Algorithm 4 `REGIME-exploration`

Input: The number of total episodes K, bonus parameter β_{ex} and regularization parameter λ_{ex}.
for $k = 1, \cdots, K$ **do**
  Initialize: $Q^k_{H+1}(\cdot,\cdot) \leftarrow 0, V^k_{H+1}(\cdot) \leftarrow 0$.
  for $h = H, \cdots, 1$ **do**
    Compute the covariance matrix: $\Lambda^k_h \leftarrow \sum_{i=1}^{k-1} \phi_h(s^i_h, a^i_h)\phi_h(s^i_h, a^i_h)^\top + \lambda_{\text{ex}} I$.
    Compute the bonus and reward:
    $b^k_h(\cdot,\cdot) \leftarrow \min\left\{\beta_{\text{ex}}\|\phi_h(\cdot,\cdot)\|_{(\Lambda^k_h)^{-1}}, H-h+1\right\}$ and $r^k_h = b^k_h/H$.
    Compute Q function:

$$Q^k_h(\cdot,\cdot) \leftarrow \text{Clip}_{[0,H-h+1]}\Big(\text{Clip}_{[0,H-h+1]}((w^k_h)^\top \phi_h(\cdot,\cdot) + r^k_h(\cdot,\cdot)) + b^k_h(\cdot,\cdot)\Big),$$

    where $w^k_h = (\Lambda^k_h)^{-1}\sum_{i=1}^{k-1}\phi_h(s^i_h, a^i_h)\cdot V^k_{h+1}(s^i_{h+1})$.
    Compute value function and policy:

$$V^k_h(\cdot) \leftarrow \max_{a\in\mathcal{A}} Q^k_h(\cdot, a), \pi^k_h(\cdot) \leftarrow \arg\max_{a\in\mathcal{A}} Q^k_h(\cdot, a).$$

  end for
  Collect a trajectory $\tau^k = (s^k_h, a^k_h, s^k_{h+1})_{h=1}^H$ by running $\pi^k = \{\pi^k_h\}_{h=1}^H$ and add τ^k into $\mathcal{D}_{\text{ex}}$.
end for
Sample K states from the initial states $\{s^{i,\text{in}}_1\}_{i=1}^K$ and add them to $\mathcal{D}_{\text{in}}$.
Return $\mathcal{D}_{\text{ex}}, \mathcal{D}_{\text{in}}$.

A Discussion

Validity of Linear Parametrization. In this work we consider linear reward parametrization, or more generally, linear trajectory embeddings. Such assumptions are commonly used in the theoretical works of PbRL (Pacchiano et al., 2021; Zhu et al., 2023) and relevant examples can be borrowed from the practical works (Pacchiano et al., 2020; Parker-Holder et al., 2020) like the behavior guided class of algorithms for policy optimization. We admit that our analysis cannot cover all kinds of reward parametrization, but we think it is a reasonable starting point to consider linear function approximation for theoretical analysis.

Applicability of `REGIME`. Our results are not restricted to linear MDPs and low-rank MDPs. Our main result (Theorem 1) only requires a suitable reward-free oracle for dynamics learning and in practice there exist some oracles ready for use, even when the MDPs are very complicated (Xu et al., 2022). This implies that by plugging in these general reward-free oracles, we are able to deal with general MDPs as well.

Relationship to Active Learning. The trajectory collection process in Step 1 of `REGIME` utilizes a similar idea to active learning. In active learning, people choose a trajectory pair to query human feedback which can maximize the information gain in each iteration. Similarly, in our algorithm, the estimated covariance matrix $\widehat{\Sigma}_n$ can be regarded as the current information in n-th iteration. Then we choose a pair of policies $(\pi^{n,0}, \pi^{n,1})$ which can maximize the information gain approximately (line 5, it is approximate because we are using an approximate dynamics $\widehat{P}$ for evaluation).

B Omit details in Section 4

In this section we present the details of Algorithm 4 and Algorithm 5. Here $\text{Clip}_{[a,b]}(x)$ means $\min\{\max\{a, x\}, b\}$. In particular, when estimating $(\phi(\pi))_{h,j}$, we use the reward function $r^{h,j}_{h'}(s,a) = \phi_{h'}(s,a)^\top \theta^{h,j}_{h'}$ for all $h' \in [H]$ (up to an R factor) where

$$\theta^{h,j}_{h'} = \begin{cases} \frac{1}{R}\cdot e_j, & \text{if} \quad h' = h, \\ 0, & \text{otherwise.} \end{cases}$$

Algorithm 5 `REGIME-planning`

Input: Dataset $\mathcal{D}_{\text{ex}} = \{(s_h^i, a_h^i, s_{h+1}^i)\}_{i=1,h=1}^{K,H}$, $\mathcal{D}_{\text{in}} = \{s_1^{i,\text{in}}\}_{i=1}^K$, bonus parameter β_{pl} and regularization parameter λ_{pl}, policy π, reward function $(r_h)_{h=1}^H$.
for $h' = H, \cdots, 1$ **do**
  Compute the covariance matrix: $\Lambda_{h'} \leftarrow \sum_{i=1}^K \phi_{h'}(s_{h'}^i, a_{h'}^i)\phi_{h'}(s_{h'}^i, a_{h'}^i)^\top + \lambda_{\text{pl}} I$.
  Compute the bonus: $b_{h'}(\cdot,\cdot) \leftarrow \min\big\{\beta_{\text{pl}}\|\phi_{h'}(\cdot,\cdot)\|_{(\Lambda_{h'})^{-1}}, 2(H-h+1)\big\}$.
end for
Initialize: $\widehat{Q}_{H+1}^{r,\pi}(\cdot,\cdot) \leftarrow 0, \widehat{V}_{H+1}^{r,\pi}(\cdot) \leftarrow 0$.
for $h' = H, \cdots, 1$ **do**
  Compute Q function:

$$\widehat{Q}_{h'}^{r,\pi}(\cdot,\cdot) \leftarrow \text{Clip}_{[-(H-h+1),H-h+1]}\Big(\text{Clip}_{[-(H-h+1),H-h+1]}((w_{h'}^{r,\pi})^\top \phi_{h'}(\cdot,\cdot) + r_{h'}(\cdot,\cdot)) + b_{h'}(\cdot,\cdot)\Big),$$

  where $w_{h'}^{r,\pi} = (\Lambda_{h'})^{-1}\sum_{i=1}^K \phi_{h'}(s_{h'}^i, a_{h'}^i)\cdot \widehat{V}_{h'+1}^{r,\pi}(s_{h'+1}^i)$.
  Compute value function: $\widehat{V}_{h'}^{r,\pi}(\cdot) \leftarrow \mathbb{E}_{a\sim\pi_{h'}}\widehat{Q}_{h'}^{r,\pi}(\cdot,a)$.
end for
Compute $\widehat{V}^\pi(r) \leftarrow \frac{1}{K}\sum_{i=1}^K \widehat{V}_1^{r,\pi}(s_1^{i,\text{in}})$.
Return $\widehat{V}^\pi(r)$.

Here e_j is the one-hot vector whose j-th entry is 1. For simplicity, we denote $\widehat{V}^{r^{h,j},\pi}, \widehat{Q}^{r^{h,j},\pi}, w^{r^{h,j},\pi}$ by $V^{h,j,\pi}, Q^{h,j,\pi}, w^{h,j,\pi}$ and let the estimation $(\widehat{\phi}(\pi))_{h,j}$ be $R\widehat{V}^\pi(r^{h,j})$.

Then we have the following formal theorem characterizing the sample complexity of Algorithm 2:

Theorem 4. *Let*

$$\lambda_{\text{ex}} = \lambda_{\text{pl}} = R^2,$$
$$\beta_{\text{ex}} = C_\beta dHR\sqrt{\log(dKHR/\delta)}, \beta_{\text{pl}} = C_\beta dHR\sqrt{\log(dKHR\mathcal{N}_\Pi(\epsilon')/\delta)}$$
$$\lambda \geq 4HR^2, K \geq \widetilde{\mathcal{O}}\Big(\frac{H^8B^2R^4d^4\log(\mathcal{N}_\Pi(\epsilon')/\delta)}{\epsilon^2}\Big), N \geq \widetilde{\mathcal{O}}\Big(\frac{\lambda\kappa^2B^2R^2H^4d^2\log(1/\delta)}{\epsilon^2}\Big),$$

where $\epsilon' = \frac{\epsilon}{72BR^2H\sqrt{d^HK^{H-1}}}$, $C_\beta > 0$ *is a universal constant and* $\kappa = 2 + \exp(2r_{\max}) + \exp(-2r_{\max})$. *Then under Assumption 1 and 3, with probability at least* $1-\delta$, *we have*

$$V^{r^*,\hat{\pi}} \geq V^{r^*,*} - \epsilon.$$

The proof is deferred to Appendix E.1.

B.1 Log-Linear Policy Class

The sample complexity in Theorem 2 depends on the covering number of the policy class Π. Therefore we want to find a policy class for linear MDPs that is rich enough (i.e., contains near-global-optimal policies) while retains a small covering number at the same time. Indeed, the log-linear policy class (Agarwal et al., 2020b) satisfies this requirement, which is defined as follows:

$$\Pi = \Big\{\pi : \pi_h^\zeta(a|s) = \frac{\exp(\zeta_h^\top\phi_h(s,a))}{\sum_{a'\in\mathcal{A}}\exp(\zeta_h^\top\phi_h(s,a'))}, \zeta_h \in \mathbb{B}(d,W), \forall s\in\mathcal{S}, a\in\mathcal{A}, h\in[H]\Big\}$$

Here $\mathbb{B}(d,W)$ is the d-dimensional ball centered at the origin with radius W. The following proposition characterizes the covering number of such log-linear policy class:

Proposition 1. *Let* Π *be the log-linear policy class. Then under Assumption 1, for any* $\epsilon \leq 1$, *we have* $\log\mathcal{N}_\Pi(\epsilon) \leq Hd\log\Big(\frac{12WR}{\epsilon}\Big)$.

