\pi(s_i^j, a_i^j) + c_i^j,
\]

where \( c_i^j = \mathbb{P}_i(V_i^j - V_i^\pi)(s_i^j, a_i^j) - (V_{i+1}^j(s_{i+1}^j) - V_{i+1}^\pi(s_{i+1}^j)) \). It is easy to verify that \( |c_i^j| \leq \mathcal{M} \) when \( h \leq i \leq H \).

Taking the telescoping summation on (5.5) over \( h \leq i \leq H \), \( j \in \{k_1, \ldots, k_{k'}\} \) using the fact that \( V_i^\star(s_i^j, a_i^j) = Q_i^\star(s_i^j, a_i^j) \) and \( V_i^\pi(s_i^j) = Q_i^\pi(s_i^j, a_i^j) \) we have

\[
\sum_{j=1}^{k'} Q_h^\pi(s_{\bar{k}}^j, a_{\bar{k}}^j) \leq Q_h^\pi(s_{\bar{k}}^j, a_{\bar{k}}^j) + I_1 + I_2,
\]

where
where

$$I_1 = \sum_{j=1}^{k'} \sum_{i=h}^{H} 2C_{\phi} H \beta_{k_j, \phi} \phi^T (s_{k_j}^i, a_{k_j}^i) (U_{i, \phi}^{k_j})^{-1} \phi(s_{k_j}^i, a_{k_j}^i)$$

$$I_2 = \sum_{j=1}^{k'} \sum_{i=h}^{H} \epsilon_i^{k_j}.$$

To bound $I_1$, by Cauchy-Schwarz inequality,

$$I_1 = \sum_{j=1}^{k'} \sum_{i=h}^{H} 2C_{\phi} H \beta_{k_j, \phi} \phi^T (s_{k_j}^i, a_{k_j}^i) (U_{i, \phi}^{k_j})^{-1} \phi(s_{k_j}^i, a_{k_j}^i)$$

$$\leq 2C_{\phi} H \beta_{k, \phi} \sum_{i=h}^{H} \sum_{j=1}^{k'} \phi^T (s_{k_j}^i, a_{k_j}^i) (U_{i, \phi}^{k_j})^{-1} \phi(s_{k_j}^i, a_{k_j}^i)$$

$$\leq 2C_{\phi} H^2 \sqrt{\beta_{k, \phi} k'} \sqrt{2d\phi \log(1 + C_{\phi} d)}$$

$$\leq 2C_{\phi} H^2 \sqrt{2\beta_{k, \phi} d \phi k'} \log(1 + C_{\phi} kd),$$

where the second inequality 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

$$\Delta k' \leq 2C_{\phi} H^2 \sqrt{2\beta_{k, \phi} d \phi k'} \log(1 + C_{\phi} kd) + \sqrt{2k' H^3 \log(k/\delta)}.$$  \hspace{1cm} (5.7)

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

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

Since event $\mathcal{E}_k^K$ directly implies $\mathcal{E}_k$ for all $k \leq K$, we can get the claimed result (5.8) holds for all $k \leq K$ with probability $1 - \delta$. Replace $\delta$ 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 $k$.

$$\sum_{j=1}^{k} 1[ V_h^*(s_{k_j}^i) - Q_h^*(s_{k_j}^i, a_{k_j}^i) ] \leq \frac{16 C_{\phi}^2 H^4 d \phi \beta_{k, \phi} \log(1 + C_{\phi} kd) + 4H^3 \log(k^2(k + 1)/\delta)}{\Delta^2}$$

$$\leq \frac{16 C_{\phi}^2 H^4 d \phi \beta_{k, \phi} \log(1 + C_{\phi} kd) + 12H^3 \log(2k/\delta)}{\Delta^2}.$$
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 \(Z_{h, \phi}\) where \((s, a) \in Z_{h, \phi}\) and the corresponding representation \(\phi\). By Lemma 3.4, we denote \(B\) as

\[
B := (k - 1)A_{h, \phi} - \mathbf{l}_{d_{\phi}} = U_{k, \phi}^h.
\]

