A EXAMPLES

This section gives additional examples of Low-rank Multi-task Bilinear class and demonstrates the sample complexity of multi-task learning in these examples.

Low Occupancy Complexity Low occupancy complexity model is proposed in Du et al. (2021). We consider the low occupancy complexity model with known feature β_h, as defined in the following.

Definition A.1 (Low Occupancy Complexity)**.** An MDP $\mathcal{M}$ and hypothesis class $\mathcal{H}$ has low occupancy complexity with respect to an unknown feature mapping $\phi_h : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ if there exists a known $\beta_h : \mathcal{H} \mapsto \mathbb{R}^d, h \in [H]$ such that for all $f \in \mathcal{H}$ and $(s_h, a_h) \in \mathcal{S} \times \mathcal{A}$ we have that:

$$d^{\pi_f}(s_h, a_h) = \langle \beta_h(f), \phi_h(s_h, a_h) \rangle.$$

Here d^{π_f} denotes the occupancy measure under policy π_f.

It is known that low occupancy complexity model is Bilinear class with

$$\begin{aligned} X_h(g) &= \beta_h(g) \\ W_h(g) &= \sum_{s\in\mathcal{S}, a\in\mathcal{A}} \phi_h(s,a)(Q_{h,g}(s,a) - r(s,a) - \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)}[V_{h+1,g}(s')]. \end{aligned}$$

The estimation policies can be chosen as $\pi_{\text{est}}(f) = \pi_f$ and the discrepancy measure is the Bellman error:

$$l_f(o_h, g) = Q_{h,g}(s_h, a_h) - r_h - V_{h+1,g}(s_{h+1}).$$

We consider Low rank Multi-task Bilinear class where each MDP $\mathcal{M}_m$ is low occupancy complexity model, i.e.

$$d_m^{\pi_{m,f}}(s_h, a_h) = \langle \beta_{m,h}(f), \phi_{m,h}(s_{m,h}, a_{m,h}) \rangle,$$

and there exist $B_h \in \mathbb{R}^{d\times k}$ and features $\nu_{m,h} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ such that $\phi_{m,h} = B_h \nu_{m,h}$. The hypothesis set is chosen as $\mathcal{G} = \mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_M$. Notice that

$$W_{m,h}(g) = B_h \bigg(\sum_{s\in\mathcal{S}, a\in\mathcal{A}} \nu_h(s,a)(Q_{h,g}(s,a) - r(s,a) - \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)}[V_{h+1,g}(s')] \bigg).$$

Then we use the fixed feature map $X_{m,h} = \beta_{m,h}$. Denoting $\mathcal{V}_{m,h}$ as a feature class to capture $\nu_{m,h}$ and $\mathcal{B} = \mathcal{N}(\mathbb{R}^{d\times k}, \epsilon)$ where $\mathcal{N}(\mathbb{R}^{d\times k}, \epsilon)$ denotes the ϵ-covering of $\mathbb{R}^{d\times k}$, we set the feature class $\mathcal{V}$ induced by $\otimes_{m\in[M],h\in[H]}\mathcal{V}_{m,h}$ in which $\nu_{m,h}$ defines the map $v_{m,h} : \mathcal{G} \mapsto \mathbb{R}^k$ in the following way:

$$v_{m,h}(g) = \sum_{s\in\mathcal{S}_m, a\in\mathcal{A}_m} \nu_{m,h}(s,a)(Q_{m,h,g}(s,a) - r(s,a) - \mathbb{E}_{s'\sim\mathbb{P}_{m,h}(\cdot|s,a)}[V_{m,h+1,g}(s')].$$

Therefore $W_{m,h} = B_h v_{m,h}$. We thus have the following result.