Meanwhile, we can quantify the bias of such log-linear policy class as follows:

Proposition 2. *Let* $W = \frac{\left(B+(H+\epsilon)\sqrt{d}\right)H\log|\mathcal{A}|}{\epsilon}$, *then under Assumption 1 and 3, we have*

$$V^{r^*,\pi_{\mathrm{g}}} - \max_{\pi\in\Pi} V^{r^*,\pi} \le \epsilon,$$

where π_{g} *is the global optimal policy.*

Combining Theorem 4, Proposition 1 and Proposition 2, we know that the returned policy $\widehat{\pi}$ by Algorithm 2 with log-linear policy classes can indeed compete against the global optimal policy with the following sample complexities:

$$N_{\mathrm{tra}} = K + N = \widetilde{\mathcal{O}}\left(\frac{d^5+\kappa^2 d^2}{\epsilon^2}\right), N_{\mathrm{hum}} = \widetilde{\mathcal{O}}\left(\frac{\kappa^2 d^2}{\epsilon^2}\right).$$

C Proof of Theorem 1 With Known Transitions

In this section, we consider the proof of Theorem 1 when transitions are known, i.e., $\epsilon' = 0$ and $\widehat{P} = P^*$. In this case we have $\widehat{\phi}(\pi) = \phi(\pi)$. We will deal with the unknown transition in Appendix D.1.

First, note that from the definition of $\widehat{\pi}$, we have

$$V^{\widehat{r},\widehat{\pi}} \ge V^{\widehat{r},\pi^*},$$

where π^* is the optimal policy with respect to the ground-truth reward r^*, i.e., $\pi^* = \arg\max_{\pi\in\Pi} V^{r^*,\pi}$. Therefore we can expand the suboptimality as follows:

$$\begin{aligned}
V^{r^*,*} - V^{r^*,\widehat{\pi}} &= (V^{r^*,*} - V^{\widehat{r},\pi^*}) + (V^{\widehat{r},\pi^*} - V^{\widehat{r},\widehat{\pi}}) + (V^{\widehat{r},\widehat{\pi}} - V^{r^*,\widehat{\pi}}) \\
&\le (V^{r^*,*} - V^{\widehat{r},\pi^*}) + (V^{\widehat{r},\widehat{\pi}} - V^{r^*,\widehat{\pi}}) \\
&= \mathbb{E}_{\tau\sim(\pi^*,P^*)}[\langle\phi(\tau),\theta^* - \widehat{\theta}\rangle] - \mathbb{E}_{\tau\sim(\widehat{\pi},P^*)}[\langle\phi(\tau),\theta^* - \widehat{\theta}\rangle] \\
&= \langle\phi(\pi^*) - \phi(\widehat{\pi}),\theta^* - \widehat{\theta}\rangle \\
&\le \|\phi(\pi^*) - \phi(\widehat{\pi})\|_{\Sigma_{N+1}^{-1}} \cdot \|\theta^* - \widehat{\theta}\|_{\Sigma_{N+1}}, \qquad (3)
\end{aligned}$$

where $\Sigma_n := \lambda I + \sum_{i=1}^{n-1}(\phi(\pi^{i,0}) - \phi(\pi^{i,1}))(\phi(\pi^{i,0}) - \phi(\pi^{i,1}))^\top$ for all $n \in [N+1]$. Here the third step is due to the definition of value function and the last step comes from Cauchy-Schwartz inequality. Next we will bound $\|\phi(\pi^*) - \phi(\widehat{\pi})\|_{\Sigma_{N+1}^{-1}}$ and $\|\theta^* - \widehat{\theta}\|_{\Sigma_{N+1}}$ respectively.

First for $\|\phi(\pi^*) - \phi(\widehat{\pi})\|_{\Sigma_{N+1}^{-1}}$, notice that $\Sigma_{N+1} \succeq \Sigma_n$ for all $n \in [N+1]$, which implies

$$\begin{aligned}
\|\phi(\pi^*) - \phi(\widehat{\pi})\|_{\Sigma_{N+1}^{-1}} &\le \frac{1}{N}\sum_{n=1}^{N}\|\phi(\pi^*) - \phi(\widehat{\pi})\|_{\Sigma_n^{-1}} \le \frac{1}{N}\sum_{n=1}^{N}\|\phi(\pi^{n,0}) - \phi(\pi^{n,1})\|_{\Sigma_n^{-1}} \\
&\le \frac{1}{\sqrt{N}}\sqrt{\sum_{n=1}^{N}\|\phi(\pi^{n,0}) - \phi(\pi^{n,1})\|_{\Sigma_n^{-1}}^2}, \qquad (4)
\end{aligned}$$

where the second step comes from the definition of $\pi^{n,0}$ and $\pi^{n,1}$ and the last step is due to Cauchy-Schwartz inequality. To bound the right hand side of (4), we utilize the following Elliptical Potential Lemma:

Lemma 1 (Elliptical Potential Lemma)**.** *For any* $\lambda \ge R_x^2$ *and* $d \ge 1$, *consider a sequence of vectors* $\{x^n \in \mathbb{R}^d\}_{n=1}^N$ *where* $\|x^n\| \le R_x$ *for all* $n \in [N]$. *Let* $\Sigma_n = \lambda I + \sum_{i=1}^{n-1} x^n (x^n)^\top$, *then we have*

$$\sum_{n=1}^{N}\|x^n\|_{\Sigma_n^{-1}}^2 \le 2d\log\left(1 + \frac{N}{d}\right).$$

The proof is deferred to Appendix C.1. Since we have $\lambda \ge 4HR^2$, by Lemma 1 we know

$$\sqrt{\sum_{n=1}^{N}\|\phi(\pi^{n,0}) - \phi(\pi^{n,1})\|_{\Sigma_n^{-1}}^2} \le \sqrt{2HdN\log(1+N/(Hd))}.$$

Combining the above inequality with (4), we have

$$\|\phi(\pi^*) - \phi(\widehat{\pi})\|_{\Sigma_{N+1}^{-1}} \leq \sqrt{\frac{2Hd\log(1+N/(Hd))}{N}}. \tag{5}$$

For $\|\theta^* - \widehat{\theta}\|_{\Sigma_{N+1}}$, first note that $\widehat{\theta}$ is the MLE estimator. Let $\widetilde{\Sigma}_n$ denote the empirical cumulative covariance matrix $\lambda I + \sum_{i=1}^{n-1}(\phi(\tau^{i,0}) - \phi(\tau^{i,1}))(\phi(\tau^{i,0}) - \phi(\tau^{i,1}))^\top$, then from the literature (Zhu et al., 2023), we know that MLE has the following guarantee:

Lemma 2 (MLE guarantee). *For any $\lambda > 0$ and $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\|\hat{\theta} - \theta^*\|_{\widetilde{\Sigma}_{N+1}} \leq C_{\mathrm{MLE}} \cdot \sqrt{\kappa^2(Hd + \log(1/\delta)) + \lambda H B^2}, \tag{6}$$

where $\kappa = 2 + \exp(2r_{\max}) + \exp(-2r_{\max})$ and $C_{\mathrm{MLE}} > 0$ is a universal constant.

The proof is deferred to Appendix C.2. With Lemma 2, to $\|\theta^* - \widehat{\theta}\|_{\Sigma_{N+1}}$ we only need to show $\widetilde{\Sigma}_{N+1}$ is close to Σ_{N+1}. This can be achieved by the following concentration result from the literature:

Lemma 3 (Pacchiano et al. (2021)[Lemma 7]). *For any $\lambda > 0$ and $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\|\theta^* - \hat{\theta}\|_{\Sigma_{N+1}}^2 \leq 2\|\theta^* - \hat{\theta}\|_{\widetilde{\Sigma}_{N+1}}^2 + C_{\mathrm{CON}} H^3 d R^2 B^2 \log(N/\delta), \tag{7}$$

where $C_{\mathrm{CON}} > 0$ is a universal constant.

Therefore, combining (7) and (6), by union bound with probability at least $1-\delta$, we have that

$$\|\theta^* - \widehat{\theta}\|_{\Sigma_{N+1}} \leq C_1 \cdot \kappa B R \sqrt{\lambda H^3 d \log(N/\delta)}, \tag{8}$$

where C_1 is a universal constant.