Decomposing \(A_{h, \phi} = Q^T DQ\) where \(Q \in \mathbb{R}^{d_{\phi} \times d_{\phi}}\) is the orthogonal matrix and \(D\) is the diagonal matrix, we have

\[
B = Q^T((k - 1)D - \mathbf{l}_{d_{\phi}})Q.
\]

We first prove the non-singular property of \(B\). Considering the zero diagonal element \(D_{[ii]}\), we have

\[
((k - 1)D - \mathbf{l}_{d_{\phi}})_{[ii]} = -\mathbf{l} - C_{\phi}d_{\phi}\sqrt{32H(k - 1)\log(d_{\phi}/\Phi H(k + 1)/\delta)} + 1,
\]

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 \(k \geq K_1\), \(((k - 1)D - \mathbf{l}_{d_{\phi}})_{[ii]} < 0\) for all zero diagonal element \(D_{[ii]}\) in \(D\). Next we consider the non-zero diagonal value \(D_{[jj]}\). By Assumption 4.2, \(D_{[jj]} \geq \sigma_{\phi}\). Therefore, the corresponding diagonal value \(((k - 1)D - \mathbf{l}_{d_{\phi}})_{[jj]}\) could be bounded by

\[
((k - 1)D - \mathbf{l}_{d_{\phi}})_{[jj]} \geq \sigma_{\phi}(k - 1) - \mathbf{l}.
\]

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

\[
((k - 1)D - \mathbf{l}_{d_{\phi}})_{[jj]} \geq 1 + \sigma_{\phi}(k - 1) - C_{\phi}d_{\phi}\sqrt{32H(k - 1)\log(d_{\phi}/\Phi H(k + 1)/\delta)} - \frac{64C_{\phi}^2H^2d_{\phi}^2\beta_{k, \phi}\log(1 + C_{\phi}k\delta_{\phi}) + 48H^3\log(2k + 1 + \log(H/\text{gap}_{\min})/\delta)}{\text{gap}_{\min}}.
\]

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

\[
\tau_{\phi} = \text{poly}(d_{\phi}, \sigma_{\phi}^{-1}, H, \log(\Phi/\delta), \text{gap}_{\min}, C_{\phi}, C_{\psi}, C_{\bar{M}}, C_{\Phi})
\]

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

By the definition of \(Z_{h, \phi}\) in Assumption 4.2, there exists a vector \(x \in \mathbb{R}^{d_{\phi}}\) such that \(Ax = \phi(s, a)/\|\phi(s, a)\|_2\). Since \(U_{k, \phi}^h \geq B\) and \(U_{k, \phi}^h, B\) are both invertible, it follows

\[
\phi^T(s, a)(U_{k, \phi}^h)^{-1}\phi(s, a) \leq \|\phi(s, a)\|_2^2 \frac{\phi^T(s, a)}{\|\phi(s, a)\|_2}B^{-1}\frac{\phi(s, a)}{\|\phi(s, a)\|_2},
\]

where \(I_1\) could be rewrote by

\[
I_1 = x^T\Lambda B^{-1}Ax
= x^TQ^TDDQ^T((k - 1)D - \mathbf{l}_{d_{\phi}})^{-1}QQ^TDQx
= x^TQ^TD((k - 1)D - \mathbf{l}_{d_{\phi}})^{-1}DQx.
\]

Since \(\|Ax\|_2^2 = 1\), it is easy to verify that \(x^TQ^TDQx = x^TAx = \|Ax\|_2^2 = 1\). We hereby denote \(y\) as \(DQx\) and we have \(\|y\|_2^2 = 1\). Furthermore, it is obvious that \(y_{[i]} = 0\) as long as \(D_{[ii]} = 0\). Therefore, \(\sum_{i=1}^{d_{\phi}}D_{[ii]}y_{[i]}^2 = 1\). Then plugging the notation of \(y\) into \((5.10)\) yields

\[
I_1 = y^T((k - 1)D - \mathbf{l}_{d_{\phi}})^{-1}y = \sum_{i=1, D_{[ii]} \neq 0}^{d_{\phi}} \frac{y_{[i]}^2}{((k - 1)D - \mathbf{l}_{d_{\phi}})_{[ii]}},
\]

