A. Proofs of Propositions 1,2

Proof of Proposition 1. Let v^π be the value function of π. Since $M \in \mathcal{M}^{trans}(\mathcal{S}, \mathcal{A}, \gamma, \phi)$, we have $P(s'|s,a) = \sum_{k\in[K]} \psi_k(s')\phi_k(s,a)$ for some ψ_k's. We have

$$\begin{aligned} Q^\pi(s,a) &= r(s,a) + \gamma \sum_{s'\in\mathcal{S}} P(s'|s,a)v^\pi(s') = r(s,a) + \gamma \sum_{k\in[K]} \phi_k(s,a) \sum_{s'\in\mathcal{S}} \psi_k(s')v^\pi(s') \\ &= r(s,a) + \gamma \sum_{k\in[K]} \phi_k(s,a) w^\pi(k) \end{aligned}$$

where vector $w^\pi \in \mathbb{R}^K$ is specified by

$$\forall k \in [K] : w^\pi(k) = \sum_{s'\in\mathcal{S}} \psi_k(s')v^\pi(s').$$

Therefore $Q^\pi \in \text{Span}(r, \phi)$. □

Proof of Proposition 2. "If" direction: Since $M \in \mathcal{M}^{trans}$, we have from the proof of Proposition 1 that for any $Q \in \mathcal{F}$, $\mathcal{T}Q \in \mathcal{F}$.

"Only if" direction: If $d(\mathcal{T}\mathcal{F}, \mathcal{F}) = 0$, then for any $Q \in \mathcal{F}$ We have

$$\mathcal{T}Q = r + \gamma PV(Q) \in \mathcal{F}.$$

We can then pick a maximum-sized set $\{Q_1, Q_2, \dots Q_k\} \subset \mathcal{F}$ such that $V(Q_1), V(Q_2), \dots V(Q_k)$ are linear independent. Note that $k \le K$. Denote $A = [V(Q_1), V(Q_2), \dots V(Q_k)]$, $B = [\mathcal{T}Q_1, \mathcal{T}Q_2, \dots, \mathcal{T}Q_k]$ and $R = [r, r, r \dots, r]$ (with k columns). We then have

$$B = R + \gamma PA.$$

Hence we have

$$P = \gamma^{-1}(B - R)A^\top(AA^\top)^{-1}.$$

Since each column of $B - R$ is a vector in $\mathcal{F}$, we conclude that each column of P is a vector in $\mathcal{F}$. □

B. Proof of Theorem 1

Proof of Theorem 1. Let $\mathcal{M}'$ be the class of all tabular DMDPs with state space $\mathcal{S}'$, action space $\mathcal{A}'$, and discount factor γ. Let $\mathcal{K}'$ be an algorithm for such a class of DMDPs with a generative model. Let

$$N = O\left(\frac{|\mathcal{S}'||\mathcal{A}'|}{(1-\gamma)^3 \cdot \epsilon^2 \cdot \log \epsilon^{-1}}\right).$$

For each $M' \in \mathcal{M}'$, let $\pi^{\mathcal{K}', M', N}$ be the policy returned by $\mathcal{K}'$ with querying at most N samples from the generative model. The lower bound in Theorem B.3 in Sidford et al. (2018a)(which is derived from Theorem 3 in Azar et al. (2013)) states that

$$\inf_{\mathcal{K}'} \sup_{M'\in\mathcal{M}'} \mathbb{P}\left[\sup_{s\in\mathcal{S}} (v^{*,M'}(s) - v^{\pi^{\mathcal{K}',M',N}}(s)) \ge \epsilon\right] \ge 1/3,$$

where $v^{*,M'}$ is the optimal value function of M'. Suppose, without loss of generality, $K = |\mathcal{S}'||\mathcal{A}'| + 1$. We prove Theorem 1 by showing that every DMDP instance $M' \in \mathcal{M}'$ can be converted to an instance $M \in \mathcal{M}_K^{trans}(\mathcal{S}, \mathcal{A}, \gamma)$ such that any algorithm $\mathcal{K}$ for $\mathcal{M}_K^{trans}(\mathcal{S}, \mathcal{A}, \gamma)$ can be used to solve M'.

For a DMDP instance $M' = (\mathcal{S}', \mathcal{A}', P', r', \gamma) \in \mathcal{M}'$, we construct a corresponding DMDP instance $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma) \in \mathcal{M}_K^{trans}(\mathcal{S}, \mathcal{A}, \gamma)$ with a feature representation ϕ. We pick $\mathcal{S}$ and $\mathcal{S}$ to be supersets of $\mathcal{S}$ and $\mathcal{A}'$ respectively, so that the transition distributions and rewards remain unchanged on $\mathcal{S}' \times \mathcal{A}'$, i.e., $P(\cdot \mid s,a) = P'(\cdot \mid s,a)$ and $r(s,a) = r'(s,a)$ for $s \in \mathcal{S}', a \in \mathcal{A}'$. From $(s,a) \in (\mathcal{S} \times \mathcal{A})/(\mathcal{S}' \times \mathcal{A}')$, the process transitions to an absorbing state $s^0 \in \mathcal{S}/\mathcal{S}'$ with probability 1 and stays there with reward 0.

Now we show that M admits a feature representation $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^K$ as follows. Say (s, a) is the k-th element in $\mathcal{S}' \times \mathcal{A}$, we let $\phi(s, a) = \mathbf{1_k}$, which is the unit vector whose kth entry equals one. For $(s, a) \notin \mathcal{S}' \times \mathcal{A}'$, we let $\phi(s, a) = \mathbf{1_K}$. Then we can verify that $P(s' \mid s, a) = \sum_{k\in[K]} \phi_k(s, a)\psi_k(s')$ for some ψ_k's. Thus we have constructed an MDP instance $M' \in \mathcal{M}_K^{trans}(\mathcal{S}, \mathcal{A}, \gamma)$ with feature representation ϕ .

