Appendices

A. Omitted Proof in Section 4

A.1. Proof of Lemma 1

Proof. By the optimality of $\hat{\boldsymbol{B}}_t$ and $\hat{\boldsymbol{W}}_t = [\hat{\boldsymbol{w}}_{t,1}, \cdots, \hat{\boldsymbol{w}}_{t,M}]$, we know that $\sum_{i=1}^{M} \left\| \boldsymbol{y}_{t-1,i} - \boldsymbol{X}_{t-1,i}^\top \hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} \right\|_2^2 \leq \sum_{i=1}^{M} \left\| \boldsymbol{y}_{t-1,i} - \boldsymbol{X}_{t-1,i}^\top \boldsymbol{B}\boldsymbol{w}_i \right\|_2^2$. Since $\boldsymbol{y}_{t-1,i} = \boldsymbol{X}_{t-1,i}^\top \boldsymbol{B}\boldsymbol{w}_i + \boldsymbol{\eta}_{t-1,i}$, we have

$$\sum_{i=1}^{M} \left\| \boldsymbol{X}_{t-1,i}^\top \left(\hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i \right) \right\|_2^2 \leq 2 \sum_{i=1}^{M} \boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \left(\hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i \right). \tag{28}$$

We firstly analyse the non-trivial setting where $d \geq 2k$. Note that both $\boldsymbol{\Theta} = \boldsymbol{B}\boldsymbol{W}$ and $\hat{\boldsymbol{\Theta}}_t = \hat{\boldsymbol{B}}_t \hat{\boldsymbol{W}}_t$ are low-rank matrix with rank upper bounded by k, which indicates that $\operatorname{rank}\left(\hat{\boldsymbol{\Theta}}_t - \boldsymbol{\Theta} \right) \leq 2k$. In that case, we can write $\hat{\boldsymbol{\Theta}}_t - \boldsymbol{\Theta} = \boldsymbol{U}_t \boldsymbol{R}_t = [\boldsymbol{U}_t \boldsymbol{r}_{t,1}, \boldsymbol{U}_t \boldsymbol{r}_{t,2}, \cdots, \boldsymbol{U}_t \boldsymbol{r}_{t,M}]$, where $\boldsymbol{U}_t \in \mathbb{R}^{d \times 2k}$ is an orthonormal matrix with $\|\boldsymbol{U}_t\|_F = \sqrt{2k}$, and $\boldsymbol{R}_t \in \mathbb{R}^{2k \times M}$ satisfies $\|\boldsymbol{r}_{t,i}\|_2 \leq \sqrt{k}$. In other words, we can write $\hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i = \boldsymbol{U}_t \boldsymbol{r}_{t,i}$ for certain $\boldsymbol{U}_t$ and $\boldsymbol{r}_{t,i}$.

Define $\boldsymbol{V}_{t-1,i}(\lambda) \overset{\text{def}}{=} \left(\boldsymbol{U}_t^\top \boldsymbol{X}_{t-1,i} \right) \left(\boldsymbol{U}_t^\top \boldsymbol{X}_{t-1,i} \right)^\top + \lambda \boldsymbol{I}$. We have:

$$\begin{aligned}
&\sum_{i=1}^{M} \left\| \hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i \right\|_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)}^2 && (29)\\
=&\sum_{i=1}^{M} \left\| \boldsymbol{X}_{t-1,i}^\top \left(\hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i \right) \right\|_2^2 + \sum_{i=1}^{M} \lambda \left\| \hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i \right\|_2^2 && (30)\\
\leq& 2 \sum_{i=1}^{M} \boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \left(\hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i \right) + 4M\lambda && (31)\\
=& 2 \sum_{i=1}^{M} \boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \boldsymbol{U}_t \boldsymbol{r}_{t,i} + 4M\lambda && (32)\\
\leq& 2 \sum_{i=1}^{M} \left\| \boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \boldsymbol{U}_t \right\|_{\boldsymbol{V}_{t-1,i}^{-1}(\lambda)} \left\| \boldsymbol{r}_{t,i} \right\|_{\boldsymbol{V}_{t-1,i}(\lambda)} + 4M\lambda && (33)\\
\leq& 2 \sqrt{\sum_{i=1}^{M} \left\| \boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \boldsymbol{U}_t \right\|_{\boldsymbol{V}_{t-1,i}^{-1}(\lambda)}^2} \sqrt{\sum_{i=1}^{M} \left\| \boldsymbol{r}_{t,i} \right\|_{\boldsymbol{V}_{t-1,i}(\lambda)}^2} + 4M\lambda && (34)\\
=& 2 \sqrt{\sum_{i=1}^{M} \left\| \boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \boldsymbol{U}_t \right\|_{\boldsymbol{V}_{t-1,i}^{-1}(\lambda)}^2} \sqrt{\sum_{i=1}^{M} \left\| \hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i \right\|_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)}^2} + 4M\lambda && (35)
\end{aligned}$$

Eqn 31 is due to Eqn 28, $\left\| \hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} \right\| \leq 1$ and $\|\boldsymbol{B}\boldsymbol{w}_i\| \leq 1$. Eqn 34 is due to Cauchy-Schwarz inequality. Eqn 35 is from

$$\sum_{i=1}^{M} \left\| \hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i \right\|_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)}^2 = \sum_{i=1}^{M} \left\| \boldsymbol{U}_t \boldsymbol{r}_{t,i} \right\|_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)}^2 = \sum_{i=1}^{M} \left\| \boldsymbol{r}_{t,i} \right\|_{\boldsymbol{U}_t^\top \tilde{\boldsymbol{V}}_{t-1,i}(\lambda) \boldsymbol{U}_t}^2 = \sum_{i=1}^{M} \left\| \boldsymbol{r}_{t,i} \right\|_{\boldsymbol{V}_{t-1,i}(\lambda)}^2.$$

The main problem is how to bound $\left\| \boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \boldsymbol{U}_t \right\|_{\boldsymbol{V}_{t-1,i}^{-1}(\lambda)} = \left\| \sum_{n=1}^{t-1} \eta_{n,i} \boldsymbol{U}_t^\top x_{n,i} \right\|_{\boldsymbol{V}_{t-1,i}^{-1}(\lambda)}$. Note that for a fixed $\boldsymbol{U}_t = \bar{\boldsymbol{U}}$, we can regard $\bar{\boldsymbol{U}}^\top \boldsymbol{x}_{n,i} \in \mathbb{R}^k$ as the corresponding "action" chosen in step t. With this observation, if $\boldsymbol{U}_t$ is fixed, we can bound this term following the arguments of the self-normalized bound for vector-valued martingales (Abbasi-Yadkori et al., 2011).

Lemma 2. *For a fixed $\bar{\boldsymbol{U}}$, define $\bar{\boldsymbol{V}}_{t,i}(\lambda) \overset{def}{=} \left(\bar{\boldsymbol{U}}^\top \boldsymbol{X}_{t,i}\right)\left(\bar{\boldsymbol{U}}^\top \boldsymbol{X}_{t,i}\right)^\top + \lambda \boldsymbol{I}$, then any $\delta > 0$, with probability at least $1-\delta$, for all $t \geq 0$,*

$$\sum_{i=1}^{M} \left\|\bar{\boldsymbol{U}}^\top \boldsymbol{X}_{t,i}\boldsymbol{\eta}_{t,i}\right\|^2_{\bar{\boldsymbol{V}}_{t,i}^{-1}} \tag{36}$$

$$\leq 2\log\left(\frac{\prod_{i=1}^{M}\left(\det(\bar{\boldsymbol{V}}_{t,i})^{1/2}\det(\lambda\boldsymbol{I})^{-1/2}\right)}{\delta}\right). \tag{37}$$

We defer the proof of Lemma 2 to Appendix A.3. We set $\lambda = 1$. By Lemma 2, we know that for a fixed $\bar{\boldsymbol{U}}$, with probability at least $1-\delta_1$,

$$\sum_{i=1}^{M}\left\|\sum_{n=1}^{t-1}\eta_{n,i}\bar{\boldsymbol{U}}^\top x_{n,i}\right\|^2_{\bar{\boldsymbol{V}}_{t,i}^{-1}(\lambda)} \leq 2\log\left(\frac{\prod_{i=1}^{M}\det(\bar{\boldsymbol{V}}_{t,i}(\lambda))^{1/2}\det(\lambda\boldsymbol{I})^{-1/2}}{\delta_1}\right) \leq 2Mk + 2\log(1/\delta_1). \tag{38}$$

The above analysis shows that we can bound $\left\|\boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \boldsymbol{U}_t\right\|_{\boldsymbol{V}_{t-1,i}^{-1}(\lambda)}$ if $\boldsymbol{U}_t$ is fixed as $\bar{\boldsymbol{U}}$. Following this idea, we prove the lemma by the construction of ϵ-net over all possible $\boldsymbol{U}_t$. To apply the trick of ϵ-net, we need to slightly modify the derivation of Eqn 29. For a fixed matrix $\bar{\boldsymbol{U}} \in \mathbb{R}^{d\times 2k}$, we have

$$\sum_{i=1}^{M}\left\|\hat{\boldsymbol{B}}_t\hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i\right\|^2_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)} \tag{39}$$

$$\leq 2\sum_{i=1}^{M}\boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \boldsymbol{U}_t \boldsymbol{r}_{t,i} + 4M\lambda \tag{40}$$

$$= 2\sum_{i=1}^{M}\boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \bar{\boldsymbol{U}} \boldsymbol{r}_{t,i} + 2\sum_{i=1}^{M}\boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \left(\boldsymbol{U}_t - \bar{\boldsymbol{U}}\right)\boldsymbol{r}_{t,i} + 4M\lambda \tag{41}$$

$$\leq 2\sum_{i=1}^{M}\left\|\boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \bar{\boldsymbol{U}}\right\|_{\bar{\boldsymbol{V}}_{t-1,i}^{-1}(\lambda)}\left\|\boldsymbol{r}_{t,i}\right\|_{\bar{\boldsymbol{V}}_{t-1,i}(\lambda)} + 2\sum_{i=1}^{M}\boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \left(\boldsymbol{U}_t - \bar{\boldsymbol{U}}\right)\boldsymbol{r}_{t,i} + 4M\lambda \tag{42}$$

$$= 2\sum_{i=1}^{M}\left\|\boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \bar{\boldsymbol{U}}\right\|_{\bar{\boldsymbol{V}}_{t-1,i}^{-1}(\lambda)}\left\|\boldsymbol{r}_{t,i}\right\|_{\boldsymbol{V}_{t-1,i}(\lambda)} + 2\sum_{i=1}^{M}\boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \left(\boldsymbol{U}_t - \bar{\boldsymbol{U}}\right)\boldsymbol{r}_{t,i} \tag{43}$$

$$+ 2\sum_{i=1}^{M}\left\|\boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \bar{\boldsymbol{U}}\right\|_{\bar{\boldsymbol{V}}_{t-1,i}^{-1}(\lambda)}\left(\left\|\boldsymbol{r}_{t,i}\right\|_{\bar{\boldsymbol{V}}_{t-1,i}(\lambda)} - \left\|\boldsymbol{r}_{t,i}\right\|_{\boldsymbol{V}_{t-1,i}(\lambda)}\right) + 4M\lambda \tag{44}$$

$$\leq 2\sqrt{\sum_{i=1}^{M}\left\|\boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \bar{\boldsymbol{U}}\right\|^2_{\boldsymbol{V}_{t-1,i}^{-1}(\lambda)}}\sqrt{\sum_{i=1}^{M}\left\|\hat{\boldsymbol{B}}_t\hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i\right\|^2_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)}} + 2\sum_{i=1}^{M}\boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \left(\boldsymbol{U}_t - \bar{\boldsymbol{U}}\right)\boldsymbol{r}_{t,i} \tag{45}$$

$$+ 2\sum_{i=1}^{M}\left\|\boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \bar{\boldsymbol{U}}\right\|_{\bar{\boldsymbol{V}}_{t-1,i}^{-1}(\lambda)}\left(\left\|\boldsymbol{r}_{t,i}\right\|_{\bar{\boldsymbol{V}}_{t-1,i}(\lambda)} - \left\|\boldsymbol{r}_{t,i}\right\|_{\boldsymbol{V}_{t-1,i}(\lambda)}\right) + 4M\lambda \tag{46}$$

Eqn 40, 42 and 45 follow the same idea of Eqn 32, 33 and 35.

