Supplement to “Sample-Efficient Reinforcement Learning for Linearly-Parameterized MDPs with a Generative Model”

Bingyan Wang*
Princeton University
bingyanw@princeton.edu

Yuling Yan*
Princeton University
yulingy@princeton.edu

Jianqing Fan
Princeton University
jqfan@princeton.edu

A Notations

In this section we gather the notations that will be used throughout the appendix.

For any vectors $\boldsymbol{u} = [u_i]_{i=1}^n \in \mathbb{R}^n$ and $\boldsymbol{v} = [u_i]_{i=1}^n \in \mathbb{R}^n$, let $\boldsymbol{u} \circ \boldsymbol{v} = [u_i v_i]_{i=1}^n$ denote the Hadamard product of $\boldsymbol{u}$ and $\boldsymbol{v}$. We slightly abuse notations to use $\sqrt{\cdot}$ and $|\cdot|$ to define entry-wise operation, i.e. for any vector $\boldsymbol{v} = [v_i]_{i=1}^n$ denote $\sqrt{\boldsymbol{v}} := [\sqrt{v_i}]_{i=1}^n$ and $|\boldsymbol{v}| := [|v_i|]_{i=1}^n$. Furthermore, the binary notations $\leq$ and $\geq$ are both defined in entry-wise manner, i.e. $\boldsymbol{u} \leq \boldsymbol{v}$ (resp. $\boldsymbol{u} \geq \boldsymbol{v}$) means $u_i \leq v_i$ (resp. $u_i \geq v_i$) for all $1 \leq i \leq n$. For a collection of vectors $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_m \in \mathbb{R}^n$ with $\boldsymbol{v}_i = [v_{i,j}]_{j=1}^n \in \mathbb{R}^n$, we define the max operator to be $\max_{1\leq i\leq m} \boldsymbol{v}_i := [\max_{1\leq i\leq m} v_{i,j}]_{j=1}^n$.

For any matrix $\boldsymbol{M} \in \mathbb{R}^{m\times n}$, $\|\boldsymbol{M}\|_1$ is defined as the largest row-wise ℓ_1 norm of $\boldsymbol{M}$, i.e. $\|\boldsymbol{M}\|_1 := \max_i \sum_j |M_{i,j}|$. In addition, we define $\mathbf{1}$ to be a vector with all the entries being 1, and $\boldsymbol{I}$ be the identity matrix. To express the probability transition function P in matrix form, we define the matrix $\boldsymbol{P} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$ to be a matrix whose (s,a)-th row $\boldsymbol{P}_{s,a}$ corresponds to $P(\cdot|s,a)$. In addition, we define $\boldsymbol{P}^\pi$ to be the probability transition matrix induced by policy π, i.e. $\boldsymbol{P}^\pi_{(s,a),(s',a')} = \boldsymbol{P}_{s,a}(s')\,\mathbb{1}_{\pi(s')=a'}$ for all state-action pairs (s,a) and (s',a'). We define π_t to be the policy induced by Q_t, i.e. $Q_t(s,\pi_t(s)) = \max_a Q_t(s,a)$ for all $s \in \mathcal{S}$. Furthermore, we denote the reward function r by vector $\boldsymbol{r} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, i.e. the (s,a)-th element of $\boldsymbol{r}$ equals $r(s,a)$. In the same manner, we define $\boldsymbol{V}^\pi \in \mathbb{R}^{|\mathcal{S}|}$, $\boldsymbol{V}^\star \in \mathbb{R}^{|\mathcal{S}|}$, $\boldsymbol{V}_t \in \mathbb{R}^{|\mathcal{S}|}$, $\boldsymbol{Q}^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $\boldsymbol{Q}^\star \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and $\boldsymbol{Q}_t \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ to represent V^π, $V^\star$, V_t, Q^π, $Q^\star$ and Q_t respectively. By using these notations, we can rewrite the Bellman equation as

$$\boldsymbol{Q}^\pi = \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{V}^\pi = \boldsymbol{r} + \gamma \boldsymbol{P}^\pi \boldsymbol{Q}^\pi. \tag{11}$$

Further, for any vector $\boldsymbol{V} \in \mathbb{R}^{|\mathcal{S}|}$, let $\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be

$$\mathsf{Var}_{\boldsymbol{P}}\left(\boldsymbol{V}\right) := \boldsymbol{P}\left(\boldsymbol{V}\circ\boldsymbol{V}\right) - \left(\boldsymbol{P}\boldsymbol{V}\right)\circ\left(\boldsymbol{P}\boldsymbol{V}\right), \tag{12}$$

and define $\mathsf{Var}_{\boldsymbol{P}_{s,a}}(\boldsymbol{V}) \in \mathbb{R}$ to be

$$\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\boldsymbol{V}\right) := \boldsymbol{P}_{s,a}\left(\boldsymbol{V}\circ\boldsymbol{V}\right) - \left(\boldsymbol{P}_{s,a}\boldsymbol{V}\right)^2, \tag{13}$$

where $\boldsymbol{P}_{s,a}$ is the (s,a)-th row of $\boldsymbol{P}$.

Next, we reconsider Assumption 1. For any state-action pair (s,a), we define vector $\boldsymbol{\lambda}(s,a) \in \mathbb{R}^K$ (resp. $\boldsymbol{\phi}(s,a) \in \mathbb{R}^K$) with $\boldsymbol{\lambda}(s,a) = [\lambda_i(s,a)]_{i=1}^K$ (resp. $\boldsymbol{\phi}(s,a) = [\phi_i(s,a)]_{i=1}^K$) and matrix

*Equal contribution.

$\mathbf{\Lambda} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times K}$ (resp. $\mathbf{\Phi} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times K}$) whose (s,a)-th row corresponds to $\boldsymbol{\lambda}(s,a)^\top$ (resp. $\boldsymbol{\phi}(s,a)^\top$). Define vector $\boldsymbol{\psi}(s,a) \in \mathbb{R}^K$ with $\boldsymbol{\psi}(s,a) = [\psi_i(s,a)]_{i=1}^K$ and matrix $\mathbf{\Psi} \in \mathbb{R}^{K \times |\mathcal{S}|}$ whose (s,a)-th column corresponds to $\boldsymbol{\psi}(s,a)^\top$. Further, let $\boldsymbol{P}_{\mathcal{K}} \in \mathbb{R}^{K \times |\mathcal{S}|}$ (resp. $\mathbf{\Phi}_{\mathcal{K}} \in \mathbb{R}^{K \times K}$) to be a submatrix of $\boldsymbol{P}$ (resp. $\mathbf{\Phi}$) formed by concatenating the rows $\{\boldsymbol{P}_{s,a}, (s,a) \in \mathcal{K}\}$ (resp. $\{\mathbf{\Phi}_{s,a}, (s,a) \in \mathcal{K}\}$). By using the previous notations, we can express the relations in Definition 1 and Assumption 1 as $\boldsymbol{P}_{\mathcal{K}} = \mathbf{\Phi}_{\mathcal{K}} \mathbf{\Psi}$, $\boldsymbol{P} = \mathbf{\Phi} \mathbf{\Psi}$ and $\mathbf{\Phi} = \mathbf{\Lambda} \mathbf{\Phi}_{\mathcal{K}}$. Note that Assumption 1 suggests $\mathbf{\Phi}_{\mathcal{K}}$ is invertible. Taking these equations collectively yields

$$\boldsymbol{P} = \mathbf{\Phi} \mathbf{\Psi} = \mathbf{\Phi} \mathbf{\Phi}_{\mathcal{K}}^{-1} \boldsymbol{P}_{\mathcal{K}} = \mathbf{\Lambda} \mathbf{\Phi}_{\mathcal{K}} \mathbf{\Phi}_{\mathcal{K}}^{-1} \boldsymbol{P}_{\mathcal{K}} = \mathbf{\Lambda} \boldsymbol{P}_{\mathcal{K}}, \tag{14}$$

which is reminiscent of the anchor word condition in topic modelling [2]. In addition, for each iteration t, we denote the collected samples as $\{s_t(s,a)\}_{(s,a) \in \mathcal{K}}$ and define a matrix $\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} \in \{0,1\}^{K \times |\mathcal{S}|}$ to be

$$\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)}\left((s,a), s'\right) := \begin{cases} 1, & \text{if } s' = s_t\left(s,a\right) \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

for any $(s,a) \in \mathcal{K}$ and $s' \in \mathcal{S}$. Further, we define $\widehat{\boldsymbol{P}}_t = \mathbf{\Lambda} \widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)}$. Then it is obvious to see that $\widehat{\boldsymbol{P}}_t$ has nonnegative entries and unit ℓ_1 norm for each row due to Assumption 1, i.e. $\|\widehat{\boldsymbol{P}}_t\|_1 = 1$.

B Analysis of model-based RL (Proof of Theorem 1)

In this section, we will provide complete proof for Theorem 1. As a matter of fact, our proof strategy here justifies a more general version of Theorem 1 that accounts for model misspecification, as stated below.

Theorem 3. *Suppose that $\delta > 0$ and $\varepsilon \in (0, (1-\gamma)^{-1/2}]$. Assume that there exists a probability transition model $\widetilde{\boldsymbol{P}}$ obeying Definition 1 and Assumption 1 with feature vectors $\{\phi(s,a)\}_{(s,a) \in \mathcal{S} \times \mathcal{A}} \subset \mathbb{R}^K$ and anchor state-action pairs $\mathcal{K}$ such that*

$$\|\widetilde{\boldsymbol{P}} - \boldsymbol{P}\|_1 \leq \xi$$

for some $\xi \geq 0$. Let $\widehat{\pi}$ be the policy returned by Algorithm 1. Assume that

$$N \geq \frac{C \log\left(K / \left((1-\gamma)\,\delta\right)\right)}{\left(1-\gamma\right)^3 \varepsilon^2} \tag{16}$$

for some sufficiently large constant $C > 0$. Then with probability exceeding $1 - \delta$,

$$Q^\star\left(s,a\right) - Q^{\widehat{\pi}}\left(s,a\right) \leq \varepsilon + \frac{4\varepsilon_{\mathsf{opt}}}{1-\gamma} + \frac{22\xi}{(1-\gamma)^2}, \tag{17}$$

for every state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Theorem 3 subsumes Theorem 1 as a special case with $\xi = 0$. The remainder of this section is devoted to proving Theorem 3.

B.1 Proof of Theorem 3

The error $\boldsymbol{Q}^{\widehat{\pi}} - \boldsymbol{Q}^\star$ can be decomposed as

$$\begin{aligned} \boldsymbol{Q}^{\widehat{\pi}} - \boldsymbol{Q}^\star &= \boldsymbol{Q}^{\widehat{\pi}} - \widehat{\boldsymbol{Q}}^{\widehat{\pi}} + \widehat{\boldsymbol{Q}}^{\widehat{\pi}} - \widehat{\boldsymbol{Q}}^\star + \widehat{\boldsymbol{Q}}^\star - \boldsymbol{Q}^\star \\ &\geq \boldsymbol{Q}^{\widehat{\pi}} - \widehat{\boldsymbol{Q}}^{\widehat{\pi}} + \widehat{\boldsymbol{Q}}^{\widehat{\pi}} - \widehat{\boldsymbol{Q}}^\star + \widehat{\boldsymbol{Q}}^{\pi^\star} - \boldsymbol{Q}^\star \\ &\geq -\left(\left\|\boldsymbol{Q}^{\widehat{\pi}} - \widehat{\boldsymbol{Q}}^{\widehat{\pi}}\right\|_\infty + \left\|\widehat{\boldsymbol{Q}}^{\widehat{\pi}} - \widehat{\boldsymbol{Q}}^\star\right\|_\infty + \left\|\widehat{\boldsymbol{Q}}^{\pi^\star} - \boldsymbol{Q}^\star\right\|_\infty\right) \mathbf{1}. \end{aligned} \tag{18}$$

For policy $\widehat{\pi}$ satisfying the condition in Theorem 1, we have $\|\widehat{\boldsymbol{Q}}^{\widehat{\pi}} - \widehat{\boldsymbol{Q}}^\star\|_\infty \leq \varepsilon_{\mathsf{opt}}$. It boils down to control $\|\boldsymbol{Q}^{\widehat{\pi}} - \widehat{\boldsymbol{Q}}^{\widehat{\pi}}\|_\infty$ and $\|\widehat{\boldsymbol{Q}}^{\pi^\star} - \boldsymbol{Q}^\star\|_\infty$.