Thus substituting (5) and (8) into (3), we have $V^*(r^*) - V(r^*, \hat{\pi}) \leq \epsilon$ with probability at least $1-\delta$ as long as

$$N \geq \widetilde{\mathcal{O}}\Big(\frac{\lambda\kappa^2 B^2 R^2 H^4 d^2 \log(1/\delta)}{\epsilon^2}\Big).$$

C.1 Proof of Lemma 1

Note that when $\lambda \geq R_x^2$, we have $\|x^n\|_{\Sigma_n^{-1}} \leq 1$ for all $n \in [N]$, which implies that for all $n \in [N]$, we have

$$\|x^n\|_{\Sigma_n^{-1}}^2 \leq \log\left(1 + \|x^n\|_{\Sigma_n^{-1}}^2\right).$$

On the other hand, let w^n denote $\|x^n\|_{\Sigma_n^{-1}}$, then we know for any $n \in [N-1]$

$$\begin{aligned}
\log\det\Sigma_{n+1} &= \det(\Sigma_n + x^n(x^n)^\top) = \log\det(\Sigma_n^{1/2}(I + \Sigma_n^{-1/2}x^n(x^n)^\top\Sigma_n^{-1/2})\Sigma_n^{1/2}) \\
&= \log\det(\Sigma_n) + \log\det(I + (\Sigma_n^{-1/2}x^n)(\Sigma_n^{-1/2}x^n)^\top) \\
&= \log\det(\Sigma_n) + \log\det(I + (\Sigma_n^{-1/2}x^n)^\top(\Sigma_n^{-1/2}x^n)) \\
&= \log\det(\Sigma_n) + \log\left(1 + \|x^n\|_{\Sigma_n^{-1}}^2\right),
\end{aligned}$$

where the fourth step is due to the property of determinants. Therefore we have

$$\begin{aligned}
\sum_{n=1}^{N}\log\left(1 + \|x^n\|_{\Sigma_n^{-1}}^2\right) &= \log\det\Sigma_{N+1} - \log\det\Sigma_1 = \log(\det\Sigma_{N+1}/\det\Sigma_1) \\
&= \log\det\left(I + \frac{1}{\lambda}\sum_{n=1}^{N}x^n(x^n)^\top\right).
\end{aligned}$$

Now let $\{\lambda_i\}_{i=1}^d$ denote the eigenvalues of $\sum_{n=1}^N x^n(x^n)^\top$, then we know

$$\log\det\left(I+\frac{1}{\lambda}\sum_{n=1}^N x^n(x^n)^\top\right)=\log\left(\prod_{i=1}^d(1+\lambda_i/\lambda)\right)$$
$$\leq d\log\left(\frac{1}{d}\sum_{i=1}^d(1+\lambda_i/\lambda)\right)\leq d\log\left(1+\frac{NR_x^2}{d\lambda}\right)\leq d\log\left(1+\frac{N}{d}\right),$$

where the third step comes from $\sum_{i=1}^d\lambda_i=\mathrm{Tr}\left(\sum_{n=1}^N x^n(x^n)^\top\right)=\sum_{n=1}^N\|x^n\|^2\leq NR_x^2$ and the last step is due to the fact that $\lambda\geq R_x^2$. This concludes our proof.

C.2 Proof of Lemma 2

First note that we have the following lemma from literature:

Lemma 4 ((Zhu et al., 2023)[Lemma 3.1]). *For any $\lambda'>0$, with probability at least $1-\delta$, we have*

$$\|\widehat{\theta}-\theta^*\|_{D+\lambda' I}\leq O\left(\sqrt{\frac{\kappa^2(Hd+\log(1/\delta))}{N}+\lambda' HB^2}\right),$$

where $D=\frac{1}{N}\sum_{i=1}^N(\phi(\tau^{i,0})-\phi(\tau^{i,1}))(\phi(\tau^{i,0})-\phi(\tau^{i,1}))^\top$.

Therefore let $\lambda'=\frac{\lambda}{N}$ and from the above lemma we can obtain

$$\|\widehat{\theta}-\theta^*\|_{\frac{\widetilde{\Sigma}_{N+1}}{N}}\leq O\left(\sqrt{\frac{\kappa^2(Hd+\log(1/\delta))}{N}+\frac{\lambda HB^2}{N}}\right),$$

which is equivalent to

$$\|\widehat{\theta}-\theta^*\|_{\widetilde{\Sigma}_{N+1}}\leq O\left(\sqrt{\kappa^2(Hd+\log(1/\delta))+\lambda HB^2}\right).$$

This concludes our proof.

D Proofs in Section 3

D.1 Proof of Theorem 1

Note that from the proof of Theorem 1 with known transition dynamics, we have:

$$V^{r^*,*}-V^{r^*,\widehat{\pi}}\leq\langle\phi(\pi^*)-\phi(\widehat{\pi}),\theta^*-\widehat{\theta}\rangle+(V^{\widehat{r},\pi^*}-V^{\widehat{r},\widehat{\pi}}),\tag{9}$$

Then we have

$$V^{r^*,*}-V^{r^*,\widehat{\pi}}\leq\langle\phi(\pi^*)-\widehat{\phi}(\pi^*),\theta^*-\hat{\theta}\rangle+\langle\widehat{\phi}(\hat{\pi})-\phi(\hat{\pi}),\theta^*-\hat{\theta}\rangle$$
$$+\langle\widehat{\phi}(\pi^*)-\widehat{\phi}(\hat{\pi}),\theta^*-\hat{\theta}\rangle+(V^{\widehat{r},\pi^*}-V^{\widehat{r},\widehat{\pi}}).\tag{10}$$

Now we only need to bound the three terms in the RHS of (10). For the first and second term, we need to utilize the following lemma:

Lemma 5. *Let $d_h^\pi(s,a)$ and $\widehat{d}_h^\pi(s,a)$ denote the visitation measure of policy π under P^* and $\hat{P}$. Then with probability at least $1-\delta/4$, we have for all $h\in[H]$ and $\pi\in\Pi$,*

$$\|d_h^\pi-\widehat{d}_h^\pi\|_1\leq h\epsilon'.\tag{11}$$

Let $\mathcal{E}_1$ denote the event when (11) holds. Then under event $\mathcal{E}_1$, we further have the following lemma:

Lemma 6. *Under event $\mathcal{E}_1$, for all policy $\pi\in\Pi$ and vector $v=[v_1,\cdots,v_H]$ where $v_h\in\mathbb{R}^d$ and $\|v_h\|\leq 2B$ for all $h\in[H]$ we have,*

$$|\langle\phi(\pi)-\widehat{\phi}(\pi),v\rangle|\leq BRH^2\epsilon'.$$

Substitute Lemma 6 into (10), we have

$$V^{r^*,*} - V^{r^*,\widehat{\pi}} \leq \langle \widehat{\phi}(\pi^*) - \widehat{\phi}(\widehat{\pi}), \theta^* - \widehat{\theta} \rangle + 2BRH^2\epsilon' + (V^{\widehat{r},\pi^*} - V^{\widehat{r},\widehat{\pi}}).$$

Then by Cauchy-Schwartz inequality, we have under event $\mathcal{E}_1$,

$$V^{r^*,*} - V^{r^*,\widehat{\pi}} \leq \|\widehat{\phi}(\pi^*) - \widehat{\phi}(\widehat{\pi})\|_{\widehat{\Sigma}_{N+1}^{-1}} \cdot \|\theta^* - \widehat{\theta}\|_{\widehat{\Sigma}_{N+1}} + 2BRH^2\epsilon' + (V^{\widehat{r},\pi^*} - V^{\widehat{r},\widehat{\pi}}). \quad (12)$$

Following the same analysis in the proof of Theorem 1 with known transition, we know

$$\|\widehat{\phi}(\pi^*) - \widehat{\phi}(\widehat{\pi})\|_{\widehat{\Sigma}_{N+1}^{-1}} \leq \sqrt{\frac{2Hd\log(1+N/(Hd))}{N}}. \quad (13)$$

Now we only need to bound $\|\theta^* - \hat{\theta}\|_{\widehat{\Sigma}_{N+1}}$. Similar to the proof of Theorem 1 with known transition, we use Σ_n and $\widetilde{\Sigma}_n$ to denote $\lambda I + \sum_{i=1}^{n-1}(\phi(\pi^{i,0}) - \phi(\pi^{i,1}))(\phi(\pi^{i,0}) - \phi(\pi^{i,1}))^\top$ and $\lambda I + \sum_{i=1}^{n-1}(\phi(\tau^{i,0}) - \phi(\tau^{i,1}))(\phi(\tau^{i,0}) - \phi(\tau^{i,1}))^\top$ respectively. Then under event $\mathcal{E}_1$, we have the following connection between $\widehat{\Sigma}_{N+1}$ and Σ_{N+1}:

Lemma 7. *Under event $\mathcal{E}_1$, we have*

$$\|\theta^* - \widehat{\theta}\|_{\widehat{\Sigma}_{N+1}} \leq \sqrt{2}\|\theta^* - \widehat{\theta}\|_{\Sigma_{N+1}} + 2\sqrt{2}BRH^2\epsilon'.$$

Combining Lemma 7 with Lemma 2 and Lemma 3, we have under event $\mathcal{E}_1 \cap \mathcal{E}_2$,

$$\begin{aligned}\|\theta^* - \widehat{\theta}\|_{\widehat{\Sigma}_{N+1}} &\leq \sqrt{2}\|\theta^* - \widehat{\theta}\|_{\Sigma_{N+1}} + 2\sqrt{2}BRH^2\epsilon' \\ &\leq C_2 \cdot \kappa BR\sqrt{\lambda H^3 d\log(N/\delta)} + 2\sqrt{2}BRH^2\epsilon', \end{aligned} \quad (14)$$

where $\Pr(\mathcal{E}_2) \geq 1 - \delta/2$ and $C_2 > 0$ is a universal constant.

Now we only need to bound $(V^{\widehat{r},\pi^*} - V^{\widehat{r},\widehat{\pi}})$, which can be achieved with Lemma 6:

$$\begin{aligned} V^{\widehat{r},\pi^*} - V^{\widehat{r},\widehat{\pi}} &= \langle \phi(\pi^*), \widehat{\theta} \rangle - \langle \phi(\widehat{\pi}), \widehat{\theta} \rangle \\ &= \langle \phi(\pi^*) - \widehat{\phi}(\pi^*), \widehat{\theta} \rangle + \langle \widehat{\phi}(\pi^*) - \widehat{\phi}(\widehat{\pi}), \widehat{\theta} \rangle + \langle \widehat{\phi}(\widehat{\pi}) - \phi(\widehat{\pi}), \widehat{\theta} \rangle \leq 2BRH^2\epsilon', \end{aligned} \quad (15)$$

where the last step comes from Lemma 6 and the definition of $\widehat{\pi}$.