Suppose that $\mathcal{K}$ is an algorithm that applies to M using N samples. Based on the reduction, we immediately obtained an algorithm $\mathcal{K}'$ that applies to M' using N samples and the feature map ϕ: $\mathcal{K}'$ works by applying $\mathcal{K}$ to M and outputs the restricted policy on $\mathcal{S}' \times \mathcal{A}'$. It can be easily verified that if π is an ϵ-optimal policy for M then the reduction gives an ϵ-optimal policy for M'. By virtue of the reduction, one gets

$$\inf_{\mathcal{K}} \sup_{M\in\mathcal{M}_K^{trans}(\mathcal{S},\mathcal{A},\gamma)} \mathbb{P}\bigg(\sup_{s\in\mathcal{S}}(v^*(s) - v^{\pi^{\mathcal{K},M,N}}(s)) \geq \epsilon\bigg) \geq \inf_{\mathcal{K}'} \sup_{M'\in\mathcal{M}'} \mathbb{P}\bigg(\sup_{s\in\mathcal{S}}(v^{*,M'}(s) - v^{\pi^{\mathcal{K}',M',N}}(s)) \geq \epsilon\bigg)$$
$$\geq 1/3,$$

This completes the proof. □

C. Proof of Theorem 2.

Proof. Recall that $P_{\mathcal{K}}$ is a submatrix of P formed by the rows indexed by $\mathcal{K}$. We denote $\widetilde{P}_{\mathcal{K}}$ in the same manner for $\widetilde{P}$. Recall that $\|P - \widetilde{P}\|_{1,\infty} \leq \xi$. Let $\widehat{P}_{\mathcal{K}}^{(t)}$ be the matrix of empirical transition probabilities based on $m := N/(KR)$ sample transitions per $(s, a) \in \mathcal{K}$ generated at iteration k. It can be viewed as an estimate of $P_{\mathcal{K}}$ at iteration t. Since $\widetilde{P}$ admits a context representation, it can be written as

$$\widetilde{P} = \Phi\Psi \quad \text{where} \quad \Psi = \Phi_{\mathcal{K}}^{-1}\widetilde{P}_{\mathcal{K}}.$$

Let $\widehat{\Psi}^{(t)} = \Phi_{\mathcal{K}}^{-1}\widehat{P}_{\mathcal{K}}^{(t)}$ be the estimate of Ψ at iteration t. We can view $\Phi\widehat{\Psi}^{(t)}$ as an estimate of P.

We will show that each iteration of the algorithm is an approximate value iteration. We first define the approximate Bellman operator, $\widehat{\mathcal{T}}$ as, $\forall v \in \mathbb{R}^{\mathcal{S}}$:

$$[\widehat{\mathcal{T}}^{(t)}v](s) = \max_a \Big[r(s, a) + \gamma\phi(s, a)^\top\Phi_{\mathcal{K}}^{-1}\widehat{P}_{\mathcal{K}}^{(t)}v\Big].$$

Notice that, by definition of the algorithm,

$$V_{w^{(t)}} \leftarrow \widehat{\mathcal{T}}^{(t)}\Pi_{[0,H]}[V_{w^{(t-1)}}],$$

where $w^{(0)} = 0 \in \mathbb{R}^K$ and $w^{(t)}$ is the w at the end of the t-th iteration of the algorithm and $H = (1 - \gamma)^{-1}$ and $\Pi_{[0,H]}(\cdot)$ denotes entrywise projection to $[0, H]$. For the rest of the proof, we denote

$$\widehat{V}_{w^{(t-1)}} = \Pi_{[0,H]}[V_{w^{(t-1)}}].$$

We now show the approximation quality of $\widehat{\mathcal{T}}$, i.e., estimate $\|\widehat{\mathcal{T}}^{(t)}\widehat{V}_{w^{(t-1)}} - \mathcal{T}\widehat{V}_{w^{(t-1)}}\|_\infty$, where $\mathcal{T}$ is the exact Bellman operator. Notice that

$$\forall s: \quad |[\widehat{\mathcal{T}}^{(t)}\widehat{V}_{w^{(t-1)}}](s) - [\mathcal{T}\widehat{V}_{w^{(t-1)}}](s)| \leq \gamma \max_a \big|\phi(s, a)^\top\Phi_{\mathcal{K}}^{-1}\widehat{P}_{\mathcal{K}}^{(t)}\widehat{V}_{w^{(t-1)}} - P(\cdot|s, a)^\top\widehat{V}_{w^{(t-1)}}\big|.$$

It remains to show the right hand side of the above inequality is small.

Denote $\mathcal{F}_t$ to be the filtration defined by the samples up to iteration t. Then, by the Hoeffding inequality and the fact that the samples at iteration t are independent with that from iteration $t - 1$, we have

$$\Pr\bigg[\|\widehat{P}_{\mathcal{K}}^{(t)}\widehat{V}_{w^{(t-1)}} - P_{\mathcal{K}}\widehat{V}_{w^{(t-1)}}\|_\infty \leq \epsilon_1\bigg|\mathcal{F}_{t-1}\bigg] \geq 1 - \delta/R$$

where we denote

$$\epsilon_1 = cH \cdot \sqrt{\frac{\log(KR\delta^{-1})}{m}}$$

for some generic constant c. Next, let $\mathcal{E}_t$ be the event that,

$$\|\widehat{P}_{\mathcal{K}}^{(t)}\widehat{V}_{w^{(t-1)}} - P_{\mathcal{K}}\widehat{V}_{w^{(t-1)}}\|_\infty \le \epsilon_1.$$

We thus have $\Pr[\mathcal{E}_t|\mathcal{F}_{t-1}] \ge 1-\delta/R$ and $\Pr[\mathcal{E}_t|\mathcal{E}_1, \mathcal{E}_2, \ldots \mathcal{E}_{t-1}] \ge 1-\delta/R$ since $\mathcal{E}_1, \mathcal{E}_2, \ldots \mathcal{E}_{t-1}$ are adapted to $\mathcal{F}_{t-1}$. This lead to

$$\Pr[\mathcal{E}_1 \cap \mathcal{E}_2 \cap \ldots \cap \mathcal{E}_R] = \Pr[\mathcal{E}_1]\Pr[\mathcal{E}_2|\mathcal{E}_1]\ldots \ge 1-\delta.$$

Now we consider event $\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_2 \cap \ldots \cap \mathcal{E}_R$, on which we have, for all $t \in [R]$,

$$|\phi(s,a)^\top \Phi_{\mathcal{K}}^{-1}\widehat{P}_{\mathcal{K}}^{(t)}\widehat{V}_{w^{(t-1)}} - \phi(s,a)^\top\Phi_{\mathcal{K}}^{-1}P_{\mathcal{K}}\widehat{V}_{w^{(t-1)}}| \le \|\phi(s,a)^\top\Phi_{\mathcal{K}}^{-1}\|_1 \cdot \epsilon_1 \le L\epsilon_1.$$

Note that, $\|P_{\mathcal{K}} - \widetilde{P}_{\mathcal{K}}\|_{1,\infty} \le \xi$, we thus have

$$|\phi(s,a)^\top \Phi_{\mathcal{K}}^{-1}\widehat{P}_{\mathcal{K}}^{(t)}\widehat{V}_{w^{(t-1)}} - \phi(s,a)^\top\Phi_{\mathcal{K}}^{-1}\widetilde{P}_{\mathcal{K}}\widehat{V}_{w^{(t-1)}}| \le L\epsilon_1 + |\phi(s,a)^\top\Phi_{\mathcal{K}}^{-1}(P_{\mathcal{K}} - \widetilde{P}_{\mathcal{K}})\widehat{V}_{w^{(t-1)}}| \le L\epsilon_1 + LH\xi,$$