To begin with, we can use (11) to further decompose $\|\boldsymbol{Q}^{\widehat{\pi}} - \widehat{\boldsymbol{Q}}^{\widehat{\pi}}\|_\infty$ as

$$
\begin{aligned}
\left\|\boldsymbol{Q}^{\widehat{\pi}} - \widehat{\boldsymbol{Q}}^{\widehat{\pi}}\right\|_\infty &= \left\|\left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widehat{\pi}}\right)^{-1} \boldsymbol{r} - \left(\boldsymbol{I} - \gamma \widehat{\boldsymbol{P}}^{\widehat{\pi}}\right)^{-1} \boldsymbol{r}\right\|_\infty \\
&= \left\|\left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widehat{\pi}}\right)^{-1} \left[\left(\boldsymbol{I} - \gamma \widehat{\boldsymbol{P}}^{\widehat{\pi}}\right) - \left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widehat{\pi}}\right)\right] \widehat{\boldsymbol{Q}}^{\widehat{\pi}}\right\|_\infty \\
&= \left\|\gamma \left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widehat{\pi}}\right)^{-1} \left(\boldsymbol{P} - \widehat{\boldsymbol{P}}\right) \widehat{\boldsymbol{V}}^{\widehat{\pi}}\right\|_\infty \\
&\leq \left\|\gamma \left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widehat{\pi}}\right)^{-1} \left(\boldsymbol{P} - \widehat{\boldsymbol{P}}\right) \widehat{\boldsymbol{V}}^\star\right\|_\infty + \left\|\gamma \left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widehat{\pi}}\right)^{-1} \left(\boldsymbol{P} - \widehat{\boldsymbol{P}}\right) \left(\widehat{\boldsymbol{V}}^{\widehat{\pi}} - \widehat{\boldsymbol{V}}^\star\right)\right\|_\infty \\
&\leq \left\|\gamma \left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widehat{\pi}}\right)^{-1} \left|\left(\boldsymbol{P} - \widehat{\boldsymbol{P}}\right) \widehat{\boldsymbol{V}}^\star\right|\right\|_\infty + \frac{2\gamma\varepsilon_{\text{opt}}}{1-\gamma}.
\end{aligned} \tag{19}
$$

Here the last inequality is due to

$$
\begin{aligned}
&\left\|\gamma \left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widehat{\pi}}\right)^{-1} \left(\boldsymbol{P} - \widehat{\boldsymbol{P}}\right) \left(\widehat{\boldsymbol{V}}^{\widehat{\pi}} - \widehat{\boldsymbol{V}}^\star\right)\right\|_\infty \\
&\quad \leq \gamma \left\|\left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widehat{\pi}}\right)^{-1}\right\|_1 \left\|\left(\boldsymbol{P} - \widehat{\boldsymbol{P}}\right) \left(\widehat{\boldsymbol{V}}^{\widehat{\pi}} - \widehat{\boldsymbol{V}}^\star\right)\right\|_\infty \\
&\quad \leq \gamma \left\|\left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widehat{\pi}}\right)^{-1}\right\|_1 \left(\|\boldsymbol{P}\|_1 + \left\|\widehat{\boldsymbol{P}}\right\|_1\right) \left\|\widehat{\boldsymbol{V}}^{\widehat{\pi}} - \widehat{\boldsymbol{V}}^\star\right\|_\infty \\
&\quad \leq \frac{2\gamma\varepsilon_{\text{opt}}}{1-\gamma},
\end{aligned}
$$

where we use the fact that $\|(\boldsymbol{I} - \gamma \boldsymbol{P}^{\widehat{\pi}})^{-1}\|_1 \leq 1/(1-\gamma)$ and $\|\boldsymbol{P}\|_1 = \|\widehat{\boldsymbol{P}}\|_1 = 1$.

Similarly, for the term $\|\widehat{\boldsymbol{Q}}^{\pi^\star} - \boldsymbol{Q}^\star\|_\infty$ in (18), we have

$$
\begin{aligned}
\left\|\widehat{\boldsymbol{Q}}^{\pi^\star} - \boldsymbol{Q}^\star\right\|_\infty &= \left\|\gamma \left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^\star}\right)^{-1} \left(\boldsymbol{P} - \widehat{\boldsymbol{P}}\right) \widehat{\boldsymbol{V}}^{\pi^\star}\right\|_\infty \\
&\leq \left\|\gamma \left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^\star}\right)^{-1} \left|\left(\boldsymbol{P} - \widehat{\boldsymbol{P}}\right) \widehat{\boldsymbol{V}}^{\pi^\star}\right|\right\|_\infty.
\end{aligned} \tag{20}
$$

As can be seen from (19) and (20), it boils down to bound $|(\boldsymbol{P} - \widehat{\boldsymbol{P}})\widehat{\boldsymbol{V}}^\star|$ and $|(\boldsymbol{P} - \widehat{\boldsymbol{P}})\widehat{\boldsymbol{V}}^{\pi^\star}|$. We have the following lemma.

Lemma 1. *With probability exceeding* $1-\delta$, *one has*

$$
\begin{aligned}
\left|\left(\boldsymbol{P} - \widehat{\boldsymbol{P}}\right)_{s,a} \widehat{\boldsymbol{V}}^\star\right| &\leq \frac{10\xi}{1-\gamma} + 4\sqrt{\frac{2\log(4K/\delta)}{N}} + \frac{4\log(8K/((1-\gamma)\delta))}{(1-\gamma)N} \\
&\quad + \sqrt{\frac{4\log(8K/((1-\gamma)\delta))}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^\star\right)},
\end{aligned} \tag{21}
$$

$$
\begin{aligned}
\left|\left(\boldsymbol{P} - \widehat{\boldsymbol{P}}\right)_{s,a} \widehat{\boldsymbol{V}}^{\pi^\star}\right| &\leq \frac{10\xi}{1-\gamma} + 4\sqrt{\frac{2\log(4K/\delta)}{N}} + \frac{4\log(8K/((1-\gamma)\delta))}{(1-\gamma)N} \\
&\quad + \sqrt{\frac{4\log(8K/((1-\gamma)\delta))}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\pi^\star}\right)}.
\end{aligned} \tag{22}
$$

Proof. See Appendix B.2. ◻

Applying (21) to (19) reveals that

$$\begin{aligned}\left\|\boldsymbol{Q}^{\widehat{\pi}}-\widehat{\boldsymbol{Q}}^{\widehat{\pi}}\right\|_{\infty} \leq & \sqrt{\frac{4\log\left(8K/\left((1-\gamma)\,\delta\right)\right)}{N}}\left\|\gamma\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\widehat{\pi}}\right)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\star}\right)}\right\|_{\infty} \\ & +\frac{\gamma}{1-\gamma}\left[4\sqrt{\frac{2\log\left(4K/\delta\right)}{N}}+\frac{4\log\left(8K/\left((1-\gamma)\,\delta\right)\right)}{(1-\gamma)\,N}\right] \\ & +\frac{10\gamma\xi}{(1-\gamma)^{2}}+\frac{2\gamma\varepsilon_{\text{opt}}}{1-\gamma}. \end{aligned} \tag{23}$$

For the first term, one has

$$\begin{aligned}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\star}\right)} &\leq \sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\boldsymbol{V}^{\widehat{\pi}}\right)}+\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\boldsymbol{V}^{\widehat{\pi}}-\widehat{\boldsymbol{V}}^{\widehat{\pi}}\right)}+\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\widehat{\pi}}-\widehat{\boldsymbol{V}}^{\star}\right)} \\ &\leq \sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\boldsymbol{V}^{\widehat{\pi}}\right)}+\left\|\boldsymbol{V}^{\widehat{\pi}}-\widehat{\boldsymbol{V}}^{\widehat{\pi}}\right\|_{\infty}+\varepsilon_{\text{opt}} \\ &\leq \sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\boldsymbol{V}^{\widehat{\pi}}\right)}+\left\|\boldsymbol{Q}^{\widehat{\pi}}-\widehat{\boldsymbol{Q}}^{\widehat{\pi}}\right\|_{\infty}+\varepsilon_{\text{opt}}, \end{aligned}$$

where the first inequality comes from the fact that $\sqrt{\mathsf{Var}(X+Y)} \leq \sqrt{\mathsf{Var}(X)}+\sqrt{\mathsf{Var}(Y)}$ for any random variables X and Y. It follows that

$$\begin{aligned}&\left\|\gamma\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\widehat{\pi}}\right)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\star}\right)}\right\|_{\infty} \\ &\quad\leq \left\|\gamma\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\widehat{\pi}}\right)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\boldsymbol{V}^{\widehat{\pi}}\right)}\right\|_{\infty}+\frac{\gamma}{1-\gamma}\left(\left\|\boldsymbol{Q}^{\widehat{\pi}}-\widehat{\boldsymbol{Q}}^{\widehat{\pi}}\right\|_{\infty}+\varepsilon_{\text{opt}}\right) \\ &\quad\leq \gamma\sqrt{\frac{2}{(1-\gamma)^{3}}}+\frac{\gamma}{1-\gamma}\left(\left\|\boldsymbol{Q}^{\widehat{\pi}}-\widehat{\boldsymbol{Q}}^{\widehat{\pi}}\right\|_{\infty}+\varepsilon_{\text{opt}}\right), \end{aligned} \tag{24}$$

where the second inequality utilizes [3, Lemma 7].

Plugging (24) into (23) yields

$$\begin{aligned}\left\|\boldsymbol{Q}^{\widehat{\pi}}-\widehat{\boldsymbol{Q}}^{\widehat{\pi}}\right\|_{\infty} \leq & \sqrt{\frac{4\log\left(8K/\left((1-\gamma)\,\delta\right)\right)}{N}}\left[\gamma\sqrt{\frac{2}{(1-\gamma)^{3}}}+\frac{\gamma}{1-\gamma}\left(\left\|\boldsymbol{Q}^{\widehat{\pi}}-\widehat{\boldsymbol{Q}}^{\widehat{\pi}}\right\|_{\infty}+\varepsilon_{\text{opt}}\right)\right] \\ & +\frac{\gamma}{1-\gamma}\left[4\sqrt{\frac{2\log\left(4K/\delta\right)}{N}}+\frac{4\log\left(8K/\left((1-\gamma)\,\delta\right)\right)}{(1-\gamma)\,N}\right]+\frac{10\gamma\xi}{(1-\gamma)^{2}}+\frac{2\gamma\varepsilon_{\text{opt}}}{1-\gamma}. \end{aligned}$$

Then we can rearrange terms to obtain

$$\left\|\boldsymbol{Q}^{\widehat{\pi}}-\widehat{\boldsymbol{Q}}^{\widehat{\pi}}\right\|_{\infty} \leq 10\gamma\sqrt{\frac{\log\left(8K/\left((1-\gamma)\,\delta\right)\right)}{N\,(1-\gamma)^{3}}}+\frac{11\gamma\xi}{(1-\gamma)^{2}}+\frac{3\gamma\varepsilon_{\text{opt}}}{1-\gamma} \tag{25}$$

as long as $N \geq \widetilde{C}\log(8K/((1-\gamma)\delta))/(1-\gamma)^{2}$ for some sufficiently large constant $\widetilde{C}>0$.