Corollary A.2. *Consider Low-rank Multi-task Bilinear class where each MDP $\mathcal{M}_m$ is low occupancy complexity model with*

$$d_m^{\pi_{m,f}}(s_h, a_h) = \langle \beta_{m,h}(f), \phi_{m,h}(s_{m,h}, a_{m,h}) \rangle,$$

and there exist $B_h \in \mathbb{R}^{d\times k}$ and features $\nu_{m,h} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ such that $\phi_{m,h} = B_h \nu_{m,h}$. Denote $\mathcal{V}_{m,h}$ as a feature class to capture $\nu_{m,h}$ and set the feature class $\mathcal{V}$ induced by $\otimes_{m\in[M],h\in[H]}\mathcal{V}_{m,h}$. Under Assumption 5.2, there exists an algorithm that with probability at least $1-\delta$ finds a set of policies $\{\pi_m\}_{m\in[M]}$ such that

$$\sum_{m=1}^{M} V_{m,1}^{\pi_m^*}(s_1) - \sum_{m=1}^{M} V_{m,1}^{\pi_m}(s_1) \le \epsilon$$

with

$$O\left(\frac{H^6 M^2 d(Mk + dk\log(1/\epsilon) + \log(|\mathcal{V}||\mathcal{G}|/\delta))}{\epsilon^2} \right)$$

trajectories.

Remark A.3. Using Bilin-UCB to learn each task individually, it takes

$$O\left(\frac{H^6M^3d^2\log(|\mathcal{G}|/\delta))}{\epsilon^2}\right)$$

trajectories to learn a set of policies $\{\pi_m\}_{m\in[M]}$ such that

$$\sum_{m=1}^{M} V_{m,1}^{\pi_m^*}(s_1) - \sum_{m=1}^{M} V_{m,1}^{\pi_m}(s_1) \leq \epsilon.$$

Therefore, if the cardinality of the feature set $\mathcal{V}$ is not greater than $e^{O(Md)}$, Algorithm 1 achieves sample complexity improvement comparing to single-task learning.

FLAMBE Feature selection Agarwal et al. (2020) setting is an extension of linear MDP where the features are all unknown.

Definition A.4 (Feature Selection)**.** An MDP $\mathcal{M}$ is feature selection model if there exist unknown functions $\mu_h^* : \mathcal{S} \mapsto \mathbb{R}^d$ and unknown features $\phi_h^* : \mathcal{S}\times\mathcal{A} \mapsto \mathbb{R}^d$ such that for all $h \in [H]$ and $(s, a, s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$

$$\mathbb{P}_h(s'|s, a) = \langle\mu_h^*(s'), \phi^*(s, a)\rangle.$$

Using the function class Φ to capture ϕ^* and function class $\mathcal{H} = \mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_H$ to capture the target Q-function Q^* where specifically

$$\mathcal{H}_h = \{w^\top\phi(s, a) : \|w\|_1 \leq C_W, \phi \in \Phi\},$$

it is known that the feature selection model is an instance of Bilinear class with

$$X_h(g) = \mathbb{E}_{\pi_f}[\phi^*(s_{h-1}, a_{h-1})]$$
$$W_h(g) = \sum_{s\in\mathcal{S}} \mu_h^*(s, a)\big(V_{h,g}(s) - r(s, \pi_g(s)) - \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,\pi_g(s))}[V_{h+1,g}(s')]\big).$$

The estimation policy is uniform over the action set $\pi_{\text{est}} = \text{Unif}(\mathcal{A})$ and the discrepancy measure is defined as

$$l_f(o_h, g) = \frac{\mathbb{1}(a_h = \pi_g(s_h))}{1/A}\big(Q_{h,g}(s_h, a_h) - r_h - V_{h+1,g}(s_{h+1})\big).$$

We consider Low rank Multi-task Bilinear class where each MDP M_m is feature selection model with $\mathbb{P}_{m,h} = \langle\mu_{m,h}^*(s'), \phi^*(s, a)\rangle$ (ϕ^* is shared among tasks) and there exists $B_h^* : \mathcal{S} \mapsto \mathbb{R}^{d\times k}$ and $\nu_{m,h}^* : \mathcal{S} \mapsto \mathbb{R}^k$ such that $\mu_{m,h}^* = B_h^*\nu_{m,h}^*$. We can therefore use the hypothesis class $\mathcal{G} = \mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_M$ where $\mathcal{H}_m$ is defined similarly to capture the Q-function Q^*. In addition, consider function class Φ to capture ϕ^*, $\mathcal{B}$ to capture B_h^*, and $\mathcal{V}$ to capture $\nu_{m,h}^*$. Notice that every $\nu_{m,h} \in \mathcal{V}$ defines the map $v_{m,h} : g \mapsto \mathbb{R}^k$ in the following way:

$$v_{m,h}(g) = \sum_{s\in\mathcal{S}_m, a\in\mathcal{A}_m} \nu_{m,h}(s, a)\big(V_{h,g}(s) - r(s, \pi_g(s)) - \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,\pi_g(s))}[V_{h+1,g}(s')]\big).$$