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

$$\{ \text{tr} \left[ (M_{h, \phi} - M_{h, \phi})^T U_{h, \phi} (M_{h, \phi} - M_{h, \phi}^*) \right] \leq \beta_\phi, \forall h \in [H] \} =: \mathcal{E}_\phi^k.$$

With $\beta_\phi = C d_\phi \log(K H/\delta)$ for some absolute constant $C > 0$, we have $\Pr(\mathcal{E}_\phi^k) \geq 1 - \delta$ for all $\phi \in \Phi$.

**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 essentially the first part of Theorem 5.4 which provides an upper bound of the sub-optimality planned by Algorithm 2.

**Lemma 6.2.** Let $\beta$ set as Lemma 6.1. If the event $\mathcal{E}_{\phi}$ in Lemma 6.1 holds for all $\phi \in \Phi$, for all state $s \in S$ and $h \in [H]$, then the next lemma suggests that the uncertainty $\| \phi \|_{U^{-1}}$ is bounded by $\mathcal{O}(1/\sqrt{K})$ where the $K$ is the size of the offline dataset.

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

$$K > \frac{32 C_\phi^2 d_\phi^2 \log(H d_\phi \| \Phi \| / \delta)}{\sigma_{h, \phi}^2} \left( 1 + \frac{C_\phi^2 H^4 \beta_\phi \sigma_{h, \phi}}{4 \epsilon_{\min}^2 C_\phi^2 d_\phi \log(H d_\phi \| \Phi \| / \delta)} \right),$$

(6.1)

we have $\| \phi(s, a) \|_{U^{-1}_{h, \phi}} < \epsilon_{\min} / (2 H^2 C_\phi \sqrt{\sigma_{h, \phi}})$. Here $\sigma_{h, \phi}$ is the minimum non-zero eigen value of expected offline matrix $E_{d_\phi^2} [\phi \phi^T]$ 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 $\mathcal{E}_\phi$ 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.

$$K > \max_{\phi \in \Phi, h \in [H]} \left\{ \frac{32 C_\phi^2 d_\phi^2 \log(H d_\phi \| \Phi \| / \delta)}{\sigma_{h, \phi}^2} \left( 1 + \frac{C_\phi^2 H^4 \beta_\phi \sigma_{h, \phi}}{4 \epsilon_{\min}^2 C_\phi^2 d_\phi \log(H d_\phi \| \Phi \| / \delta)} \right) \right\}.$$

(6.2)

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

$$V^*_h(s) - V^*_h(s) \leq 2 C_\phi H \sum_{h' = h}^H \mathbb{E}_{\pi^*} \left[ \min_{\phi \in \Phi} \left\{ \sqrt{\beta_\phi} \| \phi(s, a) \|_{U_{h', \phi}^{-1}} \right\} \right] s_{h'} = s$$

$$< \sum_{h' = h}^H \mathbb{E}_{\pi^*} \left[ \frac{\epsilon_{\min}}{H} \right] s_{h'} = s$$

$$< \epsilon_{\min}.$$
On the other hand, since \( Q^\pi_h(s, \pi(s)) \geq V^\pi_h(s) \), by the definition of sub-optimality gap in Definition 3.2
\[
V^\pi_h(s) - V^\pi_h(s) \geq V^\pi_h(s) - Q^\pi_h(s, \pi(s)) \geq \mathbb{1}[\pi_h(s) \neq \pi^*_h(s)]\text{gap}_{\text{min}},
\] (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),
\[
\mathbb{1}[\pi_h(s) \neq \pi^*_h(s)]\text{gap}_{\text{min}} < \text{gap}_{\text{min}},
\]
which yields that \( \pi_h(s) = \pi^*_h(s) \). Applying this to all \((s, h) \in S \times [H]\) and replacing \( \delta \) with \( \delta/(2|\Phi|) \), we can get the claimed result in Theorem 5.4 by union bound. \( \square \)