Further using

$$|(\phi(s,a)^\top\Phi_{\mathcal{K}}^{-1}\widetilde{P}_{\mathcal{K}}^{(t)} - P(\cdot|s,a)^\top)\widehat{V}_{w^{(t-1)}}| \le H\xi,$$

we thus have

$$\begin{aligned}|\phi(s,a)^\top \Phi_{\mathcal{K}}^{-1}\widehat{P}_{\mathcal{K}}^{(t)}\widehat{V}_{w^{(t-1)}} - P(\cdot|s,a)^\top\widehat{V}_{w^{(t-1)}}| &\le |\phi(s,a)^\top(\Phi_{\mathcal{K}}^{-1}\widehat{P}_{\mathcal{K}}^{(t)} - \Phi_{\mathcal{K}}^{-1}\widetilde{P}_{\mathcal{K}}^{(t)} + \Phi_{\mathcal{K}}^{-1}\widetilde{P}_{\mathcal{K}}^{(t)})\widehat{V}_{w^{(t-1)}} \\ &\quad - P(\cdot|s,a)^\top\widehat{V}_{w^{(t-1)}}| \\ &\le L\epsilon_1 + LH\xi + H\xi.\end{aligned}$$

Further notice that $\Pi_{[0,H]}$ can only makes error smaller. Therefore, we have shown that the $\widehat{V}_{w^{(t)}}$s follow an approximate value iteration with error $\gamma[L\epsilon_1 + (L+1)H\xi]$ with probability at least $1-\delta$. Because of the contraction of the operator $\mathcal{T}$, we have, after R iterations,

$$\|\widehat{V}_{w^{(R-1)}} - v^*\|_\infty \le \gamma^{R-1}H + \gamma R[L\epsilon_1 + (L+1)H\xi] \le \gamma R[2L\epsilon_1 + (L+1)H\xi]$$

for appropriately chosen $R = \Theta(\log(NH)/(1-\gamma))$. Since $Q_{w^{(R)}}(s,a) = r(s,a) + \gamma\phi(s,a)^\top\Phi_{\mathcal{K}}^{-1}\widehat{P}_{\mathcal{K}}^{(R)}\widehat{V}_{w^{(R-1)}}$, we have,

$$\|Q_{w^{(R)}} - Q^*\|_\infty \le 2\gamma R[2L\epsilon_1 + (L+1)H\xi]$$

happens with probability at least $1-\delta$. It follows that (see, e.g., Proposition 2.1.4 of (Bertsekas, 2005)),

$$\|v^{\pi_{w^{(R)}}} - v^*\|_\infty \le 2\gamma RH[2L\epsilon_1 + (L+1)H\xi],$$

with probability at least $1-\delta$. Plugging the values of H, ϵ_1 and m, we have

$$\|v^{\pi_{w^{(R)}}} - v^*\|_\infty \le C\gamma \cdot \frac{\log(NH)}{1-\gamma}\cdot\frac{1}{1-\gamma}\cdot L\cdot\sqrt{\frac{K\log(KR\delta^{-1})}{(1-\gamma)^2\cdot N}\cdot\frac{\log(NH)}{1-\gamma}} + C\gamma\cdot\frac{\log(NH)}{1-\gamma}\cdot\frac{L}{(1-\gamma)^2}\cdot\xi$$

for some generic constant $C > 0$. This completes the proof.

□

D. Proof of Theorem 3

According to the discussions following Assumption 2, we assume without loss of generality:

- For each anchor $(s_k, a_k) \in \mathcal{K}$, $\phi(s_k, a_k)$ is a vector with ℓ_1-norm 1.

Then Assumption 2 further implies

- $\phi(s,a)$ is a vector of probabilities for all (s,a).
- For each (s,a), $P(\cdot|s,a) = \sum_k \phi_k(s,a)P(\cdot \mid s_k, a_k)$.

D.1. Notations

$\mathcal{T}$-operator For any value function $V:\mathcal{S}\to\mathbb{R}$ and policy $\pi:\mathcal{S}\to\mathcal{A}$, we denote the Bellman operators as

$$\mathcal{T}V[s]=\max_{a\in\mathcal{A}}\left[r(s,a)+\gamma P(\cdot|s,a)^\top V\right]\quad\text{and}\quad \mathcal{T}_\pi V[s]=r(s,\pi(s))+\gamma P(\cdot|s,\pi(s))^\top V$$

The key properties, e.g. monotonicity and contraction, of the $\mathcal{T}$-operator can be found in Puterman (2014). For completeness, we state them here.

Fact 4 (Bellman Operator)**.** *For any value function $V,V':\mathcal{S}\to\mathbb{R}$, if $V\le V'$ entry-wisely, we then have,*

$$\begin{aligned}\mathcal{T}V\le\mathcal{T}V' \quad &\textit{and}\quad \mathcal{T}_\pi V\le\mathcal{T}_\pi V',\\ \|\mathcal{T}V-v^*\|_\infty\le\gamma\|V-v^*\|_\infty\quad &\textit{and}\quad \|\mathcal{T}_\pi V-v^\pi\|_\infty\le\gamma\|V-v^\pi\|_\infty,\\ \lim_{t\to\infty}\mathcal{T}^tV=v^*\quad &\textit{and}\quad \lim_{t\to\infty}\mathcal{T}_\pi^tV=v^\pi.\end{aligned}$$

Q-function We let, for any (s,a),

$$\begin{aligned}Q_{\theta^{(i,j)}}(s,a)&=r(s,a)+\gamma\phi(s,a)^\top\overline{w}^{(i,j)},\\ \overline{Q}_{\theta^{(i,j)}}(s,a)&=r(s,a)+\gamma P(\cdot|s,a)^\top V_{\theta^{(i,j-1)}}(\cdot).\end{aligned}$$

Variance of value function For (s,a), we denote the variance of a function (or a vector) $V:\mathcal{S}\to\mathbb{R}$ as,

$$\sigma_{s,a}[V]:=\sum_{s'}P(s'|s,a)V^2(s')-\Big(\sum_{s'}P(s'|s,a)V(s')\Big)^2,$$

we also denote $\sigma_k(\cdot)=\sigma_{s_k,a_k}(\cdot)$ for $(s_k,a_k)\in\mathcal{K}$.