We construct an ϵ-net $\mathcal{E}$ in Frobenius norm over the matrix set $\left\{\boldsymbol{U} \in \mathbb{R}^{d\times 2k} : \|\boldsymbol{U}\|_F \leq k\right\}$. It is not hard to see that $|\mathcal{E}| \leq \left(\frac{6\sqrt{2k}}{\epsilon}\right)^{2kd}$. By the union bound over all possible $\bar{\boldsymbol{U}} \in \mathcal{E}$, we know that with probability $1 - |\mathcal{E}|\delta_1$, Eqn 38 holds for any $\bar{\boldsymbol{U}} \in \mathcal{E}$. For each $\boldsymbol{U}_t$, we choose an $\bar{\boldsymbol{U}} \in \mathcal{E}$ with $\left\|\boldsymbol{U}_t - \bar{\boldsymbol{U}}\right\|_F \leq \epsilon$, and we have

$$2\sqrt{\sum_{i=1}^{M}\left\|\boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \bar{\boldsymbol{U}}\right\|^2_{\boldsymbol{V}_{t-1,i}^{-1}(\lambda)}} \leq 2\sqrt{2Mk + 2\log(1/\delta_1)} \tag{47}$$

Since $\left\|\boldsymbol{U}_t - \bar{\boldsymbol{U}}\right\|_F \leq \epsilon$, we have

$$2\sum_{i=1}^{M} \left\|\boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \bar{\boldsymbol{U}}\right\|_{\bar{\boldsymbol{V}}_{t-1,i}^{-1}(\lambda)} \left(\|\boldsymbol{r}_{t,i}\|_{\bar{\boldsymbol{V}}_{t-1,i}(\lambda)} - \|\boldsymbol{r}_{t,i}\|_{\boldsymbol{V}_{t-1,i}(\lambda)}\right) \leq 2\sqrt{Mk\epsilon(2Mk + 2\log(1/\delta_1))}. \tag{48}$$

For the term $2\sum_{i=1}^{M} \boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \left(\boldsymbol{U}_t - \bar{\boldsymbol{U}}\right) \boldsymbol{r}_{t,i}$, the following inequality holds for any step $t \in [T]$ with probability $1 - MT\delta_2$,

$$\begin{aligned}
2\sum_{i=1}^{M} \boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \left(\boldsymbol{U}_t - \bar{\boldsymbol{U}}\right) \boldsymbol{r}_{t,i} \leq & 2\sum_{i=1}^{M} \|\boldsymbol{\eta}_{t-1,i}\|_2 \left\|\boldsymbol{X}_{t-1,i}^\top \left(\boldsymbol{U}_t - \bar{\boldsymbol{U}}\right) \boldsymbol{r}_{t,i}\right\|_2 && (49)\\
\leq & 2\sum_{i=1}^{M} \|\boldsymbol{\eta}_{t-1,i}\|_2 \sqrt{kT\epsilon} && (50)\\
\leq & 2M\sqrt{2\log(2/\delta_2)kT^2\epsilon} && (51)
\end{aligned}$$

The last inequality follows from the fact that $|\eta_{n,i}| \leq \sqrt{2\log(2/\delta_2)}$ with probability $1 - \delta_2$ for fixed n, i, and apply a union bound over $n \in [t-1], i \in [M]$. Plugging Eqn. 47, 48 and 49 back to Eqn. 45, the following inequality holds for any $t \in [T]$ with probability at least $1 - |\mathcal{E}|\delta_1 - MT\delta_2$:

$$\begin{aligned}
& \sum_{i=1}^{M} \left\|\hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i\right\|^2_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)} && (52)\\
\leq & 2\sqrt{Mk + 2\log(1/\delta_1)} \sqrt{\sum_{i=1}^{M} \left\|\hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i\right\|^2_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)}} && (53)\\
& + 2M\sqrt{2\log(2/\delta_2)kT^2\epsilon} + 2\sqrt{Mk\epsilon(2Mk + 2\log(1/\delta_1))} + 4M\lambda && (54)
\end{aligned}$$

By solving the above inequality, we know that

$$\begin{aligned}
\sum_{i=1}^{M} \left\|\hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i\right\|^2_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)} \leq & 32\left(Mk + \log(1/\delta_1)\right) + 4M\sqrt{2\log(2/\delta_2)kT^2\epsilon} && (55)\\
& + 4\sqrt{Mk\epsilon(2Mk + 2\log(1/\delta_1))} + 8M\lambda && (56)
\end{aligned}$$

Setting $\lambda = 1$, $\epsilon = \frac{1}{kM^2T^2}$, $\delta_1 = \frac{\delta}{2\left(\frac{6\sqrt{2k}}{\epsilon}\right)^{2kd}} \leq \frac{\delta}{2|\mathcal{E}|}$, and $\delta_2 = \frac{\delta}{2MT}$, the following inequality holds with probability $1 - \delta$:

$$\sum_{i=1}^{M} \left\|\hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i\right\|^2_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)} \leq L \stackrel{\text{def}}{=} 48\left(Mk + 5kd\log(kMT)\right) + 32\log(4MT) + 76\log(1/\delta) \tag{57}$$

At last we talk about the trivial setting where $k < d < 2k$. In this case, we can write $\hat{\boldsymbol{\Theta}}_t - \boldsymbol{\Theta} = \boldsymbol{R}_t$ where $\boldsymbol{R}_t \in \mathbb{R}^{d\times M}$. The proof then follows the same framework as the case when $d \geq 2k$, except that we don't need to consider $\boldsymbol{U}_t$ and construct ϵ-net over all possible $\boldsymbol{U}_t$. It is not hard to show that $\sum_{i=1}^{M} \left\|\hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i\right\|^2_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)} \leq 24\left(Md + 2\log(Tk/\delta)\right)$ in this case, which is also less than L since $d < 2k$. □

A.2. Proof of Theorem 1

With Lemma 1, we are ready to prove Theorem 1.

Proof. Let $\tilde{\boldsymbol{V}}_{t,i}(\lambda) = \boldsymbol{X}_{t,i}\boldsymbol{X}_{t,i}^\top + \lambda \boldsymbol{I}_d$ for some $\lambda > 0$.

$$\begin{aligned}
\mathrm{Reg}(T) &= \sum_{t=1}^{T}\sum_{i=1}^{M} \left\langle \boldsymbol{\theta}_i, \boldsymbol{x}_{t,i}^* - \boldsymbol{x}_{t,i} \right\rangle && (58)\\
&\le \sum_{t=1}^{T}\sum_{i=1}^{M} \left\langle \tilde{\boldsymbol{\theta}}_{t,i} - \boldsymbol{\theta}_i, \boldsymbol{x}_{t,i} \right\rangle && (59)\\
&= \sum_{t=1}^{T}\sum_{i=1}^{M} \left\langle \tilde{\boldsymbol{\theta}}_{t,i} - \hat{\boldsymbol{\theta}}_{t,i} + \hat{\boldsymbol{\theta}}_{t,i} - \boldsymbol{\theta}_i, \boldsymbol{x}_{t,i} \right\rangle && (60)\\
&\le \sum_{t=1}^{T}\sum_{i=1}^{M} \left(\left\| \tilde{\boldsymbol{\theta}}_{t,i} - \hat{\boldsymbol{\theta}}_{t,i} \right\|_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)} + \left\| \hat{\boldsymbol{\theta}}_{t,i} - \boldsymbol{\theta}_i \right\|_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)} \right) \|\boldsymbol{x}_{t,i}\|_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)^{-1}} && (61)\\
&\le \left(\sqrt{\sum_{t=1}^{T}\sum_{i=1}^{M} \left\| \tilde{\boldsymbol{\theta}}_{t,i} - \hat{\boldsymbol{\theta}}_{t,i} \right\|^2_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)}} + \sqrt{\sum_{i=1}^{M} \left\| \hat{\boldsymbol{\theta}}_{t,i} - \boldsymbol{\theta}_i \right\|^2_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)}} \right) \cdot \sqrt{\sum_{t=1}^{T}\sum_{i=1}^{M} \|\boldsymbol{x}_{t,i}\|^2_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)^{-1}}} && (62)\\
&\le 2\sqrt{T\left(L + 4\lambda M\right)} \cdot \sqrt{\sum_{i=1}^{M}\sum_{t=1}^{T} \|\boldsymbol{x}_{t,i}\|^2_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)^{-1}}} && (63)
\end{aligned}$$

where the first inequality is due to $\sum_{i=1}^{M} \left\langle \boldsymbol{\theta}_i, \boldsymbol{x}_{t,i}^* \right\rangle \le \left\langle \tilde{\boldsymbol{\theta}}_{t,i}, \boldsymbol{x}_{t,i} \right\rangle$ from the optimistic choice of $\tilde{\boldsymbol{\theta}}_{t,i}$ and $\boldsymbol{x}_{t,i}$. By Lemma 11 of Abbasi-Yadkori et al. (2011), as long as $\lambda \ge 1$ we have

$$\sum_{t=1}^{T} \|\boldsymbol{x}_{t,i}\|^2_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda')^{-1}} \le 2\log \frac{\det(\tilde{\boldsymbol{V}}_{T,i}(\lambda'))}{\det(\lambda' \boldsymbol{I}_d)} \le 2d \log\left(1 + \frac{T}{\lambda d}\right) \tag{64}$$

Therefore, we can finally bound the regret by choosing $\lambda = 1$

$$\begin{aligned}
\mathrm{Reg}(T) &\le 2\sqrt{T(L+4M)} \cdot \sqrt{\sum_{i=1}^{M}\sum_{t=1}^{T} \|\boldsymbol{x}_{t,i}\|^2_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda')^{-1}}} && (65)\\
&\le 2\sqrt{T\left(L+4M\right)} \cdot \sqrt{Md\log\left(1 + \frac{T}{d}\right)} && (66)\\
&= \tilde{O}\left(M\sqrt{dkT} + d\sqrt{kMT}\right). && (67)
\end{aligned}$$

□

A.3. Proof of Lemma 2

The proof of Lemma 2 follows the similar idea of Theorem 1 in Abbasi-Yadkori et al. (2011). We consider the σ-algebra $F_t = \sigma\left(\{\boldsymbol{x}_{1,i}\}_{i=1}^M, \{\boldsymbol{x}_{2,i}\}_{i=1}^M, \cdots, \{\boldsymbol{x}_{t+1,i}\}_{i=1}^M, \{\eta_{1,i}\}_{i=1}^M, \{\eta_{2,i}\}_{i=1}^M, \cdots, \{\eta_{t,i}\}_{i=1}^M\right)$, then $\{\boldsymbol{x}_{t,i}\}_{i=1}^M$ is F_{t-1}-measurable, and $\{\eta_{t,i}\}_{i=1}^M$ is F_t-measurable.

Define $\bar{\boldsymbol{x}}_{t,i} = \boldsymbol{U}^\top \boldsymbol{x}_{t,i}$ and $\boldsymbol{S}_{t,i} = \sum_{n=1}^t \bar{\boldsymbol{U}}^\top \boldsymbol{x}_{t,i}\eta_{t,i}$. Let

$$M_t(\boldsymbol{Q}) = \exp\left(\sum_{n=1}^t \sum_{i=1}^M \left[\eta_{t,i}\left\langle \boldsymbol{q}_i, \bar{\boldsymbol{x}}_{t,i}\right\rangle - \frac{1}{2}\left\langle \boldsymbol{q}_i, \bar{\boldsymbol{x}}_{t,i}\right\rangle^2\right]\right), \quad \boldsymbol{Q} = [\boldsymbol{q}_1, \cdots, \boldsymbol{q}_M] \in \mathbb{R}^{2k\times M} \tag{68}$$

Lemma 3. *Let τ be a stopping time w.r.t the filtration $\{F_t\}_{t=0}^\infty$. Then $M_t(\boldsymbol{Q})$ is almost surely well-defined and $\mathbb{E}[M_t(\boldsymbol{Q})] \le 1$.*

Proof. Let $D_t(\boldsymbol{Q}) = \exp\left(\sum_{i=1}^M \left[\eta_{t,i}\left\langle \boldsymbol{q}_i, \bar{\boldsymbol{x}}_{t,i}\right\rangle - \frac{1}{2}\left\langle \boldsymbol{q}_i, \bar{\boldsymbol{x}}_{t,i}\right\rangle^2\right]\right)$. By the sub-Gaussianity of $\eta_{t,i}$, we have

$$\mathbb{E}\left[\exp\left(\left[\eta_{t,i}\left\langle \boldsymbol{q}_i, \bar{\boldsymbol{x}}_{t,i}\right\rangle - \frac{1}{2}\left\langle \boldsymbol{q}_i, \bar{\boldsymbol{x}}_{t,i}\right\rangle^2\right]\right)] \mid F_{t-1}\right] \le 1. \tag{69}$$

Then we have $\mathbb{E}\left[D_t(\boldsymbol{Q}) \mid F_{t-1}\right] \le 1$. Further,

$$\begin{aligned}\mathbb{E}\left[M_t(\boldsymbol{Q}) \mid F_{t-1}\right] &= \mathbb{E}\left[M_1(\boldsymbol{Q})\cdots D_{t-1}(\boldsymbol{Q})D_t(\boldsymbol{Q}) \mid F_{t-1}\right] &(70)\\ &= D_1(\boldsymbol{Q})\cdots D_{t-1}(\boldsymbol{Q})\mathbb{E}\left[D_t(\boldsymbol{Q}) \mid F_{t-1}\right] \le M_{t-1}(\boldsymbol{Q}) &(71)\end{aligned}$$

This shows that $\{M_t(\boldsymbol{Q})\}_{t=0}^\infty$ is a supermartingale and $\mathbb{E}\left[M_t(\boldsymbol{Q})\right] \le 1$.

Following the same argument of Lemma 8 in Abbasi-Yadkori et al. (2011), we show that $M_\tau(\boldsymbol{Q})$ is almost surely well-defined. By the convergence theorem for nonnegative supermartingales, $M_\infty(\boldsymbol{Q}) = \lim_{t\to\infty} M_t(\boldsymbol{Q})$ is almost surely well-defined. Therefore, $M_\tau(\boldsymbol{Q})$ is indeed well-defined independently of whether $\tau < \infty$ or not. Let $W_t(\boldsymbol{Q}) = M_{\min\{\tau,t\}}(\boldsymbol{Q})$ be a stopped version of $(M_t((Q)))_t$. By Fatou's Lemma, $\mathbb{E}[M_\tau(\boldsymbol{Q})] = \mathbf{E}\left[\liminf_{t\to\infty} W_t(\boldsymbol{Q})\right] \le \liminf_{t\to\infty} \mathbf{E}\left[W_t(\boldsymbol{Q})\right] \le 1$. This shows that $\mathbb{E}[M_\tau(\boldsymbol{Q})] \le 1$. □

The next lemma uses the "method of mixtures" technique to bound $\sum_{i=1}^M \|\boldsymbol{S}_{t,i}\|^2_{\bar{\boldsymbol{V}}_{t,i}^{-1}(\lambda)}$.

