A Proof of Theorem 5.1

We use $K(T)-1$ to denote the value of the counter k when Algorithm 1 finishes, and $t_{K(T)}=T+1$ for convenience. By these notation, the learning process from $t=1$ to $t=T$ can be divided into $K(T)$ episodes.

The following lemma, proved by Jaksch et al. (2010), states that EVI (Algorithm 2) always outputs a near-optimal policy and an optimistic model.

Lemma A.1 (Theorem 7 and Equation (12) in Jaksch et al. 2010)**.** Stopping the extended value iteration when

$$\max_{s\in\mathcal{S}}\{u^{(i+1)}(s)-u^{(i)}(s)\}-\min_{s\in\mathcal{S}}\{u^{(i+1)}(s)-u^{(i)}(s)\}<\epsilon,$$

the greedy policy $\widetilde{\pi}$ with respect to $u^{(i)}$ is ϵ-optimal, namely

$$\widetilde{\rho}:=\rho(\widetilde{M},\widetilde{\pi})\geq\max_{\pi,M\in\mathcal{M}}\rho(M,\pi)-\epsilon. \tag{A.1}$$

Here, $\widetilde{M}$ means the Markov Decision Process (MDP) determined by the parameterized transition probability, e.g. $\mathbb{P}_k(\cdot|s,a)=\langle\boldsymbol{\phi}(\cdot|s,a),\boldsymbol{\theta}_k(s,a)\rangle$. For each $M\in\mathcal{M}$, M is an MDP with parameter from the confidence set. $\mathcal{M}$ is assumed to contain the true transition model.

Moreover, we have $\forall s\in\mathcal{S}$,

$$|u^{(i+1)}(s)-u^{(i)}(s)-\widetilde{\rho}|\leq\epsilon. \tag{A.2}$$

The next lemma describes that indeed, the confidence sets we constructed contain the true parameter with high probability.

Lemma A.2. With probability at least $1-\delta$, for all $0\leq k\leq K(T)-1$, we have $\boldsymbol{\theta}^*\in\widehat{\mathcal{C}}_{t_k}$.

Proof. See Section D.1. □

The number of episodes in our algorithm turns out can be bounded as follows:

Lemma A.3. We have $K(T)\leq d\log[(2\lambda+2TD^2)/\lambda]$.

Proof. See Section D.2. □

The rest lemmas either is standard concentration inequalities or is from the works regarding linear bandit problems.

Lemma A.4 (Azuma–Hoeffding inequality)**.** Let $\{X_k\}_{k=0}^{\infty}$ be a discrete-parameter real-valued martingale sequence such that for every $k\in\mathbb{N}$, the condition $|X_k-X_{k-1}|\leq\mu$ holds for some non-negative constant μ. Then with probability at least $1-\delta$, we have

$$X_n-X_0\leq\mu\sqrt{2n\log 1/\delta}.$$

Lemma A.5 (Lemma 11 in Abbasi-Yadkori et al. 2011)**.** For any $\{\mathbf{x}_t\}_{t=1}^T\subset\mathbb{R}^d$ satisfying that $\|\mathbf{x}_t\|_2\leq L$, let $\mathbf{A}_0=\lambda\mathbf{I}$ and $\mathbf{A}_t=\mathbf{A}_0+\sum_{i=1}^{t-1}\mathbf{x}_i\mathbf{x}_i^\top$, then we have

$$\sum_{t=1}^{T}\min\{1,\|\mathbf{x}_t\|_{\mathbf{A}_{t-1}^{-1}}\}^2\leq 2d\log\frac{d\lambda+TL^2}{d\lambda}.$$

Lemma A.6 (Lemma 12 in Abbasi-Yadkori et al. 2011)**.** Suppose $\mathbf{A},\mathbf{B}\in\mathbb{R}^{d\times d}$ are two positive definite matrices satisfying that $\mathbf{A}\succeq\mathbf{B}$, then for any $\mathbf{x}\in\mathbb{R}^d$, $\|\mathbf{x}\|_{\mathbf{A}}\leq\|\mathbf{x}\|_{\mathbf{B}}\cdot\sqrt{\det(\mathbf{A})/\det(\mathbf{B})}$.

Proof of Theorem 5.1. We first split the regret into each episode. Denote the regret in episode k as Δ_k, and we have

$$\begin{aligned}
\Delta_k &:= \sum_{t=t_k}^{t_{k+1}-1} [\rho^* - r(s_t, a_t)] \\
&\le (t_{k+1} - t_k)\epsilon + \sum_{t=t_k}^{t_{k+1}-1} [\rho_k - r(s_t, a_t)] \\
&\le 2(t_{k+1} - t_k)\epsilon + \sum_{t=t_k}^{t_{k+1}-1} \big[\sum_{s' \in \mathcal{S}} \mathbb{P}_k(s'|s_t, a_t) u_k(s') - u_k(s_t) \big] \\
&= 2(t_{k+1} - t_k)\epsilon + \sum_{t=t_k}^{t_{k+1}-1} \big[\sum_{s' \in \mathcal{S}} \mathbb{P}_k(s'|s_t, a_t) w_k(s') - w_k(s_t) \big] \\
&= 2(t_{k+1} - t_k)\epsilon + \sum_{t=t_k}^{t_{k+1}-1} \big[[\mathbb{P}_k w_k](s_t, a_t) - w_k(s_t) \big].
\end{aligned}$$

The first inequality is due to the $\epsilon-$optimality of the EVI algorithm (Lemma A.1). The second inequality is due to (A.2) and substitute the iteration rule $u^{(i+1)}(s) \leftarrow \max_{a \in \mathcal{A}} \Big\{ r(s,a) + \max_{\boldsymbol{\theta} \in \mathcal{C} \cap \mathcal{B}} \Big\{ \big\langle \boldsymbol{\theta}, \boldsymbol{\phi}_{u^{(i)}}(s,a) \big\rangle \Big\} \Big\}$. Here, notice that we denote $\mathbb{P}_k(s'|s_t, a_t) = \langle \boldsymbol{\theta}_k(s_t, a_t), \boldsymbol{\phi}(s'|s_t, a_t) \rangle$ and $\boldsymbol{\theta}_k(s_t, a_t) = \operatorname{argmax}_{\boldsymbol{\theta} \in \mathcal{C} \cap \mathcal{B}} \Big\{ \big\langle \boldsymbol{\theta}, \boldsymbol{\phi}_{u^{(i)}}(s,a) \big\rangle \Big\}$. By the definition of π_k, a_t achieves the outer maximum in the iteration rule of $u^{(i+1)}$. The second last equality is due to the fact that adding a bias to u_k won't change the difference, as what has been done in Algorithm 1. So we subtract $(\max_s u_k(s) + \min_s u_k(s))/2$ from $u_k(s)$. The last equality is a shorthand. Notice that since the span of $u_k(s)$ is D, we have $|w_k(s)| \le D/2$.

Summing over all episodes, we further have

$$\begin{aligned}
\sum_{k=0}^{K(T)-1} \Delta_k = 2T\epsilon &+ \underbrace{\sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \big[[\mathbb{P}_k w_k](s_t, a_t) - [\mathbb{P} w_k](s_t, a_t) \big]}_{I_1} \\
&+ \underbrace{\sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \big[[\mathbb{P} w_k](s_t, a_t) - w_k(s_t) \big]}_{I_2}.
\end{aligned}$$

The first term can be controlled following the idea of bounding the regret of linear bandit. We have that with probability $1 - \delta$,

$$\begin{aligned}
I_1 &= \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \big\langle \boldsymbol{\theta}_k - \boldsymbol{\theta}^*, \boldsymbol{\phi}_{w_k}(s_t, a_t) \big\rangle \\
&\le \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \big(\|\boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_t} + \|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_t} \big) \big\| \boldsymbol{\phi}_{w_k}(s_t, a_t) \big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}} \\
&\le \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} 2\big(\|\boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_{t_k}} + \|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_{t_k}} \big) \big\| \boldsymbol{\phi}_{w_k}(s_t, a_t) \big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}} \\
&\le \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} 4\widehat{\beta}_T \big\| \boldsymbol{\phi}_{w_k}(s_t, a_t) \big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}.
\end{aligned}$$

The first inequality is due to first applying Cauchy-Schwartz inequality and then the triangle inequality. The second inequality is due to Lemma A.6 and the fact that for $t_k \le t < t_{k+1}$ $\det(\boldsymbol{\Sigma}_t) \le \det(\boldsymbol{\Sigma}_{t_{k+1}}) \le 2\det(\boldsymbol{\Sigma}_{t_k})$. The third inequality is due to Lemma A.2 and the fact that $\{\widehat{\beta}_t\}_t$ is increasing.

Meanwhile, for each term in I_1, we also have that due to the fact $|w_k(s)| \leq D/2$,

$$[\mathbb{P}_k w_k](s_t, a_t) - [\mathbb{P} w_k](s_t, a_t) \leq D.$$

Therefore, we have

$$\begin{aligned}
I_1 &\leq \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \min\Big\{D, 4\widehat{\beta}_T \big\|\boldsymbol{\phi}_{w_k}(s_t, a_t)\big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}\Big\} \\
&\leq 4\widehat{\beta}_T \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \min\Big\{1, \big\|\boldsymbol{\phi}_{w_k}(s_t, a_t)\big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}\Big\} \\
&\leq 4\widehat{\beta}_T \sqrt{T \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \min\Big\{1, \big\|\boldsymbol{\phi}_{w_k}(s_t, a_t)\big\|^2_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}\Big\}} \\
&\leq 4\widehat{\beta}_T \sqrt{T \cdot 2d \log\bigg(\frac{d\lambda + TD^2}{d\lambda}\bigg)} \\
&\leq 6\widehat{\beta}_T \sqrt{dT \log\bigg(\frac{d\lambda + TD^2}{d\lambda}\bigg)}.
\end{aligned}$$

The second inequality is due to the fact $D \leq 4\widehat{\beta}_T$. The third is due to Cauchy-Schwartz inequality. The fourth is due to Lemma A.5.

The second term, can be controlled by the concentration of martingale. With probability $1 - \delta$,

$$\begin{aligned}
I_2 &= \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \big[[\mathbb{P} w_k](s_t, a_t) - w_k(s_t)\big] \\
&= \sum_{k=0}^{K(T)-1} \bigg[\sum_{t=t_k}^{t_{k+1}-1} \big([\mathbb{P} w_k](s_t, a_t) - w_k(s_{t+1})\big) - w_k(s_{t_k}) + w_k(s_{t_{k+1}})\bigg] \\
&\leq \sum_{k=0}^{K(T)-1} \bigg[\sum_{t=t_k}^{t_{k+1}-1} \big([\mathbb{P} w_k](s_t, a_t) - w_k(s_{t+1})\big)\bigg] + D \cdot K(T) \\
&\leq D\sqrt{2T \log(1/\delta)} + D \cdot K(T),
\end{aligned}$$

where the first inequality holds because $|w_k(s)| \leq D/2$; the second inequality is due to Lemma A.4.

Therefore, the total regret is bounded by

$$\text{Regret}(T) = \sum_{k=0}^{K(T)-1} \Delta_k \leq 2T\epsilon + 6\widehat{\beta}_T \sqrt{dT \log\bigg(\frac{\lambda + TD^2}{\lambda}\bigg)} + D\sqrt{2T \log(1/\delta)} + D \cdot K(T).$$

If we set

$$\widehat{\beta}_t = D\sqrt{d \log\bigg(\frac{\lambda + tD^2}{\delta\lambda}\bigg)} + \sqrt{\lambda} B,$$

and

$$\epsilon = \frac{1}{\sqrt{T}},$$

then by taking union bound we have with probability $1 - 2\delta$,

$$\begin{aligned}
\text{Regret}(T) &\leq 2\sqrt{T} + Dd\sqrt{T} \cdot \widetilde{O}(1) + B\sqrt{\lambda dT} \cdot \widetilde{O}(1) + D\sqrt{2T \log(1/\delta)} + Dd \log\bigg(\frac{2\lambda + 2dTD^2}{\lambda}\bigg) \\
&\leq \widetilde{O}(Dd\sqrt{T}),
\end{aligned}$$

where $\widetilde{O}(1)$ hides the log factor, the last inequality holds since we set $\lambda = 1/B^2$.