In a similar vein, we can use (20) and (22) to obtain that

$$\left\|\widehat{\boldsymbol{Q}}^{\pi^{\star}}-\boldsymbol{Q}^{\star}\right\|_{\infty} \leq 10\gamma\sqrt{\frac{\log\left(8K/\left((1-\gamma)\,\delta\right)\right)}{N\,(1-\gamma)^{3}}}+\frac{11\gamma\xi}{(1-\gamma)^{2}}. \tag{26}$$

Finally, we can substitute (25) and (26) into (18) to achieve

$$\boldsymbol{Q}^{\widehat{\pi}}-\boldsymbol{Q}^{\star} \geq -\left(20\gamma\sqrt{\frac{\log\left(8K/\left((1-\gamma)\,\delta\right)\right)}{N\,(1-\gamma)^{3}}}+\frac{22\gamma\xi}{(1-\gamma)^{2}}+\frac{3\gamma\varepsilon_{\text{opt}}}{1-\gamma}+\varepsilon_{\text{opt}}\right)\boldsymbol{1}.$$

This result implies that

$$\boldsymbol{Q}^{\widehat{\pi}} \geq \boldsymbol{Q}^{\star}-\left(\varepsilon+\frac{22\xi}{(1-\gamma)^{2}}+\frac{4\varepsilon_{\text{opt}}}{1-\gamma}\right)\boldsymbol{1},$$

as long as

$$N \geq \frac{C\log\left(8K/\left((1-\gamma)\,\delta\right)\right)}{(1-\gamma)^{3}\,\varepsilon^{2}},$$

for some sufficiently large constant $C>0$.

B.2 Proof of Lemma 1

To prove this theorem, we invoke the idea of s-absorbing MDP proposed by [1]. For a state $s \in \mathcal{S}$ and a scalar u, we define a new MDP $M_{s,u}$ to be identical to M on all the other states except s; on state s, $M_{s,u}$ is absorbing such that $P_{M_{s,u}}(s|s,a) = 1$ and $r_{M_{s,u}}(s,a) = (1-\gamma)u$ for all $a \in \mathcal{A}$. More formally, we define $P_{M_{u,s}}$ and $r_{M_{u,s}}$ as

$$P_{M_{s,u}}(s|s,a) = 1, \quad r_{M_{s,u}}(s,a) = (1-\gamma)u, \qquad \text{for all } a \in \mathcal{A},$$
$$P_{M_{s,u}}(\cdot|s',a') = P(\cdot|s',a'), \quad r_{M_{s,u}}(s,a) = r(s,a), \qquad \text{for all } s' \neq s \text{ and } a' \in \mathcal{A}.$$

To streamline notations, we will use $\boldsymbol{V}^{\pi}_{s,u} \in \mathbb{R}^{|\mathcal{S}|}$ and $\boldsymbol{V}^{\star}_{s,u} \in \mathbb{R}^{|\mathcal{S}|}$ to denote the value function of $M_{s,u}$ under policy π and the optimal value function of $M_{s,u}$ respectively. Furthermore, we denote by $\widehat{M}_{s,u}$ the MDP whose probability transition kernel is identical to $\widehat{P}$ at all states except that state s is absorbing. Similar as before, we use $\widehat{\boldsymbol{V}}^{\star}_{s,u} \in \mathbb{R}^{|\mathcal{S}|}$ to denote the optimal value function under $\widehat{M}_{s,u}$. The construction of this collection of auxiliary MDPs will facilitate our analysis by decoupling the statistical dependency between $\widehat{\boldsymbol{P}}$ and $\widehat{\pi}^{\star}$.

To begin with, we can decompose the quantity of interest as

$$\begin{aligned}
\left|\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)_{s,a}\widehat{\boldsymbol{V}}^{\star}\right| &= \left|\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)_{s,a}\left(\widehat{\boldsymbol{V}}^{\star}-\widehat{\boldsymbol{V}}^{\star}_{s,u}+\widehat{\boldsymbol{V}}^{\star}_{s,u}\right)\right| \\
&\leq \left|\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)_{s,a}\widehat{\boldsymbol{V}}^{\star}_{s,u}\right| + \left|\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)_{s,a}\left(\widehat{\boldsymbol{V}}^{\star}-\widehat{\boldsymbol{V}}^{\star}_{s,u}\right)\right| \\
&\overset{\text{(i)}}{\leq} \left|\left(\boldsymbol{P}-\widetilde{\boldsymbol{P}}\right)_{s,a}\widehat{\boldsymbol{V}}^{\star}_{s,u}\right| + \left|\boldsymbol{\lambda}(s,a)\left(\widetilde{\boldsymbol{P}}_{\mathcal{K}}-\boldsymbol{P}_{\mathcal{K}}\right)\widehat{\boldsymbol{V}}^{\star}_{s,u}\right| \\
&\quad + \left|\boldsymbol{\lambda}(s,a)\left(\boldsymbol{P}_{\mathcal{K}}-\widehat{\boldsymbol{P}}_{\mathcal{K}}\right)\widehat{\boldsymbol{V}}^{\star}_{s,u}\right| + \left(\|\boldsymbol{P}_{s,a}\|_1 + \left\|\widehat{\boldsymbol{P}}_{s,a}\right\|_1\right)\left\|\widehat{\boldsymbol{V}}^{\star}-\widehat{\boldsymbol{V}}^{\star}_{s,u}\right\|_{\infty} \\
&\leq \left\|\left(\boldsymbol{P}-\widetilde{\boldsymbol{P}}\right)_{s,a}\right\|_1\left\|\widehat{\boldsymbol{V}}^{\star}_{s,u}\right\|_{\infty} + \|\boldsymbol{\lambda}(s,a)\|_1\cdot\left\|\left(\widetilde{\boldsymbol{P}}_{\mathcal{K}}-\boldsymbol{P}_{\mathcal{K}}\right)\widehat{\boldsymbol{V}}^{\star}_{s,u}\right\|_{\infty} \\
&\quad + \|\boldsymbol{\lambda}(s,a)\|_1\cdot\left\|\left(\boldsymbol{P}_{\mathcal{K}}-\widehat{\boldsymbol{P}}_{\mathcal{K}}\right)\widehat{\boldsymbol{V}}^{\star}_{s,u}\right\|_{\infty} + 2\left\|\widehat{\boldsymbol{V}}^{\star}-\widehat{\boldsymbol{V}}^{\star}_{s,u}\right\|_{\infty} \\
&\overset{\text{(ii)}}{\leq} \frac{2\xi}{1-\gamma} + \max_{(s,a)\in\mathcal{K}}\left|\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)_{s,a}\widehat{\boldsymbol{V}}^{\star}_{s,u}\right| + 2\left\|\widehat{\boldsymbol{V}}^{\star}-\widehat{\boldsymbol{V}}^{\star}_{s,u}\right\|_{\infty},
\end{aligned} \tag{27}$$

where (i) makes use of $\widetilde{\boldsymbol{P}}_{s,a} = \boldsymbol{\lambda}(s,a)\widetilde{\boldsymbol{P}}_{\mathcal{K}}$ and $\widehat{\boldsymbol{P}}_{s,a} = \boldsymbol{\lambda}(s,a)\widehat{\boldsymbol{P}}_{\mathcal{K}}$; (ii) depends on $\|\boldsymbol{P}-\widetilde{\boldsymbol{P}}\|_1 \leq \xi$, $\|\boldsymbol{\lambda}(s,a)\|_1 = 1$ and $\|\widehat{\boldsymbol{V}}^{\star}_{s,u}\|_{\infty} \leq (1-\gamma)^{-1}$. For each state s, the value of u will be selected from a set $\mathcal{U}_s$. The choice of $\mathcal{U}_s$ will be specified later. Then for some fixed u in $\mathcal{U}_s$ and fixed state-action pair $(s,a) \in \mathcal{K}$, due to the independence between $\widehat{\boldsymbol{P}}_{s,a}$ and $\widehat{\boldsymbol{V}}^{\star}_{s,u}$, we can apply Bernstein's inequality (cf. [5, Theorem 2.8.4]) conditional on $\widehat{\boldsymbol{V}}^{\star}_{s,u}$ to reveal that with probability greater than $1-\delta/2$,

$$\left|\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)_{s,a}\widehat{\boldsymbol{V}}^{\star}_{s,u}\right| \leq \sqrt{\frac{2\log(4/\delta)}{N}\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\star}_{s,u}\right)} + \frac{2\log(4/\delta)}{3(1-\gamma)N}. \tag{28}$$

Invoking the union bound over all the K state-action pairs of $\mathcal{K}$ and all the possible values of u in $\mathcal{U}_s$ demonstrate that with probability greater than $1-\delta/2$,

$$\left|\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)_{s,a}\widehat{\boldsymbol{V}}^{\star}_{s,u}\right| \leq \sqrt{\frac{2\log(4K|\mathcal{U}_s|/\delta)}{N}\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\star}_{s,u}\right)} + \frac{2\log(4K|\mathcal{U}_s|/\delta)}{3(1-\gamma)N}, \tag{29}$$

holds for all state-action pair $(s,a) \in \mathcal{K}$ and all $u \in \mathcal{U}_s$. Here, $\mathsf{Var}_{\boldsymbol{P}_{s,a}}(\cdot)$ is defined in (13). Then we observe that

$$\begin{aligned}
\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\star}_{s,u}\right)} &\leq \sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\star}-\widehat{\boldsymbol{V}}^{\star}_{s,u}\right)} + \sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\star}\right)} \\
&\leq \left\|\widehat{\boldsymbol{V}}^{\star}-\widehat{\boldsymbol{V}}^{\star}_{s,u}\right\|_{\infty} + \sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\star}\right)} \\
&\leq \left|\widehat{\boldsymbol{V}}^{\star}(s)-u\right| + \sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\star}\right)},
\end{aligned} \tag{30}$$

where (i) is due to $\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}(\boldsymbol{V}_1+\boldsymbol{V}_2)} \leq \sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}(\boldsymbol{V}_1)} + \sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}(\boldsymbol{V}_2)}$ and (ii) holds since

$$\left\|\widehat{\boldsymbol{V}}^{\star} - \widehat{\boldsymbol{V}}^{\star}_{s,u}\right\|_{\infty} = \left\|\widehat{\boldsymbol{V}}^{\star}_{s,\widehat{\boldsymbol{V}}^{\star}(s)} - \widehat{\boldsymbol{V}}^{\star}_{s,u}\right\|_{\infty} \leq \left|\widehat{\boldsymbol{V}}^{\star}(s) - u\right|, \tag{31}$$

whose proof can be found in [1, Lemma 8 and 9].

By substituting (29), (30) and (31) into (27), we arrive at

$$\begin{aligned}\left|\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)_{s,a}\widehat{\boldsymbol{V}}^{\star}\right| \leq \frac{2\xi}{1-\gamma} + \left|\widehat{\boldsymbol{V}}^{\star}(s)-u\right|\left(2+\sqrt{\frac{2\log\left(4K\left|\mathcal{U}_s\right|/\delta\right)}{N}}\right) \\ + \sqrt{\frac{2\log\left(4K\left|\mathcal{U}_s\right|/\delta\right)}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\star}\right)} + \frac{2\log\left(4K\left|\mathcal{U}_s\right|/\delta\right)}{3(1-\gamma)N}.\end{aligned} \tag{32}$$

Then it boils down to determining $\mathcal{U}_s$. The coarse bounds of $\widehat{\boldsymbol{Q}}^{\pi^\star}$ and $\widehat{\boldsymbol{Q}}^{\star}$ in the following lemma provide a guidance on the choice of $\mathcal{U}_s$.