Lemma 4. *Let τ be a stopping time w.r.t the filtration $\{F_t\}_{t=0}^\infty$. Then, for $\delta > 0$, with probability $1-\delta$,*

$$\sum_{i=1}^M \|\boldsymbol{S}_{\tau,i}\|^2_{\bar{\boldsymbol{V}}_{\tau,i}^{-1}(\lambda)} \le 2\log\left(\frac{\prod_{i=1}^M \left(\det(\bar{\boldsymbol{V}}_{\tau,i})^{1/2}\det(\lambda\boldsymbol{I})^{-1/2}\right)}{\delta}\right). \tag{72}$$

Proof. For each $i \in [M]$, let $\boldsymbol{\Lambda}_i$ be a $\mathbb{R}^{2k}$ Gaussian random variable which is independent of all the other random variables and whose covariance is $\lambda^{-1}\boldsymbol{I}$. Define $M_t = \mathbb{E}\left[M_t([\boldsymbol{\Lambda}_1, \cdots, \boldsymbol{\Lambda}_M]) \mid F_\infty\right]$. We still have $\mathbb{E}[M_\tau] = \mathbb{E}[\mathbb{E}[M_t([\boldsymbol{\Lambda}_1, \cdots, \boldsymbol{\Lambda}_M]) \mid \{\boldsymbol{\Lambda}_i\}_{i=1}^M]] \le 1$.

Now we calculate M_t. Define $M_{t,i}(\boldsymbol{q}_i) \overset{\text{def}}{=} \exp\left(\sum_{n=1}^t \left[\eta_{t,i}\left\langle \boldsymbol{q}_i, \bar{\boldsymbol{x}}_{t,i}\right\rangle - \frac{1}{2}\left\langle \boldsymbol{q}_i, \bar{\boldsymbol{x}}_{t,i}\right\rangle^2\right]\right)$, then we have $M_t = \mathbb{E}\left[\prod_{i=1}^M M_{t,i}(\boldsymbol{\Lambda_i}) \mid F_\infty\right] = \prod_{i=1}^M \mathbb{E}\left[M_{t,i}(\boldsymbol{\Lambda_i}) \mid F_\infty\right]$, where the second equality is due to the fact that $\{M_{t,i}(\boldsymbol{\Lambda}_i)\}_{i=1}^M$ are relatively independent given F_∞. We only need to calculate $\mathbb{E}\left[M_{t,i}(\boldsymbol{\Lambda}_i) \mid F_\infty\right]$ for each $i \in [M]$.

Following the proof of Lemma 9 in Abbasi-Yadkori et al. (2011), we know that

$$\mathbb{E}\left[M_{t,i}(\boldsymbol{\Lambda}_i) \mid F_\infty\right] = \left(\frac{\det(\lambda\boldsymbol{I})}{\det(\bar{\boldsymbol{V}}_{t,i})}\right)^{1/2} \exp\left(\frac{1}{2}\|\boldsymbol{S}_{t,i}\|^2_{\bar{\boldsymbol{V}}_{t,i}^{-1}(\lambda)}\right). \tag{73}$$

Then we have

$$M_t = \prod_{i=1}^M \left(\left(\frac{\det(\lambda\boldsymbol{I})}{\det(\bar{\boldsymbol{V}}_{t,i})}\right)^{1/2}\right) \exp\left(\frac{1}{2}\sum_{i=1}^M \|\boldsymbol{S}_{t,i}\|^2_{\bar{\boldsymbol{V}}_{t,i}^{-1}(\lambda)}\right). \tag{74}$$

Since $\mathbb{E}[M_\tau] \leq 1$, we have

$$\begin{aligned}
&\Pr\left[\sum_{i=1}^{M} \|\boldsymbol{S}_{\tau,i}\|^2_{\bar{\boldsymbol{V}}^{-1}_{\tau,i}(\lambda)} > 2\log\left(\frac{\prod_{i=1}^{M}\left(\det(\bar{\boldsymbol{V}}_{\tau,i})^{1/2}\det(\lambda\boldsymbol{I})^{-1/2}\right)}{\delta}\right)\right] \\
=&\Pr\left[\frac{\exp\left(\frac{1}{2}\sum_{i=1}^{M}\|\boldsymbol{S}_{\tau,i}\|^2_{\bar{\boldsymbol{V}}^{-1}_{\tau,i}(\lambda)}\right)}{\delta^{-1}\left(\prod_{i=1}^{M}\left(\det(\bar{\boldsymbol{V}}_{t,i})^{1/2}\det(\lambda\boldsymbol{I})^{-1/2}\right)\right)} > 1\right] \\
\leq&\mathbb{E}\left[\frac{\exp\left(\sum_{i=1}^{M}\|\boldsymbol{S}_{\tau,i}\|^2_{\bar{\boldsymbol{V}}^{-1}_{\tau,i}(\lambda)}\right)}{\delta^{-1}\left(\prod_{i=1}^{M}\left(\det(\bar{\boldsymbol{V}}_{\tau,i})^{1/2}\det(\lambda\boldsymbol{I})^{-1/2}\right)\right)}\right] \\
=&\mathbb{E}[M_\tau]\delta \leq \delta.
\end{aligned}$$

□

Proof. (Proof of Lemma 2) The only remaining issue is the stopping time construction. Define the bad event

$$B_t(\delta) \stackrel{\text{def}}{=} \left\{\omega \in \Omega : \sum_{i=1}^{M} \|\boldsymbol{S}_{t,i}\|^2_{\bar{\boldsymbol{V}}^{-1}_{t,i}(\lambda)} > 2\log\left(\frac{\prod_{i=1}^{M}\left(\det(\bar{\boldsymbol{V}}_{t,i})^{1/2}\det(\lambda\boldsymbol{I})^{-1/2}\right)}{\delta}\right)\right\} \tag{75}$$

Consider the stopping time $\tau(\omega) = \min\{t \geq 0 : \omega \in B_t(\delta)\}$, we have $\bigcup_{t\geq 0} B_t(\delta) = \{\omega : \tau(\omega) < \infty\}$.

By lemma 4, we have

$$\Pr\left[\bigcup_{t\geq 0} B_t(\delta)\right] = \Pr[\tau < \infty] \tag{76}$$

$$= \Pr\left[\sum_{i=1}^{M} \|\boldsymbol{S}_{\tau,i}\|^2_{\bar{\boldsymbol{V}}^{-1}_{\tau,i}(\lambda)} > 2\log\left(\frac{\prod_{i=1}^{M}\left(\det(\bar{\boldsymbol{V}}_{\tau,i})^{1/2}\det(\lambda\boldsymbol{I})^{-1/2}\right)}{\delta}\right), \tau \leq \infty\right] \tag{77}$$

$$\leq \Pr\left[\sum_{i=1}^{M} \|\boldsymbol{S}_{\tau,i}\|^2_{\bar{\boldsymbol{V}}^{-1}_{\tau,i}(\lambda)} > 2\log\left(\frac{\prod_{i=1}^{M}\left(\det(\bar{\boldsymbol{V}}_{\tau,i})^{1/2}\det(\lambda\boldsymbol{I})^{-1/2}\right)}{\delta}\right)\right] \tag{78}$$

$$\leq \delta. \tag{79}$$

□

A.4. Proof of Theorem 2

Proof. The proof follows the same idea of that for Theorem 1. The only difference is that, in our setting, we have $y_{t,i} = \boldsymbol{x}_{t,i}^\top \boldsymbol{B}\boldsymbol{w}_i + \eta_{t,i} + \Delta_{t,i}$, where $\boldsymbol{\theta}_i = \boldsymbol{B}\boldsymbol{w}_i$ is the best approximator for task $i \in [M]$ such that $\left|\mathbb{E}\left[y_i \mid \boldsymbol{x}_i\right] - \left\langle \boldsymbol{x}_i, \dot{\boldsymbol{B}}\dot{\boldsymbol{w}}_i \right\rangle\right| \leq \zeta$, and $\|\Delta_{t,i}\| \leq \zeta$. Define $\boldsymbol{\Delta}_{t,i} = [\Delta_{1,i}, \Delta_{2,i}, \cdots, \Delta_{t,i}]$. Similarly, by the optimality of $\hat{\boldsymbol{B}}_t$ and $\hat{\boldsymbol{W}}_t = [\hat{\boldsymbol{w}}_{t,1}, \cdots, \hat{\boldsymbol{w}}_{t,M}]$, we know that $\sum_{i=1}^M \left\|\boldsymbol{y}_{t-1,i} - \boldsymbol{X}_{t-1,i}^\top \hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i}\right\|_2^2 \leq \sum_{i=1}^M \left\|\boldsymbol{y}_{t-1,i} - \boldsymbol{X}_{t-1,i}^\top \boldsymbol{B}\boldsymbol{w}_i\right\|^2$. Since $\boldsymbol{y}_{t-1,i} = \boldsymbol{X}_{t-1,i}^\top \boldsymbol{B}\boldsymbol{w}_i + \boldsymbol{\eta}_{t-1,i} + \boldsymbol{\Delta}_{t,i}$, thus we have

$$
\begin{aligned}
&\sum_{i=1}^M \left\| \boldsymbol{X}_{t-1,i}^\top \left(\hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i\right)\right\|^2 && (80)\\
\leq& 2\sum_{i=1}^M \boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \left(\hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i\right) + 2\sum_{i=1}^M \boldsymbol{\Delta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \left(\hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i\right) && (81)\\
\leq& 2\sum_{i=1}^M \boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \left(\hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i\right) + 2\sum_{i=1}^M \|\boldsymbol{X}_{t-1,i}\boldsymbol{\Delta}_{t-1,i}\|_{\tilde{\boldsymbol{V}}_{t-1,i}^{-1}(\lambda)} \left\|\hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i\right\|_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)} && (82)\\
\leq& 2\sum_{i=1}^M \boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \left(\hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i\right) + 2\sum_{i=1}^M \sqrt{T}\zeta \left\|\hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i\right\|_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)} && (83)\\
\leq& 2\sum_{i=1}^M \boldsymbol{\eta}_{t-1,i}^\top \boldsymbol{X}_{t-1,i}^\top \left(\hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i\right) + 2\sqrt{MT}\zeta \sqrt{\sum_{i=1}^M \left\|\hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i\right\|_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)}^2} && (84)
\end{aligned}
$$

The third inequality follows from Projection Bound (Lemma 8) in Zanette et al. (2020a). The first term of Eqn 84 shares the same form of Eqn 28. Following the same proof idea of Lemma 1, we know that with probability $1-\delta$,

$$
\begin{aligned}
&\sum_{i=1}^M \left\|\hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i\right\|_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)}^2 && (85)\\
\leq& \left(2\sqrt{Mk + 8kd\log(kMT/\delta)} + 2\sqrt{MT}\zeta\right)\sqrt{\sum_{i=1}^M \left\|\hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i\right\|_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)}^2} + 4M + 4\sqrt{\log(4MT/\delta)} && (86)
\end{aligned}
$$

Solving for $\sum_{i=1}^M \left\|\hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i\right\|_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)}^2$, we know that the true parameter $\boldsymbol{B}\boldsymbol{W}$ is always contained in the confidence set, i.e.

$$
\sum_{i=1}^M \left\|\hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i\right\|_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)}^2 \leq L', \tag{87}
$$

where $L' = 2L + 32MT\zeta^2$.

Thus we have

$$\mathrm{Reg}(T) = \sum_{t=1}^{T}\sum_{i=1}^{M}\left(y_{t,i}^{*} - y_{t,i}\right) \tag{88}$$

$$\leq 2MT\zeta + \sum_{t=1}^{T}\sum_{i=1}^{M}\left\langle \boldsymbol{\theta}_i, \boldsymbol{x}_{t,i}^{*} - \boldsymbol{x}_{t,i}\right\rangle \tag{89}$$

$$\leq 2MT\zeta + \sum_{t=1}^{T}\sum_{i=1}^{M}\left\langle \tilde{\boldsymbol{\theta}}_{t,i} - \boldsymbol{\theta}_i, \boldsymbol{x}_{t,i}\right\rangle \tag{90}$$

$$= 2MT\zeta + \sum_{t=1}^{T}\sum_{i=1}^{M}\left\langle \tilde{\boldsymbol{\theta}}_{t,i} - \hat{\boldsymbol{\theta}}_{t,i} + \hat{\boldsymbol{\theta}}_{t,i} - \boldsymbol{\theta}_i, \boldsymbol{x}_{t,i}\right\rangle \tag{91}$$

$$\leq 2MT\zeta + \sum_{t=1}^{T}\sum_{i=1}^{M}\left(\left\|\tilde{\boldsymbol{\theta}}_{t,i} - \hat{\boldsymbol{\theta}}_{t,i}\right\|_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)} + \left\|\hat{\boldsymbol{\theta}}_{t,i} - \boldsymbol{\theta}_i\right\|_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)}\right)\|\boldsymbol{x}_{t,i}\|_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)^{-1}} \tag{92}$$

$$\leq 2MT\zeta + \left(\sqrt{\sum_{t=1}^{T}\sum_{i=1}^{M}\left\|\tilde{\boldsymbol{\theta}}_{t,i} - \hat{\boldsymbol{\theta}}_{t,i}\right\|^2_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)}} + \sqrt{\sum_{i=1}^{M}\left\|\hat{\boldsymbol{\theta}}_{t,i} - \boldsymbol{\theta}_i\right\|^2_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)}}\right)\cdot\sqrt{\sum_{t=1}^{T}\sum_{i=1}^{M}\|\boldsymbol{x}_{t,i}\|^2_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)^{-1}}} \tag{93}$$

$$\leq 2MT\zeta + 2\sqrt{T\left(L' + 4\lambda M\right)}\cdot\sqrt{\sum_{i=1}^{M}\sum_{t=1}^{T}\|\boldsymbol{x}_{t,i}\|^2_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)^{-1}}} \tag{94}$$

$$\leq 2MT\zeta + 2\sqrt{T\left(L' + 4\lambda M\right)}\sqrt{Md\log(1 + \frac{T}{d})} \tag{95}$$

$$= \tilde{O}(M\sqrt{dkT} + d\sqrt{kMT} + MT\sqrt{d}\zeta), \tag{96}$$

where the second inequality is due to $\sum_{i=1}^{M}\left\langle \boldsymbol{\theta}_i, \boldsymbol{x}_{t,i}^{*}\right\rangle \leq \left\langle \tilde{\boldsymbol{\theta}}_{t,i}, \boldsymbol{x}_{t,i}\right\rangle$ from the optimistic choice of $\tilde{\boldsymbol{\theta}}_{t,i}$ and $\boldsymbol{x}_{t,i}$. The third inequality is due to Eqn 87. The last inequality is from Eqn 64. □

A.5. Proof of Theorem 3

Since our setting is strictly harder than the setting of multi-task linear bandit with infinite arms in Yang et al. (2020), we can prove the following lemma directly from their Theorem 4 by reduction.