□

B Proof of Theorem 5.3

Most part of the proof resembles that of Theorem 5.1. The additional part is essentially about the new concentration results from variance-aware linear bandit problem. As previously defined, we use $K(T)-1$ to denote the value of the counter k when Algorithm 1 finishes, and $t_{K(T)} = T+1$ for convenience. By these notations, the learning process from $t=1$ to $t=T$ can be divided into $K(T)$ episodes.

The first lemma provides a better confidence set given the information of the noise's variance.

Lemma B.1 (Bernstein inequality for vector-valued martingales (Zhou et al., 2021a))**.** Let $\{\mathcal{G}_t\}_{t=1}^{\infty}$ be a filtration, $\{\mathbf{x}_t, \eta_t\}_{t\geq 1}$ a stochastic process so that $\mathbf{x}_t \in \mathbb{R}^d$ is $\mathcal{G}_t$-measurable and $\eta_t \in \mathbb{R}$ is $\mathcal{G}_{t+1}$-measurable. Fix $R, L, \sigma, \lambda > 0$, $\boldsymbol{\mu}^* \in \mathbb{R}^d$. For $t \geq 1$ let $y_t = \langle \boldsymbol{\mu}^*, \mathbf{x}_t \rangle + \eta_t$ and suppose that $\eta_t, \mathbf{x}_t$ also satisfy

$$|\eta_t| \leq R,\ \mathbb{E}[\eta_t|\mathcal{G}_t] = 0,\ \mathbb{E}[\eta_t^2|\mathcal{G}_t] \leq \sigma^2,\ \|\mathbf{x}_t\|_2 \leq L.$$

Then, for any $0 < \delta < 1$, with probability at least $1-\delta$ we have

$$\forall t > 0,\ \left\| \sum_{i=1}^{t} \mathbf{x}_i \eta_i \right\|_{\mathbf{Z}_t^{-1}} \leq \beta_t,\ \|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t} \leq \beta_t + \sqrt{\lambda}\|\boldsymbol{\mu}^*\|_2, \tag{B.1}$$

where for $t \geq 1$, $\boldsymbol{\mu}_t = \mathbf{Z}_t^{-1}\mathbf{b}_t$, $\mathbf{Z}_t = \lambda\mathbf{I} + \sum_{i=1}^{t} \mathbf{x}_i\mathbf{x}_i^\top$, $\mathbf{b}_t = \sum_{i=1}^{t} y_i\mathbf{x}_i$ and

$$\beta_t = 8\sigma\sqrt{d\log(1+tL^2/(d\lambda))\log(4t^2/\delta)} + 4R\log(4t^2/\delta)\,.$$

The number of episodes is bounded almost in the same way as in Lemma A.3:

Lemma B.2. Let $K(T)$ be as defined above. Then, $K(T) \leq 2d\log(1+Td/\lambda)$.

Proof. See Section D.3. □

The variance term is defined as

$$[\mathbb{V}w_k](s_t, a_t) := \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k^2(s')] - \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')]^2.$$

The following lemma states that with high probability the estimated variance is close the the true variance.

Lemma B.3. With probability $1-3\delta$, we have for all $1 \leq t \leq T$,

$$\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_t \cap \mathcal{B}, \left|[\bar{\mathbb{V}}_t w_k](s_t, a_t) - [\mathbb{V}w_k](s_t, a_t)\right| \leq E_t.$$

We denote the event above by $\mathcal{E}_0$, and $\mathbb{P}(\mathcal{E}_0) \geq 1-3\delta$.

Proof. See Section D.4. □

Now, we define other events:

$$\mathcal{E}_1 := \left\{ \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \left[\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')^2] - w_k^2(s_{t+1})\right] \leq (D^2/4)\sqrt{2T\log(1/\delta)} \right\}$$

$$\mathcal{E}_2 := \left\{ \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \left[\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')] - w_k(s_{t+1})\right] \leq (D/2)\sqrt{2T\log(1/\delta)} \right\}$$

By the Azuma-Hoeffding inequality(Lemma A.4), we have $\mathbb{P}(\mathcal{E}_1) \geq 1-\delta$ and $\mathbb{P}(\mathcal{E}_2) \geq 1-\delta$.

The next lemma characterizes the total variance.

Lemma B.4. Under the events $\mathcal{E}_0$ and $\mathcal{E}_1$, we have

$$\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}[\mathbb{V}w_k](s_t,a_t) \le (D^2/4)\sqrt{2T\log(1/\delta)} + (K(T)+1)(D^2/4) + 2DT + D^2\widehat{\beta}_T\sqrt{T2d\log(1+T/\lambda)}.$$

Proof. See Section D.5. □

The following lemma serves as a wrapper of calculating the total estimation error.

Lemma B.5. Under the event $\mathcal{E}_0$, we have

$$\sum_{t=1}^{T} E_t \le \widetilde{\beta}_T\sqrt{2Td\log(1+TD^2/4d\lambda)} + D^2\check{\beta}_T\sqrt{2Td\log(1+T/\lambda)}.$$

Proof. See Section D.6. □

Now we are ready to show the regret upper bound.

Proof. We first follow the same procedure as Jaksch et al. (2010) did to decompose the regret and tackle each term respectively.

We have

$$\begin{aligned}
\text{Regret}(T) &:= \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}[\rho^* - r(s_t,a_t)] \\
&\le T\epsilon + \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}[\rho_k - r(s_t,a_t)] \\
&\le 2T\epsilon + \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big[\mathbb{E}_{s'\sim\mathbb{P}_k(\cdot|s_t,a_t)}[u_k(s')] - u_k(s_t)\big] \\
&= 2T\epsilon + \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big[\mathbb{E}_{s'\sim\mathbb{P}_k(\cdot|s_t,a_t)}[w_k(s')] - w_k(s_t)\big] \\
&= 2T\epsilon + \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big[[\mathbb{P}_k w_k](s_t,a_t) - w_k(s_t)\big].
\end{aligned}$$

The first inequality is due to the ϵ-optimality of the EVI algorithm. The second inequality is due to (12) in Jaksch et al. (2010). The third inequality is due to the fact that add a bias to u_t won't change the difference, as done in Algorithm 1. So we subtract $(\max_s u_t(s) + \min_s u_t(s))/2$ from $u_t(s)$. The last equality is a shorthand. It can be further decomposed into:

$$\begin{aligned}
\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big[[\mathbb{P}_k w_k](s_t,a_t) - w_k(s_t)\big] = &\underbrace{\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big[[\mathbb{P}_k w_k](s_t,a_t) - [\mathbb{P}w_k](s_t,a_t)\big]}_{I_1} \\
&+ \underbrace{\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big[[\mathbb{P}w_k](s_t,a_t) - w_k(s_t)\big]}_{I_2}.
\end{aligned}$$

We deal with the second term I_2 first:

The second term, can be controlled by the concentration of martingale. In fact, $\mathcal{E}_2$ defined above exactly characterizes the concentration. Under event $\mathcal{E}_2$, we have

$$\begin{aligned}
I_2 &= \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big[[\mathbb{P}w_k](s_t,a_t) - w_k(s_t)\big] \\
&= \sum_{k=0}^{K(T)-1}\left[\sum_{t=t_k}^{t_{k+1}-1}\big[[\mathbb{P}w_k](s_t,a_t) - w_k(s_{t+1})\big] - w_k(s_{t_k}) + w_k(s_{t_{k+1}})\right] \\
&\le \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big[[\mathbb{P}w_k](s_t,a_t) - w_k(s_{t+1})\big] + K(T)\cdot D \\
&\le D\sqrt{2T\log(1/\delta)} + K(T)\cdot D \\
&= \widetilde{O}(D\sqrt{T}) + \widetilde{O}(Dd),
\end{aligned}$$

where the first inequality holds since $|w_k(\cdot)| \le D/2$, the second one holds due to the definition of $\mathcal{E}_2$. For term I_1,

$$\begin{aligned}
I_1 &= \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big[[\mathbb{P}_k w_k](s_t,a_t) - [\mathbb{P}w_k](s_t,a_t)\big] \\
&= \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big\langle \boldsymbol{\theta}_k - \boldsymbol{\theta}^*, \boldsymbol{\phi}_{w_k}(s_t,a_t)\big\rangle \\
&\le \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big(\|\boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_t} + \|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_t}\big)\big\|\boldsymbol{\phi}_{w_k}(s_t,a_t)\big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}} \\
&\le 2\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big(\|\boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_{t_k}} + \|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_{t_k}}\big)\big\|\boldsymbol{\phi}_{w_k}(s_t,a_t)\big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}} \\
&\le 4\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\widehat{\beta}_{t_k}\big\|\boldsymbol{\phi}_{w_k}(s_t,a_t)\big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}} \\
&\le 4\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\widehat{\beta}_t\bar{\sigma}_t\big\|\boldsymbol{\phi}_{w_k}(s_t,a_t)/\bar{\sigma}_t\big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}.
\end{aligned}$$

The first inequality is due to first applying Cauchy-Schwartz inequality and then the triangle inequality. The second is due to $\det(\widehat{\boldsymbol{\Sigma}}_t) \le 2\det(\widehat{\boldsymbol{\Sigma}}_{t_k})$ and Lemma A.6. The third is due to event $\mathcal{E}_0$. The last is due to the fact that $\{\widehat{\beta}_t\}_{t\ge 0}$ is increasing.

Meanwhile, for each term in I_1, we also have that due to $|w_k(s)| \le D/2$,

$$[\mathbb{P}_k w_k](s_t,a_t) - [\mathbb{P}w_k](s_t,a_t) \le D.$$

Therefore, we have

$$\begin{aligned}
I_1 &\le \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\min\left\{D, 4\widehat{\beta}_t\bar{\sigma}_t\big\|\boldsymbol{\phi}_{w_k}(s_t,a_t)/\bar{\sigma}_t\big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}\right\} \\
&\le \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}4\widehat{\beta}_t\bar{\sigma}_t\min\left\{1, \big\|\boldsymbol{\phi}_{w_k}(s_t,a_t)/\bar{\sigma}_t\big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}\right\} \\
&\le 4\widehat{\beta}_T\sqrt{\underbrace{\sum_{t=1}^{T}(\bar{\sigma}_t)^2}_{J_1}}\sqrt{\underbrace{\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\left\{1, \big\|\boldsymbol{\phi}_{w_k}(s_t,a_t)/\bar{\sigma}_t\big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}\right\}}_{J_2}}.
\end{aligned}$$

The second inequality is due to the fact $D \leq 4\widehat{\beta}_t\bar{\sigma}_t$. The third is due to Cauchy-Schwartz inequality.