Therefore we have $W_{m,h} = B_h v_{m,h}$. Notice that each $(\phi, B_h)_{h\in[H],m\in[M]}$ will define the expectation $\mathbb{E}^m_{\pi_{m,g}}$ (via $\mathbb{P}_h(s'|s, a) = \langle\mu_h(s'), \phi(s, a)\rangle$ with $\mu_{m,h} = B_h\nu_{m,h}$). Finally, we set $\mathcal{X}$ induced by $\Phi\otimes\mathcal{B}\otimes\mathcal{V}$ where the feature $X_{m,h}(g)$ can be computed for each $g \in \mathcal{G}$ as follows:

$$X_{m,h}(g) = \mathbb{E}^m_{\pi_{m,f}}[\phi(s_{h-1}, a_{h-1})].$$

We thus have the following result.

Corollary A.5. *Consider Low-rank Multi-task Bilinear class where each MDP M_m is feature selection model with $\mathbb{P}_{m,h} = \langle\mu_{m,h}^*(s'), \phi^*(s, a)\rangle$ (ϕ^* is shared among tasks) and there exists $B_h^* : \mathcal{S} \mapsto \mathbb{R}^{d\times k}$ and $\nu_{m,h}^* : \mathcal{S} \mapsto \mathbb{R}^k$ such that $\mu_{m,h}^* = B_h^*\nu_{m,h}^*$. Consider function class Φ to*

capture ϕ^*, $\mathcal{B}$ to capture B_h^*, and $\mathcal{V}$ to capture $\nu_{m,h}^*$. Under Assumption 5.2, there exists an algorithm that with probability at least $1-\delta$ finds a set of policies $\{\pi_m\}_{m\in[M]}$ such that

$$\sum_{m=1}^{M} V_{m,1}^{\pi_m^*}(s_1) - \sum_{m=1}^{M} V_{m,1}^{\pi_m}(s_1) \le \epsilon$$

using

$$O\left(\frac{H^6 M^2 d(Mk + \log(|\mathcal{B}||\Phi||\mathcal{V}||\mathcal{G}|/\delta))}{\epsilon^2}\right)$$

trajectories.

B Proof of Theorem 5.3

Proof. Fix $\lambda = R^2$. We have

$$\begin{aligned}
&\sum_{m=1}^{M} V_{m,1}^{\pi_m^*}(s_1) - \sum_{m=1}^{M} V_{m,1}^{\pi_{m,g^{(t_0)}}}(s_1) \\
\le &\sum_{m=1}^{M} V_{m,1,g^{(t)}}(s_1) - \sum_{m=1}^{M} V_{m,1}^{\pi_{m,g^{(t_0)}}}(s_1) \\
\le &\sum_{m=1}^{M}\sum_{h=1}^{H} |\langle B_h^*(g^{(t_0)}) v_{m,h}^*(g^{(t_0)}) - B_h^*(g^*) v_{m,h}^*(g^*), X_{m,h}^*(g^{(t_0)})\rangle|, \qquad (7)
\end{aligned}$$

where the first step is from Lemma C.2 and the second step is from Lemma C.3.