Lemma 5. *Under the setting of Theorem 3, the regret of any Algorithm $\mathcal{A}$ is lower bounded by* $\Omega\left(Mk\sqrt{T} + d\sqrt{kMT}\right)$.

In order to prove Theorem 3, we only need to show that the following lemma is true.

Lemma 6. *Under the setting of Theorem 3, the regret of any Algorithm $\mathcal{A}$ is lower bounded by* $\Omega\left(MT\sqrt{d\zeta}\right)$.

Proof. (Proof of Lemma 6)

To prove Lemma 6, we leverage the lower bound for misspecified linear bandits in the single-task setting. We restate the following lemma from the previous literature with a slight modification of notations.

Lemma 7. *(Proposition 6 in Zanette et al. (2020a)). There exists a feature map $\phi : \mathcal{A} \to \mathbb{R}^d$ that defines a misspecified linear bandits class $\mathcal{M}$ such that every bandit instance in that class has reward response:*

$$\mu_a = \boldsymbol{\phi}_a^\top \boldsymbol{\theta} + z_a$$

for any action a (Here $z_a \in [0, \zeta]$ is the deviation from linearity and $\mu_a \in [0, 1]$) and such that the expected regret of any algorithm on at least a member of the class up to round T is $\Omega(\sqrt{d\zeta}T)$.

Suppose M can be exactly divided by k, we construct the following instances to prove lemma 6. We divide M tasks into k groups. Each group shares the same parameter θ_i. To be more specific, we let $\boldsymbol{w}_1 = \boldsymbol{w}_2 = \cdots = \boldsymbol{w}_{M/k} = \boldsymbol{e}_1$, $\boldsymbol{w}_{M/k+1} = \boldsymbol{w}_{M/k+2} = \cdots = \boldsymbol{w}_{2M/k} = \boldsymbol{e}_2$, $\cdots$, $\boldsymbol{w}_{(k-1)M/k+1} = \boldsymbol{w}_{(k-1)M/k+2} = \cdots = \boldsymbol{w}_M = \boldsymbol{e}_k$. Under this construction, the parameters θ_i for these tasks are exactly the same in each group, but relatively independent among different groups. That is to say, the expected regret lower bound is at least the summation of the regret lower bounds in all k groups.

Now we consider the regret lower bound for group $j \in [k]$. Since the parameters are shared in the same group, the regret of running an algorithm for M/k tasks with T steps each is at least the regret of running an algorithm for single-task linear bandit with $M/k \cdot T$ steps. By Lemma 7, the regret for single-task linear bandit with MT/k steps is at least $\Omega(\sqrt{d\zeta}MT/k)$. Summing over all k groups, we can prove that the regret lower bound is $\Omega(\sqrt{d\zeta}MT)$. □

Combining Lemma 5 and Lemma 6, we complete the proof of Theorem 3.

B. Proof of Theorem 4

B.1. Definitions and First Step Analysis

Before presenting the proof of theorem 4, we will make a first step analysis on the low-rank least-square estimator in equation 15.

For any $\left\{Q_{h+1}^i\right\}_{i=1}^M \in \mathcal{Q}_{h+1}$, there exists $\left\{\dot{\boldsymbol{\theta}}_h^i\left(Q_{h+1}^i\right)\right\}_{i=1}^M \in \Theta_h$ that

$$\Delta_h^i\left(Q_{h+1}^i\right)(s,a) = \mathcal{T}_h^i\left(Q_{h+1}^i\right)(s,a) - \boldsymbol{\phi}(s,a)^\top \dot{\boldsymbol{\theta}}_h^i\left(Q_{h+1}^i\right) \tag{97}$$

where the approximation error $\left\|\Delta_h^i\left(Q_{h+1}^i\right)\right\|_\infty \leq \mathcal{I}$ is small for each $i \in [M]$. We also use $\dot{\boldsymbol{B}}_h \dot{\boldsymbol{w}}_h^i\left(Q_{h+1}^i\right)$ in place of $\dot{\boldsymbol{\theta}}_h^i\left(Q_{h+1}^i\right)$ in the following sections since we can write $\dot{\boldsymbol{\theta}}_h^i$ as $\dot{\boldsymbol{B}}_h \dot{\boldsymbol{w}}_h^i$.

In the multi-task low-rank least-square regression (equation 15), we are actually trying to recover $\dot{\boldsymbol{\theta}}_h^i$. However, due to the noise and representation error (i.e. the inherent Bellman error), we can only obtain an approximate solution $\hat{\boldsymbol{\theta}}_h^i = \hat{\boldsymbol{B}}_h \hat{\boldsymbol{w}}_h^i$ (see the global optimization problem in Definition 1).

$$\left(\hat{\boldsymbol{\theta}}_h^1, ..., \hat{\boldsymbol{\theta}}_h^M\right) = \hat{\boldsymbol{B}}_h \begin{bmatrix} \hat{\boldsymbol{w}}_h^1 & \hat{\boldsymbol{w}}_h^2 & \cdots & \hat{\boldsymbol{w}}_h^M \end{bmatrix} \tag{98}$$

$$= \underset{\left\|\boldsymbol{B}_h \boldsymbol{w}_h^i\right\|_2 \leq D}{\operatorname{argmin}} \sum_{i=1}^M \sum_{j=1}^{t-1} \left(\boldsymbol{\phi}\left(s_{hj}^i, a_{hj}^i\right)^\top \boldsymbol{B}_h \boldsymbol{w}_h^i - R\left(s_{hj}^i, a_{hj}^i\right) - \max_a Q_{h+1}^i\left(s_{h+1,j}^i\right)\right)^2 \tag{99}$$

$$= \underset{\left\|\boldsymbol{B}_h \boldsymbol{w}_h^i\right\|_2 \leq D}{\operatorname{argmin}} \sum_{i=1}^M \sum_{j=1}^{t-1} \left(\boldsymbol{\phi}\left(s_{hj}^i, a_{hj}^i\right)^\top \boldsymbol{B}_h \boldsymbol{w}_h^i - \mathcal{T}_h^i\left(Q_{h+1}^i\right)\left(s_{hj}^i, a_{hj}^i\right) - z_{hj}^i\left(Q_{h+1}^i\right)\left(s_{hj}^i, a_{hj}^i\right)\right)^2 \tag{100}$$

where $z_{hj}^i\left(Q_{h+1}^i\right)\left(s_{hj}^i, a_{hj}^i\right) \stackrel{\text{def}}{=} R\left(s_{hj}^i, a_{hj}^i\right) + \max_a Q_{h+1}^i\left(s_{h+1,j}^i, a\right) - \mathcal{T}_h^i\left(Q_{h+1}^i\right)\left(s_{hj}^i, a_{hj}^i\right)$.

Define $\boldsymbol{\Phi}_{ht}^i \in \mathbb{R}^{(t-1)\times d}$ to be the collection of linear features up to episode $t-1$ in task i, i.e. the j-th row of $\boldsymbol{\Phi}_{ht}^i$ is $\boldsymbol{\phi}\left(s_{hj}^i, a_{hj}^i\right)^\top$. Let $\boldsymbol{Y}_{ht}^i \in \mathbb{R}^{t-1}$ be a vector whose j-th dimension is $\mathcal{T}_h^i\left(Q_{h+1}^i\right)\left(s_{hj}^i, a_{hj}^i\right) + z_{hj}^i\left(Q_{h+1}^i\right)\left(s_{hj}^i, a_{hj}^i\right)$. Then the objective in (100) can be written as

$$\underset{\left\|\boldsymbol{B}_h \boldsymbol{w}_h^i\right\|_2 \leq D}{\operatorname{argmin}} \sum_{i=1}^M \left\|\boldsymbol{\Phi}_{ht}^i \boldsymbol{B}_h \boldsymbol{w}_h^i - \boldsymbol{Y}_{ht}^i\right\|_2^2 \tag{101}$$

Therefore, we have

$$\sum_{i=1}^M \left\|\boldsymbol{\Phi}_{ht}^i \hat{\boldsymbol{B}}_h \hat{\boldsymbol{w}}_h^i\left(Q_{h+1}^i\right) - Y_{ht}^i\right\|_2^2 \leq \sum_{i=1}^M \left\|\boldsymbol{\Phi}_{ht}^i \dot{\boldsymbol{B}}_h \dot{\boldsymbol{w}}_h^i\left(Q_{h+1}^i\right) - \boldsymbol{Y}_{ht}^i\right\|_2^2 \tag{102}$$

which implies

$$\sum_{i=1}^{M} \left\| \boldsymbol{\Phi}_{ht}^{i} \hat{\boldsymbol{B}}_h \hat{\boldsymbol{w}}_h^i \left(Q_{h+1}^i \right) - \boldsymbol{\Phi}_{ht}^{i} \dot{\boldsymbol{B}}_h \dot{\boldsymbol{w}}_h^i \left(Q_{h+1}^i \right) \right\|_2^2 \tag{103}$$

$$\leq 2 \sum_{i=1}^{M} \left(\boldsymbol{\Delta}_{ht}^{i} \right)^{\top} \boldsymbol{\Phi}_{ht}^{i} \left(\hat{\boldsymbol{B}}_h \hat{\boldsymbol{w}}_h^i \left(Q_{h+1}^i \right) - \dot{\boldsymbol{B}}_h \dot{\boldsymbol{w}}_h^i \left(Q_{h+1}^i \right) \right) \tag{104}$$

$$+ 2 \sum_{i=1}^{M} \left(\boldsymbol{z}_{ht}^{i} \right)^{\top} \boldsymbol{\Phi}_{ht}^{i} \left(\hat{\boldsymbol{B}}_h \hat{\boldsymbol{w}}_h^i \left(Q_{h+1}^i \right) - \dot{\boldsymbol{B}}_h \dot{\boldsymbol{w}}_h^i \left(Q_{h+1}^i \right) \right) \tag{105}$$

where $\boldsymbol{\Delta}_{ht}^{i} \stackrel{\text{def}}{=} \left[\Delta_{h1}^{i} \left(Q_{h+1}^i \right) \left(s_{h1}^i, a_{h1}^i \right) \quad \Delta_{h2}^{i} \left(Q_{h+1}^i \right) \left(s_{h2}^i, a_{h2}^i \right) \quad \cdots \quad \Delta_{h,t-1}^{i} \left(Q_{h+1}^i \right) \left(s_{h,t-1}^i, a_{h,t-1}^i \right) \right] \in \mathbb{R}^{t-1}$, and $\boldsymbol{z}_{ht}^{i} \stackrel{\text{def}}{=} \left[z_{h1}^{i} \left(Q_{h+1}^i \right) \left(s_{h1}^i, a_{h1}^i \right) \quad \cdots \quad z_{h,t-1}^{i} \left(Q_{h+1}^i \right) \left(s_{h,t-1}^i, a_{h,t-1}^i \right) \right] \in \mathbb{R}^{t-1}$.

In the next sections we will show how to bound 104 and 105.

B.2. FAILURE EVENT

Define the failure event at step h in episode t as

Definition 2 (Failure Event)**.**

$$E_{ht} \stackrel{\text{def}}{=} I\Big[\exists \left\{Q_{h+1}^i\right\}_{i=1}^M \in \mathcal{Q}_{h+1} \quad \sum_{i=1}^M \left(\boldsymbol{z}_{ht}^i\right)^\top \boldsymbol{\Phi}_{ht}^i \left(\hat{\boldsymbol{B}}_h \hat{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) - \dot{\boldsymbol{B}}_h \dot{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right)\right) > \tag{106}$$

$$F_h^1 \sqrt{\sum_{i=1}^M \left\| \hat{\boldsymbol{B}}_h \hat{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) - \dot{\boldsymbol{B}}_h \dot{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) \right\|_{\tilde{\boldsymbol{V}}_{ht}^i(\lambda)}^2} + F_h^2\Big] \tag{107}$$

where F_h^1 and F_h^2 will be specified later.

We have the following lemma to bound the probability of E_{ht}.

Lemma 8. *For the input parameter $\delta > 0$, there exists F_h^1 and F_h^2 such that*

$$\mathbb{P}\left(\bigcup_{t=1}^T \bigcup_{h=1}^H E_{ht}\right) \leq \frac{\delta}{2} \tag{108}$$

Proof. According to Lemma A.5 of Du et al. (2020), there exists an ϵ-net $\mathcal{E}_{h+1}^o$ over $\mathcal{O}^{d\times k}$ (with regards to the Frobenius norm) such that $\left|\mathcal{E}_{h+1}^o\right| \leq (6\sqrt{k}/\epsilon')^{kd}$. Moreover, there exists an ϵ-net $\mathcal{E}_{h+1}^b$ over $\mathcal{B}^k$ that $\left|\mathcal{E}_{h+1}^b\right| \leq (1+2/\epsilon')^k$. We can show a corresponding ϵ-net $\mathcal{E}_{h+1}^{\text{mul}} \stackrel{\text{def}}{=} \mathcal{E}_{h+1}^o \times \left(\mathcal{E}_{h+1}^b\right)^M$ over Θ_{h+1}.