Combining (12), (13) (14) and (15), we have $V^{r^*,*} - V^{r^*,\widehat{\pi}} \leq \epsilon$ with probability at least $1 - \delta$ as long as

$$\epsilon' \leq \frac{\epsilon}{6BRH^2}, \qquad N \geq \widetilde{\mathcal{O}}\Big(\frac{\lambda\kappa^2 B^2 R^2 H^4 d^2 \log(1/\delta)}{\epsilon^2}\Big).$$

D.2 Proof of Lemma 5

First notice that $d_h^\pi(s,a) = d_h^\pi(s)\pi(a|s)$ and $\widehat{d}_h^\pi(s,a) = \widehat{d}_h^\pi(s)\pi(a|s)$, which implies that for all $h \in [H]$

$$\begin{aligned} \big\|d_h^\pi - \widehat{d}_h^\pi\big\|_1 &= \sum_{s,a} \big|d_h^\pi(s,a) - \widehat{d}_h^\pi(s,a)\big| = \sum_{s,a} \big|d_h^\pi(s) - \widehat{d}_h^\pi(s)\big|\pi(a|s) \\ &= \sum_s \big|d_h^\pi(s) - \widehat{d}_h^\pi(s)\big| \sum_a \pi(a|s) = \sum_s \big|d_h^\pi(s) - \widehat{d}_h^\pi(s)\big|. \end{aligned}$$

Therefore we only need to prove $\sum_s \big|d_h^\pi(s) - \widehat{d}_h^\pi(s)\big| \leq h\epsilon'$ for all $h \in [H]$. We use induction to prove this. First for the base case, we have $\sum_s |d_1^\pi(s) - \widehat{d}_1^\pi(s)| = \sum_s \big|P_1^*(s) - \widehat{P}_1(s)\big| \leq \epsilon'$ according to the guarantee of the reward-free learnign oracle $\mathcal{P}$.

Now assume that $\sum_s \big|d_{h'}^\pi(s) - \widehat{d}_{h'}^\pi(s)\big| \leq h'\epsilon'$ for all $h' \in [h]$ where $h \in [H-1]$. Then we have

$$\sum_s \big|d_{h+1}^\pi(s) - \widehat{d}_{h+1}^\pi(s)\big| = \sum_s \Big|\sum_{s',a'} \widehat{d}_h^\pi(s')\pi(a'|s')\widehat{P}_h(s|s',a') - d_h^\pi(s')\pi(a'|s')P_h^*(s|s',a')\Big|$$

$$
\begin{aligned}
&\leq \Big(\sum_{s,s',a'} \Big| \widehat{d}_h^{\pi}(s') - d_h^{\pi}(s') \Big| \pi(a'|s') \widehat{P}_h(s|s',a') \Big) \\
&\quad + \Big(\sum_{s,s',a'} d_h^{\pi}(s') \pi(a'|s') \Big| \widehat{P}_h(s|s',a') - P_h^*(s|s',a') \Big| \Big) \\
&= \Big(\sum_{s'} \Big| \widehat{d}_h^{\pi}(s') - d_h^{\pi}(s') \Big| \sum_{a'} \pi(a'|s') \sum_{s} \widehat{P}_h(s|s',a') \Big) \\
&\quad + \mathbb{E}_{\pi, P^*} [\| \widehat{P}_h(\cdot|s',a') - P_h^*(\cdot|s',a') \|_1] \\
&\leq (h+1)\epsilon',
\end{aligned}
$$

where the second step comes from the triangle inequality and the last step is dueto the induction hypothesis and the guarantee of $\mathcal{P}$. Therefore, we have $\sum_s |d_{h+1}^{\pi}(s) - \widehat{d}_{h+1}^{\pi}(s)| \leq (h+1)\epsilon'$. Then by induction, we know $\sum_s |d_h^{\pi}(s) - \widehat{d}_h^{\pi}(s)| \leq h\epsilon'$ for all $h \in [H]$, which concludes our proof.

D.3 Proof of Lemma 6

Note that from the definition of $\phi(\pi)$ we have

$$
\langle \phi(\pi), v \rangle = \mathbb{E}_{\tau \sim (\pi, P^*)} \Big[\sum_{h=1}^{H} \phi_h^{\top}(s_h, a_h) v_h \Big] = \sum_{h=1}^{H} \sum_{s_h, a_h} d_h^{\pi}(s_h, a_h) \phi_h^{\top}(s_h, a_h) v_h.
$$

Similarly, we have

$$
\langle \widehat{\phi}(\pi), v \rangle = \sum_{h=1}^{H} \sum_{s_h, a_h} \widehat{d}_h^{\pi}(s_h, a_h) \phi_h^{\top}(s_h, a_h) v_h.
$$

Therefore,

$$
\begin{aligned}
|\langle \phi(\pi) - \widehat{\phi}(\pi), v \rangle| &\leq \sum_{h=1}^{H} \sum_{s_h, a_h} |\widehat{d}_h^{\pi}(s_h, a_h) - d_h^{\pi}(s_h, a_h)| \cdot |\phi_h^{\top}(s_h, a_h) v_h| \\
&\leq 2BR \sum_{h=1}^{H} \sum_{s_h, a_h} |\widehat{d}_h^{\pi}(s_h, a_h) - d_h^{\pi}(s_h, a_h)| \\
&\leq 2BR \sum_{h=1}^{H} h\epsilon' \leq BRH^2\epsilon',
\end{aligned}
$$

where the first step is due to the triangle inequality and the third step comes from Lemma 5. This concludes our proof.

D.4 Proof of Lemma 7

We use $\Delta\theta$ to denote $\theta^* - \hat{\theta}$ in this proof. From Lemma 6, we know that for any policy π,

$$
|\langle \phi(\pi) - \widehat{\phi}(\pi), \Delta\theta \rangle| \leq BRH^2\epsilon'.
$$

By the triangle inequality, this implies that for any policy π^0, π^1,

$$
|\langle \widehat{\phi}(\pi^0) - \widehat{\phi}(\pi^1), \Delta\theta \rangle| \leq |\langle \phi(\pi^0) - \phi(\pi^1), \Delta\theta \rangle| + 2BRH^2\epsilon'.
$$

Therefore we have for any policy π^0, π^1,

$$
|\langle \widehat{\phi}(\pi^0) - \widehat{\phi}(\pi^1), \Delta\theta \rangle|^2 \leq 2|\langle \phi(\pi^0) - \phi(\pi^1), \Delta\theta \rangle|^2 + 8(BRH^2\epsilon')^2. \tag{16}
$$

Note that from the definition of $\widehat{\Sigma}_{N+1}$ and Σ_{N+1}, we have

$$
\|\Delta\theta\|_{\widehat{\Sigma}_{N+1}}^2 = \Delta\theta^{\top} \Big(\lambda I + \sum_{n=1}^{N} (\widehat{\phi}(\pi^{n,0}) - \widehat{\phi}(\pi^{n,1}))(\widehat{\phi}(\pi^{n,0}) - \widehat{\phi}(\pi^{n,1}))^{\top} \Big) \Delta\theta
$$

$$
\begin{aligned}
&= \lambda\|\Delta\theta\|^2 + \sum_{n=1}^{N} |\langle \widehat{\phi}(\pi^{n,0}) - \widehat{\phi}(\pi^{n,1}), \Delta\theta\rangle|^2 \\
&\le 2\Big(\lambda\|\Delta\theta\|^2 + \sum_{n=1}^{N} |\langle \phi(\pi^{n,0}) - \phi(\pi^{n,1}), \Delta\theta\rangle|^2\Big) + 8(BRH^2\epsilon')^2 \\
&= 2\|\Delta\theta\|^2_{\Sigma_{N+1}} + 8(BRH^2\epsilon')^2,
\end{aligned}
$$

where the third step comes from (16). This implies that

$$
\|\Delta\theta\|_{\widehat{\Sigma}_{N+1}} \le \sqrt{2}\|\Delta\theta\|_{\Sigma_{N+1}} + 2\sqrt{2}BRH^2\epsilon',
$$

which concludes our proof.

E Proofs in Section 4 and Appendix B

E.1 Proof of Theorem 4

First note that Algorithm 4 provides us with the following guarantee:

Lemma 8. *We have with probability at least $1-\delta/6$ that*

$$
\mathbb{E}_{s_1\sim P_1(\cdot)}[V_1^{b/H,*}(s_1)] \le C_{\mathrm{lin}}\sqrt{d^3H^4R^2\cdot\log(\mathcal{N}_\Pi(\epsilon')dKHR/\delta)/K},
$$

where b_h is defined in Algorithm 5 and $C_{\mathrm{lin}} > 0$ is a universal constant. Here $V_1^{r,*}(s_1) := \max_{\pi\in\Pi} V_1^{r,\pi}(s_1)$.