Note that by Lemma A.5, it is clear that

$$J_2 \leq 2d\log(1+T/\lambda).$$

For term J_1,

$$\begin{aligned}
J_1 &= \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1} \max\{D^2/d, [\bar{\mathbb{V}}_t w_k](s_t,a_t) + E_t\} \\
&\leq TD^2/d + \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1} [\bar{\mathbb{V}}_t w_k](s_t,a_t) + \sum_{t=1}^{T} E_t \\
&\leq TD^2/d + \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1} [\mathbb{V} w_k](s_t,a_t) + 2\sum_{t=1}^{T} E_t \\
&\leq TD^2/d + (D^2/4)\sqrt{2T\log(1/\delta)} + (K(T)+1)(D^2/4) + 2DT + D^2\widehat{\beta}_T\sqrt{T2d\log(1+T/\lambda)} \\
&\quad + \widetilde{\beta}_T\sqrt{2Td\log(1+TD^2/4d\lambda)} + D^2\check{\beta}_T\sqrt{2Td\log(1+T/\lambda)}.
\end{aligned}$$

The second inequality uses Lemma B.1. The third uses Lemma B.4 and Lemma B.5.

Now, based on Lemma B.2 we have $K(T) = \widetilde{O}(d)$. By definition, we have

$$\begin{aligned}
\widehat{\beta}_T &= \widetilde{O}(\sqrt{d}) \\
\check{\beta}_T &= \widetilde{O}(d) \\
\widetilde{\beta}_T &= \widetilde{O}(D^2\sqrt{d}),
\end{aligned}$$

if we set $\lambda = B^{-2}$.

This means we can express I_1 in Big-O notation term by term as:

$$\begin{aligned}
J_1 &= \widetilde{O}(TD^2/d) + \widetilde{O}(D^2\sqrt{T}) + \widetilde{O}(D^2 d) + \widetilde{O}(DT) + \widetilde{O}(D^2 d\sqrt{T}) + \widetilde{O}(D^2 d\sqrt{T}) + \widetilde{O}(D^2 d^{3/2}\sqrt{T}) \\
&= \widetilde{O}(TD^2/d) + \widetilde{O}(DT) + \widetilde{O}(D^2 d^{3/2}\sqrt{T}).
\end{aligned}$$

We have

$$\begin{aligned}
I_1 &= \widetilde{O}(\sqrt{d}) \cdot \sqrt{\widetilde{O}(TD^2/d) + \widetilde{O}(DT) + \widetilde{O}(D^2 d^{3/2}\sqrt{T})} \cdot \sqrt{\widetilde{O}(d)} \\
&= \widetilde{O}(D\sqrt{dT}) + \widetilde{O}(d\sqrt{DT}) + \widetilde{O}(Dd^{7/4}T^{1/4}).
\end{aligned}$$

Finally, by setting $\epsilon = 1/\sqrt{T}$, the regret is upper bounded as

$$\begin{aligned}
\text{Regret}(T) &= \mathcal{O}(\sqrt{T}) + \widetilde{O}(D\sqrt{dT}) + \widetilde{O}(d\sqrt{DT}) + \widetilde{O}(Dd^{7/4}T^{1/4}) + \widetilde{O}(D\sqrt{T}) + \widetilde{O}(Dd) \\
&= \widetilde{O}(D\sqrt{dT}) + \widetilde{O}(d\sqrt{DT}) + \widetilde{O}(Dd^{7/4}T^{1/4}).
\end{aligned}$$

As long as $d \geq D$ and $T \geq D^2 d^3$, we have

$$\text{Regret}(T) = \widetilde{O}(d\sqrt{DT}).$$

□

C Proof of Theorem 5.5

C.1 Construction of Hard-to-learn MDPs

Here we describe the construction of the hard-to-learn MDPs $M(\mathcal{S}, \mathcal{A}, r, \mathbb{P}_{\boldsymbol{\theta}})$ for our lower bound proof (illustrated in Figure 2). The state space $\mathcal{S}$ consists of two states x_0, x_1. The action space $\mathcal{A}$ consists of 2^{d-1} vectors $\mathbf{a} \in \mathbb{R}^{d-1}$

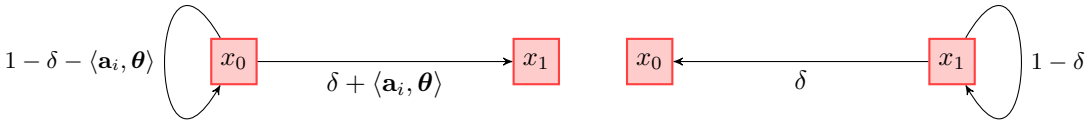

Figure 2: Illustration of the hard-to-learn linear mixture MDP considered in Theorem 5.5. The left figure demonstrates the state transition probability starting from x_0 with some specific action $\mathbf{a}_i$. The right figure demonstrates the state transition probability starting from x_1 with any action.

whose coordinates are 1 or -1. The reward function r satisfies that $r(x_0, \mathbf{a}) = 0$ and $r(x_1, \mathbf{a}) = 1$ for any $\mathbf{a} \in \mathcal{A}$. The probability transition function $\mathbb{P}_{\boldsymbol{\theta}}$ is parameterized by a $(d-1)$-dimensional vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, which is defined as

$$\mathbb{P}_{\boldsymbol{\theta}}(x_0|x_0, \mathbf{a}) = 1 - \delta - \langle \mathbf{a}, \boldsymbol{\theta}\rangle, \qquad \mathbb{P}_{\boldsymbol{\theta}}(x_1|x_0, \mathbf{a}) = \delta + \langle \mathbf{a}, \boldsymbol{\theta}\rangle,$$
$$\mathbb{P}_{\boldsymbol{\theta}}(x_0|x_1, \mathbf{a}) = \delta, \qquad \mathbb{P}_{\boldsymbol{\theta}}(x_1|x_1, \mathbf{a}) = 1 - \delta,$$
$$\boldsymbol{\Theta} = \{-\Delta/(d-1), \Delta/(d-1)\}^{d-1},$$

where δ and Δ are positive parameters that need to be determined in later proof. We set $\delta = 1/D$, and Δ as $\Delta = (1/45\sqrt{2\log 2/5})d/\sqrt{DT}$. It can be verified that M is indeed a linear kernel MDP with the feature mapping $\boldsymbol{\phi}(s'|s, a)$ defined as follows:

$$\boldsymbol{\phi}(x_0|x_0, \mathbf{a}) = \begin{pmatrix} -\alpha\mathbf{a} \\ \beta(1-\delta) \end{pmatrix}, \boldsymbol{\phi}(x_1|x_0, \mathbf{a}) = \begin{pmatrix} \alpha\mathbf{a} \\ \beta\delta \end{pmatrix}, \boldsymbol{\phi}(x_0|x_1, \mathbf{a}) = \begin{pmatrix} \mathbf{0} \\ \beta\delta \end{pmatrix}, \boldsymbol{\phi}(x_1|x_1, \mathbf{a}) = \begin{pmatrix} \mathbf{0} \\ \beta(1-\delta) \end{pmatrix},$$

where $\alpha = \sqrt{\Delta/[(d-1)(1+\Delta)]}$, $\beta = \sqrt{1/(1+\Delta)}$, and the vector $\widetilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}^\top/\alpha, 1/\beta)^\top \in \mathbb{R}^d$. We can verify that $\boldsymbol{\phi}$ and $\widetilde{\boldsymbol{\theta}}$ satisfy the requirements of B-bounded linear mixture MDP. In detail, we have

$$\|\widetilde{\boldsymbol{\theta}}\|_2^2 = \frac{\|\boldsymbol{\theta}\|_2^2}{\alpha^2} + \frac{1}{\beta^2} = (1+\Delta)^2 \le B^2,$$

as long as $\Delta \le \sqrt{B} - 1$. In addition, for any function $F : \mathcal{S} \to [0, 1]$, we have

$$\|\boldsymbol{\phi}_F(x_0, \mathbf{a})\|_2^2 = \alpha^2\|\mathbf{a}\|_2^2(F(x_1) - F(x_0))^2 + (\beta(1-\delta)F(x_0) + \beta\delta F(x_1))^2 \le (d-1)\alpha^2 + \beta^2 = 1.$$

Therefore, our defined MDP is indeed a B-bounded linear mixture MDP.

Remark C.1. Similar to Zhou et al. (2021a), our lower bound can also imply a lower bound for a related MDP class called *linear MDPs* (Yang and Wang, 2019a; Jin et al., 2019), which assumes that $\mathbb{P}(s'|s, a) = \langle \boldsymbol{\psi}(s, a), \boldsymbol{\mu}(s')\rangle$ and $r(s, a) = \langle \boldsymbol{\psi}(s, a), \boldsymbol{\xi}\rangle$. We construct $\boldsymbol{\psi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\xi}$ as follows:

$$\boldsymbol{\psi}(s, a) = \begin{cases} (\alpha\mathbf{a}^\top, \beta, 0)^\top & s = x_0 \\ (0, 0, 1) & s = x_1 \end{cases}, \boldsymbol{\mu}(s') = \begin{cases} (-\boldsymbol{\theta}^\top/\alpha, (1-\delta)/\beta), \delta)^\top & s' = x_0 \\ (\boldsymbol{\theta}^\top/\alpha, \delta/\beta, 1-\delta)^\top & s' = x_1 \end{cases}, \boldsymbol{\xi} = (\mathbf{0}^\top, 1)^\top.$$

It can be verified that such a feature mapping $\boldsymbol{\phi}, \boldsymbol{\mu}$ and parameters $\boldsymbol{\xi}$ satisfy the requirements of a linear MDP with $(d+1)$-dimension feature mapping. Meanwhile, the MDP $\langle \boldsymbol{\psi}(s, a), \boldsymbol{\mu}(s')\rangle$ has exactly the same form as the linear mixture MDPs proposed in Theorem 5.5. Therefore, the lower bound in Theorem 5.5 can also be applied to infinite-horizon average-reward linear MDPs, which are studied by Wei et al. (2020a). This also suggests that there still exists a gap between the best upper bound (Wei et al., 2020a) and lower bound in the linear MDP setting.

C.2 Proof of the Lower Bound in Theorem 5.5

Given the example we constructed above (shown in Figure 2), it is easy to see that the optimal policy is to choose the action $\mathbf{a}$ satisfying $\langle \mathbf{a}, \boldsymbol{\theta}\rangle = \Delta$, namely each coordinate of $\mathbf{a}$ has the same sign as $\boldsymbol{\theta}$'s.

Given the optimal policy, it is clear that the stationary distribution is

$$\boldsymbol{\mu} = \begin{bmatrix} \dfrac{\delta}{2\delta+\Delta} & \dfrac{\delta+\Delta}{2\delta+\Delta} \end{bmatrix},$$

and the optimal average reward is $\rho^* = (\delta+\Delta)/(2\delta+\Delta)$.

In the construction, we leave the two parameters δ and Δ unspecified. Now we set $\delta = 1/D$. From state x_1 to x_0, it is clear that any policy has only one action and the expected travel time is $1/\delta = D$. From state x_0 to x_1, there always exists an policy that chooses the action $\mathbf{a}$ that has the same sign coordinate-wise, and in that case the transition probability from x_0 to x_1 is $\delta+\Delta$, which indicates the expected travel time is smaller then D. From the argument above, we know the MDP has a diameter of D.

The choice of Δ is $\Delta = (1/45\sqrt{2\log 2/5})d/\sqrt{DT}$; the motivation will be revealed later in the proof.