For any $\left(Q_{h+1}^1\left(\boldsymbol{B}_{h+1}\boldsymbol{w}_{h+1}^1\right), \cdots, Q_{h+1}^M\left(\boldsymbol{B}_{h+1}\boldsymbol{w}_{h+1}^M\right)\right) \in \mathcal{Q}_{h+1}$, there exists $\bar{\boldsymbol{B}}_{h+1} \in \mathcal{E}_{h+1}^o$ and $\left(\bar{\boldsymbol{w}}_{h+1}^1, \cdots, \bar{\boldsymbol{w}}_{h+1}^M\right) \in \left(\mathcal{E}_{h+1}^b\right)^M$ such that

$$\left\|\boldsymbol{B}_{h+1} - \bar{\boldsymbol{B}}_{h+1}\right\|_F \leq \epsilon' \quad \left\|\boldsymbol{w}_{h+1}^i - \bar{\boldsymbol{w}}_{h+1}^i\right\|_2 \leq \epsilon', \forall i \in [M]$$

Therefore,

$$\left\|\boldsymbol{B}_{h+1}\boldsymbol{w}_{h+1}^i - \bar{\boldsymbol{B}}_{h+1}\bar{\boldsymbol{w}}_{h+1}^i\right\|_2 \leq 2\epsilon', \forall i \in [M]$$

Define $\bar{Q}_{h+1}^i$ to be $Q_{h+1}^i\left(\bar{\boldsymbol{B}}_{h+1}\bar{\boldsymbol{w}}_{h+1}^i\right)$, and let $\bar{\boldsymbol{z}}_{ht}^i \stackrel{\text{def}}{=} \left[z_{h1}^i\left(\bar{Q}_{h+1}^i\right)\left(s_{h1}^i, a_{h1}^i\right) \quad \cdots \quad z_{h,t-1}^i\left(\bar{Q}_{h+1}^i\right)\left(s_{h,t-1}^i, a_{h,t-1}^i\right)\right] \in \mathbb{R}^{t-1}$, then

$$\sum_{i=1}^M \left(\boldsymbol{z}_{ht}^i\right)^\top \boldsymbol{\Phi}_{ht}^i \left(\hat{\boldsymbol{B}}_h \hat{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) - \dot{\boldsymbol{B}}_h \dot{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right)\right) \tag{109}$$

$$= \sum_{i=1}^M \left(\bar{\boldsymbol{z}}_{ht}^i\right)^\top \boldsymbol{\Phi}_{ht}^i \left(\hat{\boldsymbol{B}}_h \hat{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) - \dot{\boldsymbol{B}}_h \dot{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right)\right) \tag{110}$$

$$+ \sum_{i=1}^M \left(\boldsymbol{z}_{ht}^i - \bar{\boldsymbol{z}}_{ht}^i\right)^\top \boldsymbol{\Phi}_{ht}^i \left(\hat{\boldsymbol{B}}_h \hat{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) - \dot{\boldsymbol{B}}_h \dot{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right)\right) \tag{111}$$

For fixed $\left\{\bar{\boldsymbol{B}}_{h+1}\bar{\boldsymbol{w}}_{h+1}^i\right\}_{i=1}^M \in \mathcal{E}_{h+1}^{\text{mul}}$, $z_{h,j}^i\left(\bar{Q}_{h+1}^i\right)\left(s_{h,j}^i, a_{h,j}^i\right)$ is zero-mean 1-subgaussian conditioned on $\mathcal{F}_{h,j}$ according to Assumption 3. Thus, we can use exactly the same argument as in Lemma 1 to show that

$$\sum_{i=1}^{M}\left(\bar{\boldsymbol{z}}_{ht}^{i}\right)^{\top}\boldsymbol{\Phi}_{ht}^{i}\left(\hat{\boldsymbol{B}}_{h}\hat{\boldsymbol{w}}_{h}^{i}\left(Q_{h+1}^{i}\right)-\dot{\boldsymbol{B}}_{h}\dot{\boldsymbol{w}}_{h}^{i}\left(Q_{h+1}^{i}\right)\right) \tag{112}$$

$$\leq\sqrt{Mk+5kd\log(kMT)+2\log(1/\delta')}\sqrt{\sum_{i=1}^{M}\left\|\hat{\boldsymbol{B}}_{h}\hat{\boldsymbol{w}}_{h}^{i}\left(Q_{h+1}^{i}\right)-\dot{\boldsymbol{B}}_{h}\dot{\boldsymbol{w}}_{h}^{i}\left(Q_{h+1}^{i}\right)\right\|_{\tilde{\boldsymbol{V}}_{ht}^{i}(\lambda)}^{2}} \tag{113}$$

$$+\sqrt{2\log(2MT/\delta')}+\sqrt{k+3kd\log(kMT)+\log(1/\delta')} \tag{114}$$

by setting $\epsilon=\frac{1}{kM^{2}T^{2}}$, $\delta_{1}=\frac{\delta'}{2\left(\frac{6\sqrt{2k}}{\epsilon}\right)^{2kd}}$, and $\delta_{2}=\frac{\delta'}{2MT}$ in equation 54. Thus, we have that with probability $1-\delta'$ the inequality above holds for any $h\in[H],t\in[T]$. Take $\delta=\frac{\delta'}{2\left|\mathcal{E}_{h+1}^{\text{mul}}\right|}$, by union bound we know the above ineqaulity holds with probability $1-\delta$ for any $\left\{\bar{\boldsymbol{B}}_{h+1}\bar{\boldsymbol{w}}_{h+1}^{i}\right\}_{i=1}^{M}\in\mathcal{E}_{h+1}^{\text{mul}}$ and any $h\in[H],t\in[T]$.

Since it holds that $\left|Q_{h+1}^{i}\left(\boldsymbol{B}_{h+1}\boldsymbol{w}_{h+1}^{i}\right)(s,a)-Q_{h+1}^{i}\left(\bar{\boldsymbol{B}}_{h+1}\bar{\boldsymbol{w}}_{h+1}^{i}\right)(s,a)\right|\leq 2\epsilon'$ for any $(s,a)\in\mathcal{S}\times\mathcal{A},i\in[M]$, we have

$$\left|z_{hj}^{i}\left(\bar{Q}_{h+1}^{i}\right)\left(s_{hj}^{i},a_{hj}^{i}\right)-z_{hj}^{i}\left(Q_{h+1}^{i}\right)\left(s_{hj}^{i},a_{hj}^{i}\right)\right|\leq 8\epsilon' \tag{115}$$

Then we have

$$\sum_{i=1}^{M}\left(\boldsymbol{z}_{ht}^{i}-\bar{\boldsymbol{z}}_{ht}^{i}\right)^{\top}\boldsymbol{\Phi}_{ht}^{i}\left(\hat{B}_{h}\hat{w}_{h}^{i}\left(Q_{h+1}^{i}\right)-\dot{\boldsymbol{B}}_{h}\dot{\boldsymbol{w}}_{h}^{i}\left(Q_{h+1}^{i}\right)\right) \tag{116}$$

$$\leq\sum_{i=1}^{M}\left\|\left(\boldsymbol{\Phi}_{ht}^{i}\right)^{\top}\left(\boldsymbol{z}_{ht}^{i}-\bar{\boldsymbol{z}}_{ht}^{i}\right)\right\|_{\tilde{\boldsymbol{V}}_{ht}^{i}(\lambda)^{-1}}\left\|\hat{\boldsymbol{B}}_{h}\hat{\boldsymbol{w}}_{h}^{i}\left(Q_{h+1}^{i}\right)-\dot{\boldsymbol{B}}_{h}\dot{\boldsymbol{w}}_{h}^{i}\left(Q_{h+1}^{i}\right)\right\|_{\tilde{\boldsymbol{V}}_{ht}^{i}(\lambda)} \tag{117}$$

$$\leq 8\epsilon'\sqrt{T}\sum_{i=1}^{M}\left\|\hat{\boldsymbol{B}}_{h}\hat{\boldsymbol{w}}_{h}^{i}\left(Q_{h+1}^{i}\right)-\dot{\boldsymbol{B}}_{h}\dot{\boldsymbol{w}}_{h}^{i}\left(Q_{h+1}^{i}\right)\right\|_{\tilde{\boldsymbol{V}}_{ht}^{i}(\lambda)} \tag{118}$$

$$\leq 8\epsilon'\sqrt{MT}\sqrt{\sum_{i=1}^{M}\left\|\hat{\boldsymbol{B}}_{h}\hat{\boldsymbol{w}}_{h}^{i}\left(Q_{h+1}^{i}\right)-\dot{\boldsymbol{B}}_{h}\dot{\boldsymbol{w}}_{h}^{i}\left(Q_{h+1}^{i}\right)\right\|_{\tilde{\boldsymbol{V}}_{ht}^{i}(\lambda)}^{2}} \tag{119}$$

for arbitrary $\{Q_{h+1}^{i}\}$ and any $h\in[H],t\in[T]$. The second inequality follows from the Projection Bound (Lemma 8) in Zanette et al. (2020a).

Take $\epsilon'=1/8\sqrt{MT}$, we finally finish the proof by setting

$$F_{h}^{1}\overset{\text{def}}{=}\sqrt{9kd\log(kMT)+5Mk\log(MT)+2\log(2/\delta)} \tag{120}$$

$$F_{h}^{2}\overset{\text{def}}{=}\sqrt{4kd\log(kMT)+5Mk\log(MT)+2\log(2/\delta)} \tag{121}$$

$$+\sqrt{k+5kd\log(kMT)+2Mk\log(MT)+\log(2/\delta)} \tag{122}$$

□

In the next sections we assume the failure event $\bigcup_{t=1}^{T}\bigcup_{h=1}^{H}E_{ht}$ won't happen.

B.3. Bellman Error

Outside the failure event, we can bound the estimation error of the least-square regression 15.

Lemma 9. *For any episode $t \in [T]$ and step $h \in [H]$, any $\left\{Q_{h+1}^i\right\}_{i=1}^M \in \mathcal{Q}_{h+1}$, we have*

$$\sum_{i=1}^{M} \left\| \hat{\boldsymbol{B}}_h \hat{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) - \dot{\boldsymbol{B}}_h \dot{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) \right\|_{\tilde{\boldsymbol{V}}_{ht}^i(\lambda)}^2 \leq \alpha_{ht} \stackrel{\text{def}}{=} \left(2\sqrt{MT\mathcal{I}} + 2F_h^1 + \sqrt{2F_h^2 + 4MD^2\lambda} \right)^2 \tag{123}$$

Proof. Recall that

$$\sum_{i=1}^{M} \left\| \boldsymbol{\Phi}_{ht}^i \hat{\boldsymbol{B}}_h \hat{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) - \boldsymbol{\Phi}_{ht}^i \dot{\boldsymbol{B}}_h \dot{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) \right\|_2^2 \tag{124}$$

$$\leq 2 \sum_{i=1}^{M} \left(\boldsymbol{\Delta}_{ht}^i\right)^\top \boldsymbol{\Phi}_{ht}^i \left(\hat{\boldsymbol{B}}_h \hat{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) - \dot{\boldsymbol{B}}_h \dot{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) \right) \tag{125}$$

$$+ 2 \sum_{i=1}^{M} \left(\boldsymbol{z}_{ht}^i\right)^\top \boldsymbol{\Phi}_{ht}^i \left(\hat{\boldsymbol{B}}_h \hat{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) - \dot{\boldsymbol{B}}_h \dot{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) \right) \tag{126}$$

For the first term, we have

$$\sum_{i=1}^{M} \left(\boldsymbol{\Delta}_{ht}^i\right)^\top \Phi_{ht}^i \left(\hat{\boldsymbol{B}}_h \hat{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) - \dot{\boldsymbol{B}}_h \dot{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) \right) \tag{127}$$

$$\leq \sum_{i=1}^{M} \left\| \left(\boldsymbol{\Phi}_{ht}^i\right)^\top \boldsymbol{\Delta}_{ht}^i \right\|_{\tilde{\boldsymbol{V}}_{ht}^i(\lambda)^{-1}} \left\| \hat{\boldsymbol{B}}_h \hat{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) - \dot{\boldsymbol{B}}_h \dot{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) \right\|_{\tilde{\boldsymbol{V}}_{ht}^i(\lambda)} \tag{128}$$

$$\leq \sqrt{T\mathcal{I}} \sum_{i=1}^{M} \left\| \hat{\boldsymbol{B}}_h \hat{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) - \dot{\boldsymbol{B}}_h \dot{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) \right\|_{\tilde{\boldsymbol{V}}_{ht}^i(\lambda)} \tag{129}$$

$$\leq \sqrt{MT\mathcal{I}} \sqrt{\sum_{i=1}^{M} \left\| \hat{\boldsymbol{B}}_h \hat{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) - \dot{\boldsymbol{B}}_h \dot{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) \right\|_{\tilde{\boldsymbol{V}}_{ht}^i(\lambda)}^2} \tag{130}$$

The second inequality follows from the Projection Bound (Lemma 8) in Zanette et al. (2020a), and the last inequality is due to Cauchy-Schwarz.