Lemma 8 is adapted from Wang et al. (2020)[Lemma 3.2] and we highlight the difference of the proof in Appendix E.2. Then we consider a ϵ'-covering for Π, denoted by $\mathcal{C}(\Pi,\epsilon')$. Following the similar analysis in Wang et al. (2020)[Lemma 3.3], we have the following lemma:

Lemma 9. *With probability $1-\delta/6$, for all $h'\in[H]$, policy $\pi\in\mathcal{C}(\Pi,\epsilon')$ and linear reward function r with $r_h\in[-1,1]$, we have*

$$
Q_{h'}^{r,\pi}(\cdot,\cdot) \le \widehat{Q}_{h'}^{r,\pi}(\cdot,\cdot) \le r_{h'}(\cdot,\cdot) + \sum_{s'} P_{h'}^{*}(s'|\cdot,\cdot)\widehat{V}_{h'+1}^{r,\pi}(s') + 2b_{h'}(\cdot,\cdot).
$$

The proof of Lemma 9 is deferred to Appendix E.3. Denote the event in Lemma 8 and Lemma 9 by $\mathcal{E}_4$ and $\mathcal{E}_5$ respectively. Then under event $\mathcal{E}_4\cap\mathcal{E}_5$, we have for all policy $\pi\in\mathcal{C}(\Pi,\epsilon')$ and all linear reward function r with $r_h\in[-1,1]$,

$$
\begin{aligned}
0 &\le \mathbb{E}_{s_1\sim P_1^*(\cdot)}[\widehat{V}_1^{r,\pi}(s_1) - V_1^{r,\pi}(s_1)] \le 2\mathbb{E}_{s_1\sim P_1^*(\cdot)}[V_1^{b,\pi}(s_1)] \\
&\le 2H\mathbb{E}_{s_1\sim P_1(\cdot)}[V_1^{b/H,*}(s_1)] \le 2C_{\mathrm{lin}}\sqrt{\frac{d^3H^6R^2\cdot\log(dKHR\mathcal{N}_\Pi(\epsilon')/\delta)}{K}} \le \epsilon_0, \qquad (17)
\end{aligned}
$$

where $\epsilon_0 = \frac{\epsilon}{72BR\sqrt{Hd}}$. Here the first step comes from the left part of Lemma 9 and the second step is due to the right part of Lemma 9.

Note that in the proof of Lemma 13, we calculate the covering number of the function class $\{\widehat{V}_1^{r,\pi}: r \text{ is linear and } r_h\in[-1,1]\}$ for any fixed π in (24). Then by Azuma-Hoeffding's inequality and (24), we have with probability at least $1-\delta/6$ that for all policy $\pi\in\mathcal{C}(\Pi,\epsilon')$ and all linear reward function r with $r_h\in[-1,1]$ that

$$
\Big|\mathbb{E}_{s_1\sim P_1^*(\cdot)}[\widehat{V}_1^{r,\pi}(s_1)] - \frac{1}{K}\sum_{i=1}^{K}\widehat{V}_1^{r,\pi}(s_1^{i,\mathrm{in}})\Big| \le C_3H\cdot\sqrt{\frac{\log(\mathcal{N}_\Pi(\epsilon')HKdR/\delta)}{K}} \le \epsilon_0, \qquad (18)
$$

where $C_3>0$ is a universal constant.

Combining (17) and (18), we have with probability at least $1-\delta/2$ that for all policy $\pi\in\mathcal{C}(\Pi,\epsilon')$ and all linear reward function r with $r_h\in[-1,1]$

$$
|\widehat{V}^{\pi}(r) - V^{r,\pi}| \le 2\epsilon_0. \qquad (19)
$$

This implies that we can estimate the value function for all $\pi \in \mathcal{C}(\Pi, \epsilon')$ and linear reward function r with $r_h \in [-1, 1]$ up to estimation error $2\epsilon_0$.

Now we consider any policy $\pi \in \Pi$. Suppose that $\pi' \in \mathcal{C}(\Pi, \epsilon')$ satisfies that

$$\max_{s \in \mathcal{S}, h \in [H]} \|\pi_h(\cdot|s) - \pi'_h(\cdot|s)\|_1 \leq \epsilon'. \tag{20}$$

Then we can bound $|\widehat{V}^{\pi}(r) - \widehat{V}^{\pi'}(r)|$ and $|V^{r,\pi} - V^{r,\pi'}|$ for all linear reward function r with $r_h \in [-1, 1]$ respectively.

For $|V^{r,\pi} - V^{r,\pi'}|$, note that we have the following performance difference lemma:

Lemma 10. *For any policy π, π' and reward function r, we have*

$$V^{r,\pi'} - V^{r,\pi} = \sum_{h=1}^{H} \mathbb{E}_{\pi', P^*} \Big[\langle Q_h^{r,\pi}(s_h, \cdot), \pi'_h(\cdot|s) - \pi_h(\cdot|s) \rangle \Big].$$

The proof is deferred to Appendix E.4. Therefore from Lemma 10 we have

$$\begin{aligned} |V^{r,\pi'} - V^{r,\pi}| &\leq \sum_{h'=1}^{H} \mathbb{E}_{\pi, P^*} \Big[\big| \langle Q_{h'}^{r^{h,j},\pi'}(s_{h'}, \cdot), \pi_{h'}(\cdot|s) - \pi'_{h'}(\cdot|s) \rangle \big| \Big] \\ &\leq \sum_{h'=1}^{H} \mathbb{E}_{\pi, P^*} \Big[\|\pi_{h'}(\cdot|s) - \pi'_{h'}(\cdot|s)\|_1 \Big] \leq H\epsilon'. \end{aligned} \tag{21}$$

On the other hand, we have the following lemma to bound $|\widehat{V}^{\pi}(r) - \widehat{V}^{\pi'}(r)|$:

Lemma 11. *Suppose (20) holds and $\widehat{V}^{\pi}(r), \widehat{V}^{\pi'}(r)$ are calculated as in Algorithm 5. Then for all linear reward function r with $0 \leq r(\tau) \leq r_{\max}$, we have*

$$|\widehat{V}^{\pi}(r) - \widehat{V}^{\pi'}(r)| \leq \epsilon_{\text{cover}} := \frac{H\epsilon'}{\sqrt{dK} - 1} \cdot (dK)^{\frac{H}{2}}. \tag{22}$$

The proof is deferred to Appendix E.5.

Combining (19),(21) and (22), we have for all policy $\pi \in \Pi$ and linear reward function r with $0 \leq r(\tau) \leq r_{\max}$,

$$|\widehat{V}^{\pi}(r) - V^{r,\pi}| \leq 2\epsilon_0 + H\epsilon' + \epsilon_{\text{cover}}. \tag{23}$$

In particular, since $(\phi(\pi))_{h,j} = RV^{r^{h,j},\pi}$, we have for all policy $\pi \in \Pi$ and $h \in [H], j \in [d]$,

$$|(\phi(\pi))_{h,j} - (\widehat{\phi}(\pi))_{h,j}| \leq 2R\epsilon_0 + HR\epsilon' + R\epsilon_{\text{cover}}.$$

This implies that for all policy $\pi \in \Pi$ and any v defined in Lemma 6, we have

$$|\langle (\phi(\pi)) - (\widehat{\phi}(\pi)), v \rangle| \leq 2BH\sqrt{d}(2R\epsilon_0 + HR\epsilon' + R\epsilon_{\text{cover}}).$$

The rest of the proof is the same as Theorem 1 and thus is omitted here. The only difference is that we need to show $\widehat{\pi}$ is a near-optimal policy with respect to $\widehat{r}$. This can be proved as follows:

$$\begin{aligned} V^{\widehat{r},*} - V^{\widehat{r},\widehat{\pi}} &= \Big(V^{\widehat{r},*} - \widehat{V}^{\widehat{\pi}}(\widehat{r}) \Big) + \Big(\widehat{V}^{\widehat{\pi}}(\widehat{r}) - \widehat{V}^{\pi^*(\widehat{r})}(\widehat{r}) \Big) + \Big(\widehat{V}^{\pi^*(\widehat{r})}(\widehat{r}) - V^{\widehat{r},\widehat{\pi}} \Big) \\ &\leq 4\epsilon_0 + 2H\epsilon' + 2\epsilon_{\text{cover}}, \end{aligned}$$

where the last step comes from (23) and the definition of $\widehat{\pi}$.

E.2 Proof of Lemma 8

Here we outline the difference of the proof from Wang et al. (2020)[Lemma 3.2]. First, we also have the following concentration guarantee:

Lemma 12. *Fix a policy π. Then with probability at least $1-\delta$, we have for all $h \in [H]$ and $k \in [K]$,*

$$\left\| \sum_{i=1}^{k} \phi_h^i \left(V_{h+1}^k(s_{h+1}^i) - \sum_{s' \in \mathcal{S}} P_h^*(s'|s_h^i, a_h^i) V_{h+1}^k(s') \right) \right\|_{\Lambda_h^{-1}} \leq O\big(dHR\sqrt{\log(dKHR/\delta)}\big).$$

The proof is almost the same as Lemma 13 and thus is omitted here. Then following the same arguments in Wang et al. (2020), we have the following inequality under Lemma 12:

$$\left| \phi_h(s,a)^\top w_h^k - \sum_{s' \in \mathcal{S}} P_h^*(s'|s,a) V_{h+1}^k(s') \right| \leq \beta_{\text{ex}} \|\phi_h(s,a)\|_{(\Lambda_h^k)^{-1}}.$$

Note that $V_{h+1}^k(s) \in [0, H-h]$ for all $s \in \mathcal{S}$, which implies that

$$0 \leq \sum_{s' \in \mathcal{S}} P_h^*(s'|s,a) V_{h+1}^k(s') + r_h^k(s,a) \leq H - h + 1.$$

Note that Clip is a contraction operator, which implies that

$$\begin{aligned} &\left| \text{Clip}_{[0,H-h+1]}((w_h^k)^\top \phi_h(s,a) + r_h^k(s,a)) - \left(\sum_{s' \in \mathcal{S}} P_h^*(s'|s,a) V_{h+1}^k(s') + r_h^k(s,a) \right) \right| \\ &\leq \left| (w_h^k)^\top \phi_h(s,a) - \sum_{s' \in \mathcal{S}} P_h^*(s'|s,a) V_{h+1}^k(s') \right| \leq \beta_{\text{ex}} \|\phi_h(s,a)\|_{(\Lambda_h^k)^{-1}}. \end{aligned}$$

On the other hand,

$$\left| \text{Clip}_{[0,H-h+1]}((w_h^k)^\top \phi_h(s,a) + r_h^k(s,a)) - \left(\sum_{s' \in \mathcal{S}} P_h^*(s'|s,a) V_{h+1}^k(s') + r_h^k(s,a) \right) \right| \leq H - h + 1.$$

This implies that

$$\left| \text{Clip}_{[0,H-h+1]}((w_h^k)^\top \phi_h(s,a) + r_h^k(s,a)) - \left(\sum_{s' \in \mathcal{S}} P_h^*(s'|s,a) V_{h+1}^k(s') + r_h^k(s,a) \right) \right| \leq b_h^k(s,a).$$

The rest of the proof is the same as Wang et al. (2020) and thus is omitted.