$\mathcal{E}$-event In Algorithm 2, let $\mathcal{E}^{(i,0)}$ be the event that

$$\forall k\in[K]:|w^{(i,0)}(k)-P(\cdot|s_k,a_k)^\top V_{\theta^{(i,0)}}|\le\epsilon^{(i,0)}(k)\le C\left[\sqrt{\frac{\log(R'RK\delta^{-1})\sigma_k[V_{\theta^{(i,0)}}]}{m}}+\frac{\log(R'RK\delta^{-1})}{(1-\gamma)m^{3/4}}\right]$$

for some generic constant $C>0$. Let $\mathcal{E}^{(i,j)}$ be the event on which

$$\forall k\in[K]:\quad |w^{(i,j)}(k)-w^{(i,0)}(k)-P(\cdot|s_k,a_k)^\top(V_{\theta^{(i,j-1)}}-V_{\theta^{(i,0)}})|\le C(1-\gamma)^{-1}2^{-i}\sqrt{\log(R'RK\delta^{-1})/m_1},$$

where R',R,m,m_1 are parameters defined in Algorithm 2.

$\mathcal{G}$-event Let $\mathcal{G}^{(i)}$ be the event such that

$$0\le V_{\theta^{(i,0)}}(s)\le\mathcal{T}_{\pi_{\theta^{(i,0)}}}V_{\theta^{(i,0)}}[s]\le v^*(s),\qquad v^*(s)-V_{\theta^{(i,0)}}(s)\le c2^{-i}/(1-\gamma),\qquad\forall s\in\mathcal{S},$$

for some sufficiently small constant c.

D.2. Some Properties

Firstly we notice that the parameterized functions Q_θ,V_θ (eq. (5)) increase pointwisely (as index (i,j) increases).

Lemma 5 (Monotonicity of the Parametrized V)**.** *For every $(i,j),(i',j')\in[R']\times[R]$, and $s\in\mathcal{S}$, if $(i,j)\le(i',j')$ (in lexical order), we have*

$$V_{\theta^{(i,j)}}(s)\le V_{\theta^{(i',j')}}(s).$$

We note the triangle inequality of variance.

Lemma 6. *For any $V_1,V_2:\mathcal{S}\to\mathbb{R}$, we have $\sqrt{\sigma_k[V_1+V_2]}\le\sqrt{\sigma_k[V_1]}+\sqrt{\sigma_k[V_2]}$ for all $k\in[K]$.*

The next is a key lemma showing a property of the convex combination of the standard deviations, which relies on the anchor condition.

Lemma 7. *For any $V : \mathcal{S} \to \mathbb{R}$ and $s, a \in \mathcal{S} \times \mathcal{A}$:*

$$\sum_{k \in [K]} \phi_k(s,a)\sqrt{\sigma_k[V]} \le \sqrt{\sigma_{s,a}(V)}.$$

Proof. Since $[\phi_1(s,a), \ldots, \phi_K(s,a)]$ is a vector of probability distribution (due to Assumption 2 without loss of generality), by Jensen's inequality we have,

$$\begin{aligned}\sum_k \phi_k(s,a)\sqrt{\sigma_k[V]} \le \sqrt{\sum_k \phi_k(s,a)\sigma_k[V]} &= \sqrt{\sum_k \phi_k(s,a)\Big[\sum_{s'} P(s'|s_k,a_k)V^2(s') - \Big(\sum_{s'} P(s'|s_k,a_k)V(s')\Big)^2\Big]} \\ &= \sqrt{\sum_{s'} P(s'|s,a)V^2(s') - \sum_k \phi_k(s,a)\Big[\Big(\sum_{s'} P(s'|s_k,a_k)V(s')\Big)^2\Big]}.\end{aligned}$$

By the Jensen's inequality again, we have

$$\sum_k \phi_k(s,a)\Big(\sum_{s'} P(s'|s_k,a_k)V(s')\Big)^2 \ge \Big(\sum_k \phi_k(s,a)\sum_{s'} P(s'|s_k,a_k)V(s')\Big)^2 = \Big(\sum_{s'} P(s'|s,a)V(s')\Big)^2.$$

Combining the above two equations, we complete the proof. □

D.3. Monotonicity Preservation

The next lemma illustrates, conditioning on $\mathcal{E}^{(i,j)}$ and $\mathcal{G}^{(i)}$, a monotonicity property is preserved throughout the inner loop.

Lemma 8 (Preservation of Monotonicity Property)**.** *Conditioning on the events $\mathcal{G}^{(i)}$, $\mathcal{E}^{(i,0)}, \mathcal{E}^{(i,1)}, \ldots, \mathcal{E}^{(i,j)}$, we have for all $s \in \mathcal{S}, j' \in [0,j]$,*

$$V_{\theta^{(i,j')}}(s) \le \mathcal{T}_{\pi_{\theta^{(i,j')}}} V_{\theta^{(i,j')}}[s] \le \mathcal{T} V_{\theta^{(i,j')}}[s] \le v^*(s). \tag{6}$$

Moreover, for any fixed policy π^*, we have, for $j' \in [j]$,

$$v^*(s) - V_{\theta^{(i,j')}}(s) \le \gamma P(\cdot|s,\pi^*(s))^\top (v^* - V_{\theta^{(i,j'-1)}}) + 2\gamma \sum_k \phi_k(s,\pi^*(s))\epsilon^{(i,j')}(k). \tag{7}$$

Proof.
Proof of (6) **by Induction**: We first prove the inequalities in (6) by induction on j'. The base case of $j' = 0$ holds by definition of $\mathcal{G}^{(i)}$.

Now assuming it holds for $j' - 1 \ge 0$, let us verify that (6) holds for j'. For any $s \in \mathcal{S}$, we rewrite the corresponding value function defined in (5) as follows:

$$V_{\theta^{(i,j')}}(s) = \max\big\{\max_a Q_{\theta^{(i,j')}}(s,a), V_{\theta^{(i,j'-1)}}(s)\big\}.$$

For any $s \in \mathcal{S}$, there are only two cases to make the above equation hold:

1. $V_{\theta^{(i,j')}}(s) = V_{\theta^{(i,j'-1)}}(s) \Rightarrow \max_a Q_{\theta^{(i,j')}}(s,a) < V_{\theta^{(i,j'-1)}}(s)$ and $\pi_{\theta^{(i,j')}}(s) = \pi_{\theta^{(i,j'-1)}}(s)$;
2. $V_{\theta^{(i,j')}}(s) = \max_a Q_{\theta^{(i,j')}}(s,a) \Rightarrow \max_a Q_{\theta^{(i,j')}}(s,a) \ge V_{\theta^{(i,j'-1)}}(s)$ and $\pi_{\theta^{(i,j')}}(s) = \arg\max_a Q_{\theta^{(i,j')}}(s,a)$.