Outside the failure event, we have

$$\sum_{i=1}^{M} \left\| \hat{\boldsymbol{B}}_h \hat{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) - \dot{\boldsymbol{B}}_h \dot{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) \right\|_{\tilde{\boldsymbol{V}}_{ht}^i(\lambda)}^2 \tag{131}$$

$$\leq \sum_{i=1}^{M} \left\| \boldsymbol{\Phi}_{ht}^i \hat{\boldsymbol{B}}_h \hat{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) - \boldsymbol{\Phi}_{ht}^i \dot{\boldsymbol{B}}_h \dot{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) \right\|_2^2 + 4MD^2\lambda \tag{132}$$

$$\leq \left(2\sqrt{MT\mathcal{I}} + 2F_h^1 \right) \sqrt{\sum_{i=1}^{M} \left\| \hat{\boldsymbol{B}}_h \hat{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) - \dot{\boldsymbol{B}}_h \dot{\boldsymbol{w}}_h^i \left(Q_{h+1}^i\right) \right\|_{\tilde{\boldsymbol{V}}_{ht}^i(\lambda)}^2} + 2F_h^2 + 4MD^2\lambda \tag{133}$$

which implies

$$\sum_{i=1}^{M}\left\|\hat{\boldsymbol{B}}_h\hat{\boldsymbol{w}}_h^i\left(Q_{h+1}^i\right)-\dot{\boldsymbol{B}}_h\dot{\boldsymbol{w}}_h^i\left(Q_{h+1}^i\right)\right\|_{\tilde{\boldsymbol{V}}_{ht}^i(\lambda)}^2 \tag{134}$$

$$\leq\left(2\sqrt{MT\mathcal{I}}+2F_h^1\right)^2+2F_h^2+4MD^2\lambda+\left(2\sqrt{MT\mathcal{I}}+2F_h^1\right)\sqrt{2F_h^2+4MD^2\lambda} \tag{135}$$

$$\leq\left(2\sqrt{MT\mathcal{I}}+2F_h^1+\sqrt{2F_h^2+4MD^2\lambda}\right)^2 \tag{136}$$

□

Lemma 10 (Bound on Bellman Error)**.** *Outside the failure event, for any feasible solution* $\left\{Q_h^i\left(\bar{\theta}_h^i\right)\right\}_h^i$ *(*$\bar{Q}_h^i$ *for short, with a little abuse of notations) of the global optimization procedure in definition 1, for any* $(s,a)\in\mathcal{S}\times\mathcal{A}$*, any* $h\in[H]$*,* $t\in[T]$

$$\sum_{i=1}^{M}\left|\bar{Q}_h^i(s,a)-\mathcal{T}_h^i\bar{Q}_{h+1}^i(s,a)\right|\leq M\mathcal{I}+2\sqrt{\alpha_{ht}\cdot\sum_{i=1}^{M}\|\boldsymbol{\phi}(s,a)\|_{\tilde{\boldsymbol{V}}_{ht}^i(\lambda)^{-1}}^2} \tag{137}$$

Proof.

$$\sum_{i=1}^{M}\left|\bar{Q}_h^i(s,a)-\mathcal{T}_h^i\bar{Q}_{h+1}^i(s,a)\right|=\sum_{i=1}^{M}\left|\boldsymbol{\phi}(s,a)^\top\bar{\boldsymbol{\theta}}_h^i-\boldsymbol{\phi}(s,a)^\top\dot{\boldsymbol{\theta}}_h^i\left(\bar{Q}_{h+1}^i\right)-\Delta_h^i\left(\bar{Q}_{h+1}^i\right)(s,a)\right| \tag{138}$$

$$\leq M\mathcal{I}+\sum_{i=1}^{M}\left|\boldsymbol{\phi}(s,a)^\top\bar{\boldsymbol{\theta}}_h^i-\boldsymbol{\phi}(s,a)^\top\dot{\boldsymbol{\theta}}_h^i\left(\bar{Q}_{h+1}^i\right)\right| \tag{139}$$

$$\leq M\mathcal{I}+\sum_{i=1}^{M}\left(\left|\boldsymbol{\phi}(s,a)^\top\dot{\boldsymbol{\theta}}_h^i\left(\bar{Q}_{h+1}^i\right)-\boldsymbol{\phi}(s,a)^\top\hat{\boldsymbol{\theta}}_h^i\right|+\left|\boldsymbol{\phi}(s,a)^\top\hat{\boldsymbol{\theta}}_h^i-\boldsymbol{\phi}(s,a)^\top\bar{\boldsymbol{\theta}}_h^i\right|\right) \tag{140}$$

$$\leq M\mathcal{I}+\sum_{i=1}^{M}\|\boldsymbol{\phi}(s,a)\|_{\tilde{\boldsymbol{V}}_{ht}^i(\lambda)^{-1}}\left(\left\|\dot{\boldsymbol{\theta}}_h^i\left(\bar{Q}_{h+1}^i\right)-\hat{\boldsymbol{\theta}}_h^i\right\|_{\tilde{\boldsymbol{V}}_{ht}^i(\lambda)}+\left\|\hat{\boldsymbol{\theta}}_h^i-\bar{\boldsymbol{\theta}}_h^i\right\|_{\tilde{\boldsymbol{V}}_{ht}^i(\lambda)}\right) \tag{141}$$

$$\leq M\mathcal{I}+2\sqrt{\alpha_{ht}\cdot\sum_{i=1}^{M}\|\boldsymbol{\phi}(s,a)\|_{\tilde{\boldsymbol{V}}_{ht}^i(\lambda)^{-1}}^2} \tag{142}$$

The first equality is due to the definition of $\Delta_h^i\left(\bar{Q}_{h+1}^i\right)(s,a)$. The last inequality is due to lemma 9. □

B.4. OPTIMISM

We can find the "best" approximator of optimal value functions in our function class recursively defined as

$$\left(\boldsymbol{\theta}_h^{1*}, \boldsymbol{\theta}_h^{2*}, \cdots, \boldsymbol{\theta}_h^{M*}\right) \stackrel{\text{def}}{=} \underset{\left(\boldsymbol{\theta}_h^1, \boldsymbol{\theta}_h^2, \cdots, \boldsymbol{\theta}_h^M\right) \in \Theta_h}{\operatorname{argmin}} \sup_{s,a,i} \left| \left(\boldsymbol{\phi}(s,a)^\top \boldsymbol{\theta}_h^i - \mathcal{T}_h^i Q_{h+1}^i \left(\boldsymbol{\theta}_{h+1}^{i*}\right)\right)(s,a) \right| \tag{143}$$

with $\boldsymbol{\theta}_{H+1}^{i*} = \mathbf{0}, \forall i \in [M]$

For the accuracy of this best approximator, we have

Lemma 11. *For any $h \in [H]$,*

$$\sup_{(s,a) \in \mathcal{S} \times \mathcal{A}, i \in [M]} \left| Q_h^{i*}(s,a) - \boldsymbol{\phi}(s,a)^\top \boldsymbol{\theta}_h^* \right| \leq (H - h + 1)\mathcal{I}$$

where Q_h^{i*} is the optimal value function for task i. This lemma is derived directly from Lemma 6 in Zanette et al. (2020a).

For our solution of the problem in Definition 1 in episode t, we have the following lemma:

Lemma 12. *$\left\{\left(\boldsymbol{\theta}_h^{1*}, \boldsymbol{\theta}_h^{2*}, \cdots, \boldsymbol{\theta}_h^{M*}\right)\right\}_{h=1}^H$ is a feasible solution of the problem in Definition 1. Moreover, denote the solution of the problem in Definition 1 in episode t by $\bar{\boldsymbol{\theta}}_{ht}^i$ for $h \in [H], i \in [M]$, it holds that*

$$\sum_{i=1}^M V_1^i \left(\bar{\boldsymbol{\theta}}_{1t}^i\right)\left(s_{1t}^i\right) \geq \sum_{i=1}^M V_1^{i*}\left(s_{1t}^i\right) - MH\mathcal{I} \tag{144}$$

Proof. First we show that $\left\{\left(\boldsymbol{\theta}_h^{1*}, \boldsymbol{\theta}_h^{2*}, \cdots, \boldsymbol{\theta}_h^{M*}\right)\right\}_{h=1}^H$ is a feasible solution. We can construct $\left\{\bar{\boldsymbol{\xi}}_h^i\right\}_{i=1}^M$ so that $\bar{\boldsymbol{\theta}}_h^i = \boldsymbol{\theta}_h^{i*}$ and no other constraints are violated. We use an inductive construction, and the base case when $\bar{\boldsymbol{\theta}}_{H+1}^i = \boldsymbol{\theta}_{H+1}^{i*} = 0$ is trivial.

Now suppose we have $\left\{\bar{\boldsymbol{\xi}}_y^i\right\}_{i=1}^M$ for $y = h+1, ..., H$ such that $\bar{\boldsymbol{\theta}}_y^i = \boldsymbol{\theta}_y^{i*}$ for $y = h+1, ..., H$ and $i \in [M]$, we show we can find $\left\{\bar{\boldsymbol{\xi}}_h^i\right\}_{i=1}^M$ so $\bar{\boldsymbol{\theta}}_h^i = \boldsymbol{\theta}_h^{i*}$ for $i \in [M]$, and no constraints are violated. From the definition of $\boldsymbol{\theta}_h^{i*}$ we can set (with a little abuse of notations)

$$\dot{\boldsymbol{\theta}}_h^i \left(\boldsymbol{\theta}_{h+1}^{i*}\right) = \boldsymbol{\theta}_h^{i*} \tag{145}$$

According to lemma 9 we have

$$\sum_{i=1}^M \left\| \hat{\boldsymbol{\theta}}_h^i \left(\boldsymbol{\theta}_{h+1}^{i*}\right) - \dot{\boldsymbol{\theta}}_h^i \left(\boldsymbol{\theta}_{h+1}^{i*}\right) \right\|_{\tilde{\boldsymbol{V}}_{ht}^i(\lambda)}^2 \leq \alpha_{ht} \tag{146}$$

Therefore, set $\bar{\boldsymbol{\xi}}_h^i = \dot{\boldsymbol{\theta}}_h^i \left(\boldsymbol{\theta}_{h+1}^{i*}\right) - \hat{\boldsymbol{\theta}}_h^i \left(\boldsymbol{\theta}_{h+1}^{i*}\right)$, then

$$\begin{aligned}
\bar{\boldsymbol{\theta}}_h^i &= \hat{\boldsymbol{\theta}}_h^i \left(\bar{\boldsymbol{\theta}}_{h+1}^i\right) + \bar{\boldsymbol{\xi}}_h^i && (147)\\
&= \hat{\boldsymbol{\theta}}_h^i \left(\boldsymbol{\theta}_{h+1}^{i*}\right) + \dot{\boldsymbol{\theta}}_h^i \left(\boldsymbol{\theta}_{h+1}^{i*}\right) - \hat{\boldsymbol{\theta}}_h^i \left(\boldsymbol{\theta}_{h+1}^{i*}\right) && (148)\\
&= \boldsymbol{\theta}_h^{i*} && (149)
\end{aligned}$$

Finally, we can verify $\left(\bar{\boldsymbol{\theta}}_h^1, ..., \bar{\boldsymbol{\theta}}_h^M\right) \in \Theta_h$ from $\left(\boldsymbol{\theta}_h^{1*}, \cdots, \boldsymbol{\theta}_h^{M*}\right) \in \Theta_h$.

Since $\bar{\boldsymbol{\theta}}_{1t}^i$ is the optimal solution, we can finish the proof by showing

$$\sum_{i=1}^{M} V_1^i\left(\bar{\boldsymbol{\theta}}_{1t}^i\right)\left(s_{1t}^i\right) = \sum_{i=1}^{M} \max_a \boldsymbol{\phi}\left(s_{1t}^i, a\right)^\top \bar{\boldsymbol{\theta}}_{1t}^i \tag{150}$$

$$\geq \sum_{i=1}^{M} \max_a \boldsymbol{\phi}\left(s_{1t}^i, a\right)^\top \boldsymbol{\theta}_1^{i*} \qquad \text{(since } \theta_1^{i*} \text{ is the feasible solution)} \tag{151}$$

$$\geq \sum_{i=1}^{M} \boldsymbol{\phi}\left(s_{1t}^i, \pi_1^{i*}\left(s_{1t}^i\right)\right)^\top \boldsymbol{\theta}_1^{i*} \tag{152}$$

$$\geq \sum_{i=1}^{M} Q_h^{i*}\left(s_{1t}^i, \pi_1^{i*}\left(s_{1t}^i\right)\right) - MH\mathcal{I} \qquad \text{(by Lemma 11)} \tag{153}$$

$$\geq \sum_{i=1}^{M} V_h^{i*}\left(s_{1t}^i\right) - MH\mathcal{I} \tag{154}$$

□

B.5. Regret Bound

We are ready to present the proof of our regret bound.

From Lemma 8 we know that the failure event $\bigcup_{t=1}^{T}\bigcup_{h=1}^{H} E_{ht}$ happens with probability at most $\delta/2$, so we assume it does not happen. Then we can decompose the regret as

$$\begin{aligned}
\mathrm{Reg}(T) &= \sum_{t=1}^{T}\sum_{i=1}^{M}\left(V_1^{i*} - V_1^{\pi_t^i}\right)\left(s_{1t}^i\right) && (155)\\
&= \sum_{t=1}^{T}\sum_{i=1}^{M}\left(V_1^{i*} - V_1^{i}\left(\bar{\boldsymbol{\theta}}_{1t}^i\right)\right)\left(s_{1t}^i\right) + \sum_{t=1}^{T}\sum_{i=1}^{M}\left(V_1^{i}\left(\bar{\boldsymbol{\theta}}_{1t}^i\right) - V_1^{\pi_t^i}\right)\left(s_{1t}^i\right) && (156)\\
&\le \sum_{t=1}^{T}\sum_{i=1}^{M}\left(V_1^{i}\left(\bar{\boldsymbol{\theta}}_{1t}^i\right) - V_1^{\pi_t^i}\right)\left(s_{1t}^i\right) + MHT\mathcal{I} \qquad \text{(by Lemma 12)} && (157)
\end{aligned}$$