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 Cai et al. [2020], Jin et al. [2021]

**Lemma 7.1 (Extended value difference)** Let \( \{\pi\}_h, \{\pi'\}_h \) by any two policies and let \( \{\hat{Q}\}_h \) be any estimated Q-function. For any \( h \in [H] \), define the estimated value function as \( V_h(s) = \hat{Q}_h(s, \pi_h(s)) \). For all \( s \in S \) we have
\[
\hat{V}_h(s) - V^\pi_h(s) = \sum_{h'=h}^H \mathbb{E}_{\pi'} \left[ \hat{Q}_{h'}(s_{h'}, \pi_{h'}(s_{h'})) - \hat{Q}_{h'}(s_{h'}, \pi^*_h(s_{h'})) \right] |s_h = s] + \sum_{h'=h}^H \mathbb{E}_\pi \left[ \hat{Q}_{h'}(s_{h'}, a_{h'}) - r(s_{h'}, a_{h'}) - [\mathbb{P}\hat{V}_{h'+1}](s_{h'}, a_{h'}) |s_h = s] \right.,
\]
where \( \mathbb{E}_{\pi'} \) is taken with respect to the trajectory generated by \( \pi' \) 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 \( h \). \( \square \)

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

**Lemma 7.2 (Lemma 5 on B^{(2)}_h, pp. 23, Yang and Wang [2020])** Suppose \( \mathbb{E}_\phi \) holds, then for any \((s, a) \in S \times A\), we have
\[
\|\phi(s, a)^\top (M_{h, \phi} - M_{h, \phi}^*) \|_2 \leq \sqrt{\beta_\phi \phi^\top (s, a) U_{h, \phi}^{-1} \phi^\top (s, a)}.
\]

**Proof.** The proof is similar with Yang and Wang [2020] by fixing \( k \) to \( K \). \( \square \)

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

**Proof.** First it is obvious that \( V^*_h(s) - V^\pi_h(s) = (V^*_h(s) - V_h(s)) - (V^\pi_h(s) - V_h(s)) \) where \( V_h \) is the estimated value function in Line 6 in Algorithm 2. Note that \( V(s) = Q(s, \pi(s)) \) where \( \pi \) 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
\[
V_h(s) - V^*_h(s) = \sum_{h'=h}^H \mathbb{E}_{\pi'} \left[ Q_{h'}(s_{h'}, \pi_{h'}(s_{h'})) - Q_{h'}(s_{h'}, \pi^*_h(s_{h'})) \right] |s_h = s] + \sum_{h'=h}^H \mathbb{E}_{\pi} \left[ Q_{h'}(s_{h'}, a_{h'}) - r(s_{h'}, a_{h'}) - [\mathbb{P}V_{h'+1}](s_{h'}, a_{h'}) |s_h = s] \right. \geq \sum_{h'=h}^H \mathbb{E}_{\pi} \left[ Q_{h'}(s_{h'}, a_{h'}) - r(s_{h'}, a_{h'}) - [\mathbb{P}V_{h'+1}](s_{h'}, a_{h'}) |s_h = s] \right.,
\] (7.1)
where the last inequality is due to the fact that we are executing the greedy policy i.e. \( \pi_h(s) = \arg\max_{\pi} Q_h(s, a) \) thus \( Q_h(s, \pi_h(s)) \geq Q_h(s, \pi_h(s)) \). Meanwhile, letting \( \pi = \pi' = \pi \), Lemma 7.1 suggests that

\[
V_h(s) - V_h^\pi(s) = \sum_{h' = h}^H E_{\pi} [Q_h(s_{h'}, a_{h'}) - r(s_{h'}, a_{h'}) - [P V_{h+1}](s_{h'}, a_{h'}) | s_h = s].
\] (7.2)