E.3 Proof of Lemma 9

In the following discussion we will use ϕ_h^i to denote $\phi_h(s_h^i, a_h^i)$. First we need the following concentration lemma which is similar to Jin et al. (2020b)[Lemma B.3]:

Lemma 13. *Fix a policy π. Then with probability at least $1-\delta$, we have for all $h \in [H]$ and linear reward functions r with $r_h \in [-1, 1]$,*

$$\left\| \sum_{i=1}^{K} \phi_h^i \left(\widehat{V}_{h+1}^{r,\pi}(s_{h+1}^i) - \sum_{s' \in \mathcal{S}} P_h^*(s'|s_h^i, a_h^i) \widehat{V}_{h+1}^{r,\pi}(s') \right) \right\|_{\Lambda_h^{-1}} \leq O\big(dHR\sqrt{\log(dKHR/\delta)}\big).$$

The proof is deferred to Appendix E.6. Then by union bound, we know with probability $1-\delta/6$, we have for all policy $\pi \in \mathcal{C}(\Pi, \epsilon')$, $h \in [H]$ and linear reward functions r with $r_h \in [-1, 1]$ that

$$\left\| \sum_{i=1}^{K} \phi_h^i \left(\widehat{V}_{h+1}^{r,\pi}(s_{h+1}^i) - \sum_{s' \in \mathcal{S}} P_h^*(s'|s_h^i, a_h^i) \widehat{V}_{h+1}^{r,\pi}(s') \right) \right\|_{\Lambda_h^{-1}} \leq O\big(dHR\sqrt{\log(dKHR\mathcal{N}_\Pi(\epsilon')/\delta)}\big).$$

Let $\mathcal{E}_6$ denote the event thar the above inequality holds. Then under $\mathcal{E}_6$, following the same analysis in (Wang et al., 2020)[Lemma 3.1], we have for all policy $\pi \in \mathcal{C}(\Pi, \epsilon')$, $(s,a) \in \mathcal{S} \times \mathcal{A}$, $h \in [H]$ and linear reward functions r with $r_h \in [-1, 1]$ that

$$\left| \phi_h(s,a)^\top w_h^{r,\pi} - \sum_{s' \in \mathcal{S}} P_h^*(s'|s,a) \widehat{V}_{h+1}^{r,\pi}(s') \right| \leq \beta_{\text{pl}} \|\phi_h(s,a)\|_{\Lambda_h^{-1}}.$$

Form the contraction property of Clip and the fact that $\sum_{s'\in\mathcal{S}} P_h^*(s'|s,a)\widehat{V}_{h+1}^{r,\pi}(s') + r_h(s,a) \in [-(H-h+1), H-h+1]$, we know

$$\left|\text{Clip}_{[-(H-h+1),H-h+1]}((w_h^{r,\pi})^\top\phi_h(s,a) + r_h(s,a)) - \sum_{s'\in\mathcal{S}} P_h^*(s'|s,a)\widehat{V}_{h+1}^{r,\pi}(s') - r_h(s,a)\right| \le b_h(s,a)$$

Therefore, under $\mathcal{E}_6$ we have

$$\widehat{Q}_h^{r,\pi}(s,a) \le r_h(s,a) + \sum_{s'} P_h^*(s'|s,a)\widehat{V}_{h+1}^{r,\pi}(s') + 2b_h(s,a).$$

Now we only need to prove under $\mathcal{E}_6$, for all policy $\pi \in \mathcal{C}(\Pi,\epsilon')$, $(s,a)\in\mathcal{S}\times\mathcal{A}$, $h\in[H]$ and linear reward function r with $r_h \in [-1,1]$, we have $Q_h^{r,\pi}(s,a) \le \widehat{Q}_h^{r,\pi}(s,a)$. We use induction to prove this. The claim holds obviously for $h = H+1$. Then we suppose for some $h\in[H]$, we have $Q_{h+1}^{r,\pi}(s,a) \le \widehat{Q}_{h+1}^{r,\pi}(s,a)$ for all policy $\pi\in\mathcal{C}(\Pi,\epsilon')$, $(s,a)\in\mathcal{S}\times\mathcal{A}$ and linear reward function r with $r_h\in[-1,1]$. Then we have:

$$V_{h+1}^{r,\pi}(s) = \mathbb{E}_{a\sim\pi_{h+1}(\cdot|s)}\left[Q_{h+1}^{r,\pi}(s,a)\right] \le \widehat{V}_{h+1}^{r,\pi}(s) = \mathbb{E}_{a\sim\pi_{h+1}(\cdot|s)}\left[\widehat{Q}_{h+1}^{r,\pi}(s,a)\right].$$

This implies that

$$\text{Clip}_{[-(H-h+1),H-h+1]}((w_h^{r,\pi})^\top\phi_h(s,a) + r_h(s,a)) + b_h(s,a) \ge \sum_{s'\in\mathcal{S}} P_h^*(s'|s,a)V_{h+1}^{r,\pi}(s') + r_h(s,a) = Q_h^{r,\pi}(s,a).$$

On the other hand we have

$$Q_h^{r,\pi}(s,a) \le H-h+1.$$

Therefore we have

$$Q_h^{r,\pi}(s,a) \le \widehat{Q}_h^{r,\pi}(s,a).$$

By induction we can prove the lemma.

E.4 Proof of Lemma 10

For any two policies π' and π, it follows from the definition of $V^{r,\pi'}$ and $V^{r,\pi}$ that

$$\begin{aligned}
&V^{r,\pi'} - V^{r,\pi}\\
=&\mathbb{E}_{\pi',P^*}\left[r_1(s_1,a_1) + V_2^{r,\pi'}(s_2)\right] - \mathbb{E}_{\pi',P^*}\left[V_1^{r,\pi}(s_1)\right]\\
=&\mathbb{E}_{\pi',P^*}\left[V_2^{r,\pi'}(s_2) - (V_1^{r,\pi}(s_1) - r_1(s_1,a_1))\right]\\
=&\mathbb{E}_{\pi',P^*}\left[V_2^{r,\pi'}(s_2) - V_2^{r,\pi}(s_2)\right] + \mathbb{E}_{\pi',P^*}\left[Q_1^{r,\pi}(s_1,a_1) - V_1^{r,\pi}(s_1)\right]\\
=&\mathbb{E}_{\pi',P^*}\left[V_2^{r,\pi'}(s_2) - V_2^{r,\pi}(s_2)\right] + \mathbb{E}_{\pi',P^*}\left[\langle Q_1^{r,\pi}(s_1,\cdot), \pi_1'(\cdot|s_1) - \pi_1(\cdot|s_1)\rangle\right]\\
=&\cdots = \sum_{h=1}^{H}\mathbb{E}_{\pi',P^*}\left[\langle Q_h^{r,\pi}(s_h,\cdot), \pi_h'(\cdot|s) - \pi_h(\cdot|s)\rangle\right].
\end{aligned}$$

This concludes our proof.

E.5 Proof of Lemma 11

For any $h'\in[H]$, suppose $\max_{s\in\mathcal{S}}|\widehat{V}_{h'+1}^{r,\pi}(s) - \widehat{V}_{h'+1}^{r,\pi'}(s)| \le \epsilon_{h'+1}$, then for any $s\in\mathcal{S}, a\in\mathcal{A}$, we have

$$\begin{aligned}
|\widehat{Q}_{h'}^{r,\pi}(s,a) - \widehat{Q}_{h'}^{r,\pi'}(s,a)| &\le |(w_{h'}^{r,\pi} - w_{h'}^{r,\pi'})^\top\phi_{h'}(s,a)|\\
&\le \epsilon_{h'+1}\sum_{i=1}^{K}|\phi_{h'}(s,a)^\top(\Lambda_{h'})^{-1}\phi_{h'}(s_{h'}^i, a_{h'}^i)|
\end{aligned}$$

$$\leq \epsilon_{h'+1}\sqrt{\Big[\sum_{i=1}^{K}\|\phi_{h'}(s,a)\|^2_{(\Lambda_{h'})^{-1}}\Big]\cdot\Big[\sum_{i=1}^{K}\|\phi_{h'}(s^i_{h'},a^i_{h'})\|^2_{(\Lambda_{h'})^{-1}}\Big]} \leq \epsilon_{h'+1}\sqrt{dK}.$$

Here the final step is comes from the auxiliary Lemma 14 and the fact that $\Lambda_{h'} \geq R^2 I$ and thus $\sum_{i=1}^{K}\|\phi_{h'}(s,a)\|^2_{(\Lambda_{h'})^{-1}} \leq \sum_{i=1}^{K} 1 \leq K$.

Therefore we have

$$\epsilon_{h'} := \max_{s\in\mathcal{S}} |\widehat{V}^{r,\pi}_{h'}(s) - \widehat{V}^{r,\pi'}_{h'}(s)| \leq H\epsilon' + \sqrt{dK}\epsilon_{h'+1}.$$

Note that $\epsilon_{H+1} = 0$, therefore we have

$$\epsilon_1 \leq \frac{H\epsilon'}{\sqrt{dK}-1}\cdot(dK)^{\frac{H}{2}},$$

This concludes our proof.