Lemma 2. *For $\delta \in (0,1)$, with probability exceeding $1-\delta/2$ one has*

$$\left\|\boldsymbol{Q}^{\star} - \widehat{\boldsymbol{Q}}^{\pi^\star}\right\|_{\infty} \leq \frac{\gamma}{1-\gamma}\sqrt{\frac{\log\left(4K/\delta\right)}{2N(1-\gamma)^2}} + \frac{2\gamma\xi}{(1-\gamma)^2}, \tag{33}$$

$$\left\|\boldsymbol{Q}^{\star} - \widehat{\boldsymbol{Q}}^{\star}\right\|_{\infty} \leq \frac{\gamma}{1-\gamma}\sqrt{\frac{\log\left(4K/\delta\right)}{2N(1-\gamma)^2}} + \frac{2\gamma\xi}{(1-\gamma)^2}. \tag{34}$$

Proof. See Appendix B.3. □

This inspires us to choose $\mathcal{U}_s$ to be the set consisting of equidistant points in $[\boldsymbol{V}^\star(s) - R(\delta), \boldsymbol{V}^\star(s) + R(\delta)]$ with $|U_s| = \lceil 1/(1-\gamma)^2 \rceil$ and

$$R(\delta) \coloneqq \frac{\gamma}{1-\gamma}\sqrt{\frac{\log\left(4K/\delta\right)}{2N(1-\gamma)^2}} + \frac{2\gamma\xi}{(1-\gamma)^2}.$$

Since $\|\boldsymbol{V}^\star - \widehat{\boldsymbol{V}}^\star\|_\infty \leq \|\boldsymbol{Q}^\star - \widehat{\boldsymbol{Q}}^\star\|_\infty$, Lemma 2 implies that $\widehat{\boldsymbol{V}}^\star(s) \in [\boldsymbol{V}^\star(s) - R(\delta), \boldsymbol{V}^\star(s) + R(\delta)]$ with probability over $1-\delta/2$. Hence, we have

$$\min_{u\in\mathcal{U}_s}\left|\widehat{\boldsymbol{V}}^{\star}(s) - u\right| \leq \frac{2R(\delta)}{|U_s|+1} \leq 2\gamma\sqrt{\frac{2\log\left(4K/\delta\right)}{N}} + 4\gamma\xi. \tag{35}$$

Consequently, with probability exceeding $1-\delta$, one has

$$\begin{aligned}\left|\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)_{s,a}\widehat{\boldsymbol{V}}^{\star}\right| &\overset{\text{(i)}}{\leq} \frac{2\xi}{1-\gamma} + \min_{u\in\mathcal{U}_s}\left|\widehat{\boldsymbol{V}}^{\star}(s)-u\right|\left(2+\sqrt{\frac{2\log\left(4K\left|\mathcal{U}_s\right|/\delta\right)}{N}}\right) \\ &\quad + \sqrt{\frac{2\log\left(4K\left|\mathcal{U}_s\right|/\delta\right)}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\star}\right)} + \frac{2\log\left(4K\left|\mathcal{U}_s\right|/\delta\right)}{3(1-\gamma)N} \\ &\overset{\text{(ii)}}{\leq} \frac{2\xi}{1-\gamma} + \left(2\gamma\sqrt{\frac{2\log\left(4K/\delta\right)}{N}} + 4\gamma\xi\right)\left(2+\sqrt{\frac{4\log\left(8K/\left((1-\gamma)\delta\right)\right)}{N}}\right) \\ &\quad + \sqrt{\frac{4\log\left(8K/\left((1-\gamma)\delta\right)\right)}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\star}\right)} + \frac{2\log\left(8K/\left((1-\gamma)\delta\right)\right)}{3(1-\gamma)N} \\ &\leq \frac{10\xi}{1-\gamma} + 4\sqrt{\frac{2\log\left(4K/\delta\right)}{N}} + \frac{4\log\left(8K/\left((1-\gamma)\delta\right)\right)}{(1-\gamma)N} \\ &\quad + \sqrt{\frac{4\log\left(8K/\left((1-\gamma)\delta\right)\right)}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{s,a}}\left(\widehat{\boldsymbol{V}}^{\star}\right)},\end{aligned}$$

where (i) follows from (32) and (ii) utilizes (35). This finishes the proof for the first inequality. The second inequality can be proved in a similar way and is omitted here for brevity.

B.3 Proof of Lemma 2

To begin with, one has

$$\begin{aligned}\left\|\left(\widehat{\boldsymbol{P}}-\boldsymbol{P}\right)\boldsymbol{V}^{\star}\right\|_{\infty} &\leq \left\|\boldsymbol{\Lambda}\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}-\boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}^{\star}\right\|_{\infty}+\left\|\boldsymbol{\Lambda}\left(\boldsymbol{P}_{\mathcal{K}}-\widetilde{\boldsymbol{P}}_{\mathcal{K}}\right)\boldsymbol{V}^{\star}\right\|_{\infty}+\left\|\left(\widetilde{\boldsymbol{P}}-\boldsymbol{P}\right)\boldsymbol{V}^{\star}\right\|_{\infty}\\ &\leq \left\|\boldsymbol{\Lambda}\right\|_{1}\left\|\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}-\boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}^{\star}\right\|_{\infty}+\left\|\boldsymbol{\Lambda}\right\|_{1}\left\|\left(\boldsymbol{P}_{\mathcal{K}}-\widetilde{\boldsymbol{P}}_{\mathcal{K}}\right)\boldsymbol{V}^{\star}\right\|_{\infty}+\left\|\widetilde{\boldsymbol{P}}-\boldsymbol{P}\right\|_{1}\left\|\boldsymbol{V}^{\star}\right\|_{\infty}\\ &\leq \left\|\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}-\boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}^{\star}\right\|_{\infty}+\frac{2\xi}{1-\gamma}, \end{aligned} \tag{36}$$

where the first line uses $\widehat{\boldsymbol{P}}=\boldsymbol{\Lambda}\widehat{\boldsymbol{P}}_{\mathcal{K}}$ and $\widetilde{\boldsymbol{P}}=\boldsymbol{\Lambda}\widetilde{\boldsymbol{P}}_{\mathcal{K}}$; the last inequality comes from the facts that $\|\widetilde{\boldsymbol{P}}-\boldsymbol{P}\|_1\leq\xi$, $\|\boldsymbol{\Lambda}\|_1=1$ and $\|\boldsymbol{V}^{\star}\|_{\infty}\leq(1-\gamma)^{-1}$. Then we turn to bound $\|(\widehat{\boldsymbol{P}}_{\mathcal{K}}-\boldsymbol{P}_{\mathcal{K}})\boldsymbol{V}^{\star}\|_{\infty}$. In view of (4), Hoeffding's inequality (cf. [5, Theorem 2.2.6]) implies that for $(s,a)\in\mathcal{K}$,

$$\mathbb{P}\left(\left|\left(\widehat{\boldsymbol{P}}-\boldsymbol{P}\right)_{s,a}\boldsymbol{V}^{\star}\right|\geq t\right)\leq 2\exp\left(-\frac{2t^2}{\left\|\boldsymbol{V}^{\star}\right\|_{\infty}^{2}/N}\right).$$

Hence by the standard union bound argument we have

$$\left\|\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}-\boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}^{\star}\right\|_{\infty}\leq\sqrt{\frac{\left\|\boldsymbol{V}^{\star}\right\|_{\infty}^{2}\log\left(4K/\delta\right)}{2N}}\leq\sqrt{\frac{\log\left(4K/\delta\right)}{2N\left(1-\gamma\right)^{2}}}, \tag{37}$$

with probability over $1-\delta/2$.

1. Now we are ready to bound $\boldsymbol{Q}^{\pi^{\star}}-\widehat{\boldsymbol{Q}}^{\pi^{\star}}$. One has

$$\begin{aligned}\boldsymbol{Q}^{\pi^{\star}}-\widehat{\boldsymbol{Q}}^{\pi^{\star}} &= \left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\pi^{\star}}\right)^{-1}\boldsymbol{r}-\left(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}^{\pi^{\star}}\right)^{-1}\boldsymbol{r}\\ &= \left(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}^{\pi^{\star}}\right)^{-1}\left(\left(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}^{\pi^{\star}}\right)-\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\pi^{\star}}\right)\right)\boldsymbol{Q}^{\pi^{\star}}\\ &= \gamma\left(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}^{\pi^{\star}}\right)^{-1}\left(\boldsymbol{P}^{\pi^{\star}}-\widehat{\boldsymbol{P}}^{\pi^{\star}}\right)\boldsymbol{Q}^{\pi^{\star}}\\ &= \gamma\left(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}^{\pi^{\star}}\right)^{-1}\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)\boldsymbol{V}^{\pi^{\star}},\end{aligned}$$

   where the first equality makes use of (11). Then we take (36) and (37) collectively to achieve

$$\begin{aligned}\left\|\gamma\left(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}^{\pi^{\star}}\right)^{-1}\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)\boldsymbol{V}^{\star}\right\|_{\infty} &\leq \gamma\sum_{i=0}^{\infty}\left\|\gamma^{i}\left(\widehat{\boldsymbol{P}}^{\pi^{\star}}\right)^{i}\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)\boldsymbol{V}^{\star}\right\|_{\infty}\\ &\leq \gamma\sum_{i=0}^{\infty}\gamma^{i}\left\|\left(\widehat{\boldsymbol{P}}^{\pi^{\star}}\right)^{i}\right\|_{1}\left\|\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)\boldsymbol{V}^{\star}\right\|_{\infty}\\ &\leq \frac{\gamma}{1-\gamma}\sqrt{\frac{\log\left(4K/\delta\right)}{2N\left(1-\gamma\right)^{2}}}+\frac{2\gamma\xi}{\left(1-\gamma\right)^{2}},\end{aligned}$$

   where the last line comes from the fact that for all $i\geq1$, $(\widehat{\boldsymbol{P}}^{\pi^{\star}})^{i}$ is a probability transition matrix so that $\|(\widehat{\boldsymbol{P}}^{\pi^{\star}})^{i}\|_1=1$. This justifies the first inequality (33).

2. In terms of the second one, [1, Section A.4] implies that

$$\left\|\boldsymbol{Q}^{\star}-\widehat{\boldsymbol{Q}}^{\star}\right\|_{\infty}\leq\frac{\gamma}{1-\gamma}\left\|\left(\boldsymbol{P}-\widehat{\boldsymbol{P}}\right)\boldsymbol{V}^{\star}\right\|_{\infty}.$$

   Substitution of (36) and (37) into the above inequality yields

$$\left\|\boldsymbol{Q}^{\star}-\widehat{\boldsymbol{Q}}^{\star}\right\|_{\infty}\leq\frac{\gamma}{1-\gamma}\sqrt{\frac{\log\left(4K/\delta\right)}{2N\left(1-\gamma\right)^{2}}}+\frac{2\gamma\xi}{\left(1-\gamma\right)^{2}}.$$

C Analysis of Q-learning (Proof of Theorem 2)

In this section, we will provide complete proof for Theorem 2. We actually prove a more general version of Theorem 2 that takes model misspecification into consideration, as stated below.