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

\[
Q_h(s, a) = r(s, a) - [P V_{h+1}](s, a)
\]

where the last inequality utilize the notation \( \Psi = (\psi(s_1), \psi(s_2), \ldots, \psi(s_{|S|}))^\top \in \mathbb{R}^{|S| 	imes d' \prime} \) and \( v_{h+1} = (V_{h+1}(s_1), V_{h+1}(s_2), \ldots, V_{h+1}(s_{|S|}))^\top \in \mathbb{R}^{|S|} \). For each \( \phi \in \Phi \), lemma 7.2 suggests that

\[
|\phi(s, a)^\top (M_{h, \phi} - M_{h, \phi}^\ast) v_{h+1}| \leq \|(M_{h, \phi} - M_{h, \phi}^\ast) \phi(s, a)\|_2 \| \Psi v_{h+1} \|_2 \\
\leq C_H H \sqrt{\beta_0} \| \phi(s, a)\|_{U_{h, \phi}^{-1}},
\]

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

\[
-2C_H H \sqrt{\beta_0} \| \phi(s, a)\|_{U_{h, \phi}^{-1}} \leq \phi(s, a)^\top (M_{h, \phi} - M_{h, \phi}^\ast) v_{h+1} - C_H \Psi \sqrt{\beta_0} \| \phi(s, a)\|_{U_{h, \phi}^{-1}} \leq 0.
\]

Plugging this into (7.3) yields

\[
Q_h(s, a) - r(s, a) - [P V_{h+1}](s, a)
\]

\[
= \max_{\phi \in \Phi} \left\{ \phi(s, a)^\top (M_{h, \phi} - M_{h, \phi}^\ast) v_{h+1} - C_H \Psi \sqrt{\beta_0} \| \phi(s, a)\|_{U_{h, \phi}^{-1}} \right\}
\]

\[
\geq \max_{\phi \in \Phi} \left\{ -2C_H H \sqrt{\beta_0} \| \phi(s, a)\|_{U_{h, \phi}^{-1}} \right\}
\]

\[
= -2C_H H \min_{\phi \in \Phi} \left\{ \sqrt{\beta_0} \| \phi(s, a)\|_{U_{h, \phi}^{-1}} \right\}
\]

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

\[
V_h(s) - V_h^\pi(s) \geq \sum_{h' = h}^H E_{\pi} [Q_h(s_{h'}, a_{h'}) - r(s_{h'}, a_{h'}) - [P V_{h+1}](s_{h'}, a_{h'}) | s_h = s]
\]

\[
\geq -2C_H H \sum_{h' = h}^H E_{\pi} \left[ \min_{\phi \in \Phi} \left\{ \sqrt{\beta_0} \| \phi(s, a)\|_{U_{h', \phi}^{-1}} \right\} \right] s_h = s \] (7.4)

and back to (7.2) yields

\[
V_h(s) - V_h^\pi(s) = \sum_{h' = h}^H E_{\pi} [Q_h(s_{h'}, a_{h'}) - r(s_{h'}, a_{h'}) - [P V_{h+1}](s_{h'}, a_{h'}) | s_h = s] \leq 0.
\] (7.5)

Substituting (7.4) from (7.3) yields the claimed result:

\[
V_h^\ast(s) - V_h^\ast(s) \leq 2C_H H \sum_{h' = h}^H E_{\pi} \left[ \min_{\phi \in \Phi} \left\{ \sqrt{\beta_0} \| \phi(s, a)\|_{U_{h', \phi}^{-1}} \right\} \right] s_h = s
\]

\[
\square
7.2 PROOF OF LEMMA 6.3

**Proof.** First we show that the covariance matrix $U_{\phi,h}$ is almost linearly growth with respect to the expectation $E_{d_h}[\phi^T\phi^T]$ and the size of offline data. Considering the formalization of covariance matrix

$$U_{\phi,h} = I + \sum_{(s,a,s') \in D_h} \phi(s,a)\phi^T(s,a)$$

$$= I - \sum_{(s,a,s') \in D_h} \mathbb{E}_{d_h}[\phi(s,a)\phi^T(s,a)] - \phi(s,a)\phi^T(s,a) + |D_h|\mathbb{E}_{d_h}[\phi(s,a)\phi^T(s,a)].$$