Fix $h \in [H]$. We use Hölder's inequality,

$$\begin{aligned}
&\sum_{m=1}^{M} |\langle B_h^*(g^{(t_0)}) v_{m,h}^*(g^{(t_0)}) - B_h^*(g^*) v_{m,h}^*(g^*), X_{m,h}^*(g^{(t_0)})\rangle| \\
\le &\left(\sum_{m=1}^{M} \|B_h^*(g^{(t_0)}) v_{m,h}^*(g^{(t_0)}) - B_h^*(g^*) v_{m,h}^*(g^*)\|_{\Sigma_{m,h}^{(t_0)}}^2\right)^{1/2} \cdot \left(\sum_{m=1}^{M} \|X_{m,h}^*(g^{(t_0)})\|_{(\Sigma_{m,h}^{(t_0)})^{-1}}^2\right)^{1/2}.
\end{aligned}$$

By the confidence set in Line 3 and Lemma C.1, we have

$$\begin{aligned}
&\sum_{m=1}^{M} \|B_h^*(g^{(t_0)}) v_{m,h}^*(g^{(t_0)}) - B_h^*(g^*) v_{m,h}^*(g^*)\|_{\Sigma_{m,h}^{(t_0)}}^2 \\
\le &\, 2\sum_{\tau=1}^{t_0-1}\sum_{m=1}^{M} \left(\langle B_h^{(g^{(t_0)})}(g^{(t_0)}) v_{m,h}^{(g^{(t_0)})}(g^{(t_0)}) - B_h^{(g^{(t_0)})}(\widetilde{g^{(t_0)}}) v_{m,h}^{(g^{(t_0)})}(\widetilde{g^{(t_0)}}), X_{m,h}^{(g^{(t_0)})}(g^{(\tau)})\rangle\right)^2 \\
&+ 2\sum_{\tau=1}^{t_0-1}\sum_{m=1}^{M} \Big(\langle B_h^*(g^{(t_0)}) v_{m,h}^*(g^{(t_0)}) - B_h^*(g^*) v_{m,h}^*(g^*), X_{m,h}^*(g^{(\tau)})\rangle \\
&- \langle B_h^{(g^{(t_0)})}(g^{(t_0)}) v_{m,h}^{(g^{(t_0)})}(g^{(t_0)}) - B_h^{(g^{(t)})}(\widetilde{g^{(t_0)}}) v_{m,h}^{(g^{(t_0)})}(\widetilde{g^{(t_0)}}), X_{m,h}^{(g^{(t)})}(g^{(\tau)})\rangle\Big)^2 \\
\le &\, 4R^2.
\end{aligned}$$

By Lemma C.4, we have

$$\sum_{m=1}^{M} \|X_{m,h}^*(g^{(t_0)})\|_{(\Sigma_{m,h}^{(t_0)})^{-1}}^2 \le 2.$$

Therefore, we combine the above to get

$$\sum_{m=1}^{M} |\langle B_h^*(g^{(t_0)})v_{m,h}^*(g^{(t_0)}) - B_h^*(g^*)v_{m,h}^*(g^*), X_{m,h}^*(g^{(t_0)})\rangle| \le 16R.$$

Plugging this into Eq. (7) completes the proof.

□

C SUPPORTING LEMMA

The following Lemma is the key concentration result on $(v_{m,h}^{(g)}, B_h^{(g)}, X_{m,h}^{(g)})$.

Lemma C.1. *With probability at least $1-\delta$, the following holds for all $h \in [H]$ and $g \in \mathcal{G}$*

$$\sum_{\tau=1}^{t-1}\sum_{m=1}^{M}\Big(\langle B_h^*(g)v_{m,h}^*(g) - B_h^*(g^*)v_{m,h}^*(g^*), X_{m,h}^*(g^{(\tau)})\rangle - \langle B_h^{(g)}(g)v_{m,h}^{(g)}(g) - B_h^{(g)}(\widetilde{g})v_{m,h}^{(g)}(\widetilde{g}), X_{m,h}^{(g)}(g^{(\tau)})\rangle\Big)^2 \le R^2,$$

where for each $g \in \mathcal{G}$, $(v_{1:M,1:H}^{(g)}, B_{1:H}^{(g)}, X_{1:M,1:H}^{(g)}, \widetilde{g})$ are defined by