We investigate the consequences of case 1. Since (6) holds for $j'-1$, we have $V_{\theta^{(i,j')}}(s) = V_{\theta^{(i,j'-1)}}(s) \le v^*(s)$. Moreover, since (6) holds for $j'-1$ and $\pi_{\theta^{(i,j')}}(s) = \pi_{\theta^{(i,j'-1)}}(s)$, we have

$$\begin{aligned}V_{\theta^{(i,j')}}(s) = V_{\theta^{(i,j'-1)}}(s) &\le \mathcal{T}_{\pi_{\theta^{(i,j')}}} V_{\theta^{(i,j'-1)}}[s] && \triangleright \text{ by induction hypothesis} \\ &\le \mathcal{T}_{\pi_{\theta^{(i,j')}}} V_{\theta^{(i,j')}}[s] && \triangleright \text{ by Lemma 5 and the monotonicity of } \mathcal{T}_\pi \\ &\le \mathcal{T} V_{\theta^{(i,j')}}[s].\end{aligned}$$

We now investigate the consequences of case 2. Notice that conditioning on $\mathcal{E}^{(i,0)}, \mathcal{E}^{(i,1)} \ldots, \mathcal{E}^{(i,j')}$ (by specifying the constant C appropriately), we can verify that,

$$\forall k \in [K]: \quad \overline{w}^{(i,j')}(k) := \Pi_{[0,H]}(w^{(i,j')}(k) - \epsilon^{(i,j')}(k)) \leq P(\cdot|s_k, a_k)^\top V_{\theta^{(i,j'-1)}},$$

where $H = (1-\gamma)^{-1}$. Thus, for any $a \in \mathcal{A}$,

$$Q_{\theta^{(i,j')}}(s,a) = r(s,a) + \gamma\phi(s,a)^\top \overline{w}^{(i,j')} \leq r(s,a) + \gamma \sum_{k\in[K]} \phi_k(s,a) P(\cdot|s_k,a_k)^\top V_{\theta^{(i,j'-1)}} = \overline{Q}_{\theta^{(i,j')}}(s,a).$$

Then we have

$$\begin{aligned} 0 \leq \max_a Q_{\theta^{(i,j')}}(s,a) &= Q_{\theta^{(i,j')}}(s, \pi_{\theta^{(i,j')}}(s)); \\ \max_a Q_{\theta^{(i,j')}}(s,a) \leq \overline{Q}_{\theta^{(i,j')}}(s, \pi_{\theta^{(i,j')}}(s)) &= \mathcal{T}_{\pi_{\theta^{(i,j')}}} V_{\theta^{(i,j'-1)}}[s]; \\ \max_a Q_{\theta^{(i,j')}}(s,a) \leq \max_a \overline{Q}_{\theta^{(i,j'-1)}}(s,a) &= \mathcal{T} V_{\theta^{(i,j'-1)}}[s]. \end{aligned} \tag{8}$$

As a result, we obtain

$$\begin{aligned} 0 \leq V_{\theta^{(i,j')}}(s) &= \max_a Q_{\theta^{(i,j')}}(s,a) \leq \mathcal{T}_{\pi_{\theta^{(i,j')}}(s)} V_{\theta^{(i,j'-1)}}[s] \\ &\leq \mathcal{T}_{\pi_{\theta^{(i,j')}}} V_{\theta^{(i,j')}}[s] \qquad \triangleright \text{ by Lemma 5 and the monotonicity of } \mathcal{T}_\pi \\ &\leq \mathcal{T} V_{\theta^{(i,j')}}[s]. \end{aligned}$$

We see that $0 \leq V_{\theta^{(i,j')}}(s) \leq \mathcal{T}_{\pi_{\theta^{(i,j')}}} V_{\theta^{(i,j')}}[s] \leq \mathcal{T} V_{\theta^{(i,j')}}[s]$ holds in both cases 1 and 2. Also note that since (6) holds for $j'-1$, we have $V_{\theta^{(i,j'-1)}} \leq v^*$. It follows from the monotonicity of the Bellman operator that

$$0 \leq V_{\theta^{(i,j')}}(s) \leq \mathcal{T}_{\pi_{\theta^{(i,j')}}} V_{\theta^{(i,j'-1)}}[s] \leq \mathcal{T}_{\pi_{\theta^{(i,j')}}} v^*[s] \leq v^*(s).$$

This completes the induction.

Proof of (7): Let π^* be some fixed optimal policy. For each $j' \in [j]$, by (5), we have

$$V_{\theta^{(i,j')}}(s) \geq \max_{a\in\mathcal{A}} Q_{\theta^{(i,j')}}(s,a) := \max_{a\in\mathcal{A}} \big[r(s,a) + \gamma\phi(s,a)^\top \overline{w}^{(i,j')}\big].$$

By definition of $\mathcal{E}^{(i,j')}$, we have

$$\forall k \in [K]: \quad \overline{w}^{(i,j')}(k) \geq w^{(i,j')}(k) - \epsilon^{(i,j')}(k) \geq P(\cdot|s_k,a_k)^\top V_{\theta^{(i,j'-1)}} - 2\epsilon^{(i,j')}(k).$$

Therefore,

$$V_{\theta^{(i,j')}}(s) \geq \max_a \Big[r(s,a) + \gamma \sum_k \phi_k(s,a)\big(P(\cdot|s_k,a_k)^\top V_{\theta^{(i,j'-1)}} - 2\epsilon^{(i,j')}(k)\big)\Big].$$

Hence,

$$\begin{aligned} v^*(s) - V_{\theta^{(i,j')}}(s) &\leq r^{\pi^*}(s) + \gamma P^{\pi^*}(\cdot|s)^\top v^* - \max_a \Big[r(s,a) + \gamma \sum_k \phi_k(s,a)\big(P(\cdot|s_k,a_k)^\top V_{\theta^{(i,j'-1)}} - 2\epsilon^{(i,j')}(k)\big)\Big] \\ &\leq r^{\pi^*}(s) + \gamma P^{\pi^*}(\cdot|s)^\top v^* - \Big[r(s,\pi^*(s)) + \gamma \sum_k \phi_k(s,\pi^*(s))\big(P(\cdot|s_k,a_k)^\top V_{\theta^{(i,j'-1)}} - 2\epsilon^{(i,j')}(k)\big)\Big] \\ &= \gamma P^{\pi^*}(\cdot|s)^\top (v^* - V_{\theta^{(i,j'-1)}}) + 2\gamma \sum_k \phi_k(s,\pi^*(s))\epsilon^{(i,j')}(k), \end{aligned}$$

where $P^{\pi^*}(\cdot|s) = P(\cdot|s,\pi^*(s))$ and we use the fact that $P^{\pi^*}(\cdot|s) = \sum_k \phi_k(s,\pi^*(s))P(\cdot|s_k,a_k)$ in the last equality. $\square$

D.4. Accuracy of Confidence Bounds

We show that the mini-batch sample sizes picked in Algorithm 2 are sufficient to control the error occurred in the inner-loop iterations, such that the events $\mathcal{E}^{(i,0)}, \mathcal{E}^{(i,1)}, \ldots, \mathcal{E}^{(i,R)}$ jointly happen with close-to-1 probability.