Theorem 4. *Consider any $\delta \in (0,1)$ and $\varepsilon \in (0,1]$. Suppose that there exists a probability transition model $\widetilde{\boldsymbol{P}}$ obeying Definition 1 and Assumption 1 with feature vectors $\{\phi(s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}} \subset \mathbb{R}^K$ and anchor state-action pairs $\mathcal{K}$ such that*

$$\|\widetilde{\boldsymbol{P}} - \boldsymbol{P}\|_1 \le \xi$$

for some $\xi \ge 0$. Assume that the initialization obeys $0 \le Q_0(s,a) \le \frac{1}{1-\gamma}$ for any $(s,a) \in \mathcal{S}\times\mathcal{A}$ and for any $0 \le t \le T$, the learning rates satisfy

$$\frac{1}{1+\frac{c_1(1-\gamma)T}{\log^2 T}} \le \eta_t \le \frac{1}{1+\frac{c_2(1-\gamma)t}{\log^2 T}}, \tag{38}$$

for some sufficiently small universal constants $c_1 \ge c_2 > 0$. Suppose that the total number of iterations T exceeds

$$T \ge \frac{C_3 \log(KT/\delta)\log^4 T}{(1-\gamma)^4\varepsilon^2}, \tag{39}$$

for some sufficiently large universal constant $C_3 > 0$. If there exists a linear probability transition model $\widetilde{\boldsymbol{P}}$ satisfying Assumption 1 with feature vectors $\{\phi(s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ such that $\|\widetilde{\boldsymbol{P}} - \boldsymbol{P}\|_1 \le \xi$, then with probability exceeding $1-\delta$, the output Q_T of Algorithm 2 satisfies

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} |Q_T(s,a) - Q^\star(s,a)| \le \varepsilon + \frac{6\gamma\xi}{(1-\gamma)^2}, \tag{40}$$

for some constant $C_4 > 0$. In addition, let π_T (resp. V_T) to be the policy (resp. value function) induced by Q_T, then one has

$$\max_{s\in\mathcal{S}} |V^{\pi_T}(s) - V^\star(s)| \le \frac{2\gamma}{1-\gamma}\left(\varepsilon + \frac{6\gamma\xi}{(1-\gamma)^2}\right). \tag{41}$$

Theorem 4 subsumes Theorem 2 as a special case with $\xi = 0$. The remainder of this section is devoted to proving Theorem 4.

C.1 Proof of Theorem 4

First we show that (41) can be easily obtained from (40). Since [49] gives rise to

$$\|V^{\pi_T} - V^\star\|_\infty \le \frac{2\gamma\|V_T - V^\star\|_\infty}{1-\gamma},$$

we have

$$\|V^{\pi_T} - V^\star\|_\infty \le \frac{2\gamma\|Q_T - Q^\star\|_\infty}{1-\gamma},$$

due to $\|V_T - V^\star\|_\infty \le \|Q_T - Q^\star\|_\infty$. Then (41) follows directly from (40).

Therefore, we are left to justify (40). To start with, we consider the update rule

$$\boldsymbol{Q}_t = (1-\eta_t)\boldsymbol{Q}_{t-1} + \eta_t\left(\boldsymbol{r} + \gamma\widehat{\boldsymbol{P}}_t\boldsymbol{V}_{t-1}\right).$$

By defining the error term $\boldsymbol{\Delta}_t \coloneqq \boldsymbol{Q}_t - \boldsymbol{Q}^\star$, we can decompose $\boldsymbol{\Delta}_t$ into

$$\begin{aligned}
\boldsymbol{\Delta}_t &= (1-\eta_t)\boldsymbol{Q}_{t-1} + \eta_t\left(\boldsymbol{r} + \gamma\widehat{\boldsymbol{P}}_t\boldsymbol{V}_{t-1}\right) - \boldsymbol{Q}^\star \\
&= (1-\eta_t)(\boldsymbol{Q}_{t-1} - \boldsymbol{Q}^\star) + \eta_t\left(\boldsymbol{r} + \gamma\widehat{\boldsymbol{P}}_t\boldsymbol{V}_{t-1} - \boldsymbol{Q}^\star\right) \\
&= (1-\eta_t)(\boldsymbol{Q}_{t-1} - \boldsymbol{Q}^\star) + \gamma\eta_t\left(\widehat{\boldsymbol{P}}_t\boldsymbol{V}_{t-1} - \boldsymbol{P}\boldsymbol{V}^\star\right) \\
&= (1-\eta_t)\boldsymbol{\Delta}_{t-1} + \gamma\eta_t\boldsymbol{\Lambda}\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} - \boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}_{t-1} + \gamma\eta_t\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}}(\boldsymbol{V}_{t-1} - \boldsymbol{V}^\star) \\
&\quad + \gamma\eta_t(\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}} - \boldsymbol{P})\boldsymbol{V}^\star.
\end{aligned} \tag{42}$$

Here in the penultimate equality, we make use of $\boldsymbol{Q}^\star = \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{V}^\star$; and the last equality comes from $\widehat{\boldsymbol{P}}_t = \boldsymbol{\Lambda}\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)}$ which is defined in (15). It is straightforward to check that $\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}}$ is also a probability transition matrix. We denote by $\overline{\boldsymbol{P}} = \boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}}$ hereafter. The third term in the decomposition above can be upper and lower bounded by

$$\overline{\boldsymbol{P}}\left(\boldsymbol{V}_{t-1} - \boldsymbol{V}^\star\right) = \overline{\boldsymbol{P}}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} - \overline{\boldsymbol{P}}^{\pi^\star}\boldsymbol{Q}^\star \le \overline{\boldsymbol{P}}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} - \overline{\boldsymbol{P}}^{\pi_{t-1}}\boldsymbol{Q}^\star = \overline{\boldsymbol{P}}^{\pi_{t-1}}\boldsymbol{\Delta}_{t-1},$$

and

$$\overline{\boldsymbol{P}}\left(\boldsymbol{V}_{t-1} - \boldsymbol{V}^\star\right) = \overline{\boldsymbol{P}}^{\pi_{t-1}}\boldsymbol{Q}_{t-1} - \overline{\boldsymbol{P}}^{\pi^\star}\boldsymbol{Q}^\star \ge \overline{\boldsymbol{P}}^{\pi^\star}\boldsymbol{Q}_{t-1} - \overline{\boldsymbol{P}}^{\pi^\star}\boldsymbol{Q}^\star = \overline{\boldsymbol{P}}^{\pi^\star}\boldsymbol{\Delta}_{t-1}.$$

Plugging these bounds into (42) yields

$$\begin{aligned}
\boldsymbol{\Delta}_t &\le (1-\eta_t)\,\boldsymbol{\Delta}_{t-1} + \gamma\eta_t\boldsymbol{\Lambda}\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} - \boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}_{t-1} + \gamma\eta_t\overline{\boldsymbol{P}}^{\pi_{t-1}}\boldsymbol{\Delta}_{t-1} + \gamma\eta_t\left(\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}} - \boldsymbol{P}\right)\boldsymbol{V}^\star,\\
\boldsymbol{\Delta}_t &\ge (1-\eta_t)\,\boldsymbol{\Delta}_{t-1} + \gamma\eta_t\boldsymbol{\Lambda}\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} - \boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}_{t-1} + \gamma\eta_t\overline{\boldsymbol{P}}^{\pi^\star}\boldsymbol{\Delta}_{t-1} + \gamma\eta_t\left(\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}} - \boldsymbol{P}\right)\boldsymbol{V}^\star.
\end{aligned}$$

Repeatedly invoking these two recursive relations leads to

$$\boldsymbol{\Delta}_t \le \eta_0^{(t)}\boldsymbol{\Delta}_0 + \sum_{i=1}^{t}\eta_i^{(t)}\gamma\left(\overline{\boldsymbol{P}}^{\pi_{t-1}}\boldsymbol{\Delta}_{t-1} + \boldsymbol{\Lambda}\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} - \boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}_{t-1} + \left(\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}} - \boldsymbol{P}\right)\boldsymbol{V}^\star\right), \tag{43}$$

$$\boldsymbol{\Delta}_t \ge \eta_0^{(t)}\boldsymbol{\Delta}_0 + \sum_{i=1}^{t}\eta_i^{(t)}\gamma\left(\overline{\boldsymbol{P}}^{\pi^\star}\boldsymbol{\Delta}_{t-1} + \boldsymbol{\Lambda}\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} - \boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}_{t-1} + \left(\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}} - \boldsymbol{P}\right)\boldsymbol{V}^\star\right), \tag{44}$$

where

$$\eta_i^{(t)} := \begin{cases} \prod_{j=1}^{t}(1-\eta_j)\,, & \text{if } i = 0,\\ \eta_i\prod_{j=i+1}^{t}(1-\eta_j)\,, & \text{if } 0 < i < t,\\ \eta_t, & \text{if } i = t.\end{cases}$$

Here we adopt the same notations as [4].

To begin with, we consider the upper bound (43). It can be further decomposed as

$$\begin{aligned}
\boldsymbol{\Delta}_t \le\ & \underbrace{\eta_0^{(t)}\boldsymbol{\Delta}_0 + \sum_{i=1}^{(1-\alpha)t}\eta_i^{(t)}\gamma\left(\overline{\boldsymbol{P}}^{\pi_{t-1}}\boldsymbol{\Delta}_{t-1} + \boldsymbol{\Lambda}\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} - \boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}_{t-1}\right)}_{=:\boldsymbol{\theta}_t}\\
& + \underbrace{\sum_{i=(1-\alpha)t+1}^{t}\eta_i^{(t)}\gamma\boldsymbol{\Lambda}\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} - \boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}_{i-1}}_{=:\boldsymbol{\nu}_t}\\
& + \underbrace{\sum_{i=1}^{t}\eta_i^{(t)}\gamma\left(\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}} - \boldsymbol{P}\right)\boldsymbol{V}^\star}_{=:\boldsymbol{\omega}_t} + \sum_{i=(1-\alpha)t+1}^{t}\eta_i^{(t)}\gamma\overline{\boldsymbol{P}}^{\pi_{t-1}}\boldsymbol{\Delta}_{i-1},
\end{aligned} \tag{45}$$

where we define $\alpha := C_4(1-\gamma)/\log T$ for some constant $C_4 > 0$. Next, we turn to bound $\boldsymbol{\theta}_t$ and $\boldsymbol{\nu}_t$ respectively for any t satisfying $\frac{T}{c_2\log\frac{1}{1-\gamma}} \le t \le T$ with stepsize choice (8).

Bounding $\boldsymbol{\omega}_t$. It is straightforward to bound

$$\begin{aligned}
\|\boldsymbol{\omega}_t\|_\infty &\overset{\text{(i)}}{=} \left\|\gamma\left(\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}} - \boldsymbol{P}\right)\boldsymbol{V}^\star\right\|_\infty\\
&\overset{\text{(ii)}}{\le} \gamma\left(\|\boldsymbol{\Lambda}\|_1\left\|\left(\boldsymbol{P}_{\mathcal{K}} - \widetilde{\boldsymbol{P}}_{\mathcal{K}}\right)\boldsymbol{V}^\star\right\|_\infty + \left\|\left(\widetilde{\boldsymbol{P}} - \boldsymbol{P}\right)\boldsymbol{V}^\star\right\|_\infty\right)\\
&\overset{\text{(iii)}}{\le} \frac{2\gamma\xi}{1-\gamma},
\end{aligned}$$

where the first equality comes from the fact that $\sum_{i=1}^{t}\eta_i^{(t)} = 1$ [4, Equation (40)]; the second inequality utilizes $\widetilde{\boldsymbol{P}} = \boldsymbol{\Lambda}\widetilde{\boldsymbol{P}}_{\mathcal{K}}$; the last line uses the facts that $\|\boldsymbol{\Lambda}\|_1 = 1$, $\|\boldsymbol{V}^\star\|_\infty \le (1-\gamma)^{-1}$ and $\|\widetilde{\boldsymbol{P}}_{\mathcal{K}} - \boldsymbol{P}_{\mathcal{K}}\|_1 \le \|\widetilde{\boldsymbol{P}} - \boldsymbol{P}\|_1 \le \xi$.