$$(v_{1:M,h}^{(g)}, B_h^{(g)}, X_{1:M,h}^{(g)}, \widetilde{g}) = \underset{v_{1:M,h}\in\mathcal{V}_h, B_h\in\mathcal{B}_h, X_{1:M,h}\in\mathcal{X}_h, \widetilde{g}\in\mathcal{G}}{\arg\min}$$
$$\sum_{\tau=1}^{t-1}\sum_{m=1}^{M}\Big(\mathbb{E}_{(s,a,s')\sim\mathcal{D}_{m,h}^{(\tau)}}[l_{m,h,g^{(\tau)}}(s,a,s',g)] - \langle B_h(g)v_{m,h}(g) - B_h(\widetilde{g})v_{m,h}(\widetilde{g}), X_{m,h}(g^{(\tau)})\rangle\Big)^2\Big\}, \forall h \in [H].$$

Proof. Fix $h \in [H]$, $g \in \mathcal{G}$, and $t \in [T]$. Let $\alpha_{m,h} = (\alpha_{m,h,1}, \ldots, \alpha_{m,h,t-1})^\top \in \mathbb{R}^{t-1}$ and $\widehat{\alpha}_{m,h} = (\widehat{\alpha}_{m,h,1}, \ldots, \widehat{\alpha}_{m,h,t-1})^\top \in \mathbb{R}^{t-1}$ where

$$\alpha_{m,h,\tau} = \langle B_h^*(g)v_{m,h}^*(g) - B_h^*(g^*)v_{m,h}^*(g^*), X_{m,h}^*(g^{(\tau)})\rangle$$
$$\widehat{\alpha}_{m,h,\tau} = \langle B_h^{(g)}(g)v_{m,h}^{(g)}(g) - B_h^{(g)}(\widetilde{g})v_{m,h}^{(g)}(\widetilde{g}), X_{m,h}^{(g)}(g^{(\tau)})\rangle$$

and let $\xi_{m,h} = (\xi_{m,h,1}, \ldots, \xi_{m,h,t-1})^\top \in \mathbb{R}^{t-1}$ where

$$\xi_{m,h,\tau} = \mathbb{E}_{(s,a,s')\sim\mathcal{D}_{m,h}^{(\tau)}}[l_{m,h,g^{(\tau)}}(s,a,s',g)] - \mathbb{E}_{a_{0:h}\sim\pi_{m,g^{(\tau)}}}[l_{m,h,g^{(\tau)}}(s,a,s',g)].$$

We know that $\xi_{m,h,1}, \ldots, \xi_{m,h,t-1}$ is a stochastic process adapted to filtration $\{\mathcal{H}_\tau\}_{\tau=1}^{t-1}$ and $\xi_{m,h,\tau}$ is conditionally σ-sub-Gaussian with variance $\sigma^2 \le H^2 \cdot n_0^{-1}$. We have the basic inequality

$$\sum_{m=1}^{M} \|\alpha_{m,h} - \widehat{\alpha}_{m,h}\|^2 \le \sum_{m=1}^{M} \xi_{m,h}^\top(\alpha_{m,h} - \widehat{\alpha}_{m,h}). \tag{8}$$

Notice that the matrix $\big(B_h^*(g)v_{1,h}^*(g) - B_h^*(g^*)v_{1,h}^*(g^*), \cdots, B_h^*(g)v_{M,h}^*(g) - B_h^*(g^*)v_{M,h}^*(g^*)\big) \in \mathbb{R}^{d\times M}$ has rank at most $2k$. We can rewrite it as $\widetilde{B}_h \cdot (\widetilde{v}_{1,h}, \cdots, \widetilde{v}_{M,h})$ where $\widetilde{B}_h \in \mathbb{R}^{d\times 2k}$ is an orthonormal matrix with and $\widetilde{v}_{m,h} \in \mathbb{R}^{2k}$. Similarly we rewrite the matrix $\Big(B_h^{(g)}(g)v_{1,h}^{(g)}(g) - B_h^{(g)}(\widetilde{g})v_{1,h}^{(g)}(\widetilde{g}), \cdots, B_h^{(g)}(g)v_{M,h}^{(g)}(g) - B_h^{(g)}(\widetilde{g})v_{M,h}^{(g)}(\widetilde{g})\Big) \in \mathbb{R}^{d\times M}$ as $\widehat{B}_h \cdot (\widehat{v}_{1,h}, \cdots, \widehat{v}_{M,h})$ where $\widehat{B}_h \in \mathbb{R}^{d\times 2k}$ is an orthonormal matrix with and $\widehat{v}_{m,h} \in \mathbb{R}^{2k}$. Then we have