Lemma 9. *For $i = 0, 1, 2, \ldots, R'$,*

$$\Pr[\mathcal{E}^{(i,0)}, \mathcal{E}^{(i,1)}, \ldots, \mathcal{E}^{(i,R)} | \mathcal{G}^{(i)}] \geq 1 - \delta/R'.$$

Proof. We analyze each event separately.

Probability of $\mathcal{E}^{(i,0)}$: We first show that $\Pr[\mathcal{E}^{(i,0)}|\mathcal{G}^{(i)}] \geq 1 - \delta/(RR')$. Note that $V_{\theta^{(i,0)}}(s) \in [0, \frac{1}{1-\gamma}]$ is determined by the samples obtained before the outer-iteration i starts, therefore samples obtained in iteration (i,j) for $j \geq 0$ are independent with $V_{\theta^{(i,0)}}$. Hence, conditioning on $\mathcal{G}^{(i)}$, for a fixed $\delta \in (0,1)$ and $k \in [K]$, by the Bernstein's and the Hoeffding's inequalities, for some constant $c_1 > 0$, the following two inequalities hold with probability at least $1-\delta$,

$$\left|w^{(i,0)}(k) - P(\cdot|s_k, a_k)^\top V_{\theta^{(i,0)}}\right| \leq \min\left\{c_1\sqrt{\frac{\log[\delta^{-1}]\sigma_k[V_{\theta^{(i,0)}}]}{m}} + \frac{c_1 \log\delta^{-1}}{(1-\gamma)m}, \quad c_1(1-\gamma)^{-1}\cdot\sqrt{\frac{\log[\delta^{-1}]}{m}}\right\}$$
$$\left|z^{(i,0)}(k) - P(\cdot|s_k, a_k)^\top V^2_{\theta^{(i,0)}}\right| \leq c_1(1-\gamma)^{-2}\cdot\sqrt{\frac{\log[\delta^{-1}]}{m}},$$

where we recall the notation $\sigma_k[V_{\theta^{(i,0)}}] = P(\cdot|s_k,a_k)^\top V^2_{\theta^{(i,0)}} - [P(\cdot|s_k,a_k)^\top V_{\theta^{(i,0)}}]^2 \leq (1-\gamma)^{-2}$ (see D.1). Conditioning on the preceding two inequalities, we have

$$\left|\sigma_k[V_{\theta^{(i,0)}}] - \sigma^{(i,0)}(k)\right| = \left|\sigma_k[V_{\theta^{(i,0)}}] - \left(z^{(i,0)}(k) - w^{(i,0)}(k)^2\right)\right| \leq c_1'(1-\gamma)^{-2}\cdot\sqrt{\frac{\log[\delta^{-1}]}{m}}$$

for some constant c_1', where $\sigma^{(i,0)}(k) := z^{(i,0)} - (w^{(i,0)}(k))^2$ according to tep 13 of Alg. 2. Thus, $\sigma_k[V_{\theta^{(i,0)}}] \leq \sigma^{(i,0)}(k) + c_1'(1-\gamma)^{-2}\cdot\sqrt{\frac{\log[\delta^{-1}]}{m}}$. We further obtain,

$$\sqrt{\sigma^{(i,0)}(k) + c_1'(1-\gamma)^{-2}\cdot\sqrt{\frac{\log[\delta^{-1}]}{m}}} \leq \sqrt{\sigma^{(i,0)}(k)} + \left(c_1'^2(1-\gamma)^{-4}\frac{\log[\delta^{-1}]}{m}\right)^{1/4}.$$

By plugging in $\delta \leftarrow \delta/(KR'R)$, we have,

$$\begin{aligned}\left|w^{(i,0)}(k) - P(\cdot|s_k,a_k)^\top V_{\theta^{(i,0)}}\right| &\leq c_1\sqrt{\frac{\log[KRR'\delta^{-1}]\sigma_k[V_{\theta^{(i,0)}}]}{m}} + \frac{c_1\log(KRR'\delta^{-1})}{(1-\gamma)m}\\ &\leq \Theta\left[\sqrt{\frac{\log[R'RK\delta^{-1}]\cdot\sigma^{(i,0)}(k)}{m}} + \frac{\log[R'RK\delta^{-1}]}{(1-\gamma)m^{3/4}}\right]\\ &= \epsilon^{(i,0)}(k)\end{aligned}$$

with probability at least $1-\delta/(KR'R)$, where $\epsilon^{(i,0)}(k)$ is defined in Step 13 of Algorithm 2. Since $\sigma^{(i,0)}(k) \leq \sigma_k[V_{\theta^{(i,0)}}] + c_1'(1-\gamma)^{-2}\cdot\sqrt{\frac{\log[\delta^{-1}]}{m}}$, we further have

$$\epsilon^{(i,0)}(k) \leq \Theta\left[\sqrt{\log(RR'K\delta^{-1})\sigma_k[V_{\theta^{(i,0)}}]/m} + \left((1-\gamma)^{-4}\frac{\log[RR'K\delta^{-1}]^4}{m^3}\right)^{1/4}\right].$$

Therefore, by applying an union bound over all $k \in [K]$, we have

$$\Pr[\mathcal{E}^{(i,0)}|\mathcal{G}^{(i)}] \geq 1 - \delta/(RR').$$

Reminder that if $\mathcal{E}^{(i,0)}$ happens, then $w^{(i,0)} - \epsilon^{(i,0)} \leq P(\cdot|s_k,a_k)^\top V_{\theta^{(i,0)}}$.