Bounding $\boldsymbol{\theta}_t$. By similar derivation as Step 1 in [4, Appendix A.2], we have

$$\begin{aligned}
\|\boldsymbol{\theta}_t\|_\infty &\le \eta_0^{(t)} \|\boldsymbol{\Delta}_0\|_\infty + t \max_{1\le i\le(1-\alpha)t} \eta_i^{(t)} \max_{1\le i\le(1-\alpha)t} \left(\left\| \overline{\boldsymbol{P}}^{\pi_{t-1}} \boldsymbol{\Delta}_{i-1} \right\|_\infty + \left\| \boldsymbol{\Lambda} \widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} \boldsymbol{V}_{i-1} \right\|_\infty + \|\boldsymbol{\Lambda} \boldsymbol{P}_{\mathcal{K}} \boldsymbol{V}_{i-1}\|_\infty \right) \\
&\overset{\text{(i)}}{\le} \eta_0^{(t)} \|\boldsymbol{\Delta}_0\|_\infty + t \max_{1\le i\le(1-\alpha)t} \eta_i^{(t)} \max_{1\le i\le(1-\alpha)t} \left(\|\boldsymbol{\Delta}_{i-1}\|_\infty + 2\|\boldsymbol{V}_{i-1}\|_\infty \right) \\
&\overset{\text{(ii)}}{\le} \frac{1}{2T^2} \cdot \frac{1}{1-\gamma} + \frac{1}{2T^2} \cdot t \cdot \frac{3}{1-\gamma} \\
&\le \frac{2}{(1-\gamma)T},
\end{aligned} \tag{46}$$

where (i) is due to the fact that $\|\overline{\boldsymbol{P}}^{\pi_{t-1}}\|_1 = \|\boldsymbol{\Lambda}\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)}\|_1 = \|\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}}\|_1 = 1$ and (ii) comes from [4, Equation (39a)].

Bounding $\boldsymbol{\nu}_t$. To control the second term, we apply the following Freedman's inequality.

Lemma 3 (Freedman's Inequality)**.** *Consider a real-valued martingale $\{Y_k : k = 0, 1, 2, \cdots\}$ with difference sequence $\{X_k : k = 1, 2, 3, \cdots\}$. Assume that the difference sequence is uniformly bounded:*

$$|X_k| \le R \qquad \text{and} \qquad \mathbb{E}\left[X_k | \{X_j\}_{j=1}^{k-1}\right] = 0 \qquad \text{for all } k \ge 1.$$

Let

$$S_n \coloneqq \sum_{k=1}^n X_i, \qquad T_n \coloneqq \sum_{k=1}^n \mathsf{Var}\left\{X_k | \{X_j\}_{j=1}^{k-1}\right\}.$$

Then for any given $\sigma^2 \ge 0$, one has

$$\mathbb{P}\left(|S_n| \ge \tau \text{ and } T_n \le \sigma^2\right) \le 2\exp\left(-\frac{\tau^2/2}{\sigma^2 + R\tau/3}\right).$$

In addition, suppose that $W_n \le \sigma^2$ holds deterministically. For any positive integer $K \ge 1$, with probability at least $1-\delta$ one has

$$|S_n| \le \sqrt{8\max\left\{T_n, \frac{\sigma^2}{2^K}\right\}\log\frac{2K}{\delta}} + \frac{4}{3}R\log\frac{2K}{\delta}.$$

Proof. See [4, Theorem 4]. ☐

To apply this inequality, we can express $\boldsymbol{\nu}_t$ as

$$\boldsymbol{\nu}_t \coloneqq \sum_{i=(1-\alpha)t+1}^{t} \boldsymbol{x}_i,$$

with

$$\boldsymbol{x}_i \coloneqq \eta_i^{(t)} \gamma \boldsymbol{\Lambda} \left(\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} - \boldsymbol{P}_{\mathcal{K}}\right) \boldsymbol{V}_{i-1}, \quad \text{and} \quad \mathbb{E}\left[\boldsymbol{x}_i | \boldsymbol{V}_{i-1}, \cdots, \boldsymbol{V}_0\right] = \boldsymbol{0}. \tag{47}$$

1. In order to calculate bound R in Lemma 3, one has

$$\begin{aligned}
B &\coloneqq \max_{(1-\alpha)t<t\le t} \|\boldsymbol{x}_i\|_\infty \le \max_{(1-\alpha)t<t\le t} \left\| \eta_i^{(t)} \boldsymbol{\Lambda} \left(\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} - \boldsymbol{P}_{\mathcal{K}}\right) \boldsymbol{V}_{i-1} \right\|_\infty \\
&\le \max_{(1-\alpha)t<t\le t} \eta_i^{(t)} \left(\left\| \boldsymbol{\Lambda} \widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} \right\|_1 + \|\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}}\|_1 \right) \|\boldsymbol{V}_{i-1}\|_\infty \\
&\le \max_{(1-\alpha)t<t\le t} \eta_i^{(t)} \cdot \frac{2}{1-\gamma} \le \frac{4\log^4 T}{(1-\gamma)^2 T},
\end{aligned}$$

where the last inequality comes from [4, Eqn (39b)] and the fact that $\|\boldsymbol{V}_{i-1}\|_\infty \le \frac{1}{1-\gamma}$.

2. Then regarding the variance term, we claim for the moment that

$$\begin{aligned}\boldsymbol{W}_t &\coloneqq \sum_{i=(1-\alpha)t+1}^{t} \mathrm{diag}\left(\mathsf{Var}\left(\boldsymbol{x}_i | \boldsymbol{V}_{i-1}, \cdots, \boldsymbol{V}_0\right)\right) \\ &\leq \gamma^2 \sum_{i=(1-\alpha)t+1}^{t} \left(\eta_i^{(t)}\right)^2 \mathsf{Var}_{\overline{\boldsymbol{P}}}\left(\boldsymbol{V}_{i-1}\right). \end{aligned} \tag{48}$$

Then we have

$$\begin{aligned}\boldsymbol{W}_t &\leq \max_{(1-\alpha)t\leq i\leq t} \eta_i^{(t)} \left(\sum_{i=(1-\alpha)t+1}^{t} \eta_i^{(t)}\right) \max_{(1-\alpha)t\leq i<t} \mathsf{Var}_{\overline{\boldsymbol{P}}}\left(\boldsymbol{V}_i\right) \\ &\leq \frac{2\log^4 T}{(1-\gamma)T} \max_{(1-\alpha)t\leq i<t} \mathsf{Var}_{\overline{\boldsymbol{P}}}\left(\boldsymbol{V}_i\right), \end{aligned} \tag{49}$$

where the second line comes from [4, Eqns (39b), (40)]. A trivial upper bound for $\boldsymbol{W}_t$ is

$$|\boldsymbol{W}_t| \leq \frac{2\log^4 T}{(1-\gamma)T} \cdot \frac{1}{(1-\gamma)^2}\mathbf{1} = \frac{2\log^4 T}{(1-\gamma)^3 T}\mathbf{1},$$

which uses the fact that $\mathsf{Var}_{\boldsymbol{P}}(\boldsymbol{V}_i) \leq \|\boldsymbol{V}_i\|_\infty^2 \leq 1/(1-\gamma)^2$.

Then, we invoke Lemma 3 with $K = \left\lceil 2\log_2 \frac{1}{1-\gamma}\right\rceil$ and apply the union bound argument over $\mathcal{K}$ to arrive at

$$\begin{aligned}|\boldsymbol{\nu}_t| &\leq \sqrt{8\left(\boldsymbol{W}_t + \frac{\sigma^2}{2^K}\mathbf{1}\right)\log\frac{8KT\log\frac{1}{1-\gamma}}{\delta}} + \frac{4}{3}B\log\frac{8KT\log\frac{1}{1-\gamma}}{\delta}\mathbf{1} \\ &\leq \sqrt{8\left(\boldsymbol{W}_t + \frac{2\log^4 T}{(1-\gamma)T}\mathbf{1}\right)\log\frac{8KT}{\delta}} + \frac{4}{3}B\log\frac{8KT\log\frac{1}{1-\gamma}}{\delta}\mathbf{1} \\ &\leq \sqrt{\frac{32\log^4 T}{(1-\gamma)T}\log\frac{8KT}{\delta}\left(\max_{(1-\alpha)t\leq i<t}\mathsf{Var}_{\boldsymbol{\Lambda P}_{\mathcal{K}}}\left(\boldsymbol{V}_i\right) + \mathbf{1}\right)} + \frac{12\log^4 T}{(1-\gamma)^2 T}\log\frac{8KT}{\delta}\mathbf{1}. \end{aligned} \tag{50}$$

Hence if we define

$$\boldsymbol{\varphi}_t \coloneqq 64\frac{\log^4 T \log\frac{KT}{\delta}}{(1-\gamma)T}\left(\max_{\frac{t}{2}\leq i\leq t}\mathsf{Var}_{\overline{\boldsymbol{P}}}\left(\boldsymbol{V}_i\right) + \mathbf{1}\right),$$

then (46) and (50) implies that

$$|\boldsymbol{\theta}_t| + |\boldsymbol{\nu}_t| + |\boldsymbol{\omega}_t| \leq \sqrt{\boldsymbol{\varphi}_t} + \frac{2\gamma\xi}{1-\gamma}\mathbf{1}, \tag{51}$$

with probability over $1-\delta$ for all $2t/3 \leq k \leq t$, as long as $T \gg \log^4 T \log\frac{KT}{\delta}/(1-\gamma)^3$. Therefore, plugging (51) into (45), we arrive at the recursive relationship

$$\boldsymbol{\Delta}_t \leq \sqrt{\boldsymbol{\varphi}_t} + \frac{2\gamma\xi}{1-\gamma}\mathbf{1} + \sum_{i=(1-\alpha)k+1}^{k} \eta_i^{(k)}\gamma\overline{\boldsymbol{P}}^{\pi_{i-1}}\boldsymbol{\Delta}_{i-1} = \sqrt{\boldsymbol{\varphi}_t} + \frac{2\gamma\xi}{1-\gamma}\mathbf{1} + \sum_{i=(1-\alpha)k}^{k-1} \eta_i^{(k)}\gamma\overline{\boldsymbol{P}}^{\pi_{i-1}}\boldsymbol{\Delta}_i.$$

This recursion is expressed in a similar way as [4, Eqn. (46)] so we can invoke similar derivation in [4, Appendix A.2] to obtain that

$$\boldsymbol{\Delta}_t \leq 30\sqrt{\frac{\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}\left(1 + \max_{\frac{t}{2}\leq i<t}\|\boldsymbol{\Delta}_i\|_\infty\right)}\mathbf{1} + \frac{2\gamma\xi}{(1-\gamma)^2}\mathbf{1}. \tag{52}$$

Then we turn to (44). Applying a similar argument, we can deduce that

$$\boldsymbol{\Delta}_t \geq -30\sqrt{\frac{\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}\left(1 + \max_{\frac{t}{2}\leq i<t}\|\boldsymbol{\Delta}_i\|_\infty\right)}\mathbf{1} - \frac{2\gamma\xi}{(1-\gamma)^2}\mathbf{1}. \tag{53}$$

For any t satisfying $\frac{T}{c_2 \log \frac{1}{1-\gamma}} \le t \le T$, taking (52) and (53) collectively gives rise to

$$\|\mathbf{\Delta}_t\|_\infty \le 30 \sqrt{\frac{\log^4 T \log \frac{KT}{\delta}}{(1-\gamma)^4 T} \left(1 + \max_{\frac{t}{2} \le i < t} \|\mathbf{\Delta}_i\|_\infty\right)} + \frac{2\gamma\xi}{(1-\gamma)^2}. \tag{54}$$

Let

$$u_k := \max\left\{\|\mathbf{\Delta}_t\|_\infty : 2^k \frac{T}{c_2 \log \frac{1}{1-\gamma}} \le t \le T\right\}.$$

By taking supremum over $t \in \{\lceil 2^k T/(c_2 \log \frac{1}{1-\gamma})\rceil, \ldots, T\}$ on both sides of (54), we have

$$u_k \le 30 \sqrt{\frac{\log^4 T \log \frac{KT}{\delta}}{(1-\gamma)^4 T} (1 + u_{k-1})} + \frac{2\gamma\xi}{(1-\gamma)^2} \qquad \forall\ 1 \le k \le \log\left(c_2 \log \frac{1}{1-\gamma}\right). \tag{55}$$

It is straightforward to bound $u_0 \le \frac{1}{1-\gamma}$. For $k \ge 1$, it is straightforward to obtain from (55) that

$$u_k \le 3 \max\left\{30 \sqrt{\frac{\log^4 T \log \frac{KT}{\delta}}{(1-\gamma)^4 T}}, 30 \sqrt{\frac{\log^4 T \log \frac{KT}{\delta}}{(1-\gamma)^4 T} u_{k-1}}, \frac{2\gamma\xi}{(1-\gamma)^2}\right\}, \tag{56}$$

for $1 \le k \le \log(c_2 \log \frac{1}{1-\gamma})$. We analyze (56) under two different cases:

1. If there exists some integer k_0 with $1 \le k_0 < \lceil \log(c_2 \log \frac{1}{1-\gamma})\rceil$, such that

   $$u_{k_0} \le \max\left\{1, \frac{6\gamma\xi}{(1-\gamma)^2}\right\},$$

   then it is straightforward to check from (56) that

   $$u_{k_0+1} \le 3 \max\left\{30 \sqrt{\frac{\log^4 T \log \frac{KT}{\delta}}{(1-\gamma)^4 T}}, \frac{2\gamma\xi}{(1-\gamma)^2}\right\} \tag{57}$$

   as long as $T \ge C_3 (1-\gamma)^{-4} \log^4 T \log(KT/\delta)$ for some sufficiently large constant $C_3 > 0$.