Let $a_{ht}^i = \pi_t^i\left(s_{ht}^i\right)$, and denote $Q_h^i\left(\bar{\boldsymbol{\theta}}_{ht}^i\right)(V_h^i\left(\bar{\boldsymbol{\theta}}_{ht}^i\right))$ by $\bar{Q}_{ht}^i(\bar{V}_{ht}^i)$ for short, we have

$$\begin{aligned}
\sum_{i=1}^{M}\left(\bar{V}_{ht}^i - V_h^{\pi_t^i}\right)\left(s_{ht}^i\right) &= \sum_{i=1}^{M}\left(\bar{Q}_{ht}^i - Q_h^{\pi_t^i}\right)\left(s_{ht}^i, a_{ht}^i\right) && (158)\\
&= \sum_{i=1}^{M}\left(\bar{Q}_{ht}^i - \mathcal{T}_h^i\bar{Q}_{h+1,t}^i\right)\left(s_{ht}^i, a_{ht}^i\right) + \sum_{i=1}^{M}\left(\mathcal{T}_h^i\bar{Q}_{h+1,t}^i - Q_h^{\pi_t^i}\right)\left(s_{ht}^i, a_{ht}^i\right) && (159)\\
&\le M\mathcal{I} + 2\sqrt{\alpha_{ht}\cdot\sum_{i=1}^{M}\left\|\boldsymbol{\phi}\left(s_{ht}^i, a_{ht}^i\right)\right\|^2_{\tilde{\boldsymbol{V}}_{ht}^i(\lambda)^{-1}}} + \sum_{i=1}^{M}\mathbb{E}_{s'\sim p_h^i\left(s_{ht}^i, a_{ht}^i\right)}\left[\left(\bar{V}_{h+1,t}^i - V_{h+1}^{\pi_t^i}\right)(s')\right] && (160)\\
&\le \sum_{i=1}^{M}\left(\bar{V}_{h+1,t}^i - V_{h+1}^{\pi_t^i}\right)\left(s_{h+1,t}^i\right) + M\mathcal{I} + 2\sqrt{\alpha_{ht}\cdot\sum_{i=1}^{M}\left\|\boldsymbol{\phi}\left(s_{ht}^i, a_{ht}^i\right)\right\|^2_{\tilde{\boldsymbol{V}}_{ht}^i(\lambda)^{-1}}} + \sum_{i=1}^{M}\zeta_{ht}^i && (161)
\end{aligned}$$

where ζ_{ht}^i is a martingale difference with regards to the filtration $\mathcal{F}_{h,t}$ defined as

$$\zeta_{ht}^i \stackrel{\text{def}}{=} \left(\bar{V}_{h+1,t}^i - V_{h+1}^{\pi_t^i}\right)\left(s_{h+1,t}^i\right) - \mathbb{E}_{s'\sim p_h^i\left(s_{ht}^i, a_{ht}^i\right)}\left[\left(\bar{V}_{h+1,t}^i - V_{h+1}^{\pi_t^i}\right)(s')\right] \tag{162}$$

According to assumption 3 we know $\left|\zeta_{ht}^i\right| \le 4$, so we can apply Azuma-Hoeffding's inequality that with probability $1-\delta/2$ for any $t\in[T]$ and $i\in[M]$

$$\sum_{j=1}^{t}\zeta_{ht}^i \le 4\sqrt{2t\ln\left(\frac{2T}{\delta}\right)} \tag{163}$$

By applying inequality 161 recursively, we can bound the regret as

$$\text{Reg}(T) \leq \sum_{t=1}^{T}\sum_{i=1}^{M}\left(\bar{V}_{1t}^{i} - V_{1}^{\pi_t^i}\right)\left(s_{1t}^{i}\right) + MHT\mathcal{I} \tag{164}$$

$$\leq 2MHT\mathcal{I} + \sum_{t=1}^{T}\sum_{h=1}^{H} 2\sqrt{\alpha_{ht}\cdot\sum_{i=1}^{M}\left\|\boldsymbol{\phi}\left(s_{ht}^{i}, a_{ht}^{i}\right)\right\|^2_{\tilde{\boldsymbol{V}}_{ht}^{i}(\lambda)^{-1}}} + \sum_{i=1}^{M}\sum_{h=1}^{H}\sum_{t=1}^{T}\zeta_{ht}^{i} \tag{165}$$

The last inequality is due to $\bar{V}_{H+1}^{i}(s) = \max_a \boldsymbol{\phi}(s,a)^\top \bar{\boldsymbol{\theta}}_{H+1,t}^{i} = 0, V_{H+1}^{\pi_t^i}(s) = 0$.

The Lemma 11 of Abbasi-Yadkori et al. (2011) gives that for any $i \in [M]$ and $h \in [H]$

$$\sum_{t=1}^{T}\left\|\boldsymbol{\phi}\left(s_{ht}^{i}, a_{ht}^{i}\right)\right\|^2_{\tilde{\boldsymbol{V}}_{ht}^{i}(\lambda)^{-1}} = \tilde{O}\left(d\right) \tag{166}$$

Moreover, by the definition of α_{ht} (see Lemma 9) we know that for any $h \in [H]$ and $t \in [T]$

$$\alpha_{ht} = \tilde{O}\left(Mk + kd + MT\mathcal{I}^2\right) \tag{167}$$

Take all of above we can show the final regret bound.

$$\text{Reg}(T) \leq 2MHT\mathcal{I} + \sum_{t=1}^{T}\sum_{h=1}^{H} 2\sqrt{\alpha_{ht}\cdot\sum_{i=1}^{M}\left\|\boldsymbol{\phi}\left(s_{ht}^{i}, a_{ht}^{i}\right)\right\|^2_{\tilde{\boldsymbol{V}}_{ht}^{i}(\lambda)^{-1}}} + \sum_{i=1}^{M}\sum_{h=1}^{H}\sum_{t=1}^{T}\zeta_{ht}^{i} \tag{168}$$

$$= \tilde{O}\left(MHT\mathcal{I} + \tilde{O}\left(\sqrt{Mk + kd + MT\mathcal{I}^2}\right)\sum_{h=1}^{H}\sum_{t=1}^{T}\sqrt{\sum_{i=1}^{M}\left\|\boldsymbol{\phi}\left(s_{ht}^{i}, a_{ht}^{i}\right)\right\|^2_{\tilde{\boldsymbol{V}}_{ht}^{i}(\lambda)^{-1}}} + MH\sqrt{T}\right) \tag{169}$$

$$= \tilde{O}\left(MHT\mathcal{I} + \tilde{O}\left(\sqrt{Mk + kd + MT\mathcal{I}^2}\right)\sum_{h=1}^{H}\sqrt{T}\cdot\sqrt{\sum_{t=1}^{T}\sum_{i=1}^{M}\left\|\boldsymbol{\phi}\left(s_{ht}^{i}, a_{ht}^{i}\right)\right\|^2_{\tilde{\boldsymbol{V}}_{ht}^{i}(\lambda)^{-1}}} + MH\sqrt{T}\right) \tag{170}$$

$$= \tilde{O}\left(MHT\mathcal{I} + \tilde{O}\left(\tilde{O}\left(\sqrt{Mk + kd + MT\mathcal{I}^2}\right)\cdot H\sqrt{MTd}\right) + MH\sqrt{T}\right) \tag{171}$$

$$= \tilde{O}\left(HM\sqrt{dkT} + Hd\sqrt{MkT} + HMT\sqrt{d\mathcal{I}}\right) \tag{172}$$

C. Proof of Theorem 5

To prove the lower bound for multi-task RL, our idea is to connect the lower bound for the multi-task learning problem to the lower bound in the single-task LSVI setting (Zanette et al., 2020a). in the paper of Zanette et al. (2020a), they assumed the feature dimension d can be varied among different steps, which is denoted as d_h for step h. They proved the lower bound for linear RL in this setting is $\Omega\left(\sum_{h=1}^{H} d_h\sqrt{T} + \sum_{h=1}^{H}\sqrt{d_h\mathcal{I}T}\right)$. However, this lower bound is derived by the hard instance with $d_1 = \sum_{h=2}^{H} d_h$. If we set $d_1 = d_2 = \cdots = d_H = d$ like our setting, we can only obtain the lower bound of $\Omega\left(d\sqrt{T} + \sqrt{d\mathcal{I}T}\right)$ following their proof idea. In fact, the dependence on H in this lower bound can be further improved. In order to obtain a tighter lower bound, we consider the lower bound for single-task misspecified linear MDP. This setting can be proved to be strictly simpler than the LSVI setting following the idea of Proposition 3 in Zanette et al. (2020a). The lower bound for misspecified linear MDP can thus be applied to LSVI setting.

C.1. Lower Bounds for single-task RL

This subsection focus on the lower bound for misspecifed linear MDP setting, in which the transition kernel and the reward function are assume to be approximately linear.

Assumption 5. *(Assumption B in Jin et al. (2020)) For any $\zeta \leq 1$, we say that $\mathrm{MDP}(\mathcal{S}, \mathcal{A}, p, r, H)$ is a ζ-approximate linear MDP with a feature map $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, if for any $h \in [H]$, there exist d unknown measures $\boldsymbol{\theta}_h = (\theta_h^{(1)}, \cdots, \theta_h^{(d)})$ over $\mathcal{S}$ and an unknown vector $\boldsymbol{\nu}_h \in \mathbb{R}^d$ such that for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have*

$$\|p_h(\cdot|s,a) - \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h(\cdot)\rangle\|_{\mathrm{TV}} \leq \zeta \tag{173}$$

$$|r_h(s,a) - \langle \boldsymbol{\phi}(s,a), \boldsymbol{\nu}_h\rangle| \leq \zeta \tag{174}$$

For regularity, we assume that Assumption 3 still holds, and we also assume that there exists a constant D such that $\|\boldsymbol{\theta}_h(s)\| \leq D$ for all $s \in \mathcal{S}, h \in [H]$, $\|\boldsymbol{\nu}_h\| \leq D$ for all $h \in [H]$. $D \geq 4$ suffices in our hard instance construction.

For misspecifed linear MDP, we can prove the following lower bound.

Proposition 1. *Suppose $T \geq \frac{d^2 H}{4}$, $d \geq 10$, $H \geq 10$ and $\zeta \leq \frac{1}{4H}$, there exist a ζ-approximate linear MDP class such that the expected regret of any algorithm on at least a member of the MDP class is at least $\Omega\left(d\sqrt{HT} + HT\mathcal{I}\sqrt{d}\right)$.*

To prove the lower bound, our basic idea is to connect the problem to $\frac{H}{2}$ linear bandit problems. Similar hard instance construction has been used in Zhou et al. (2020a;b). In our construction, the state space $\mathcal{S}$ consists of $H + 2$ states, which is denoted as $x_1, x_2, \cdots, x_{H+2}$. The agent starts the episode in state x_1. In x_h, it can either transits to x_{h+1} or x_{H+2} with certain transition probability. If the agent enters x_{H+2}, it will stay in this state in the remaining steps, i.e. x_{H+2} is an absorbing state. For each state, there are 2^{d-4} actions and $\mathcal{A} = \{-1, 1\}^{d-4}$. Suppose the agent takes action $\boldsymbol{a} \in \{-1, 1\}^{d-4}$ in state s_h, the transition probability to state s_{h+1} and s_{H+2} is $1 - \zeta_h(\boldsymbol{a}) - \delta - \boldsymbol{\mu}_h^\top \boldsymbol{a}$ and $\delta + \zeta_h(\boldsymbol{a}) + \boldsymbol{\mu}_h^\top \boldsymbol{a}$ respectively. Here $|\zeta_h(\boldsymbol{a})| \leq \zeta$ denotes the approximation error of linear representation, $\delta = 1/H$ and $\boldsymbol{\mu}_h \in \{-\Delta, \Delta\}^{d-4}$ with $\Delta = \sqrt{\delta/T}/(4\sqrt{2})$ so that the probability is well-defined. The reward can only be obtained in x_{H+2}, with $r_h(x_{H+2,a}) = 1/H$ for any h, a. We assume the reward to be deterministic.

We can check that this construction satisfies Assumption 5 with $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ defined in the following way:

$$\boldsymbol{\phi}(s, \boldsymbol{a}) = \begin{cases} \left(0, \alpha, \alpha\delta, 0, \beta \boldsymbol{a}^\top\right)^\top & s = x_1, x_2, \cdots, x_H \\ \left(0, 0, 0, \alpha, \mathbf{0}^\top\right)^\top & s = x_{H+1} \\ \left(\alpha, 0, 0, \alpha, \mathbf{0}^\top\right)^\top & s = x_{H+2} \end{cases}$$

$$\boldsymbol{\theta}_h(s') = \begin{cases} \left(0, \frac{1}{\alpha}, -\frac{1}{\alpha}, 0, -\frac{\boldsymbol{\mu}_h^\top}{\beta}\right)^\top & s' = x_{h+1} \\ \left(0, 0, \frac{1}{\alpha}, \frac{1}{\alpha}, \frac{\boldsymbol{\mu}_h^\top}{\beta}\right)^\top & s = x_{H+2} \\ \mathbf{0} & \text{otherwise} \end{cases}$$

$\boldsymbol{\nu}_h$ is defined to be $(\frac{1}{H\alpha}, \mathbf{0}^\top)^\top$, and $\alpha = \sqrt{1/(2 + \Delta(d-4))}$, $\beta = \sqrt{\Delta/(2 + \Delta(d-4))}$. Note that $\|\boldsymbol{\phi}(s,a)\| \leq 1$, $\|\boldsymbol{\theta}_h(s')\| \leq D$ and $\|\boldsymbol{\nu}_h\| \leq D$ hold for any s, a, s', h when $T \geq d^2 H/4$.

Since the rewarding state is only x_{H+2}, the optimal strategy in state x_h ($h \leq H$) is to take an action that maximizes the probability of entering x_{H+2}, i.e., to maximize $\boldsymbol{\mu}_h^\top \boldsymbol{a} + \zeta(\boldsymbol{a})$. That is to say, we can regard the problem of finding the optimal action in state s_h and step h as finding the optimal arm for a $d-4$-dimensional approximately (misspecified) linear bandits problem. Thanks to the choice of δ such that $(1-\delta)^{H/2}$ is a constant, there is sufficiently high probability of entering state x_h for any $h \leq H/2$. Therefore, we can show that this problem is harder than solving $H/2$ misspecified linear bandit problems. This following lemma characterizes this intuition. The lemma follows the same idea of Lemma C.7 in Zhou et al. (2020a), though our setting is more difficult since we consider misspecified case.