In the following, we use $\text{Regret}_{\boldsymbol{\theta}}(T)$ to denote the regret $\text{Regret}(M_{\boldsymbol{\theta}}, A, s, T)$, where A is a deterministic algorithm. As argued in Auer et al. (2002), it is sufficient to only consider deterministic policies. Let $\mathcal{P}_{\boldsymbol{\theta}}(\cdot)$ denote the distribution over $\mathcal{S}^T$, where $s_1 = x_0$, $s_{t+1} \sim \mathbb{P}_{\boldsymbol{\theta}}(\cdot|s_t, a_t)$, a_t is decided by A. Let $\mathbb{E}_{\boldsymbol{\theta}}$ denote the expectation w.r.t. distribution $\mathcal{P}_{\boldsymbol{\theta}}$. Denote $N_1, N_0, N_0^{\mathbf{a}}$ as the random variables of the times state x_1 is visited, the times state x_0 is visited and the times state x_0 is visited and $\mathbf{a}$ is chosen. We further define $N_0^{\mathcal{V}}$ for some subset $\mathcal{V} \subset \mathcal{A}$ as the random variable of the times state x_0 is visited, and the action $\mathbf{a}$ belongs to $\mathcal{V}$.

Lemma C.2. Suppose $2\Delta < \delta$ and $(1-\delta)/\delta < T/5$, then for $\mathbb{E}_{\boldsymbol{\theta}} N_1$ and $\mathbb{E}_{\boldsymbol{\theta}} N_0$, we have

$$\mathbb{E}_{\boldsymbol{\theta}} N_1 \le \frac{T}{2} + \frac{1}{2\delta}\sum_{\mathbf{a}} \langle \mathbf{a}, \boldsymbol{\theta}\rangle \mathbb{E}_{\boldsymbol{\theta}} N_0^{\mathbf{a}}, \ \ \mathbb{E}_{\boldsymbol{\theta}} N_0 \le 4T/5.$$

Proof. See Section D.7. □

Lemma C.3 (Pinsker's inequality, in Jaksch et al. (2010))**.** Denote $\mathbf{s} = \{s_1, \ldots, s_T\} \in \mathcal{S}^T$ as the observed states from step 1 to T. Then for any two distributions $\mathcal{P}_1$ and $\mathcal{P}_2$ over $\mathcal{S}^T$ and any bounded function $f : \mathcal{S}^T \to [0, B]$, we have

$$\mathbb{E}_1 f(\mathbf{s}) - \mathbb{E}_2 f(\mathbf{s}) \le \sqrt{\log 2/2} B \sqrt{\text{KL}(\mathcal{P}_2 \| \mathcal{P}_1)},$$

where $\mathbb{E}_1$ and $\mathbb{E}_2$ are expectations with respect to $\mathcal{P}_1$ and $\mathcal{P}_2$.

Lemma C.4. Suppose that $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ only differs from j-th coordinate, $2\Delta < \delta \le 1/3$. Then we have the following bound for the KL divergence between $\mathcal{P}_{\boldsymbol{\theta}}$ and $\mathcal{P}_{\boldsymbol{\theta}'}$:

$$\text{KL}(\mathcal{P}_{\boldsymbol{\theta}'} \| \mathcal{P}_{\boldsymbol{\theta}}) \le \frac{16\Delta^2}{(d-1)^2\delta} \mathbb{E}_{\boldsymbol{\theta}} N_0.$$

Proof. See Section D.8. □

Proof of Theorem 5.5. We have

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}}[\text{Regret}_{\boldsymbol{\theta}}(T)] &:= T\rho^* - \mathbb{E}_{\boldsymbol{\theta}}\bigg[\sum_{t=1}^{T} r(s_t, a_t)\bigg] \\ &= T\rho^* - \mathbb{E}_{\boldsymbol{\theta}}[N_1]. \end{aligned}$$

Averaging over all possible choice of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we have

$$\frac{1}{|\boldsymbol{\Theta}|}\sum_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\theta}}[\text{Regret}_{\boldsymbol{\theta}}(T)] = T\rho^* - \frac{1}{|\boldsymbol{\Theta}|}\sum_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\theta}}[N_1].$$

Following Lemma C.2, we first have

$$\begin{aligned}\frac{1}{|\boldsymbol{\Theta}|}\sum_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\theta}}[N_1] &\le \frac{T}{2}+\frac{1}{2\delta|\boldsymbol{\Theta}|}\sum_{\boldsymbol{\theta}}\sum_{\mathbf{a}}\langle \mathbf{a},\boldsymbol{\theta}\rangle \mathbb{E}_{\boldsymbol{\theta}}N_0^{\mathbf{a}}\\ &= \frac{T}{2}+\frac{1}{2\delta|\boldsymbol{\Theta}|}\sum_{\boldsymbol{\theta}}\sum_{\mathbf{a}}\frac{\Delta}{d-1}\sum_{j=1}^{d-1}\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)=\mathrm{sign}(\boldsymbol{\theta}_j)\}\mathbb{E}_{\boldsymbol{\theta}}N_0^{\mathbf{a}}\\ &= \frac{T}{2}+\frac{1}{2\delta|\boldsymbol{\Theta}|}\frac{\Delta}{d-1}\sum_{j=1}^{d-1}\sum_{\boldsymbol{\theta}}\sum_{\mathbf{a}}\mathbb{E}_{\boldsymbol{\theta}}\big[\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)=\mathrm{sign}(\boldsymbol{\theta}_j)\}N_0^{\mathbf{a}}\big]. \end{aligned} \tag{C.1}$$

For a fixed coordinate j, consider $\boldsymbol{\theta}'$ that only differs with $\boldsymbol{\theta}$ at its j-th coordinate. We have

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta}}\big[\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)=\mathrm{sign}(\boldsymbol{\theta}_j)\}N_0^{\mathbf{a}}\big]+\mathbb{E}_{\boldsymbol{\theta}'}\big[\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)=\mathrm{sign}(\boldsymbol{\theta}'_j)\}N_0^{\mathbf{a}}\big]\\ &=\mathbb{E}_{\boldsymbol{\theta}'}\big[N_0^{\mathbf{a}}\big]+\mathbb{E}_{\boldsymbol{\theta}}\big[\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)=\mathrm{sign}(\boldsymbol{\theta}_j)\}N_0^{\mathbf{a}}\big]-\mathbb{E}_{\boldsymbol{\theta}'}\big[\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)=\mathrm{sign}(\boldsymbol{\theta}_j)\}N_0^{\mathbf{a}}\big],\end{aligned}$$

since $\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)=\mathrm{sign}(\boldsymbol{\theta}'_j)\}=1-\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)\neq\mathrm{sign}(\boldsymbol{\theta}_j)\}$.

Summing the equation above over $\boldsymbol{\Theta}$ and $\mathcal{A}$, we have

$$\begin{aligned}&2\sum_{\boldsymbol{\theta}}\sum_{\mathbf{a}}\mathbb{E}_{\boldsymbol{\theta}}\big[\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)=\mathrm{sign}(\boldsymbol{\theta}_j)\}N_0^{\mathbf{a}}\big]\\ &=\sum_{\boldsymbol{\theta}}\sum_{\mathbf{a}}\mathbb{E}_{\boldsymbol{\theta}'}\big[N_0^{\mathbf{a}}\big]+\sum_{\boldsymbol{\theta}}\Big[\mathbb{E}_{\boldsymbol{\theta}}\Big[\sum_{\mathbf{a}}\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)=\mathrm{sign}(\boldsymbol{\theta}_j)\}N_0^{\mathbf{a}}\Big]-\mathbb{E}_{\boldsymbol{\theta}'}\Big[\sum_{\mathbf{a}}\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)=\mathrm{sign}(\boldsymbol{\theta}_j)\}N_0^{\mathbf{a}}\Big]\Big]\\ &=\sum_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\theta}'}\big[N_0\big]+\sum_{\boldsymbol{\theta}}\Big[\mathbb{E}_{\boldsymbol{\theta}}\Big[N_0^{\mathcal{A}_j}\Big]-\mathbb{E}_{\boldsymbol{\theta}'}\Big[N_0^{\mathcal{A}_j}\Big]\Big]\\ &\le\sum_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\theta}'}\big[N_0\big]+\sum_{\boldsymbol{\theta}}\sqrt{\log 2/2}T\sqrt{\mathrm{KL}(\mathcal{P}_{\boldsymbol{\theta}'}\|\mathcal{P}_{\boldsymbol{\theta}})}\\ &\le\sum_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\theta}'}\big[N_0\big]+\sum_{\boldsymbol{\theta}}2\sqrt{2\log 2}\frac{T\Delta}{d\sqrt{\delta}}\sqrt{\mathbb{E}_{\boldsymbol{\theta}}[N_0]},\end{aligned} \tag{C.2}$$

where $\mathcal{A}_j$ is the set of all $\mathbf{a}$ which satisfy $\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)=\mathrm{sign}(\boldsymbol{\theta}_j)\}$. The first equality is by matching each $\boldsymbol{\theta}$ with $\boldsymbol{\theta}'$ that differs from $\boldsymbol{\theta}$ in its j-th coordinate, and moving $\sum_{\mathbf{a}}$ inside. The second equality applies the shorthand $\mathcal{A}_j$. The first inequality is due to Lemma C.3. The last is due to Lemma C.4.

Substituting (C.2) into (C.1), we have

$$\begin{aligned}\frac{1}{|\boldsymbol{\Theta}|}\sum_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\theta}}[N_1] &\le \frac{T}{2}+\frac{1}{4\delta|\boldsymbol{\Theta}|}\frac{\Delta}{d-1}\sum_{j=1}^{d-1}\sum_{\boldsymbol{\theta}}\Big[\mathbb{E}_{\boldsymbol{\theta}'}\big[N_0\big]+2\sqrt{2\log 2}\frac{T\Delta}{d\sqrt{\delta}}\sqrt{\mathbb{E}_{\boldsymbol{\theta}}[N_0]}\Big]\\ &= \frac{T}{2}+\frac{\Delta}{4\delta|\boldsymbol{\Theta}|}\sum_{\boldsymbol{\theta}}\Big[\mathbb{E}_{\boldsymbol{\theta}'}\big[N_0\big]+2\sqrt{2\log 2}\frac{T\Delta}{d\sqrt{\delta}}\sqrt{\mathbb{E}_{\boldsymbol{\theta}}[N_0]}\Big]\\ &\le \frac{T}{2}+\frac{\Delta}{4\delta}\Big[\frac{4T}{5}+2\sqrt{2\log 2}\frac{T\Delta}{d\sqrt{\delta}}\frac{2\sqrt{T}}{\sqrt{5}}\Big]\\ &= \frac{T}{2}+\frac{\Delta T}{5\delta}+\sqrt{2\log 2/5}\frac{\Delta^2T^{3/2}}{d\delta^{3/2}},\end{aligned}$$

where the second inequality is due to Lemma C.2.