2. Otherwise we have $u_k > \max\{1, \frac{6\gamma\xi}{(1-\gamma)^2}\}$ for all $1 \le k < \lceil \log(c_2 \log \frac{1}{1-\gamma})\rceil$. This together with (56) suggests that

   $$\max\left\{1, \frac{6\gamma\xi}{(1-\gamma)^2}\right\} < 3 \max\left\{30 \sqrt{\frac{\log^4 T \log \frac{KT}{\delta}}{(1-\gamma)^4 T}}, 30 \sqrt{\frac{\log^4 T \log \frac{KT}{\delta}}{(1-\gamma)^4 T} u_{k-1}}, \frac{2\gamma\xi}{(1-\gamma)^2}\right\},$$

   and therefore

   $$\max\left\{30 \sqrt{\frac{\log^4 T \log \frac{KT}{\delta}}{(1-\gamma)^4 T}}, 30 \sqrt{\frac{\log^4 T \log \frac{KT}{\delta}}{(1-\gamma)^4 T} u_{k-1}}, \frac{2\gamma\xi}{(1-\gamma)^2}\right\} = 30 \sqrt{\frac{\log^4 T \log \frac{KT}{\delta}}{(1-\gamma)^4 T} u_{k-1}}$$

   for all $1 \le k \le \log(c_2 \log \frac{1}{1-\gamma})$. Let

   $$v_k := 90 \sqrt{\frac{\log^4 T \log \frac{KT}{\delta}}{(1-\gamma)^4 T} u_{k-1}}.$$

   Then we know from (55) that

   $$u_k \le v_k \qquad \forall\ 1 \le k \le \log\left(c_2 \log \frac{1}{1-\gamma}\right).$$

By applying the above two inequalities recursively, we know that

$$\begin{aligned}
u_k \le v_k &= \left(\frac{8100\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}\right)^{1/2} u_{k-1}^{1/2} \le \left(\frac{8100\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}\right)^{1/2} v_{k-1}^{1/2} \\
&\le \left(\frac{8100\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}\right)^{1/2+1/4} u_{k-2}^{1/4} \le \left(\frac{8100\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}\right)^{1/2+1/4} v_{k-2}^{1/4} \\
&\le \cdots \le \left(\frac{8100\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}\right)^{1-1/2^k} u_0^{1/2^k} \le \sqrt{\frac{8100\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}}\left(\frac{1}{1-\gamma}\right)^{1/2^k},
\end{aligned}$$

where the last inequality holds as long as $T \ge C_3 \log^4 T \log(KT/\delta)(1-\gamma)^{-4}$ for some sufficiently large constant $C_3 > 0$. Let $k_0 = \widetilde{c}\log\log\frac{1}{1-\gamma}$ for some properly chosen constant $\widetilde{c} > 0$ such that k_0 is an integer between 1 and $\log(c_2 \log\frac{1}{1-\gamma})$, we have

$$u_{k_0} \le \sqrt{\frac{8100\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}}\left(\frac{1}{1-\gamma}\right)^{1/2^{k_0}} = O\left(\sqrt{\frac{\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}}\right).$$

When $T \ge C_3 \log^4 T \log(KT/\delta)(1-\gamma)^{-4}$ for some sufficiently large constant $C_3 > 0$, this implies that $u_{k_0} < 1$, which contradicts with the preassumption that $u_k > \max\{1, \frac{6\gamma\xi}{(1-\gamma)^2}\}$ for all $1 \le k \le c_2 \log\frac{1}{1-\gamma}$.

Consequently, (57) must hold true and then the definition of u_k immediately leads to

$$\|\boldsymbol{\Delta}_T\|_\infty \le 90\sqrt{\frac{\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}} + \frac{6\gamma\xi}{(1-\gamma)^2}.$$

Then for any $\varepsilon \in (0,1]$, one has

$$\|\boldsymbol{\Delta}_T\|_\infty \le \varepsilon + \frac{6\gamma\xi}{(1-\gamma)^2},$$

as long as

$$90\sqrt{\frac{\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4 T}} \le \varepsilon.$$

Hence, if the total number of iterations T satisfies

$$T \ge C_3\frac{\log^4 T\log\frac{KT}{\delta}}{(1-\gamma)^4\varepsilon^2}$$

for some sufficiently large constant $C_3 > 0$, (10) would hold for Algorithm 1 with probability over $1-\delta$.

Finally, we are left to justify (48). Recall the definition of $\boldsymbol{x}_i$ (cf. (47)), one has

$$\begin{aligned}
\mathrm{diag}\left(\mathsf{Var}\left(\boldsymbol{x}_i|\boldsymbol{V}_{i-1},\cdots,\boldsymbol{V}_0\right)\right) &= \gamma^2\left(\eta_i^{(t)}\right)^2 \mathrm{diag}\left(\mathsf{Var}\left(\boldsymbol{\Lambda}\left(\widehat{\boldsymbol{P}}_{\mathcal{K}}^{(t)} - \boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}_{i-1}|\boldsymbol{V}_{i-1}\right)\right) \\
&= \gamma^2\left(\eta_i^{(t)}\right)^2 \mathrm{diag}\left(\boldsymbol{\Lambda}\mathsf{Var}\left(\left(\hat{\boldsymbol{P}}_{\mathcal{K}}^{(i)} - \boldsymbol{P}_{\mathcal{K}}\right)\boldsymbol{V}_{i-1}|\boldsymbol{V}_{i-1}\right)\boldsymbol{\Lambda}^\top\right) \\
&= \gamma^2\left(\eta_i^{(t)}\right)^2 \left\{\boldsymbol{\lambda}(s,a)^2\,\mathsf{Var}_{\boldsymbol{P}_{\mathcal{K}}}\left(\boldsymbol{V}_{i-1}\right)\right\}_{s,a},
\end{aligned}$$

where the notation $\mathsf{Var}_{\boldsymbol{P}_{\mathcal{K}}}(\boldsymbol{V}_{i-1})$ is defined in (12). Plugging this into the definition of $\boldsymbol{W}_t$ leads to

$$\begin{aligned}
\boldsymbol{W}_t &= \gamma^2 \sum_{i=(1-\alpha)t+1}^{t}\left(\eta_i^{(t)}\right)^2\left\{\boldsymbol{\lambda}(s,a)^2\,\mathsf{Var}_{\boldsymbol{P}_{\mathcal{K}}}\left(\boldsymbol{V}_{i-1}\right)\right\}_{s,a} \\
&= \gamma^2 \sum_{i=(1-\alpha)t+1}^{t}\left(\eta_i^{(t)}\right)^2\left\{\boldsymbol{\lambda}(s,a)^2\left(\boldsymbol{P}_{\mathcal{K}}\left(\boldsymbol{V}_{i-1}\circ\boldsymbol{V}_{i-1}\right) - \left(\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}_{i-1}\right)\circ\left(\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}_{i-1}\right)\right)\right\}_{s,a}. \quad (58)
\end{aligned}$$

Then we introduce a useful claim as follows. The proof is deferred to Appendix C.2.

Claim 1. For any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ and vector $\boldsymbol{V} \in \mathbb{R}^{|\mathcal{S}|}$, one has

$$\begin{aligned}&\boldsymbol{\lambda}(s,a)^2 \left(\boldsymbol{P}_{\mathcal{K}}\left(\boldsymbol{V} \circ \boldsymbol{V}\right) - \left(\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}\right) \circ \left(\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}\right)\right) \\ &\quad \leq \boldsymbol{\lambda}(s,a)\boldsymbol{P}_{\mathcal{K}}\left(\boldsymbol{V} \circ \boldsymbol{V}\right) - \left(\boldsymbol{\lambda}(s,a)\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}\right) \circ \left(\boldsymbol{\lambda}(s,a)\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}\right). \end{aligned} \tag{59}$$

By invoking this claim with $\boldsymbol{V} = \boldsymbol{V}^{i-1}$ and taking collectively with (58), one has

$$\begin{aligned}\boldsymbol{W}_t &\leq \gamma^2 \sum_{i=(1-\beta)t+1}^{t} \left(\eta_i^{(t)}\right)^2 \left\{\boldsymbol{\lambda}(s,a)\boldsymbol{P}_{\mathcal{K}}\left(\boldsymbol{V}_{i-1} \circ \boldsymbol{V}_{i-1}\right) - \left(\boldsymbol{\lambda}(s,a)\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}_{i-1}\right) \circ \left(\boldsymbol{\lambda}(s,a)\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}_{i-1}\right)\right\}_{s,a} \\ &= \gamma^2 \sum_{i=(1-\beta)t+1}^{t} \left(\eta_i^{(t)}\right)^2 \left[\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}}\left(\boldsymbol{V}_{i-1} \circ \boldsymbol{V}_{i-1}\right) - \left(\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}_{i-1}\right) \circ \left(\boldsymbol{\Lambda}\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}_{i-1}\right)\right] \\ &= \gamma^2 \sum_{i=(1-\beta)t+1}^{t} \left(\eta_i^{(t)}\right)^2 \mathsf{Var}_{\overline{\boldsymbol{P}}}\left(\boldsymbol{V}_{i-1}\right),\end{aligned}$$

which is the desired result.