E.6 Proof of Lemma 13

The proof is almost the same as Jin et al. (2020b)[Lemma B.3] except that the function class of $V^{r,\pi}_h$ is different. Therefore we only need to bound the covering number $\mathcal{N}_{\mathcal{V}}(\epsilon)$ of $V^{r,\pi}_h$ where the distance is defined as $\mathrm{dist}(V,V') = \sup_s |V(s)-V'(s)|$. Note that $V^{r,\pi}_h$ belongs to the following function class:

$$\mathcal{V} = \Bigg\{V_{w,A}(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\Bigg[\mathrm{Clip}_{[-(H-h+1),H-h+1]}\Bigg(\mathrm{Clip}_{[-(H-h+1),H-h+1]}(w^\top\phi_{h'}(s,a)) \\ +\mathrm{Clip}_{[0,2(H-h+1)]}(\|\phi(s,a)\|_A)\Bigg)\Bigg], \forall s\in\mathcal{S}\Bigg\},$$

where the parameters (w,A) satisfy $\|w\| \leq 2H\sqrt{dK/\lambda_{\mathrm{pl}}}$, $\|A\| \leq \beta^2_{\mathrm{pl}}\lambda^{-1}_{\mathrm{pl}}$.

Note that for any $V_{w_1,A_1}, V_{w_2,A_2} \in \mathcal{V}$, we have

$$\begin{aligned}
\mathrm{dist}(V_{w_1,A_1},V_{w_2,A_2}) &\leq \sup_{s,a}\Big|\Big[\mathrm{Clip}_{[-(H-h+1),H-h+1]}(w_1^\top\phi_{h'}(s,a)) + \mathrm{Clip}_{[0,2(H-h+1)]}(\|\phi(s,a)\|_{A_1})\Big] \\
&\quad - \Big[\mathrm{Clip}_{[-(H-h+1),H-h+1]}(w_2^\top\phi_{h'}(s,a)) + \mathrm{Clip}_{[0,2(H-h+1)]}(\|\phi(s,a)\|_{A_2})\Big]\Big| \\
&\leq \sup_{s,a}\Big|\mathrm{Clip}_{[-(H-h+1),H-h+1]}(w_1^\top\phi_{h'}(s,a)) - \mathrm{Clip}_{[-(H-h+1),H-h+1]}(w_2^\top\phi_{h'}(s,a))\Big| \\
&\quad + \sup_{s,a}\Big|\mathrm{Clip}_{[0,2(H-h+1)]}(\|\phi(s,a)\|_{A_1}) - \mathrm{Clip}_{[0,2(H-h+1)]}(\|\phi(s,a)\|_{A_2})\Big| \\
&\leq R\sup_{\|\phi\|\leq 1}\Big|(w_1-w_2)^\top\phi\Big| + R\sup_{\|\phi\|\leq 1}\sqrt{\Big|\phi^\top(A_1-A_2)\phi\Big|} \\
&\leq R(\|w_1-w_2\| + \sqrt{\|A_1-A_2\|_F}),
\end{aligned}$$

where the first and third step utilize the contraction property of Clip. Let $\mathcal{C}_w$ be the $\epsilon/(2R)$-cover of $\{w\in\mathbb{R}^d : \|w\| \leq 2r_{\max}\sqrt{dK/\lambda_{\mathrm{pl}}}\}$ w.r.t. ℓ_2-norm and $\mathcal{C}_A$ be the $(\epsilon/2R)$-cover of $\{A\in\mathbb{R}^{d\times d} : \|A\| \leq \beta^2_{\mathrm{pl}}\lambda^{-1}_{\mathrm{pl}}\}$ w.r.t. the Frobenius norm, then from the literature Jin et al. (2020b)[Lemma D.5], we have

$$\mathcal{N}_{\mathcal{V}}(\epsilon) \leq \log|\mathcal{C}_w| + \log|\mathcal{C}_A| \leq d\log\Big(1+8\sqrt{dKr^2_{\max}R^2/(\lambda_{\mathrm{pl}}\epsilon^2)}\Big) + d^2\log\Big[1+8d^{1/2}\beta^2_{\mathrm{pl}}R^2/(\lambda_{\mathrm{pl}}\epsilon^2)\Big]. \tag{24}$$

The rest of the proof follows Jin et al. (2020b)[Lemma B.3] directly so we omit it here.

E.7 Proof of Proposition 1

First consider ζ and ζ' which satisfies:

$$\|\zeta_h - \zeta'_h\| \le \epsilon_z, \forall h \in [H].$$

Then we know for any $h \in [H], s \in \mathcal{S}, a \in \mathcal{A}$,

$$|\zeta_h^\top \phi_h(s,a) - (\zeta'_h)^\top \phi_h(s,a)| \le \epsilon_z R. \tag{25}$$

Now fix any $h \in [H]$ and $s \in \mathcal{S}$. To simplify writing, we use $x(a)$ and $x'(a)$ to denote $\zeta_h^\top \phi_h(s,a)$ and $(\zeta'_h)^\top \phi_h(s,a)$ respectively. Without loss of generality, we assume $\sum_a \exp(x(a)) \le \sum_a \exp(x'(a))$. Then from (25) we have

$$\sum_a \exp(x(a)) \le \sum_a \exp(x'(a)) \le \exp(\epsilon_z R) \sum_a \exp(x(a)).$$

Note that we have

$$\begin{aligned}\|\pi_h^\zeta(\cdot|s) - \pi_h^{\zeta'}(\cdot|s)\|_1 &= \sum_a \Big| \frac{\exp(x(a))}{\sum_{a'} \exp(x(a'))} - \frac{\exp(x'(a))}{\sum_{a'} \exp(x'(a'))} \Big| \\ &= \frac{\sum_a \Big| \exp(x(a)) \sum_{a'} \exp(x'(a')) - \exp(x'(a)) \sum_{a'} \exp(x(a')) \Big|}{\sum_{a'} \exp(x(a')) \cdot \sum_{a'} \exp(x'(a'))}.\end{aligned}$$

For any $a \in \mathcal{A}$, if $\exp(x(a)) \sum_{a'} \exp(x'(a')) - \exp(x'(a)) \sum_{a'} \exp(x(a')) \ge 0$, then

$$\begin{aligned}&\Big| \exp(x(a)) \sum_{a'} \exp(x'(a')) - \exp(x'(a)) \sum_{a'} \exp(x(a')) \Big| \\ \le& \exp(\epsilon_z R) \exp(x(a)) \sum_{a'} \exp(x(a')) - \exp(-\epsilon_z R) \exp(x(a)) \sum_{a'} \exp(x(a')) \\ =& (\exp(\epsilon_z R) - \exp(-\epsilon_z R)) \exp(x(a)) \sum_{a'} \exp(x(a')).\end{aligned}$$

Otherwise, we have

$$\begin{aligned}&\Big| \exp(x(a)) \sum_{a'} \exp(x'(a')) - \exp(x'(a)) \sum_{a'} \exp(x(a')) \Big| \\ \le& \exp(\epsilon_z R) \exp(x(a)) \sum_{a'} \exp(x(a')) - \exp(x(a)) \sum_{a'} \exp(x(a')) \\ =& (\exp(\epsilon_z R) - 1) \exp(x(a)) \sum_{a'} \exp(x(a')).\end{aligned}$$

Therefore we have

$$\|\pi_h^\zeta(\cdot|s) - \pi_h^{\zeta'}(\cdot|s)\|_1 \le \frac{(\exp(\epsilon_z R) - \exp(-\epsilon_z R)) \sum_a \exp(x(a)) \sum_{a'} \exp(x(a'))}{\sum_{a'} \exp(x(a')) \cdot \sum_{a'} \exp(x'(a'))} \le \exp(2\epsilon_z R) - 1.$$

This implies that for any $\epsilon \le 1$,

$$\mathcal{N}_\Pi(\epsilon) \le \Big(\mathcal{N}_{\mathbb{B}(d,W)} \Big(\frac{\ln 2}{2R} \epsilon \Big) \Big)^H \le \Big(\frac{12WR}{\epsilon} \Big)^{Hd},$$

where the first step uses $\exp(x) - 1 \le x / \ln 2$ when $x \le \ln 2$. This concludes our proof.

E.8 Proof of Proposition 2

First we consider the following entropy-regularized RL problem where we try to maximize the following objective for some $\alpha > 0$:

$$\max_\pi V_\alpha(r^*, \pi) := \mathbb{E}_{\pi, P_*} \Big[\sum_{h=1}^H r_h^*(s_h, a_h) - \alpha \log \pi_h(a_h|s_h) \Big].$$

From the literature (Nachum et al., 2017; Cen et al., 2022), we know that we can define corresponding optimal regularized value function and Q function as follows:

$$Q^*_{\alpha,h}(s,a) = r^*_h(s,a) + \mathbb{E}_{s_{h+1}\sim P^*_h(\cdot|s,a)}\big[V^*_{\alpha,h+1}\big],$$
$$V^*_{\alpha,h}(s) = \max_{\pi_h} \mathbb{E}_{a_h\sim\pi_h(\cdot|s)}\big[Q^*_{\alpha,h}(s,a_h) - \alpha\log\pi_h(a_h|s)\big],$$

where $V^*_{\alpha,H+1}(s) = 0$ for all $s \in \mathcal{S}$. Note that we have $V^*_{\alpha,h}(s) \le H(1+\alpha\log|\mathcal{A}|)$ for all $s\in\mathcal{S}$ and $h\in[H]$. The global optimal regularized policy is therefore

$$\pi^*_{\alpha,h}(a|s) = \frac{\exp(Q^*_{\alpha,h}(s,a)/\alpha)}{\sum_{a'}\exp(Q^*_{\alpha,h}(s,a')/\alpha)}.$$

In particular, in linear MDPs, we have

$$Q^*_{\alpha,h}(s,a) = \phi_h(s,a)^\top\left(\theta^*_h + \int_{s\in\mathcal{S}} \mu^*_h(s)V^*_{\alpha,h+1}(s)ds\right).$$

Therefore, $Q^*_{\alpha,h}(s,a) = \phi_h(s,a)^\top w^*_{\alpha,h}$ where

$$\|w^*_{\alpha,h}\| \le B + H(1+\alpha\log|\mathcal{A}|)\sqrt{d}.$$

This implies that π^*_α belongs to the log-linear policy class Π with $W = (B+H(1+\alpha\log|\mathcal{A}|)\sqrt{d})/\alpha$.