This further leads to

$$\begin{aligned}\frac{1}{|\mathbf{\Theta}|}\sum_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\theta}}[\text{Regret}_{\boldsymbol{\theta}}(T)] &= T\rho^* - \frac{1}{|\mathbf{\Theta}|}\sum_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\theta}}[N_1]\\ &\geq T\cdot\frac{\delta+\Delta}{2\delta+\Delta} - \frac{T}{2} - \frac{\Delta T}{5\delta} - \sqrt{2\log 2/5}\frac{\Delta^2 T^{3/2}}{d\delta^{3/2}}\\ &= \frac{\Delta(\delta-2\Delta)}{5\delta(4\delta+2\Delta)}\cdot T - \sqrt{2\log 2/5}\frac{\Delta^2 T^{3/2}}{d\delta^{3/2}}\\ &\geq \frac{2}{45\delta}\cdot\Delta T - \sqrt{2\log 2/5}\cdot\frac{\Delta^2 T^{3/2}}{d\delta^{3/2}}\\ &= \frac{1}{2025\sqrt{2\log 2/5}}\cdot d\sqrt{DT}\\ &> \frac{1}{2025}\cdot d\sqrt{DT},\end{aligned}$$

where the second inequality requires $0 < 4\Delta \leq \delta$; the last equality is due to the setting $\delta = D^{-1}$ and $\Delta = (1/45\sqrt{2\log 2/5})d/\sqrt{DT}$. This further requires that $T \geq 16d^2D/2025$. □

D Proof of Supporting Lemmas

D.1 Proof of Lemma A.2

Proof of Lemma A.2. Recall the definition of $\boldsymbol{\theta}_k$ in Algorithm 1, we have

$$\boldsymbol{\theta}_k = \bigg(\lambda\mathbf{I} + \sum_{j=0}^{k-1}\sum_{i=t_j}^{t_{j+1}-1}\boldsymbol{\phi}_{w_j}(s_i,a_i)\boldsymbol{\phi}_{w_j}(s_i,a_i)^\top\bigg)^{-1}\bigg(\sum_{j=0}^{k-1}\sum_{i=t_j}^{t_{j+1}-1}\boldsymbol{\phi}_{w_j}(s_i,a_i)w(s_{i+1})\bigg).$$

It is worth noting that for any $0 \leq j \leq k-1$ and $t_j \leq i \leq t_{j+1}-1$,

$$\begin{aligned}[\mathbb{P}w_j](s_i,a_i) &= \int_{s'}\mathbb{P}(s'|s_i,a_i)w_j(s_i,a_i)ds'\\ &= \int_{s'}\langle\boldsymbol{\phi}(s'|s_i,a_i),\boldsymbol{\theta}^*\rangle w_j(s')ds'\\ &= \bigg\langle\int_{s'}\boldsymbol{\phi}(s'|s_i,a_i)w_j(s'),\boldsymbol{\theta}^*\bigg\rangle\\ &= \langle\boldsymbol{\phi}_{w_j}(s_i,a_i),\boldsymbol{\theta}^*\rangle,\end{aligned}\tag{D.1}$$

thus $\{w_j(s_{i+1}) - \langle\boldsymbol{\phi}_{w_j}(s_i,a_i),\boldsymbol{\theta}^*\rangle\}$ forms a martingale difference sequence. Besides, since $|w(s)| \leq D/2$ for any s, then $w_j(s_{i+1}) - \langle\boldsymbol{\phi}_{w_j}(s_i,a_i),\boldsymbol{\theta}^*\rangle$ is a sequence of D-subgaussian random variables with zero means. Meanwhile, we have $\|\boldsymbol{\phi}_{w_j}(s_i,a_i)\|_2 \leq D$ and $\|\boldsymbol{\theta}^*\|_2 \leq B$ by Definition 3.2. By Theorem 2 in Abbasi-Yadkori et al. (2011), we have that with probability at least $1-\delta$, $\boldsymbol{\theta}^*$ belongs to the following set for all $1 \leq k \leq K$:

$$\bigg\{\boldsymbol{\theta} : \Big\|\mathbf{\Sigma}_{t_k}^{1/2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_k)\Big\|_2 \leq D\sqrt{\log\bigg(\frac{\lambda + t_k D^2}{\delta\lambda}\bigg)} + \sqrt{\lambda}B\bigg\}.\tag{D.2}$$

Finally, by the definition of $\widehat{\beta}_t$ and the fact that $\langle\boldsymbol{\theta}^*,\boldsymbol{\phi}(s'|s,a)\rangle = \mathbb{P}(s'|s,a)$ for all (s,a), we draw the conclusion that $\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{t_k}$ for $1 \leq k \leq K$. □

D.2 Proof of Lemma A.3

Proof of Lemma A.3. For simplicity, we denote $K = K(T)$. Note that $\det(\mathbf{\Sigma}_0) = \lambda^d$. We further have

$$\begin{aligned}\|\mathbf{\Sigma}_T\|_2 &= \bigg\|\lambda\mathbf{I} + \sum_{k=0}^{K-1}\sum_{t=t_k}^{t_{k+1}-1}\boldsymbol{\phi}_{w_k}(s_t,a_t)\boldsymbol{\phi}_{w_k}(s_t,a_t)^\top\bigg\|_2 \\ &\le \lambda + \sum_{k=0}^{K-1}\sum_{t=t_k}^{t_{k+1}-1}\big\|\boldsymbol{\phi}_{w_k}(s_t,a_t)\big\|_2^2 \\ &\le \lambda + TD^2, \end{aligned} \tag{D.3}$$

where the first inequality holds due to the triangle inequality, the second inequality holds due to the fact $w_k(s) \le D/2$ and Definition 3.2. (D.3) suggests that $\det(\mathbf{\Sigma}_T) \le (\lambda + TdD^2)^d$. Therefore, we have

$$(\lambda + TD^2)^d \ge \det(\mathbf{\Sigma}_T) \ge \det(\mathbf{\Sigma}_{t_{K-1}-1}) \ge 2^{K-1}\det(\mathbf{\Sigma}_{t_0-1}) = 2^{K-1}\lambda^d, \tag{D.4}$$

where the second inequality holds since $\mathbf{\Sigma}_T \succeq \mathbf{\Sigma}_{t_{K-1}-1}$, the third inequality holds due to the fact that $\det(\mathbf{\Sigma}_{t_k-1}) \ge 2\det(\mathbf{\Sigma}_{t_{k-1}-1})$ by the update rule in Algorithm 1. (D.4) suggests

$$K \le d\log\frac{2\lambda + 2TD^2}{\lambda}.$$

□

D.3 Proof of Lemma B.2

Proof of Lemma B.2. For simplicity, we denote $K = K(T)$. Note that $\det(\widehat{\mathbf{\Sigma}}_1) = \lambda^d$. We further have

$$\|\widehat{\mathbf{\Sigma}}_{t_K}\|_2 = \bigg\|\lambda\mathbf{I} + \sum_{t=1}^{T}\boldsymbol{\phi}_{w_k}(s_t,a_t)\boldsymbol{\phi}_{w_k}(s_t,a_t)^\top/\bar{\sigma}_t^2\bigg\|_2 \le \lambda + \sum_{t=1}^{T}\big\|\boldsymbol{\phi}_{w_k}(s_t,a_t)/\bar{\sigma}_t\big\|_2^2 \le \lambda + Td,$$

where the first inequality holds due to the triangle inequality, the second inequality holds because $w_k(s) \le D$ and $\bar{\sigma}_t \ge D/\sqrt{d}$. This suggests that $\det(\widehat{\mathbf{\Sigma}}_{t_K}) \le (\lambda + Td)^d$. Therefore, we have

$$(\lambda + Td)^d \ge \det(\mathbf{\Sigma}_{t_K}) \ge \det(\mathbf{\Sigma}_{t_{K-1}}) \ge 2^{K-1}\det(\mathbf{\Sigma}_{t_0}) = 2^{K-1}\lambda^d,$$

where the second inequality holds since $\mathbf{\Sigma}_T \succeq \mathbf{\Sigma}_{t_{K-1}-1}$, the third inequality holds due to the fact that $\det(\mathbf{\Sigma}_{t_k-1}) \ge 2\det(\mathbf{\Sigma}_{t_{k-1}-1})$ by the update rule in Algorithm 1 with OPTION 2. This suggests

$$K \le 2d\log(1 + dT/\lambda).$$

□

D.4 Proof of Lemma B.3

Proof of Lemma B.3. In fact we are able to prove a stronger result:

$$\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_t \cap \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t \cap \mathcal{B},$$

where the two additional sets are defined as

$$\begin{aligned}\check{\mathcal{C}}_t &:= \Big\{\boldsymbol{\theta} : \big\|\check{\mathbf{\Sigma}}_t^{1/2}(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_t)\big\| \le \check{\beta}_t\Big\} \\ \widetilde{\mathcal{C}}_t &:= \Big\{\boldsymbol{\theta} : \big\|\widetilde{\mathbf{\Sigma}}_t^{1/2}(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_t)\big\| \le \widetilde{\beta}_t\Big\}.\end{aligned}$$

For any $1 \le t \le T$, we always have k such that $t_k \le t < t_{k+1}$. We start with the following inequality:

$$\begin{aligned}
\big|[\bar{\mathbb{V}}_t w_k](s_t,a_t) - [\mathbb{V}w_k](s_t,a_t)\big| &= \Big| \min\Big\{D^2/4, \langle \boldsymbol{\phi}_{w_k^2}(s_t,a_t), \widetilde{\boldsymbol{\theta}}_t\rangle\Big\} - \langle \boldsymbol{\phi}_{w_k^2}(s_t,a_t), \boldsymbol{\theta}^*\rangle \\
&\quad + \langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}^*\rangle^2 - \Big[\min\Big\{D/2, \langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}_t\rangle\Big\}\Big]^2\Big| \\
&\le \underbrace{\Big|\min\Big\{D^2/4, \langle \boldsymbol{\phi}_{w_k^2}(s_t,a_t), \widetilde{\boldsymbol{\theta}}_t\rangle\Big\} - \langle \boldsymbol{\phi}_{w_k^2}(s_t,a_t), \boldsymbol{\theta}^*\rangle\Big|}_{I_1} \\
&\quad + \underbrace{\Big|\langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}^*\rangle^2 - \Big[\min\Big\{D/2, \langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}_t\rangle\Big\}\Big]^2\Big|}_{I_2},
\end{aligned}$$

where the inequality is by the triangle inequality.

For I_1, we have

$$\begin{aligned}
I_1 &\le \Big|\langle \boldsymbol{\phi}_{w_k^2}(s_t,a_t), \widetilde{\boldsymbol{\theta}}_t\rangle - \langle \boldsymbol{\phi}_{w_k^2}(s_t,a_t), \boldsymbol{\theta}^*\rangle\Big| \\
&= \Big|\langle \boldsymbol{\phi}_{w_k^2}(s_t,a_t), \widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\rangle\Big| \\
&\le \Big\|\widetilde{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_k^2}(s_t,a_t)\Big\| \cdot \Big\|\widetilde{\boldsymbol{\Sigma}}_t^{1/2}(\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)\Big\|,
\end{aligned}$$

where the first inequality is due to $\langle \boldsymbol{\phi}_{w_k^2}(s_t,a_t), \boldsymbol{\theta}^*\rangle = \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k^2(s')] \in [0, D^2/4]$, and the last inequality is due to Cauchy-Schwartz inequality. Also, it is clear $I_1 \le D^2/4$.

Similarly, for I_2, we have

$$\begin{aligned}
I_2 &= \Big|\langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}^*\rangle + \min\Big\{D/2, \langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}_t\rangle\Big\}\Big| \\
&\quad \cdot \Big|\langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}^*\rangle - \Big[\min\Big\{D/2, \langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}_t\rangle\Big\}\Big]\Big| \\
&\le D \cdot \Big|\langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}^*\rangle - \langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}_t\rangle\Big| \\
&= D \cdot \Big|\langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}^* - \boldsymbol{\theta}_t\rangle\Big| \\
&\le D \cdot \Big\|\widehat{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_k}(s_t,a_t)\Big\| \cdot \Big\|\widehat{\boldsymbol{\Sigma}}_t^{1/2}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_t)\Big\|,
\end{aligned}$$

where the first equality is by $a^2 - b^2 = (a+b)(a-b)$, and the following reasoning is the same as I_1. The only additional fact used in the first inequality is $\langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}^*\rangle \in [0, D/2]$ and $\min\big\{D/2, \langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}_t\rangle\big\} \in [0, D/2]$. Also, it is clear $I_2 \le D^2/4$.