Probability of $\mathcal{E}^{(i,j)}$ by Induction: We now prove by induction that

$$\Pr[\mathcal{E}^{(i,j)}|\mathcal{E}^{(i,j-1)},\mathcal{E}^{(i,j-2)},\ldots,\mathcal{E}^{(i,0)},\mathcal{F}^{(i)}] \geq 1-\delta/(RR'). \tag{9}$$

For the base case $j=1$, we have

$$w^{(i,1)} = w^{(i,0)} \quad \text{and} \quad \epsilon^{(i,1)} = \epsilon^{(i,0)} + \Theta(1-\gamma)^{-1}2^{-i}\sqrt{\log(RR'K/\delta)},$$

therefore $\Pr[\mathcal{E}^{(i,1)}|\mathcal{E}^{(i,0)},\mathcal{G}^{(i)}] = 1$. Now consider j. Conditioning on $\mathcal{E}^{(i,j-1)},\mathcal{E}^{(i,j-2)},\ldots,\mathcal{E}^{(i,0)},\mathcal{F}^{(i)}$, we have with probability at least $1-\delta$,

$$\begin{aligned}
&\Big|\frac{1}{m_1}\sum_{\ell=1}^{m_1}\Big(V_{\theta^{(i,j-1)}}(x_k^{(\ell)}) - V_{\theta^{(i,0)}}(x_k^{(\ell)})\Big) - P(\cdot|s_k,a_k)^\top\Big(V_{\theta^{(i,j-1)}} - V_{\theta^{(i,0)}}\Big)\Big| \\
&\leq c_2\max_s |V_{\theta^{(i,j-1)}}(s) - V_{\theta^{(i,0)}}(s)|\cdot\sqrt{\frac{\log(\delta^{-1})}{m_1}} \\
&\leq c_2\max_s |v^*(s) - V_{\theta^{(i,0)}}(s)|\cdot\sqrt{\log(\delta^{-1})/m_1} && \triangleright\ V_{\theta^{(i,0)}} \leq V_{\theta^{(i,j-1)}} \leq v^* \\
&\leq c_2 2^{-i}(1-\gamma)^{-1}\cdot\sqrt{\log(\delta^{-1})/m_1}. && \triangleright\ \text{By definition of } \mathcal{G}^{(i)}
\end{aligned}$$

Letting $\delta \leftarrow \delta/(RR'K)$ and applying a union bound over $k\in[K]$, we obtain (9).

Probability of Joint Events: Finally, we have that

$$\begin{aligned}
\Pr[\mathcal{E}^{(i,0)}\cap\mathcal{E}^{(i,1)}\ldots\cap\mathcal{E}^{(i,R)}|\mathcal{G}^{(i)}] &= \Pr[\mathcal{E}^{(i,0)}|\mathcal{G}^{(i)}]\Pr[\mathcal{E}^{(i,1)}|\mathcal{E}^{(i,0)},\mathcal{G}^{(i)}]\ldots\Pr[\mathcal{E}^{(i,R)}|\mathcal{E}^{(i,0)},\mathcal{E}^{(i,1)},\ldots,\mathcal{E}^{(i,R-1)},\mathcal{G}^{(i)}] \\
&\geq 1-\delta/R'.
\end{aligned}$$

□

Lemma 10 (Upper Bound of $\epsilon^{(i,j)}(k)$)**.** *Conditioning on the events $\mathcal{F}^{(i)}$, $\mathcal{E}^{(i,0)},\mathcal{E}^{(i,1)},\ldots,\mathcal{E}^{(i,j)}$, we have, for all $k\in[K]$*

$$\epsilon^{(i,j)}(k) \leq C\left[\sqrt{\frac{\log(R'RK\delta^{-1})\sigma_k[v^*]}{m}} + \frac{\log(R'RK\delta^{-1})}{(1-\gamma)m^{3/4}} + 2^{-i}\sqrt{\frac{\log(R'RK\delta^{-1})}{(1-\gamma)^2 m_1}}\right]$$

for some universal constant $C>0$.

Proof. Conditioning on $\mathcal{F}^{(i)}$, $\mathcal{E}^{(i,0)},\mathcal{E}^{(i,1)},\ldots,\mathcal{E}^{(i,j)}$, we have

$$\begin{aligned}
\epsilon^{(i,0)}(k) &\leq c_1\left[\sqrt{\frac{\log(R'RK\delta^{-1})\sigma_k[V_{\theta^{(i,0)}}]}{m}} + \frac{\log(R'RK\delta^{-1})}{(1-\gamma)m^{3/4}}\right] \\
&\leq c_1'\left[\sqrt{\frac{\log(R'RK\delta^{-1})\sigma_k[v^*]}{m}} + \frac{\log(R'RK\delta^{-1})}{(1-\gamma)m^{3/4}} + 2^{-i}\sqrt{\frac{\log(R'RK\delta^{-1})}{(1-\gamma)^2 m}}\right],
\end{aligned}$$

for some generic constants c_1, c_1', where we use the fact that $\|V_{\theta^{(i,0)}} - v^*\|_\infty \leq 2^{-i}/(1-\gamma)$ and the triangle inequality. Using the definition of $\epsilon^{(i,j)}$ and the fact $m_1 \leq m$, we have

$$\begin{aligned}
\epsilon^{(i,j)}(k) &= \epsilon^{(i,0)}(k) + c_2 2^{-i}\sqrt{\frac{\log(R'RK\delta^{-1})}{(1-\gamma)^2 m_1}} \\
&\leq c_2'\left[\sqrt{\frac{\log(R'RK\delta^{-1})\sigma_k[v^*]}{m}} + \frac{\log(R'RK\delta^{-1})}{(1-\gamma)m^{3/4}} + 2^{-i}\sqrt{\frac{\log(R'RK\delta^{-1})}{(1-\gamma)^2 m_1}}\right],
\end{aligned}$$

for some generic constants c_2, c_2', where we use the fact that $m \geq m_1$. This concludes the proof. □

D.5. Error Accumulation in One Outer Iteration

Lemma 11. *For $i = 0, 1, 2, \ldots, R'$, $\Pr[\mathcal{G}^{(i+1)}|\mathcal{G}^{(i)}] \geq 1 - \delta/(R'+1)$.*