$$\alpha_{m,h} - \widehat{\alpha}_{m,h} = \begin{pmatrix} (X_{m,h}^*(g^{(1)}))^\top \\ \vdots \\ (X_{m,h}^*(g^{(t-1)}))^\top \end{pmatrix} \cdot \widetilde{B}_h \cdot \widetilde{v}_{m,h} - \begin{pmatrix} (X_{m,h}^{(g)}(g^{(1)}))^\top \\ \vdots \\ (X_{m,h}^{(g)}(g^{(t-1)}))^\top \end{pmatrix} \cdot \widehat{B}_h \cdot \widehat{v}_{m,h}.$$

Let $\beta_{m,h} = (\widetilde{v}_{m,h}^\top, \widehat{v}_{m,h}^\top)^\top \in \mathbb{R}^{4k}$ and

$$U_{m,h} = \left(\begin{pmatrix} (X_{m,h}^*(g^{(1)}))^\top \\ \vdots \\ (X_{m,h}^*(g^{(t-1)}))^\top \end{pmatrix} \cdot \widetilde{B}_h, - \begin{pmatrix} (X_{m,h}^{(g)}(g^{(1)}))^\top \\ \vdots \\ (X_{m,h}^{(g)}(g^{(t-1)}))^\top \end{pmatrix} \cdot \widehat{B}_h \right) \in \mathbb{R}^{(t-1)\times 4k},$$

then $\alpha_{m,h} - \widehat{\alpha}_{m,h} = U_{m,h}\beta_{m,h}$. Let $\Lambda_{m,h} = U_{m,h}^\top U_{m,h} + \lambda \cdot I$, then Cauchy-Schwarz inequality implies

$$\begin{aligned} \sum_{m=1}^M \xi_{m,h}^\top (\alpha_{m,h} - \widehat{\alpha}_{m,h}) &\le \sum_{m=1}^M \|\xi_{m,h}^\top U_{m,h}\|_{\Lambda_{m,h}^{-1}} \|\beta_{m,h}\|_{\Lambda_{m,h}} \\ &\le \sqrt{\left(\sum_{m=1}^M \|\xi_{m,h}^\top U_{m,h}\|_{\Lambda_{m,h}^{-1}}^2\right) \cdot \left(\sum_{m=1}^M \|\beta_{m,h}\|_{\Lambda_{m,h}}^2\right)}. \end{aligned} \tag{9}$$

For $\sum_{m=1}^M \|\beta_{m,h}\|_{\Lambda_{m,h}}^2$, notice that

$$\begin{aligned} \left(\sum_{m=1}^M \|\beta_{m,h}\|_{\Lambda_{m,h}}^2\right) &= \sum_{m=1}^M \|\alpha_{m,h} - \widehat{\alpha}_{m,h}\|^2 + \lambda \cdot \sum_{m=1}^M \|\beta_{m,h}\|^2 \\ &\le \sum_{m=1}^M \|\alpha_{m,h} - \widehat{\alpha}_{m,h}\|^2 + 4M\lambda C_W. \end{aligned} \tag{10}$$

Fix $v_{m,h}^{(g)}, B_h^{(g)}, X_{m,h}^{(g)}, \widetilde{g}$, then $\{U_{m,h,\tau}\}_{\tau=1}^{t-1}$ is a stochastic process adapted to filtration $\{\mathcal{H}_\tau\}_{\tau=1}^{t-1}$. Therefore, by Lemma 2 in Hu et al. (2021), we know that the following holds with probability at least $1-\delta_1$

$$\begin{aligned} \sum_{m=1}^M \|\xi_{m,h}^\top U_{m,h}\|_{\Lambda_{m,h}^{-1}}^2 &\le \sigma^2 \cdot \sum_{m=1}^M \log\left(\frac{\det(\Lambda_{m,h})\det(\lambda\cdot I)^{-1}}{\delta^2}\right) \\ &\le H^2 \cdot n_0^{-1} \cdot (Mk \cdot \log(C_X/\lambda) + \log(1/\delta_1)). \end{aligned}$$