On the other hand, let π_{g} denote the global unregularized optimal policy, then

$$\begin{aligned}
V^*(r^*,\pi_{\mathrm{g}}) - \max_{\pi\in\Pi} V(r^*,\pi) &\le V^*(r^*,\pi_{\mathrm{g}}) - V(r^*,\pi^*_\alpha)\\
&= \big(V^*(r^*,\pi_{\mathrm{g}}) - V^*_\alpha(r^*,\pi_{\mathrm{g}})\big) + \big(V^*_\alpha(r^*,\pi_{\mathrm{g}}) - V^*_\alpha(r^*,\pi^*_\alpha)\big) + \big(V^*_\alpha(r^*,\pi^*_\alpha) - V^*(r^*,\pi^*_\alpha)\big)\\
&\le V^*_\alpha(r^*,\pi^*_\alpha) - V^*(r^*,\pi^*_\alpha) \le \alpha H\log|\mathcal{A}|.
\end{aligned}$$

Therefore we only need to let $\alpha = \frac{\epsilon}{H\log|\mathcal{A}|}$ to ensure $V(r^*,\pi_{\mathrm{g}}) - \max_{\pi\in\Pi} V(r^*,\pi) \le \epsilon$.

F Proof of Theorem 3

First from performance difference lemma (Lemma 10), we have

$$\begin{aligned}
V^{r^*,\widehat{\pi}} - V^{r^*,*} &= \sum_{h=1}^H \mathbb{E}_{s_h\sim d^{\widehat{\pi}}_h}[Q^*_h(s_h,\widehat{\pi}) - Q^*_h(s_h,\pi^*)]\\
&= \sum_{h=1}^H \mathbb{E}_{s_h\sim d^{\widehat{\pi}}_h}[Q^*_h(s_h,\widehat{\pi}) - \widehat{A}_h(s_h,\widehat{\pi})] + \mathbb{E}_{s_h\sim d^{\widehat{\pi}}_h}[\widehat{A}_h(s_h,\widehat{\pi}) - \widehat{A}_h(s_h,\pi^*)]\\
&\qquad + \mathbb{E}_{s_h\sim d^{\widehat{\pi}}_h}[\widehat{A}_h(s_h,\pi^*) - Q^*_h(s_h,\pi^*)]\\
&\ge \sum_{h=1}^H \mathbb{E}_{s_h\sim d^{\widehat{\pi}}_h}[Q^*_h(s_h,\widehat{\pi}) - \widehat{A}_h(s_h,\widehat{\pi})] + \mathbb{E}_{s_h\sim d^{\widehat{\pi}}_h}[\widehat{A}_h(s_h,\pi^*) - Q^*_h(s_h,\pi^*)]\\
&= \sum_{h=1}^H \mathbb{E}_{s_h\sim d^{\widehat{\pi}}_h}[A^*_h(s_h,\widehat{\pi}) - \widehat{A}_h(s_h,\widehat{\pi}) + \widehat{A}_h(s_h,\pi^*) - A^*_h(s_h,\pi^*)]\\
&= \sum_{h=1}^H \mathbb{E}_{s_h\sim d^{\widehat{\pi}}_h}[\langle\phi_h(s_h,\widehat{\pi}),\xi^*_h - \widehat{\xi}_h\rangle - \langle\phi_h(s_h,\pi^*),\xi^*_h - \widehat{\xi}_h\rangle]\\
&= \sum_{h=1}^H \mathbb{E}_{s_h\sim d^{\widehat{\pi}}_h}[\langle\phi_h(s_h,\widehat{\pi}) - \phi_h(s_h,\pi^*),\xi^*_h - \widehat{\xi}_h\rangle]\\
&\ge -\sum_{h=1}^H \|\mathbb{E}_{s_h\sim d^{\widehat{\pi}}_h}[\phi_h(s_h,\widehat{\pi}) - \phi_h(s_h,\pi^*)]\|_{\Sigma^{-1}_{h,N+1}}\cdot\|\xi^*_h - \widehat{\xi}_h\|_{\Sigma_{h,N+1}}. \qquad (26)
\end{aligned}$$

Next we will bound $\|\mathbb{E}_{s_h\sim d_h^{\widehat{\pi}}}[\phi_h(s_h,\widehat{\pi})-\phi_h(s_h,\pi^*)]\|_{\Sigma_{h,N+1}^{-1}}$ and $\|\xi^*-\widehat{\xi}\|_{\Sigma_{h,N+1}}$ respectively. First for $\|\mathbb{E}_{s_h\sim d_h^{\widehat{\pi}}}[\phi_h(s_h,\widehat{\pi})-\phi_h(s_h,\pi^*)]\|_{\Sigma_{h,N+1}^{-1}}$, notice that $\Sigma_{h,N+1}\succeq\Sigma_{h,n}$ for all $n\in[N+1]$, which implies

$$\begin{aligned}
\|\mathbb{E}_{s_h\sim d_h^{\widehat{\pi}}}[\phi_h(s_h,\widehat{\pi})-\phi_h(s_h,\pi^*)]\|_{\Sigma_{h,N+1}^{-1}} &\le \frac{1}{N}\sum_{n=1}^{N}\|\mathbb{E}_{s_h\sim d_h^{\widehat{\pi}}}[\phi_h(s_h,\widehat{\pi})-\phi_h(s_h,\pi^*)]\|_{\Sigma_{h,n}^{-1}}\\
&\le \frac{1}{N}\sum_{n=1}^{N}\|\mathbb{E}_{s_h\sim\pi^{h,n,0}}[\phi_h(s_h,\pi^{h,n,0})-\phi_h(s_h,\pi^{h,n,1})]\|_{\Sigma_{h,n}^{-1}}\\
&\le \frac{1}{\sqrt{N}}\sqrt{\sum_{n=1}^{N}\|\mathbb{E}_{s_h\sim\pi^{h,n,0}}[\phi_h(s_h,\pi^{h,n,0})-\phi_h(s_h,\pi^{h,n,1})]\|_{\Sigma_{h,n}^{-1}}^2}\\
&\le \sqrt{\frac{2d\log(1+N/d)}{N}},
\end{aligned}\tag{27}$$

where the third step comes from the definition of $\pi^{h,n,0}$ and $\pi^{h,n,1}$ and the last step comes from Elliptical Potential Lemma (Lemma 1) and the fact that $\lambda\ge 4R^2$.

For $\|\xi_h^*-\widehat{\xi}_h\|_{\Sigma_{h,N+1}}$, let $\widetilde{\Sigma}_{h,n}$ denote $\lambda I+\sum_{i=1}^{n-1}(\phi_h(s^{h,n},a^{h,n,0})-\phi_h(s^{h,n},a^{h,n,1}))(\phi_h(s^{h,n},a^{h,n,0})-\phi_h(s^{h,n},a^{h,n,1}))^\top$. Then similar to Lemma 3, we have with probability at least $1-\delta/2$,

$$\|\xi_h^*-\widehat{\xi}_h\|_{\Sigma_{h,N+1}}^2\le 2\|\xi_h^*-\widehat{\xi}_h\|_{\widetilde{\Sigma}_{h,N+1}}^2+2C_{\mathrm{CON}}dR^2B^2\log(N/\delta).\tag{28}$$

On the other hand, similar to Lemma 2, MLE guarantees us that with probability at least $1-\delta/2$,

$$\|\widehat{\xi}_h-\xi_h^*\|_{\widetilde{\Sigma}_{h,N+1}}\le 2C_{\mathrm{MLE}}\cdot\sqrt{\kappa_{\mathrm{adv}}^2(d+\log(1/\delta))+\lambda B^2},\tag{29}$$

where $\kappa_{\mathrm{adv}}=2+\exp(2B_{\mathrm{adv}})+\exp(-2B_{\mathrm{adv}})$.

Therefore combining (28) and (29), we have with probability at least $1-\delta$,

$$\|\xi^*-\widehat{\xi}_h\|_{\Sigma_{h,N+1}}\le\mathcal{O}\big(\kappa_{\mathrm{adv}}BR\sqrt{\lambda d\log(N/\delta)}\big).\tag{30}$$

Thus combining (26), (27) and (30) via union bound, we have $V^*(r^*)-V(r^*,\widehat{\pi})\le\epsilon$ with probability at least $1-\delta$ as long as

$$N\ge\widetilde{\mathcal{O}}\Big(\frac{\lambda\kappa_{\mathrm{adv}}^2B^2R^2H^2d^2\log(1/\delta)}{\epsilon^2}\Big).$$

G Auxiliary Lemmas

Lemma 14 (Jin et al. (2020b)[Lemma D.1]). *Let $\Lambda=\lambda I+\sum_{i=1}^{K}\phi_i\phi_i^\top$ where $\phi_i\in\mathbb{R}^d$ and $\lambda>0$, then we have $\sum_{i=1}^{K}\phi_i^\top\Lambda^{-1}\phi_i\le d$.*