Proof of Lemma 11. Conditioning on $\mathcal{G}^{(i)}$, suppose that the events $\mathcal{E}^{(i,0)}, \mathcal{E}^{(i,1)}, \ldots, \mathcal{E}^{(i,R)}$ all happen, which has probability at least $1 - \delta/R'$ according to Lemma 9. For any $s \in \mathcal{S}$, we analyze the total error accumulated in the i-th outer iteration:

$$
\begin{aligned}
v^*(s) - V_{\theta^{(i,j)}}(s) &\leq \gamma P^{\pi^*}(\cdot|s)^\top (v^* - V_{\theta^{(i,j-1)}}) + 2\gamma \sum_k \phi_k(s, \pi^*(s))\epsilon^{(i,j)}(k) \qquad \triangleright \text{ Lemma 8} \\
&\leq \gamma^2 \sum_{s'} P^{\pi^*}(s'|s)^\top P^{\pi^*}(\cdot|s')^\top (v^* - V_{\theta^{(i,j-2)}}) + 2\gamma^2 P^{\pi^*}(\cdot|s)^\top \sum_k \phi_k(\cdot, \pi^*(\cdot))\epsilon^{(i,j-1)}(k) \\
&\quad + 2\gamma \sum_k \phi_k(s, \pi^*(s))\epsilon^{(i,j)}(k) \qquad \triangleright \text{ applying Lemma 8 again on } v^* - V_{\theta^{(i,j-1)}} \\
&\leq \ldots \qquad \triangleright \text{ applying Lemma 8 recursively} \\
&\leq \gamma^j [(P^{\pi^*})^j (v^* - V_{\theta^{(i,0)}})](s) + 2\sum_{j'=0}^{j-1} \gamma^{j'+1} \sum_{k,s'} \left(P^{\pi^*}\right)^{j'}_{s,s'} \phi_k(s', \pi^*(s'))\epsilon^{(i,j-j')}(k) \\
&\leq \gamma^j (1-\gamma)^{-1} + C\sum_{j'=0}^{j-1} \gamma^{j'+1} \left[\frac{\log(R'RK\delta^{-1})}{(1-\gamma)m^{3/4}} + 2^{-i}\sqrt{\frac{\log(R'RK\delta^{-1})}{(1-\gamma)^2 m_1}}\right] \\
&\quad + C\sum_{j'=0}^{j-1} \gamma^{j'+1} \sum_{s'} \left(P^{\pi^*}\right)^{j'}_{s,s'} \cdot \sum_k \phi_k(s', \pi^*(s'))\sqrt{\frac{\log(R'RK\delta^{-1})\sigma_k[v^*]}{m}} \\
&\qquad \triangleright \text{ using } \|v^* - V_{\theta^{(i,0)}}\|_\infty \leq \frac{1}{1-\gamma} \text{ and the upperbound of } \epsilon^{(i,j)} \text{ (Lemma 10)} \\
&\leq \gamma^j (1-\gamma)^{-1} + C\sum_{j'=0}^{j-1} \gamma^{j'+1} \left[\frac{\log(R'RK\delta^{-1})}{(1-\gamma)m^{3/4}} + 2^{-i}\sqrt{\frac{\log(R'RK\delta^{-1})}{(1-\gamma)^2 m_1}}\right] \\
&\quad + C\sum_{j'=0}^{j-1} \gamma^{j'+1} \sum_{s'} \left(P^{\pi^*}\right)^{j'}_{s,s'} \cdot \sqrt{\frac{\log(R'RK\delta^{-1})\sigma_{s',\pi^*(s')}[v^*]}{m}} \\
&\qquad \triangleright \text{ applying Lemma 7} \\
&= \gamma^j (1-\gamma)^{-1} + C\frac{1-\gamma^j}{1-\gamma} \cdot \left[\frac{\log(R'RK\delta^{-1})}{(1-\gamma)m^{3/4}} + 2^{-i}\sqrt{\frac{\log(R'RK\delta^{-1})}{(1-\gamma)^2 m_1}}\right] + \\
&\quad C\sum_{j'=0}^{j-1} \gamma^{j'+1} \sum_{s'} \left(P^{\pi^*}\right)^{j'}_{s,s'} \cdot \sqrt{\frac{\log(R'RK\delta^{-1})\sigma_{s',\pi^*(s')}[v^*]}{m}},
\end{aligned}
$$

where C is a generic constant. By Lemma C.1 of (Sidford et al., 2018a) (a form of law of total variance for the Markov chain under π^*), we have,

$$
\sum_{j'=0}^{j-1} \gamma^{j'+1} \sum_{s'} \left(P^{\pi^*}\right)^{j'}_{s,s'} \sqrt{\sigma_{s',\pi^*(s')}[v^*]} \leq C'\sqrt{(1-\gamma)^{-3}}
$$

for some generic constant C'. Combining the above equations, and setting

$$
m = C''\frac{1}{\epsilon^2} \cdot \frac{\log(R'RK\delta^{-1})^{4/3}}{(1-\gamma)^3} \quad \text{and} \quad m_1 = C'' \cdot \frac{\log(R'RK\delta^{-1})}{(1-\gamma)^2},
$$

$R \geq \Theta[i \cdot (1-\gamma)^{-1}]$ and $2^{-i}/(1-\gamma) \geq \Theta(\epsilon)$ for some generic constant C'', we can make the accumulated error as small as

$$
v^*(s) - V_{\theta^{(i,R)}}(s) \leq c2^{-i}/(1-\gamma)
$$

for some $c > 0$. Since $V_{\theta(i+1,0)}(s) = V_{\theta(i,R)}(s)$ together with the monotonicity properties shown in Lemma 8, we obtain that conditioning on $\mathcal{G}^{(i)}, \mathcal{E}^{(i,0)}, \mathcal{E}^{(i,1)}, \ldots, \mathcal{E}^{(i,R)}$, the event $\mathcal{G}^{(i+1)}$ happens with probability 1. $\square$

D.6. Proof of Theorem 3

Proof of Theorem 3. Conditioning on $\mathcal{G}^{(R')}$, we have

$$\forall s \in \mathcal{S}: \quad 0 \le v^*(s) - V_{\theta(R',R)}(s) \le 2^{-R'}/(1-\gamma).$$

Since $R' = \Theta(\log[\epsilon^{-1}(1-\gamma)^{-1}])$, we have $|v^*(s) - V_{\theta(R',R)}(s)| \le \epsilon$. Moreover, we have

$$v^*(s) - \epsilon \le V_{\theta(R',R)}(s) \le \mathcal{T}_{\pi_{\theta(R',R)}} V_{\theta(R',R)}[s] \le v^{\pi_{\theta(R',R)}}[s] \le v^*(s),$$

where the third inequality follows from monotonicity of $\mathcal{T}_{\pi(R',R)}$. Therefore $\pi_{\theta(R',R)}$ is an ϵ-optimal policy from any initial state s. Notice that $\Pr[\mathcal{G}^{(i)}|\mathcal{G}^{(i-1)}] \ge 1 - \delta/R'$, we have $\Pr[\mathcal{G}^{(R')}] \ge \Pr[\mathcal{G}^{(R')} \cap \mathcal{G}^{(R-1)} \cap \ldots \mathcal{G}^{(0)}] \ge 1 - \delta$. Finally, one can show the main result by counting the number of samples needed by the algorithm. $\square$