The two terms combined together gives

$$\begin{aligned}
\big|[\bar{\mathbb{V}}_t w_k](s_t,a_t) - [\mathbb{V}w_k](s_t,a_t)\big| &\le \min\Big\{D^2/4, \Big\|\widetilde{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_k^2}(s_t,a_t)\Big\| \cdot \Big\|\widetilde{\boldsymbol{\Sigma}}_t^{1/2}(\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)\Big\|\Big\} \\
&\quad + \min\Big\{D^2/4, D \cdot \Big\|\widehat{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_k}(s_t,a_t)\Big\| \cdot \Big\|\widehat{\boldsymbol{\Sigma}}_t^{1/2}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_t)\Big\|\Big\}. \qquad \text{(D.5)}
\end{aligned}$$

Now, we first show that with probability $1-\delta$, for all t, $\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t$. To show this, we apply Lemma B.1. By setting $\mathbf{x}_t = \bar{\sigma}_t^{-1}\boldsymbol{\phi}_{w_k}(s_t,a_t)$ and $\eta_t = \bar{\sigma}_t^{-1}w_k(s_{t+1}) - \bar{\sigma}_t^{-1}\langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}^*\rangle$, $\mathcal{G}_t = \mathcal{F}_t$, $\boldsymbol{\mu}^* = \boldsymbol{\theta}^*$, $y_t = \langle \boldsymbol{\mu}^*, \mathbf{x}_t\rangle + \eta_t$, $\mathbf{Z}_t = \lambda\mathbf{I} + \sum_{i=1}^t \mathbf{x}_i\mathbf{x}_i^\top$, $\mathbf{b}_t = \sum_{i=1}^t \mathbf{x}_i y_i$ and $\boldsymbol{\mu}_t = \mathbf{Z}_t^{-1}\mathbf{b}_t$, we have $y_t = \bar{\sigma}_t^{-1}w_k(s_{t+1})$ and $\boldsymbol{\mu}_t = \widehat{\boldsymbol{\theta}}_t$. Moreover, we have

$$\|\mathbf{x}_t\|_2 \le \sqrt{d}/2, |\eta_t| \le \sqrt{d}, \mathbb{E}[\eta_t|\mathcal{G}_t] = 0, \mathbb{E}[\eta_t^2|\mathcal{G}_t] = d.$$

Therefore, by Lemma B.1, we have with probability $1-\delta$, for all $t \in [T]$,

$$\|\widehat{\boldsymbol{\Sigma}}_t^{1/2}(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)\|_2 \le 8d\sqrt{\log(1+t/4\lambda)\log(4t^2/\delta)} + 4\sqrt{d}\log(4t^2/\delta) + \sqrt{\lambda}B = \check{\beta}_t.$$

This means that with probability $1-\delta$, for all t, $\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t$.

The same argument can be applied again, except that now we focus on the squared function w_k^2. This gives

$$\|\widetilde{\boldsymbol{\Sigma}}_t^{1/2}(\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)\|_2 \le 8(D^2/4)\sqrt{d\log(1+tD^2/4\lambda d\lambda)\log(4t^2/\delta)} + 4(D^2/4)\log(4t^2/\delta) + \sqrt{\lambda}B = \widetilde{\beta}_t.$$

This means that with probability $1-\delta$, for all t, $\boldsymbol{\theta}^* \in \widetilde{\mathcal{C}}_t$.

Now we show that $\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_t$ with high probability. Let $\mathbf{x}_t = \bar{\sigma}_t^{-1}\boldsymbol{\phi}_{w_k}(s_t, a_t)$, and

$$\eta_t = \bar{\sigma}_t^{-1}\,\mathbb{1}\{\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t\}\big[w_k(s_{t+1}) - \langle \boldsymbol{\phi}_{w_k}(s_t, a_t), \boldsymbol{\theta}^*\rangle\big].$$

In this case, it is clear that still we have $\mathbb{E}[\eta_t|\mathcal{G}_t] = 0$, $|\eta_t| \le \sqrt{d}$, $\|\mathbf{x}_t\|_2 \le \sqrt{d}$. Also,

$$\begin{aligned}
\mathbb{E}[\eta_t^2|\mathcal{G}_t] &= \bar{\sigma}_t^{-2}\,\mathbb{1}\{\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t\}[\mathbb{V}w_t](s_t, a_t) \\
&\le \bar{\sigma}_t^{-2}\,\mathbb{1}\{\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t\}\Bigg[[\bar{\mathbb{V}}_t w_t](s_t, a_t) \\
&\qquad + \min\Big\{D^2/4, \Big\|\widetilde{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_k^2}(s_t, a_t)\Big\| \cdot \Big\|\widetilde{\boldsymbol{\Sigma}}_t^{1/2}(\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)\Big\|\Big\} \\
&\qquad + \min\Big\{D^2/4, D \cdot \Big\|\widehat{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_k}(s_t, a_t)\Big\| \cdot \Big\|\widehat{\boldsymbol{\Sigma}}_t^{1/2}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_t)\Big\|\Big\}\Bigg] \\
&\le \bar{\sigma}_t^{-2}\Bigg[[\bar{\mathbb{V}}_t w_t](s_t, a_t) + \min\Big\{D^2/4, \Big\|\widetilde{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_k^2}(s_t, a_t)\Big\|\widetilde{\beta}_t\Big\} \\
&\qquad + \min\Big\{D^2/4, D\check{\beta}_t \cdot \Big\|\widehat{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_k}(s_t, a_t)\Big\|\Big\}\Bigg] \\
&= 1,
\end{aligned}$$

where the first inequality is due to (D.5) and the second inequality is due to first, the event that $\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t$. The last equality is by the definition of $\bar{\sigma}_t$.

Again by Lemma B.1, we have that for all $t \in [T]$,

$$\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t} \le 8\sqrt{d\log(1+t/4\lambda)\log(4t^2/\delta)} + 4\sqrt{d}\log(4t^2/\delta) + \sqrt{\lambda}B = \widehat{\beta}_t.$$

Now, denote the event when $\{\forall t \in [T], \boldsymbol{\theta}^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t\}$ and the inequality above holds as $\mathcal{E}_0$. By union bound, we have $\mathbb{P}(\mathcal{E}_0) \ge 1 - 3\delta$.

It is clear that under $\mathcal{E}_0$, we have $\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_t$ for all t because under event $\mathcal{E}_0$,

$$\begin{aligned}
y_t &= \langle \bar{\sigma}_t^{-1}\boldsymbol{\phi}_{w_k}(s_t, a_t), \boldsymbol{\theta}^*\rangle + \bar{\sigma}_t^{-1}\,\mathbb{1}\{\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t\}\big[w_k(s_{t+1}) - \langle \boldsymbol{\phi}_{w_k}(s_t, a_t), \boldsymbol{\theta}^*\rangle\big] \\
&= \bar{\sigma}_t^{-1} w_k(s_{t+1}),
\end{aligned}$$

so indeed we have $\|\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\|_{\widehat{\boldsymbol{\Sigma}}_t} \le \widehat{\beta}_t$.

Also, by the definition of E_t, it is clear that under event $\mathcal{E}_0$,

$$\big|[\bar{\mathbb{V}}_t w_k](s_t, a_t) - [\mathbb{V}w_k](s_t, a_t)\big| \le E_t.$$

□

D.5 Proof of Lemma B.4

Proof of Lemma B.4. Part of the proof is inspired by Fruit et al. (2020). We will use $\mathbb{V}_P(w)$ to denote $\mathbb{E}_{s'\sim P(\cdot)}[w(s')^2] - (\mathbb{E}_{s'\sim P(\cdot)}[w(s')])^2$, namely the variance of the random variable $w(s')$ where $s' \sim P(\cdot)$. Some examples are

$$\begin{aligned}
\mathbb{V}_{\mathbb{P}(\cdot|s_t,a_t)}(w_k) &= \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')^2] - (\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')])^2, \\
\mathbb{V}_{\mathbb{P}_k(\cdot|s_t,a_t)}(w_k) &= \mathbb{E}_{s'\sim\mathbb{P}_k(\cdot|s_t,a_t)}[w_k(s')^2] - (\mathbb{E}_{s'\sim\mathbb{P}_k(\cdot|s_t,a_t)}[w_k(s')])^2.
\end{aligned}$$

When the context is clear, we may also use short-hands like $\mathbb{E}_p[w(s')]$ to indicate expectation under $p(\cdot)$. The following decomposition is useful:

$$
\begin{aligned}
&\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\mathbb{V}_{\mathbb{P}(\cdot|s_t,a_t)}(w_k)\\
&=\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')^2]-(\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')])^2\\
&=\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big[\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')^2]-w_k^2(s_{t+1})\big]\\
&\quad+\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big[w_k^2(s_{t+1})-(\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')])^2\big]\\
&=\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big[\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')^2]-w_k^2(s_{t+1})\big]\\
&\quad+\sum_{k=0}^{K(T)-1}\Bigg[\sum_{t=t_k}^{t_{k+1}-1}\big[w_k^2(s_t)-(\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')])^2\big]+w_k^2(s_{t_{k+1}})-w_k^2(s_{t_k})\Bigg]\\
&\le\underbrace{\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big[\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')^2]-w_k^2(s_{t+1})\big]}_{I_1}\\
&\quad+\underbrace{\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big[w_k^2(s_t)-(\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')])^2\big]}_{I_2}+K(T)\cdot D^2/4.
\end{aligned}
$$

For term I_1, since the event $\mathcal{E}_1$ holds, we have

$$I_1\le(D^2/4)\sqrt{2T\log(1/\delta)}.$$

For term I_2, we have

$$
\begin{aligned}
I_2&=\sum_{t=1}^{T}\big[w_k^2(s_t)-(\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')])^2\big]\\
&\le\sum_{t=1}^{T}\big|w_k(s_t)-\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')]\big|\cdot\big|w_k(s_t)+\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')]\big|\\
&\le D\sum_{t=1}^{T}\big|w_k(s_t)-(\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')])\big|.
\end{aligned}
$$

Note that, w_k, as the output of the Extended Value Iteration, satisfies the following condition(Lemma A.1):

$$|r(s_t,a_t)+\mathbb{E}_{s'\sim\mathbb{P}_k(\cdot|s_t,a_t)}[w_k(s')]-w_k(s_t)-\rho_k|\le\epsilon.$$

Therefore, we can further bound each term in I_2 as follows:

$$\begin{aligned}
\left|w_k(s_t) - \mathbb{E}_{\mathbb{P}}[w_k(s')]\right| &= \left|w_k(s_t) - \mathbb{E}_{\mathbb{P}_k}[w_k(s')] + \mathbb{E}_{\mathbb{P}_k}[w_k(s')] - \mathbb{E}_{\mathbb{P}}[w_k(s')]\right| \\
&\le |r(s_t,a_t) + \mathbb{E}_{\mathbb{P}_k}[w_k(s')] - w_k(s_t) - \rho_k| + |r(s_t,a_t) - \rho_k| \\
&\quad + \left|\mathbb{E}_{\mathbb{P}_k}[w_k(s')] - \mathbb{E}_{\mathbb{P}}[w_k(s')]\right| \\
&\le r_{\max} + r_{\max} + \left|\mathbb{E}_{\mathbb{P}_k}[w_k(s')] - \mathbb{E}_{\mathbb{P}}[w_k(s')]\right| \\
&= 2r_{\max} + \left|\langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}_k - \boldsymbol{\theta}^*\rangle\right| \\
&\le 2r_{\max} + \left\|\boldsymbol{\phi}_{w_k}(s_t,a_t)\right\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}} \cdot \left\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\right\|_{\widehat{\boldsymbol{\Sigma}}_t} \\
&\le 2r_{\max} + 2\widehat{\beta}_t \left\|\boldsymbol{\phi}_{w_k}(s_t,a_t)\right\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}.
\end{aligned}$$

Here, the first inequality is due to triangle inequality. The second inequality is due to 1) the reward function(so should the average reward) should lie in $[0, r_{\max}]$ as assumed, and in this paper's setting actually $r_{\max} = 1$. The third inequality is due to Cauchy-Schwartz inequality. The last inequality is due to the assumption $\mathcal{E}_0$ holds. For the second equality, note that $\mathbb{E}_{\mathbb{P}}[w(s')] = \langle \boldsymbol{\phi}_w(s'|s_t,a_t), \boldsymbol{\theta}^*\rangle$.