Taking union bound for all $v_{1:M,h}^{(g)}, B_h^{(g)}, X_{1:M,h}^{(g)}, \widetilde{g} \in (\mathcal{V}_h, \mathcal{B}_h, \mathcal{X}_h, \mathcal{G})$, we have that with probability at least $1-\delta_0$

$$\sum_{m=1}^M \|\xi_{m,h}^\top U_{m,h}\|_{\Lambda_{m,h}^{-1}}^2 \le H^2 \cdot n_0^{-1} \cdot (Mk \cdot \log(C_X/\lambda) + \log(|\mathcal{V}||\mathcal{B}||\mathcal{X}||\mathcal{G}|/\delta_0)). \tag{11}$$

Combining Eq. (11), Eq. (10), Eq. (9), and Eq. (8), we come to the following with probability at least $1-\delta_0$

$$\begin{aligned} &\sum_{m=1}^M \|\alpha_{m,h} - \widehat{\alpha}_{m,h}\|^2 \\ &\le \sqrt{H^2 \cdot n_0^{-1} \cdot (Mk \cdot \log(C_X/\lambda) + \log(|\mathcal{V}||\mathcal{B}||\mathcal{X}||\mathcal{G}|/\delta)) \cdot \left(\sum_{m=1}^M \|\alpha_{m,h} - \widehat{\alpha}_{m,h}\|^2 + 4m\lambda C_W\right)}. \end{aligned}$$

Solving the above inequality and setting $\lambda = \frac{2H^2(Mk\cdot\log(C_X/\lambda)+\log(|\mathcal{V}||\mathcal{B}||\mathcal{X}||\mathcal{G}|/\delta))}{Mn_0C_W}$, with probability at least $1-\delta_0$,

$$\begin{aligned} &\sum_{m=1}^M \|\alpha_{m,h} - \widehat{\alpha}_{m,h}\|^2 \\ &\le \frac{2H^2(Mk\cdot\log(\frac{C_X}{\lambda}) + \log(\frac{|\mathcal{V}||\mathcal{B}||\mathcal{X}||\mathcal{G}|}{\delta}))}{n_0} + 4\sqrt{\frac{H(Mk\cdot\log(\frac{C_X}{\lambda}) + \log(\frac{|\mathcal{V}||\mathcal{B}||\mathcal{X}||\mathcal{G}|}{\delta}))m\lambda C_W}{n_0}} \\ &\le \frac{8H^2(Mk\cdot\log(Mn_0C_WC_X) + \log(|\mathcal{V}||\mathcal{B}||\mathcal{X}||\mathcal{G}|/\delta_0))}{n_0}. \end{aligned}$$

Finally, taking union bound for all $h \in [H]$, $g \in \mathcal{G}$, and $t \in [T]$ completes the proof.

□

The following Lemma establishes optimism of $g^{(t)}$ by showing that the true model g^* always lies in the confidence set.

Lemma C.2. *With probability at least $1-\delta$, g^* satisfies the constraints in Line 3. Therefore,*

$$\sum_{m=1}^{M} V_{m,1,g^{(t)}}(s_1) \geq \sum_{m=1}^{M} V_{m,1}^{\pi_m^*}(s_1)$$