C.2 Proof of Claim 1

To simplify notations in this proof, we use $[\lambda_i]_{i=1}^K$, $[P_{i,j}]_{1\leq i\leq K, 1\leq j\leq|\mathcal{S}|}$ and $[V_i]_{i=1}^{|\mathcal{S}|}$ to denote $\boldsymbol{\lambda}(s,a)$, $\boldsymbol{P}_{\mathcal{K}}$ and $\boldsymbol{V}$ respectively. Then one has

$$\begin{aligned}&\boldsymbol{\lambda}(s,a)\boldsymbol{P}_{\mathcal{K}}\left(\boldsymbol{V} \circ \boldsymbol{V}\right) - \left(\boldsymbol{\lambda}(s,a)\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}\right) \circ \left(\boldsymbol{\lambda}(s,a)\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}\right) \\ &\quad - \boldsymbol{\lambda}(s,a)^2 \left(\boldsymbol{P}_{\mathcal{K}}\left(\boldsymbol{V} \circ \boldsymbol{V}\right) - \left(\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}\right) \circ \left(\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}\right)\right) \\ &\quad = \sum_{i=1}^{K}\sum_{j=1}^{|\mathcal{S}|} \lambda_i P_{i,j} V_j^2 - \left(\sum_{i=1}^{K}\sum_{j=1}^{|\mathcal{S}|} \lambda_i P_{i,j} V_j\right)^2 - \sum_{i=1}^{K}\sum_{j=1}^{|\mathcal{S}|} \lambda_i^2 P_{i,j} V_j^2 + \sum_{i=1}^{K} \lambda_i^2 \left(\sum_{j=1}^{|\mathcal{S}|} P_{i,j} V_j\right)^2 \\ &\quad = \sum_{i=1}^{K}\sum_{j=1}^{|\mathcal{S}|} \lambda_i P_{i,j} V_j \left[\left(1-\lambda_i\right) V_j - \sum_{i'\neq i}\sum_{j'=1}^{|\mathcal{S}|} \lambda_{i'} P_{i',j'} V_{j'}\right]. \\ &\quad = \sum_{i=1}^{K}\sum_{j=1}^{|\mathcal{S}|} \lambda_i P_{i,j} V_j \left[\left(\sum_{i'=1}^{K}\sum_{j'=1}^{|\mathcal{S}|} \lambda_{i'} P_{i',j'} - \lambda_i\right) V_j - \sum_{i'\neq i}\sum_{j'=1}^{|\mathcal{S}|} \lambda_{i'} P_{i',j'} V_{j'}\right] \\ &\quad = \sum_{i=1}^{K}\sum_{j=1}^{|\mathcal{S}|}\sum_{i'\neq i}\sum_{j'=1}^{|\mathcal{S}|} \lambda_i P_{i,j} V_j \lambda_{i'} P_{i',j'} \left(V_j - V_{j'}\right)\end{aligned}$$

where in the penultimate equality, we use the fact that

$$\sum_{i'=1}^{K}\sum_{j'=1}^{|\mathcal{S}|} \lambda_{i'} P_{i',j'} = \boldsymbol{\lambda}(s,a)\boldsymbol{P}_{\mathcal{K}}\boldsymbol{1} = 1.$$

It follows that

$$
\begin{aligned}
&\boldsymbol{\lambda}(s,a)\,\boldsymbol{P}_{\mathcal{K}}(\boldsymbol{V}\circ\boldsymbol{V})-(\boldsymbol{\lambda}(s,a)\,\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V})\circ(\boldsymbol{\lambda}(s,a)\,\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V})\\
&\qquad-\boldsymbol{\lambda}(s,a)^2(\boldsymbol{P}_{\mathcal{K}}(\boldsymbol{V}\circ\boldsymbol{V})-(\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V})\circ(\boldsymbol{P}_{\mathcal{K}}\boldsymbol{V}))\\
&\quad=\sum_{i=1}^{K}\sum_{1\le i'<i}\sum_{j=1}^{|\mathcal{S}|}\sum_{j'=1}^{|\mathcal{S}|}\left[\lambda_i P_{i,j}V_j\lambda_{i'}P_{i',j'}(V_j-V_{j'})+\lambda_{i'}P_{i',j}V_j\lambda_i P_{i,j'}(V_j-V_{j'})\right]\\
&\quad=\sum_{i=1}^{K}\sum_{1\le i'<i}\lambda_i\lambda_{i'}\left[\sum_{j=1}^{|\mathcal{S}|}\sum_{j'=1}^{|\mathcal{S}|}P_{i,j}V_jP_{i',j'}(V_j-V_{j'})+\sum_{j=1}^{|\mathcal{S}|}\sum_{j'=1}^{|\mathcal{S}|}P_{i',j}V_jP_{i,j'}(V_j-V_{j'})\right]\\
&\quad\overset{\text{(i)}}{=}\sum_{i=1}^{K}\sum_{1\le i'<i}\lambda_i\lambda_{i'}\left[\sum_{j=1}^{|\mathcal{S}|}\sum_{j'=1}^{|\mathcal{S}|}P_{i,j}V_jP_{i',j'}(V_j-V_{j'})+\sum_{j=1}^{|\mathcal{S}|}\sum_{j'=1}^{|\mathcal{S}|}P_{i',j'}V_{j'}P_{i,j}(V_{j'}-V_j)\right]\\
&\quad=\sum_{i=1}^{K}\sum_{1\le i'<i}\lambda_i\lambda_{i'}\left[\sum_{j=1}^{|\mathcal{S}|}\sum_{j'=1}^{|\mathcal{S}|}P_{i,j}P_{i',j'}(V_j-V_{j'})^2\right]\\
&\quad\ge 0,
\end{aligned}
$$

where in (i), we exchange the indices j and j'.

D Feature dimension and the number of anchor state-action pairs

The assumption that the feature dimension (denoted by K_{d}) and the number of anchor state-action pairs (denoted by K_{n}) are equal is actually non-essential. In what follows, we will show that if $K_{\mathsf{d}}\neq K_{\mathsf{n}}$, then we can modify the current feature mapping $\phi:\mathcal{S}\times\mathcal{A}\to\mathbb{R}^{K_{\mathsf{d}}}$ to achieve a new feature mapping $\phi':\mathcal{S}\times\mathcal{A}\to\mathbb{R}^{K_{\mathsf{n}}}$ that does not change the transition model P. By doing so, the new feature dimension K_{n} equals to the number of anchor state-action pairs.

To begin with, we recall from Definition 1 that there exists K_{d} unknown functions $\psi_1,\cdots,\psi_{K_{\mathsf{d}}}:\mathcal{S}\to\mathbb{R}$, such that

$$P(s'|s,a)=\sum_{k=1}^{K_{\mathsf{d}}}\phi_k(s,a)\,\psi_k(s'),$$

for every $(s,a)\in\mathcal{S}\times\mathcal{A}$ and $s'\in\mathcal{S}$. In addition, we also recall from Assumption 1 that there exists $\mathcal{K}\subseteq\mathcal{S}\times\mathcal{A}$ with $|\mathcal{K}|=K_{\mathsf{n}}$ such that for any $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$\phi(s,a)=\sum_{i:(s_i,a_i)\in\mathcal{K}}\lambda_i(s,a)\,\phi(s_i,a_i)\in\mathbb{R}^{K_{\mathsf{d}}}\quad\text{for}\quad\sum_{i=1}^{K_{\mathsf{n}}}\lambda_i(s,a)=1\quad\text{and}\quad\lambda_i(s,a)\ge 0.$$

Case 1: $K_{\mathsf{d}}>K_{\mathsf{n}}$. In this case, the vectors in $\{\phi(s,a):(s,a)\in\mathcal{K}\}$ are linearly independent. For ease of presentation and without loss of generality, we assume that $K_{\mathsf{d}}=K_{\mathsf{n}}+1$. This indicates that the matrix $\mathbf{\Phi}\in\mathbb{R}^{K_{\mathsf{d}}\times(|\mathcal{S}||\mathcal{A}|)}$ whose columns are composed of the feature vectors of all state-action pairs has rank K_{n} and is hence not full row rank. This suggests that there exists K_{n} linearly independent rows (without loss of generality, we assume they are the first K_{n} rows). We can remove the last row from $\mathbf{\Phi}$ to obtain $\mathbf{\Phi}':=\mathbf{\Phi}_{1:K_{\mathsf{n}},:}\in\mathbb{R}^{K_{\mathsf{n}}\times(|\mathcal{S}||\mathcal{A}|)}$ such that $\mathbf{\Phi}'$ is full row rank. Then we show that we can actually use the columns of $\mathbf{\Phi}'$ as new feature mappings. To see why this is true, note that the last row $\Phi_{K_{\mathsf{n}}+1,:}$ can be represented as a linear combination of the first K_{n} rows, namely there must exist constants $\{c_k\}_{k=1}^{K_{\mathsf{n}}}$ such that for any $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$\phi_{K_{\mathsf{n}}+1}(s,a)=\sum_{k=1}^{K_{\mathsf{n}}}c_k\phi_k(s,a).$$

Define $\psi'_k = \psi_k + c_k \psi_{K_{\mathsf{n}}+1}$ for $k = 1, \ldots, K_{\mathsf{n}}$, we have

$$
\begin{aligned}
P\left(s'|s,a\right) &= \sum_{k=1}^{K_{\mathsf{d}}} \phi_k\left(s,a\right)\psi_k\left(s'\right) = \phi_{K_{\mathsf{n}}+1}\left(s,a\right)\psi_{K_{\mathsf{n}}+1}\left(s'\right) + \sum_{k=1}^{K_{\mathsf{n}}} \phi_k\left(s,a\right)\psi_k\left(s'\right) \\
&= \sum_{k=1}^{K_{\mathsf{n}}} \phi_k\left(s,a\right)\left[\psi_k\left(s'\right) + c_k\psi_{K_{\mathsf{n}}+1}\left(s'\right)\right] = \sum_{k=1}^{K_{\mathsf{n}}} \phi_k\left(s,a\right)\psi'_k\left(s'\right),
\end{aligned}
$$

which is linear with respect to the new K_{n} dimensional feature vectors. It is also straightforward to check that the new feature mapping satisfies Assumption 1 with the original anchor state-action pairs $\mathcal{K}$.

Case 2: $K_{\mathsf{d}} < K_{\mathsf{n}}$. For ease of presentation and without loss of generality, we assume that $K_{\mathsf{n}} = K_{\mathsf{d}} + 1$ and that the subspace spanned by the feature vectors of anchor state-action pairs is non-degenerate, i.e., has rank K_{d} (otherwise we can use similar method as in Case 1 to further reduce the feature dimension K_{d}). In this case, the matrix $\mathbf{\Phi}_{\mathcal{K}} \in \mathbb{R}^{K_{\mathsf{d}} \times K_{\mathsf{n}}}$ whose columns are composed of the feature vectors of anchor state-action pairs has rank K_{d}. We can add $K_{\mathsf{n}} - K_{\mathsf{d}} = 1$ new row to $\mathbf{\Phi}_{\mathcal{K}}$ to obtain $\mathbf{\Phi}'_{\mathcal{K}} \in \mathbb{R}^{K_{\mathsf{n}} \times K_{\mathsf{n}}}$ such that $\mathbf{\Phi}'_{\mathcal{K}}$ has full rank K_{n}. Then we let the columns of $\mathbf{\Phi}'_{\mathcal{K}} = [\phi'(s,a)]_{(s,a)\in\mathcal{K}}$ to be the new feature vectors of the anchor state-action pairs, and define the new feature vectors for all other state-action pairs $(s,a) \notin \mathcal{K}$ by

$$
\phi'\left(s,a\right) = \sum_{i:(s_i,a_i)\in\mathcal{K}} \lambda_i\left(s,a\right)\phi'\left(s_i,a_i\right).
$$

We can check that the transition model P is not changed if we let $\psi_{K_{\mathsf{n}}}(s') = 0$ for every $s' \in \mathcal{S}$. It is also straightforward to check that Assumption 1 is satisfied.

To conclude, when $K_{\mathsf{d}} \neq K_{\mathsf{n}}$, we can always construct a new set of feature mappings with dimension K_{n} such that: (i) the feature dimension equals to the number of anchor state-action pairs (they are both K_{n}); (ii) the transition model can still be linearly parameterized by this new set of feature mappings; and (iii) the anchor state-action pair assumption (Assumption 1) is satisfied with the original anchor state-action pairs.

References

[1] Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020.

[2] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In *2012 IEEE 53rd annual symposium on foundations of computer science*, pages 1–10. IEEE, 2012.

[3] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349, 2013.

[4] Gen Li, Changxiao Cai, Yuxin Chen, Yuantao Gu, Yuting Wei, and Yuejie Chi. Is q-learning minimax optimal? a tight sample complexity analysis. *arXiv preprint arXiv:2102.06548*, 2021.

[5] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.