Also, it is clear that $\left|\mathbb{E}_{\mathbb{P}_k}[w_k(s')] - \mathbb{E}_{\mathbb{P}}[w_k(s')]\right| \le D$. Therefore, term I_2 can be bounded as

$$\begin{aligned}
I_2 &\le D \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \left[2r_{\max} + \min\left\{D, \widehat{\beta}_t \left\|\boldsymbol{\phi}_{w_k}(s_t,a_t)\right\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}\right\}\right] \\
&= 2DT + D \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \min\left\{D, \widehat{\beta}_t \left\|\boldsymbol{\phi}_{w_k}(s_t,a_t)\right\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}\right\} \\
&\le 2DT + D \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \widehat{\beta}_t \bar{\sigma}_t \min\left\{1, \left\|\boldsymbol{\phi}_{w_k}(s_t,a_t)/\bar{\sigma}_t\right\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}\right\} \\
&\le 2DT + D^2\widehat{\beta}_T \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \min\left\{1, \left\|\boldsymbol{\phi}_{w_k}(s_t,a_t)/\bar{\sigma}_t\right\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}\right\} \\
&\le 2DT + D^2\widehat{\beta}_T\sqrt{T}\sqrt{\sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \min\left\{1, \left\|\boldsymbol{\phi}_{w_k}(s_t,a_t)/\bar{\sigma}_t\right\|^2_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}\right\}} \\
&\le 2DT + D^2\widehat{\beta}_T\sqrt{T 2d\log(1+T/\lambda)}.
\end{aligned}$$

The second inequality holds because $\widehat{\beta}_t \ge \sqrt{d}$ and $\bar{\sigma}_t \ge D/\sqrt{d}$.The third inequality holds because $\widehat{\beta}_t \le \widehat{\beta}_T$ and $\bar{\sigma}_t \le D$. The fourth inequality is due to Cauchy-Schwartz inequality. The last inequality is from Lemma A.6.

Collecting I_1 and I_2 gives

$$\sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} [\mathbb{V}w_k](s_t,a_t) \le (D^2/4)\sqrt{2T\log(1/\delta)} + (K(T)+1)(D^2/4) + 2DT + D^2\widehat{\beta}_T\sqrt{T2d\log(1+T/\lambda)},$$

given that $\mathcal{E}_0$ and $\mathcal{E}_1$ hold. Using big-O notation we have

$$\sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} [\mathbb{V}w_k](s_t,a_t) = \widetilde{O}(DT) + \widetilde{O}(D^2d\sqrt{T}).$$

□

D.6 Proof of Lemma B.5

Proof of Lemma B.5. Directly unroll the definition of E_t:

$$\sum_{t=1}^{T} E_t = \underbrace{\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1} \min\Big\{D^2/4, \widetilde{\beta}_t\Big\|\widetilde{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_k^2}(s_t,a_t)\Big\|\Big\}}_{I_1} + \underbrace{\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1} \min\Big\{D^2/4, D\check{\beta}_t\Big\|\widehat{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_k}(s_t,a_t)\Big\|\Big\}}_{I_2}.$$

For term I_1,

$$\begin{aligned} I_1 &\le \widetilde{\beta}_T \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1} \min\Big\{1, \Big\|\widetilde{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_k^2}(s_t,a_t)\Big\|\Big\} \\ &\le \widetilde{\beta}_T\sqrt{T}\sqrt{\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1} \min\Big\{1, \Big\|\widetilde{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_k^2}(s_t,a_t)\Big\|^2\Big\}} \\ &\le \widetilde{\beta}_T\sqrt{2Td\log(1+TD^2/4d\lambda)}, \end{aligned}$$

where the first inequality is due to $\widetilde{\beta}_t \le \widetilde{\beta}_T$ and $\widetilde{\beta}_t \ge D^2/4$. The second inequality is due to Cauchy-Schwartz inequality. The third is due to Lemma A.5.

Similarly, for I_2, we have

$$\begin{aligned} I_2 &= \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1} \min\Big\{D^2/4, D\check{\beta}_t\bar{\sigma}_t\Big\|\widehat{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_k}(s_t,a_t)/\bar{\sigma}_t\Big\|\Big\} \\ &\le D^2\check{\beta}_T \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1} \min\Big\{1, \Big\|\widehat{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_k}(s_t,a_t)/\bar{\sigma}_t\Big\|\Big\} \\ &\le D^2\check{\beta}_T\sqrt{T\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1} \min\Big\{1, \Big\|\widehat{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_k}(s_t,a_t)/\bar{\sigma}_t\Big\|^2\Big\}} \\ &\le D^2\check{\beta}_T\sqrt{2Td\log(1+T/\lambda)}, \end{aligned}$$

where the first inequality is due to $\check{\beta}_t\bar{\sigma}_t \ge D$, $\check{\beta}_t \le \check{\beta}_T$ and $\bar{\sigma}_t \le D$(all can be verified by the definitions).

To summarize,

$$\sum_{t=1}^{T} E_t \le \widetilde{\beta}_T\sqrt{2Td\log(1+TD^2/4d\lambda)} + D^2\check{\beta}_T\sqrt{2Td\log(1+T/\lambda)}.$$

We can also conclude that

$$\sum_{t=1}^{T} E_t = \widetilde{O}(D^2d^{3/2}\sqrt{T}).$$

□

D.7 Proof of Lemma C.2

Proof of Lemma C.2. We have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}} N_1 &= \sum_{t=2}^{T} \mathcal{P}_{\boldsymbol{\theta}}(s_t = x_1) \\ &= \underbrace{\sum_{t=2}^{T} \mathcal{P}_{\boldsymbol{\theta}}(s_t = x_1 | s_{t-1} = x_1)\mathcal{P}_{\boldsymbol{\theta}}(s_{t-1} = x_1)}_{I_1} + \underbrace{\sum_{t=2}^{T} \mathcal{P}_{\boldsymbol{\theta}}(s_t = x_1, s_{t-1} = x_0)}_{I_2}. \end{aligned} \tag{D.6}$$

For I_1, since $\mathcal{P}_{\boldsymbol{\theta}}(s_t = x_1 | s_{t-1} = x_1) = 1 - \delta$ no matter which action is taken, thus we have

$$I_1 = (1-\delta)\sum_{t=2}^{T} \mathcal{P}_{\boldsymbol{\theta}}(s_{t-1} = x_1) = (1-\delta)\mathbb{E}_{\boldsymbol{\theta}} N_1 - (1-\delta)\mathcal{P}_{\boldsymbol{\theta}}(s_T = x_1). \tag{D.7}$$

Next we bound I_2. We can further decompose I_2 as follows.

$$\begin{aligned} I_2 &= \sum_{t=2}^{T}\sum_{\mathbf{a}} \mathcal{P}_{\boldsymbol{\theta}}(s_t = x_1 | s_{t-1} = x_0, a_{t-1} = \mathbf{a})\mathcal{P}_{\boldsymbol{\theta}}(s_{t-1} = x_0, a_{t-1} = \mathbf{a}) \\ &= \sum_{t=2}^{T}\sum_{\mathbf{a}} (\delta + \langle \mathbf{a}, \boldsymbol{\theta} \rangle)\mathcal{P}_{\boldsymbol{\theta}}(s_{t-1} = x_0, a_{t-1} = \mathbf{a}) \\ &= \sum_{\mathbf{a}} (\delta + \langle \mathbf{a}, \boldsymbol{\theta} \rangle)\Big[\mathbb{E}_{\boldsymbol{\theta}} N_0^{\mathbf{a}} - \mathcal{P}_{\boldsymbol{\theta}}(s_T = x_0, a_T = \mathbf{a})\Big]. \end{aligned} \tag{D.8}$$

Substituting (D.7) and (D.8) into (D.6) and rearranging it, we have

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\theta}} N_1 \\ &= \sum_{\mathbf{a}} (1 + \langle \mathbf{a}, \boldsymbol{\theta} \rangle / \delta)\mathbb{E}_{\boldsymbol{\theta}} N_0^{\mathbf{a}} - \underbrace{\Big[\frac{1-\delta}{\delta}\mathcal{P}_{\boldsymbol{\theta}}(s_T = x_1) + \sum_{\mathbf{a}} (1 + \langle \mathbf{a}, \boldsymbol{\theta} \rangle / \delta)\mathcal{P}_{\boldsymbol{\theta}}(s_T = x_0, a_T = \mathbf{a})\Big]}_{\psi_{\boldsymbol{\theta}}} \\ &= \mathbb{E}_{\boldsymbol{\theta}} N_0 + \delta^{-1}\sum_{\mathbf{a}} \langle \mathbf{a}, \boldsymbol{\theta} \rangle \mathbb{E}_{\boldsymbol{\theta}} N_0^{\mathbf{a}} - \psi_{\boldsymbol{\theta}}, \end{aligned} \tag{D.9}$$

which suggests that

$$\mathbb{E}_{\boldsymbol{\theta}} N_1 \le T/2 + \delta^{-1}\sum_{\mathbf{a}} \langle \mathbf{a}, \boldsymbol{\theta} \rangle \mathbb{E}_{\boldsymbol{\theta}} N_0^{\mathbf{a}}/2. \tag{D.10}$$

We now bound $\mathbb{E}_{\boldsymbol{\theta}} N_0$. By (D.9), we have

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}} N_1 &\ge \mathbb{E}_{\boldsymbol{\theta}} N_0 + \delta^{-1}\sum_{\mathbf{a}} \langle \mathbf{a}, \boldsymbol{\theta} \rangle \mathbb{E}_{\boldsymbol{\theta}} N_0^{\mathbf{a}} - \psi_{\boldsymbol{\theta}} \\ &\ge \mathbb{E}_{\boldsymbol{\theta}} N_0 - \frac{\Delta}{\delta}\mathbb{E}_{\boldsymbol{\theta}} N_0 - \frac{1-\delta}{\delta}\mathcal{P}_{\boldsymbol{\theta}}(s_T = x_1) - \mathcal{P}_{\boldsymbol{\theta}}(s_T = x_0) - \frac{\Delta}{\delta}\mathcal{P}_{\boldsymbol{\theta}}(s_T = x_0) \\ &= (1 - \Delta/\delta)\mathbb{E}_{\boldsymbol{\theta}} N_0 - (1-\delta)/\delta + \frac{1-\Delta}{\delta}\mathcal{P}_{\boldsymbol{\theta}}(s_T = x_0) \\ &\ge (1 - \Delta/\delta)\mathbb{E}_{\boldsymbol{\theta}} N_0 - (1-\delta)/\delta, \end{aligned} \tag{D.11}$$

where the first inequality holds due to (D.9), the second inequality holds due to the fact that $\langle \mathbf{a}, \boldsymbol{\theta} \rangle \le \Delta$, the last inequality holds since $\mathcal{P}_{\boldsymbol{\theta}}(s_T = x_0) > 0$. (D.11) suggests that

$$\mathbb{E}_{\boldsymbol{\theta}} N_0 \le \frac{T + (1-\delta)/\delta}{2 - \Delta/\delta} \le \frac{4}{5}T,$$

where the last inequality holds due to the fact that $2\Delta \le \delta$ and $(1-\delta)/\delta < T/5$. □

D.8 Proof of Lemma C.4

We need the following lemma:

Lemma D.1 (Lemma 20 in Jaksch et al. (2010)). Suppose $0 \leq \delta' \leq 1/2$ and $\epsilon' \leq 1 - 2\delta'$, then

$$\delta' \log \frac{\delta'}{\delta' + \epsilon'} + (1 - \delta') \log \frac{(1 - \delta')}{1 - \delta' - \epsilon'} \leq \frac{2(\epsilon')^2}{\delta'}.$$

Proof of Lemma C.4. Let $\mathbf{s}_t$ denote $\{s_1, \ldots, s_t\}$. By the Markovnian property of MDP, we can first decompose the KL divergence as follows:

$$\mathrm{KL}(\mathcal{P}_{\boldsymbol{\theta}'} \| \mathcal{P}_{\boldsymbol{\theta}}) = \sum_{t=1}^{T-1} \mathrm{KL}\Big[\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1}|\mathbf{s}_t) \Big\| \mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}|\mathbf{s}_t)\Big],$$

where the KL divergence between $\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1}|\mathbf{s}_t), \mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}|\mathbf{s}_t)$ is defined as follows:

$$\mathrm{KL}\Big[\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1}|\mathbf{s}_t) \Big\| \mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}|\mathbf{s}_t)\Big] = \sum_{\mathbf{s}_{t+1} \in \mathcal{S}^{t+1}} \mathcal{P}_{\boldsymbol{\theta}'}(\mathbf{s}_{t+1}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1}|\mathbf{s}_t)}{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}|\mathbf{s}_t)}.$$

Now we further bound the above terms as follows:

$$\begin{aligned}
&\sum_{\mathbf{s}_{t+1} \in \mathcal{S}^{t+1}} \mathcal{P}_{\boldsymbol{\theta}'}(\mathbf{s}_{t+1}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1}|\mathbf{s}_t)}{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}|\mathbf{s}_t)} \\
&= \sum_{\mathbf{s}_t \in \mathcal{S}^t} \mathcal{P}_{\boldsymbol{\theta}'}(\mathbf{s}_t) \sum_{x \in \mathcal{S}} \mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1} = x|\mathbf{s}_t) \log \frac{\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1} = x|\mathbf{s}_t)}{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1} = x|\mathbf{s}_t)} \\
&= \sum_{\mathbf{s}_{t-1} \in \mathcal{S}^{t-1}} \mathcal{P}_{\boldsymbol{\theta}'}(\mathbf{s}_{t-1}) \sum_{x' \in \mathcal{S}, \mathbf{a} \in \mathcal{A}} \mathcal{P}_{\boldsymbol{\theta}'}(s_t = x', a_t = \mathbf{a}|\mathbf{s}_{t-1}) \\
&\quad \cdot \sum_{x \in \mathcal{S}} \mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1} = x|\mathbf{s}^{t-1}, s_t = x', a_t = \mathbf{a}) \underbrace{\log \frac{\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1} = x|\mathbf{s}^{t-1}, s_t = x', a_t = \mathbf{a})}{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1} = x|\mathbf{s}^{t-1}, s_t = x', a_t = \mathbf{a})}}_{I_1},
\end{aligned}$$

When $x' = x_1$, $\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1} = x|\mathbf{s}^{t-1}, s_t = x', a_t = \mathbf{a}) = \mathcal{P}_{\boldsymbol{\theta}}(s_{t+1} = x|\mathbf{s}^{t-1}, s_t = x', a_t = \mathbf{a})$ for all $\boldsymbol{\theta}', \boldsymbol{\theta}$ since the transition probability at x_1 is irrelevant to $\boldsymbol{\theta}$ due to the MDP we choose. That implies when $x' = x_1$, $I_1 = 0$. Therefore,

$$\begin{aligned}
&\sum_{\mathbf{s}_{t+1} \in \mathcal{S}^{t+1}} \mathcal{P}_{\boldsymbol{\theta}'}(\mathbf{s}_{t+1}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1}|\mathbf{s}_t)}{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}|\mathbf{s}_t)} \\
&= \sum_{\mathbf{s}^{t-1} \in \mathcal{S}^{t-1}} \mathcal{P}_{\boldsymbol{\theta}'}(\mathbf{s}^{t-1}) \sum_{\mathbf{a}} \mathcal{P}_{\boldsymbol{\theta}'}(s_t = x_0, a_t = \mathbf{a}|\mathbf{s}^{t-1}) \\
&\quad \cdot \sum_{x \in \mathcal{S}} \mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1} = x|\mathbf{s}^{t-1}, s_t = x_0, a_t = \mathbf{a}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1} = s|\mathbf{s}^{t-1}, s_t = x_0, a_t = \mathbf{a})}{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1} = s|\mathbf{s}^{t-1}, s_t = x_0, a_t = \mathbf{a})} \\
&= \sum_{\mathbf{a}} \mathcal{P}_{\boldsymbol{\theta}'}(s_t = x_0, a_t = \mathbf{a}) \\
&\quad \underbrace{\cdot \sum_{x \in \mathcal{S}} \mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1} = s|s_t = x_0, a_t = \mathbf{a}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1} = x|s_t = x_0, a_t = \mathbf{a})}{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1} = x|s_t = x_0, a_t = \mathbf{a})}}_{I_2}.
\end{aligned} \tag{D.12}$$

To bound I_2, due to the structure of the MDP, we know that s_{t+1} follows the Bernoulli distribution over x_0 and x_1 with probability $1 - \delta - \langle \mathbf{a}, \boldsymbol{\theta}' \rangle$ and $\delta + \langle \mathbf{a}, \boldsymbol{\theta}' \rangle$, then we have

$$I_2 = (1 - \langle \boldsymbol{\theta}', \mathbf{a} \rangle - \delta) \log \frac{1 - \langle \boldsymbol{\theta}', \mathbf{a} \rangle - \delta}{1 - \langle \boldsymbol{\theta}, \mathbf{a} \rangle - \delta} + (\langle \boldsymbol{\theta}', \mathbf{a} \rangle + \delta) \log \frac{\langle \boldsymbol{\theta}', \mathbf{a} \rangle + \delta}{\langle \boldsymbol{\theta}, \mathbf{a} \rangle + \delta} \leq \frac{2\langle \boldsymbol{\theta}' - \boldsymbol{\theta}, \mathbf{a} \rangle^2}{\langle \boldsymbol{\theta}', \mathbf{a} \rangle + \delta}, \tag{D.13}$$

where the inequality holds due to Lemma D.1 with $\delta' = \langle \boldsymbol{\theta}', \mathbf{a} \rangle + \delta$ and $\epsilon' = \langle \boldsymbol{\theta} - \boldsymbol{\theta}', \mathbf{a} \rangle$. It can be verified that

$$\delta' = \langle \boldsymbol{\theta}', \mathbf{a} \rangle + \delta \leq \Delta + \delta \leq 1/2, \tag{D.14}$$

where the first inequality holds due to the definition of $\boldsymbol{\theta}'$, the second inequality holds since $\Delta < \delta/2 \leq 1/6$. It can also be verified that

$$\epsilon' = \langle \boldsymbol{\theta} - \boldsymbol{\theta}', \mathbf{a} \rangle \leq 2\Delta \leq 1 - 2(\Delta + \delta) \leq 1 - 2\delta', \tag{D.15}$$

where the first inequality holds due to the definition of $\boldsymbol{\theta}', \boldsymbol{\theta}$, the second inequality holds since $\Delta < \delta/4 \leq 1/12$, the last inequality holds since $\delta' = \langle \boldsymbol{\theta}', \mathbf{a} \rangle + \delta \leq \Delta + \delta$ due to the definition of $\boldsymbol{\theta}'$. (D.14) and (D.15) suggest that we can apply Lemma D.1 onto (D.13). I_2 can be further bounded as follows:

$$I_2 \leq \frac{4\langle \boldsymbol{\theta}' - \boldsymbol{\theta}, \mathbf{a} \rangle^2}{\delta} = \frac{16\Delta^2}{(d-1)^2 \delta}, \tag{D.16}$$

where the inequality holds due to (D.13) and the fact that $\delta + \langle \boldsymbol{\theta}', \mathbf{a} \rangle \geq \delta - \Delta \geq \delta/2$. Substituting (D.16) into (D.12), taking summation from $t = 1$ to $T - 1$, we have

$$\begin{aligned}
\mathrm{KL}(\mathcal{P}_{\boldsymbol{\theta}'} \| \mathcal{P}_{\boldsymbol{\theta}}) &= \sum_{t=1}^{T-1} \sum_{\mathbf{s}_{t+1} \in \mathcal{S}^{t+1}} \mathcal{P}_{\boldsymbol{\theta}'}(\mathbf{s}_{t+1}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1} | \mathbf{s}_t)}{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1} | \mathbf{s}_t)} \\
&\leq \frac{16\Delta^2}{(d-1)^2 \delta} \sum_{t=1}^{T-1} \sum_{\mathbf{a}} \mathcal{P}_{\boldsymbol{\theta}'}(s_t = x_0, a_t = \mathbf{a}) \\
&= \frac{16\Delta^2}{(d-1)^2 \delta} \sum_{t=1}^{T-1} \mathcal{P}_{\boldsymbol{\theta}'}(s_t = x_0) \\
&\leq \frac{16\Delta^2}{(d-1)^2 \delta} \mathbb{E}_{\boldsymbol{\theta}'} N_0,
\end{aligned}$$

where the last inequality holds due to the definition of N_0. □

E Experiments

In this section, we conduct experiments to empirically study the performance of the proposed algorithm.

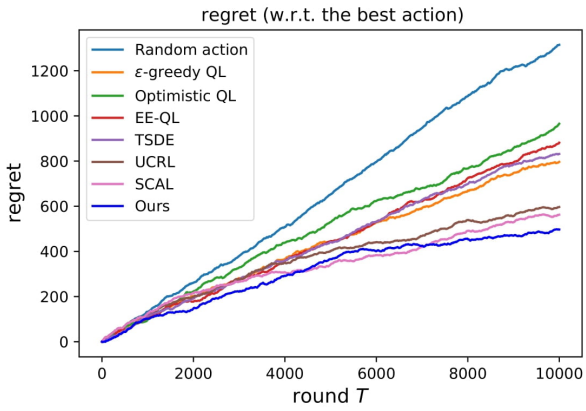

Figure 3: Regret comparison of different algorithms. UCRL2-VTR performs better than the tabular Q-learning by utilizing the given linear structure.

The MDP is constructed as described in Section C.1. We choose $d = 8$, and thus $|\mathcal{S}| = 2$ and $|\mathcal{A}| = 2^{d-1} = 128$.

We compare the following algorithms:

1. Randomly choose an action (Random action).
2. Q-learning with an ϵ-greedy, uniformly random exploration (ϵ-greedy QL).
3. Q-learning with a confidence bonus (Optimistic QL by Wei et al. (2020b)).
4. An Exploration Enhanced Q-learning algorithm (EE-QL by Jafarnia-Jahromi et al. (2020)).
5. A Thompson sampling-based algorithm (TSDE by Ouyang et al. (2017)).
6. A tabular model-based algorithm (UCRL by Jaksch et al. (2010)).
7. A tabular model-based algorithm that relies on the span of the MDP rather than the diameter (SCAL by Fruit et al. (2018b)).
8. Our algorithm with the Hoeffding bonus (Ours).

In our experiments, the parameters of each algorithm are tuned properly. For each algorithm, the experiment is replicated for 10 times and the averaged regret is plotted in Figure 3 for comparison. We can see that model-based algorithms (UCRL, SCAL, Ours) are generally better than the model-free ones (Q-learning algorithms and TSDE). Our proposed algorithm outperforms other model-based algorithms due to utilizing the linear structure of the underlying MDP.