Proof. Applying Lemma C.1 for $g = g^*$, we know that outside the failure event $\mathcal{E}$,

$$\begin{aligned}
&\sum_{\tau=1}^{t-1}\sum_{m=1}^{M}\Big(\langle B_h^{(g)}(g)v_{m,h}^{(g)}(g) - B_h^{(g)}(\widetilde{g})v_{m,h}^{(g)}(\widetilde{g}), X_{m,h}^{(g)}(g^{(\tau)})\rangle\Big)^2 \\
\leq &\sum_{\tau=1}^{t-1}\sum_{m=1}^{M}\Big(\langle B_h^{*}(g)v_{m,h}^{*}(g) - B_h^{*}(g^*)v_{m,h}^{*}(g^*), X_{m,h}^{*}(g^{(\tau)})\rangle - \langle B_h^{(g)}(g)v_{m,h}^{(g)}(g) - B_h^{(g)}(\widetilde{g})v_{m,h}^{(g)}(\widetilde{g}), X_{m,h}^{(g)}(g^{(\tau)})\rangle\Big)^2 \\
\leq &R^2.
\end{aligned}$$

This means that g^* belongs to the confidence set $\mathcal{G}^{(t)}, \forall t \in [T]$ in Line 3.

□

The following result decouples $\sum_{m=1}^{M} V_{m,1,g}(s_1) - \sum_{m=1}^{M} V_{m,1}^{\pi_{m,g}}(s_1)$ into sum of Bellman errors. This is a simple extension of Lemma 5.5 in Du et al. (2021).

Lemma C.3. *For any $g \in \mathcal{G}$ we have*

$$\begin{aligned}
&\sum_{m=1}^{M} V_{m,1,g}(s_1) - \sum_{m=1}^{M} V_{m,1}^{\pi_{m,g}}(s_1) \\
\leq &\sum_{m=1}^{M}\sum_{h=1}^{H} |\langle B_h^{*}(g)v_{m,h}^{*}(g) - B_h^{*}(g^*)v_{m,h}^{*}(g^*), X_{m,h}^{*}(g)\rangle|.
\end{aligned}$$

Finally, we show the coverage of $\Sigma_{m,h}^{(t)} = \sum_{\tau=1}^{t} X_h^*(g^{(\tau)})(X_h^*(g^{(\tau)}))^\top + \lambda I$.

Lemma C.4. *Let $\Sigma_{m,h}^{(t)} = \sum_{\tau=1}^{t} X_h^*(g^{(\tau)})(X_h^*(g^{(\tau)}))^\top + \lambda I$. There exists $t_0 \in [T]$ such that*

$$\sum_{m=1}^{M} \|X_h^*(g^{(t_0)})\|^2_{(\Sigma_{m,h}^{(t_0)})^{-1}} \leq 2.$$

Proof. We use the following Elliptic Potential Lemma.

Lemma C.5 (Elliptic Potential Lemma, Abbasi-Yadkori et al. (2011))**.** *Let $\Sigma_{m,h}^{(t)} = \sum_{\tau=1}^{t} X_h^*(g^{(\tau)})(X_h^*(g^{(\tau)}))^\top + \lambda I$. Then we have*

$$\sum_{t=1}^{T} \log(1 + \|X_h^*(g^{(t)})\|^2_{(\Sigma_{m,h}^{(t)})^{-1}}) \leq \frac{1}{T} \cdot \log\frac{\det(\Sigma_{m,h}^{(T)})}{\det(\lambda I)} \leq d \cdot \log(1 + \frac{TC_X^2}{\lambda d})$$

Set $T = 8HMd\log(1 + \frac{MC_X^2}{\lambda})$, then

$$\min_{t\in[T]} \sum_{m=1}^{M} \log(1 + \|X_h^*(g^{(t)})\|^2_{(\Sigma_{m,h}^{(t)})^{-1}}) \leq \frac{1}{T} \cdot M \cdot d \cdot \log(1 + \frac{TC_X^2}{\lambda d}) \leq \log 2.$$

For this t_0 where the above holds, $\max_{m\in[M]} \|X_h^*(g^{(t_0)})\|^2_{(\Sigma_{m,h}^{(t_0)})^{-1}} \leq 1$, which implies

$$\sum_{m=1}^{M} \|X_h^*(g^{(t_0)})\|^2_{(\Sigma_{m,h}^{(t_0)})^{-1}} \leq 2\sum_{m=1}^{M} \log(1 + \|X_h^*(g^{(t_0)})\|^2_{(\Sigma_{m,h}^{(t_0)})^{-1}}) \leq 2.$$

□