Lemma 13. *Suppose $H \geq 10$, $d \geq 10$ and $(d-4)\Delta \leq \frac{1}{2H}$. We define $r_h^b(\boldsymbol{a}) = \boldsymbol{\mu}^\top \boldsymbol{a} + \zeta_h(\boldsymbol{a})$, which can be regarded as the corresponding reward for the equivalent linear bandit problem in step h. Fix $\boldsymbol{\mu} \in (\{-\Delta, \Delta\}^{d-4})^H$. Fix a possibly history dependent policy π. Letting $V^\star$ and V^π be the optimal value function and the value function of policy π respectively, we have*

$$V_1^\star(s_1) - V_1^\pi(s_1) \geq 0.02 \sum_{h=1}^{H/2} \left(\max_{\boldsymbol{a} \in \mathcal{A}} r_h^b(\boldsymbol{a}) - \sum_{\boldsymbol{a} \in \mathcal{A}} \pi_h(\boldsymbol{a}|s_h) r_h^b(\boldsymbol{a}) \right) \tag{175}$$

Proof. Note that the only rewarding state is x_{H+2} with $r_h(x_{H+2}, \boldsymbol{a}) = \frac{1}{H}$. Therefore, the value function of a certain policy π can be calculated as:

$$V_1^\pi(x_1) = \sum_{h=1}^{H-1} \frac{H-h}{H} \mathbb{P}(N_h|\pi) \tag{176}$$

where N_h denotes the event of visiting state x_h in step h and then transits to x_{H+2}, i.e. $N_h = \{s_h = x_h, s_{h+1} = x_{H+2}\}$. Suppose $\omega_h^\pi = \sum_{\boldsymbol{a} \in \mathcal{A}} \pi_h(\boldsymbol{a}|s_h) r_h^b(\boldsymbol{a})$ and $\omega_h^\star = \max_{\boldsymbol{a} \in \mathcal{A}} r_h^b(\boldsymbol{a})$. By the law of total probability and the Markov property, we have

$$\mathbb{P}(N_h|\pi) = (\delta + \omega_h^\pi) \prod_{j=1}^{h-1} (1 - \delta - \omega_h^\pi) \tag{177}$$

Thus we have

$$V_1^\pi(x_1) = \sum_{h=1}^{H-1} \frac{H-h}{H} (\delta + \omega_h^\pi) \prod_{j=1}^{h-1} (1 - \delta - \omega_h^\pi) \tag{178}$$

Similarly, for the value function of the optimal policy, we have

$$V_1^\star(x_1) = \sum_{h=1}^{H-1} \frac{H-h}{H} (\delta + \omega_h^\star) \prod_{j=1}^{h-1} (1 - \delta - \omega_h^\star) \tag{179}$$

Define $S_i = \sum_{h=i}^{H-1} \frac{H-h}{H} (\delta + \omega_h^\pi) \prod_{j=i}^{h-1} (1 - \delta - \omega_h^\pi)$ and $T_i = \sum_{h=i}^{H-1} \frac{H-h}{H} (\delta + \omega_h^\star) \prod_{j=i}^{h-1} (1 - \delta - \omega_h^\star)$. Then we have $V_1^\star(x_1) - V_1^\pi(x_1) = T_1 - S_1$. Notice that

$$S_i = \frac{H-i}{H} (\omega_i^\pi + \delta) + S_{i+1}(1 - \omega_i^\pi - \delta) \tag{180}$$

$$T_i = \frac{H-i}{H} (\omega_i^\star + \delta) + T_{i+1}(1 - \omega_i^\star - \delta) \tag{181}$$

Thus we have

$$T_i - S_i = \left(\frac{H-i}{H} - T_{i+1} \right) (\omega_i^\star - \omega_i^\pi) + (T_{i+1} - S_{i+1})(1 - \omega_i^\pi - \delta) \tag{182}$$

By induction, we get

$$T_1 - S_1 = \sum_{h=1}^{H-1} (\omega_i^\star - \omega_i^\pi) (\frac{H-h}{H} - T_{h+1}) \prod_{j=1}^{h-1} (1 - \omega_j^\pi - \delta) \tag{183}$$

Since the reward is non-negative and only occurs in x_{H+2}, we know that $V_1^\star(x_1) \geq V_2^\star(x_2) \geq \cdots \geq V_1^\star(x_H)$. Thus we have $T_h \leq T_1 = V_1^\star(x_1) \leq \sum_{h=1}^H \mathbb{P}(N_h|\pi^\star)$. If N_h doesn't happen for any $h \in [H]$, then the agent must enter x_{H+1}. The

probability of this event has the following form:

$$\mathbb{P}\left(\neg\left(\cup_{h\in[H]}N_h|\pi^\star\right)\right) =1-\prod_{h=1}^{H}\mathbb{P}(N_h|\pi^\star) \tag{184}$$

$$=\prod_{h\in[H]}(1-\delta-\omega_h^\star) \tag{185}$$

$$\geq\prod_{h\in[H]}(1-\frac{1}{H}+\frac{1}{2H}) \tag{186}$$

$$=(1-\frac{1}{2H})^H \tag{187}$$

$$\geq 0.6 \tag{188}$$

The fist inequality is due to $\delta=\frac{2}{H}$ and $|\omega_h^\star|\leq\frac{1}{H}$. The above discussion indicates that $T_h\leq 0.4$, thus $\frac{H-h}{H}-T_{h+1}\geq 0.1$ for $h\leq H/2$. Similarly, $\prod_{j=1}^{h-1}(1-\omega_j^\pi-\delta)\geq(1-\frac{3}{2H})^{H-1}\geq 0.2$. Combining with Eqn 183, we have

$$T_1-S_1\geq 0.02\sum_{h=1}^{\frac{H}{2}}(\omega_h^\star-\omega_h^\pi)=0.02\sum_{h=1}^{H/2}\left(\max_{\boldsymbol{a}\in\mathcal{A}}r_h^b(\boldsymbol{a})-\sum_{\boldsymbol{a}\in\mathcal{A}}\pi_h(\boldsymbol{a}|s_h)r_h^b(\boldsymbol{a})\right) \tag{189}$$

Combining with the definition of T_1 and S_1, we can prove the lemma. □

After proving Lemma 13, we are ready to prove Proposition 1.

Proof. (proof of Proposition 1) By Lemma 13, we know that we can decompose the sub-optimality gap of a policy π in the following way:

$$V_1^\star(s_1)-V_1^\pi(s_1)\geq 0.02\sum_{h=1}^{H/2}\left(\max_{\boldsymbol{a}\in\mathcal{A}}r_h^b(\boldsymbol{a})-\sum_{\boldsymbol{a}\in\mathcal{A}}\pi_h(\boldsymbol{a}|s_h)r_h^b(\boldsymbol{a})\right) \tag{190}$$

where $r_h^b(\boldsymbol{a})=\boldsymbol{\mu}^\top\boldsymbol{a}+\zeta_h(\boldsymbol{a})$, which can be regarded as a reward function for misspecified linear bandit. To prove Theorem 1, the only remaining problem is to derive the lower bound for misspecified linear bandits. We directly apply the following two lower bounds for linear bandits.

Lemma 14. *(Lemma C.8 in Zhou et al. (2020a)) Fix a positive real $0<\delta\leq 1/3$, and positive integers T, d and assume that $T\geq d^2/(2\delta)$ and consider the linear bandit problem $\mathcal{L}_{\boldsymbol{\mu}}$ parametrized with a parameter vector $\boldsymbol{\mu}\in\{-\Delta,\Delta\}^d$ and action set $\mathcal{A}=\{-1,1\}^d$ so that the reward distribution for taking action $\boldsymbol{a}\in\mathcal{A}$ is a Bernoulli distribution $B(\delta+(\boldsymbol{\mu}^\star)^\top\boldsymbol{a})$. Then for any bandit algorithm $\mathcal{B}$, there exists a $\mu^*\in\{-\Delta,\Delta\}^d$ such that the expected pseudo-regret of $\mathcal{B}$ over T steps on bandit $\mathcal{L}_{\boldsymbol{\mu}^\star}$ is lower bounded by $\frac{d\sqrt{T\delta}}{8\sqrt{2}}$.*

Lemma 15. *(Proposition 6 in Zanette et al. (2020a)) There exists a feature map $\phi:\mathcal{A}\to\mathbb{R}^d$ that defines a misspecified linear bandits class $\mathcal{M}$ such that every bandit instance in that class has reward response:*

$$\mu_a=\phi_a^\top\theta+z_a$$

for any action a (Here $z_a\in[0,\zeta]$ is the deviation from linearity and $\mu_a\in[0,1]$) and such that the expected regret of any algorithm on at least a member of the class up to round T is $\Omega(\sqrt{d}\zeta T)$.

Lemma 14 is used to prove the lower bound for linear mixture MDPs in Zhou et al. (2020a), which states that the lower bound for linear bandits with approximation error $\zeta=0$, while Lemma 15 mainly consider the influence of ζ to the lower bound. Combining these two lemmas, the regret lower bound for misspecifid linear bandit is $\Omega(\max(d\sqrt{T\delta},\sqrt{d}\zeta T))=\Omega(d\sqrt{T\delta}+\sqrt{d}\zeta T)$. Since here our problem can reduce from $H/2$ misspecified linear bandit, we know that the regret lower bound is $\Omega(Hd\sqrt{T\delta}+H\sqrt{d}\zeta T)=\Omega(d\sqrt{HT}+H\sqrt{d}\zeta T)$ □

Now we obtain the regret lower bound for misspecified linear MDP. We can prove the corresponding lower bound for the LSVI setting Zanette et al. (2020a) since LSVI setting is strictly harder than linear MDP setting. The following lemma states this relation between two settings.

Lemma 16. *If an MDP$(\mathcal{S}, \mathcal{A}, p, r, H)$ is a misspecifed linear MDP with approximation error ζ, then this MDP satisfies the low inherent Bellman error assumption with $\mathcal{I} = 2\zeta$.*

Proof. If an MDP is an ζ-approximate linear MDP, then we have

$$\|p_h(\cdot|s,a) - \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h(\cdot)\rangle \|_{\mathrm{TV}} \leq \zeta \tag{191}$$

$$|r_h(s,a) - \langle \boldsymbol{\phi}(s,a), \boldsymbol{\nu}_h\rangle | \leq \zeta \tag{192}$$

For any $\theta_{h+1} \in \mathbb{R}^d$, we have $\mathcal{T}_h\left(Q_{h+1}(\theta_{h+1})\right)(s,a) = r_h(s,a) + \mathbb{E}_{s'\sim p_h(\cdot|s,a)} V_{h+1}(\theta_{h+1})(s')$. Since $V_{h+1}(\theta_{h+1})(s') \leq 1$, plugging the approximately linear form of $r_h(s,a)$ and $p_h(\cdot|s,a)$, we have

$$|\mathcal{T}_h\left(Q_{h+1}(\theta_{h+1})\right)(s,a) - \left\langle \boldsymbol{\phi}(s,a), \sum_{s'} \boldsymbol{\theta}_h(s') V_{h+1}(\theta_{h+1})(s') + \boldsymbol{\nu}_h \right\rangle | \leq 2\zeta \tag{193}$$

□

By lemma 16, we can directly apply the hard instance construction and the lower bound for misspecified linear MDP to LSVI setting.

Proposition 2. *There exist function feature maps $\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_H$ that define an MDP class $\mathcal{M}$ such that every MDP in that class satisfies low inherent Bellman error at most $\mathcal{I}$ and such that the expected reward on at least a member of the class (for $|\mathcal{A}| \geq 3, d, k, H \geq 10, T = \Omega(d^2 H), \mathcal{I} \leq \frac{1}{4H}$) is $\Omega(d\sqrt{HT} + \sqrt{d}H\mathcal{I}T)$.*

C.2. Lower Bound for Multi-task RL

In order to prove Theorem 5, we need to prove and then combine the following two lemmas.

Lemma 17. *Under the setting of Theorem 5, the expected regret of any algorithm $\mathcal{A}$ is lower bounded by $\Omega(Mk\sqrt{HT})$.*

Lemma 18. *Under the setting of Theorem 5, the expected regret of any algorithm $\mathcal{A}$ is lower bounded by $\Omega\left(d\sqrt{kMHT} + HMT\sqrt{d\mathcal{I}}\right)$.*

These two lemmas are proved by reduction from Proposition 2, which is a lower bound we proved for the single-task LSVI setting.

Proof. (Proof of Lemma 17) The lemma is proved by contradiction. Suppose there is an algorithm $\mathcal{A}$ that achieves $\sup_{M\in\mathcal{M}} \mathbb{E}[Reg(T)] \leq CMk\sqrt{HT}$ for a constant C. Then there must exists a task $i \in [M]$, such that the expected regret for this single task is at most $Ck\sqrt{HT}$. However, by Proposition 2, the expected regret for MDPs with dimension k in horizon h is at least $\Omega(k\sqrt{HT} + \sqrt{k}H\mathcal{I}T)$. This leads to a contradiction. □

Proof. (Proof of Lemma 18) The hard instance construction follows the same idea of the proof for our Lemma 6, as well as the hard instance to prove Lemma 19 in Yang et al. (2020). Without loss of generality, we assume that M can be exactly divided by k.

We divide M tasks into k groups. Each group shares the same parameter $\{\boldsymbol{\theta}_h^i\}_{h=1}^H$. To be more specific, we let $\boldsymbol{w}_h^1 = \boldsymbol{w}_h^2 = \cdots = \boldsymbol{w}_h^{M/k} = \boldsymbol{e}_h^1$, $\boldsymbol{w}_h^{M/k+1} = \boldsymbol{w}_h^{M/k+2} = \cdots = \boldsymbol{w}_h^{2M/k} = \boldsymbol{e}_h^2$, $\cdots$, $\boldsymbol{w}_h^{(k-1)M/k+1} = \boldsymbol{w}_h^{(k-1)M/k+2} = \cdots = \boldsymbol{w}_h^{M} = \boldsymbol{e}_h^k$. Under this construction, the parameters $\boldsymbol{\theta}_h^i$ for these tasks are exactly the same in each group, but relatively independent among different groups. That is to say, the expected regret lower bound is at least the summation of the regret lower bounds in all k groups.

Now we consider the regret lower bound for group $j \in [k]$. Since the parameters are shared in the same group, the regret of running an algorithm for M/k tasks with T episodes each is at least the regret of running an algorithm for single-task linear bandit with $M/k \cdot T$ episodes. By Proposition 2, the regret for single-task linear bandit with MT/k episodes is at least $\Omega(d\sqrt{MHT/k} + \sqrt{d\mathcal{I}}HMT/k)$. Summing over all k groups, we can prove that the regret lower bound is $\Omega(d\sqrt{kHMT} + \sqrt{d\mathcal{I}}HMT)$. □