One can verify that $\mathbb{E}[\epsilon_h] = 0$ where the expectation is taken with respect to the randomness in the generation of the offline data. Since we have $\|\phi(s,a)\|_2^2 \leq C_\phi d_\phi$, it is obvious that $\|\epsilon_h\|_2 \leq \|\phi\| + \mathbb{E}\|\phi^T\|_2 \leq 2\|\phi\|_2 \leq 2C_\phi d_\phi$ by triangle’s inequality. Then by Lemma 4.6, with probability at least $1 - \delta$, for $|D_h| = K$ data,

$$\lambda_{\text{max}} \left( \sum_{(s,a,s') \in D_h} \epsilon_h \right) \leq 4C_\phi d_\phi \sqrt{2K \log(d_\phi/\delta)}.$$

Therefore, it’s suffice to show that

$$U_{\phi,h} \succeq KE_{d_h}^{\phi}[\phi(s,a)\phi^T(s,a)] + \left( 1 - 4C_\phi d_\phi \sqrt{2K \log(d_\phi/\delta)} \right) I$$

(7.6)

Furthermore, noticing that $E_{d_h}^{\phi}[\phi(s,a)\phi^T(s,a)]$ can be always written as

$$E_{d_h}^{\phi}[\phi(s,a)\phi^T(s,a)] = Q^T \text{diag}(d_r, 0_{d_\phi-r}) Q,$$

(7.7)

where $Q$ is an orthogonal matrix, $d_r \in \mathbb{R}^r$ is the non-zero eigen values of $E_{d_h}^{\phi}[\phi(s,a)\phi^T(s,a)]$ its minimal element as $\sigma_{h,\phi} r = \text{rank}(E_{d_h}^{\phi}[\phi(s,a)\phi^T(s,a)])$. Then (7.6) can be formalized as

$$QU_{\phi,h}Q^T \succeq \text{diag} \left( Kd_r - a1_r + 1_r, a1_{d_\phi-r} - 1_{d_\phi-r} \right),$$

(7.8)

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

$$QU_{\phi,h}Q^T \succeq \text{diag} \left( Kd_r - a1_r + 1_r, 1_{d_\phi-r} \right).$$

(7.9)

Noticing the minimal element of $d_r$ is $\sigma_{h,\phi}$, then when $K\sigma_{h,\phi} \geq a$, which we will verify later, the RHS of (7.9) is positive definite, which implies that

$$QU_{\phi,h}^{-1}Q^T \succeq \text{diag} \left( Kd_r - a1_r + 1_r, 1_{d_\phi-r} \right)^{-1}.$$ 

(7.10)

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

$$\phi^T(s,a)U_{\phi,h}^{-1}\phi(s,a)$$

$$= y^T E_{d_h}^{\phi}[\phi(s,a)\phi^T(s,a)] U_{\phi,h}^{-1} E_{d_h}^{\phi}[\phi(s,a)\phi^T(s,a)]y$$

$$= y^T Q^T \text{diag}(d_r, 0_{d_\phi-r}) Q U_{\phi,h}^{-1} Q^T \text{diag}(d_r, 0_{d_\phi-r}) Qy$$

$$\leq y^T Q^T \text{diag}(d_r, 0_{d_\phi-r}) \text{diag}(Kd_r - a1_r + 1_r, 1_{d_\phi-r})^{-1} \text{diag}(d_r, 0_{d_\phi-r}) Qy,$$

(7.11)

where the last inequality follows (7.10). Noticing that $\text{diag}(d_r, 0_{d_\phi-r}) Qy$ can be written as $\text{diag}(d_r, 0_{d_\phi-r}) Qy = \left(z_r^T, 0_{d_\phi-r}^T \right)^T$. Therefore (7.11) becomes

$$\phi^T(s,a)U_{\phi,h}^{-1}\phi(s,a) \leq \left(z_r^T, 0_{d_\phi-r}^T \right)^T \text{diag}(Kd_r - a1_r + 1_r, 1_{d_\phi-r})^{-1} \left(z_r^T, 0_{d_\phi-r}^T \right)^T$$

$$= z_r^T \text{diag}(Kd_r - a1_r + 1_r, 1_{d_\phi-r})^{-1} z_r.$$ 

(7.12)
Noticing that
\[ \| \mathbf{z}_r \|_2 = \| \text{diag}(\mathbf{d}_r, \mathbf{0}_{d_r - r}) \mathbf{Q} \mathbf{y} \| = \| \mathbf{Q} \phi(s, a) \|_2 = \| \phi(s, a) \|_2 \leq \sqrt{C_{\phi} d_{\phi}}, \]

where the last equality comes from Definition 3.2, (7.12) finally becomes
\[ \phi^\top(s, a) \mathbf{U}_{h, \phi}^{-1} \phi(s, a) \leq z_r^\top \text{diag}(K \mathbf{d}_r - a \mathbf{1}_r + 1, r)^{-1} z_r \leq \frac{C_{\phi} d_{\phi}}{K \sigma_{h, \phi} - a + 1} \leq \frac{C_{\phi} d_{\phi}}{K \sigma_{h, \phi} - a}, \quad (7.13) \]

where the second last inequality is due to \( \lambda_{\max} \left( \text{diag}(K \mathbf{d}_r - a \mathbf{1}_r + 1, r)^{-1} \right) = (K \sigma_{h, \phi} - a + 1)^{-1} \) and the last inequality utilizes the assumption that \( K \sigma_{h, \phi} \geq a \).

Next, to control \( \| \phi(s, a) \|_{\mathbf{U}_{h, \phi}^{-1}} \leq \Delta/(2C_{\phi} H^2 \sqrt{\beta_{\phi}}) =: B, \) by (7.13), it suffices to control
\[ \frac{C_{\phi} d_{\phi}}{K \sigma_{h, \phi} - 4C_{\phi} d_{\phi} \sqrt{2K \log(d_{\phi}/\delta)}} \leq B^2. \quad (7.14) \]

Denoting \( \sqrt{K} = 4C_{\phi} d_{\phi} \sqrt{2 \log(d_{\phi}/\delta)\sigma_{h, \phi}^{-1} x}, \) then the constrain \( K \sigma_{h, \phi} \geq 4C_{\phi} d_{\phi} \sqrt{2K \log(d_{\phi}/\delta)} \) is equivalent with \( x \geq 1, \) and (7.14) becomes
\[ \frac{C_{\phi} d_{\phi}}{B^2} \leq \frac{32C_{\phi}^2 d_{\phi}^2 \log(d_{\phi}/\delta)}{\sigma_{h, \phi}^2} (x^2 - x). \quad (7.15) \]

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
\[ x > \sqrt{1 + \frac{C_{\phi} d_{\phi} \sigma_{h, \phi}}{16B^2C_{\phi}^2 d_{\phi}^2 \log(d_{\phi}/\delta)}}, \quad (7.16) \]

which can imply the constrain that \( x \geq 1. \) Plugging the notations \( B := \Delta/(2C_{\phi} H^2 \sqrt{\beta_{\phi}}) \) and \( \sqrt{K} = 4C_{\phi} d_{\phi} \sqrt{2 \log(d_{\phi}/\delta)\sigma_{h, \phi}^{-1} x} \) back into (7.16) yields for all \( (s, a, h) \in S \times A \times [H], \) with probability at least \( 1 - H|\Phi| \delta, \) there exists \( \phi \in \Phi \) such that when
\[ K > \frac{32C_{\phi}^2 d_{\phi}^2 \log(d_{\phi}/\delta)}{\sigma_{h, \phi}^2} \left( 1 + \frac{C_{\phi}^2 H^4 \beta_{\phi} d_{\phi} \sigma_{h, \phi}}{4\Delta^2 C_{\phi}^2 d_{\phi} \log(d_{\phi}/\delta)} \right), \]

\[ \| \phi(s, a) \|_{\mathbf{U}_{h, \phi}} < \text{gap}_{\min}/(2C_{\phi} H^2 \sqrt{\beta_{\phi}}). \]

Replacing \( \delta \) by \( \delta/(H|\Phi|) \) we can get the claimed result. \( \square \)

References


