Supplementary Material

A Proof of Theorem 1

In this section, we present the proof of Theorem 1. We first introduce and recall some necessary notations and assumptions. Then, we present some auxiliary lemmas and their proofs. Finally, we combine the lemmas to prove the main result.

Notations: Define $P(\nu)$ as distribution over states such that $(s', x') \sim P(\nu) \Leftrightarrow (s, x, a) \sim \nu, s' \sim P(s'|s, a), x' = g(s, a, s')$. In other words, it is the distribution of the next state if the state action pair follows ν. For $f : \mathcal{S} \times \mathcal{X} \times \mathcal{A} \to \mathbb{R}, \nu \in \Delta(\mathcal{S} \times \mathcal{X} \times \mathcal{A})$, where $\Delta(\cdot)$ is the probability simplex, and $p > 1$, define $\|f\|_{p,\nu} = (\mathbb{E}_{(s,x,a)\sim\nu}[|f(s, x, a)|^p])^{1/p}$. Define $\pi_{f,f'}(s, x) := \arg\min_{a\in\mathcal{A}} \min\{f(s, x, a), f'(s, x, a)\}$.

Recall that in Alg. 2, at iteration k,

$$\mathcal{L}_{\hat{D}_k}(f; f_{k-1}) = \frac{1}{|\hat{D}_k|} \sum_{(s,x,a,r,s',x')\in\hat{D}_k} (f(s, x, a) - r - \gamma V_{f_{k-1}}(s', x'))^2,$$

where

$$f_k := \arg\min_{f\in\mathcal{F}} \mathcal{L}_{\hat{D}_k}(f; f_{k-1}) \quad \text{and} \quad V_{f_k}(s, x) := \min_{a\in\mathcal{A}} f_k(s, x, a).$$

Assumptions:

Assumption A.1. *(Realizability) For the optimal policy, $Q^* \in \mathcal{F}$.*

Assumption A.2. *(Completeness) For the policy π to be evaluated, $\forall f \in \mathcal{F}, \mathcal{T}f \in \mathcal{F}$, where $\mathcal{T} : \mathbb{R}^{\mathcal{S}\times\mathcal{X}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{X}\times\mathcal{A}}$ is the Bellman update operator, $\forall f$:*

$$(\mathcal{T}f)(s, x, a) := R(s, x, a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[V_f(s', x' = g(s, x, a, s'))].$$

We say a distribution $\nu \in \Delta(\mathcal{S} \times \mathcal{A})$ is admissible in MDP $(\mathcal{S}, \mathcal{X}, \mathcal{A}, P, R, \gamma)$, if there exist $h \geq 0$ and a policy π such that $\nu(s, a) = \sum_{x\in\mathcal{X}} \Pr(s_h = s, x_h = x, a_h = a|s_0, x_0, \pi)$. The following assumption is imposed to limit the distribution shift. Note that "admissible" is defined on the stochastic state and action. Later, we will also abuse the notation and call $\nu \in \Delta(\mathcal{S} \times \mathcal{X} \times \mathcal{A})$ is admissible if $\nu(s, a) = \sum_x \nu(s, x, a)$ is admissible.

Assumption A.3. *For a data distribution μ, we assume that there exists $C < \infty$ such that for any admissible v and any $(s, a) \in \mathcal{S} \times \mathcal{A}$,*

$$\frac{\nu(s, a)}{\mu(s, a)} \leq C,$$

where $\nu(s, a) = \sum_{x\in\mathcal{X}} \nu(s, x, a)$ and $\mu(s, a) = \sum_{x\in\mathcal{X}} \mu(s, x, a)$.

Assumptions A.2-A.3 are standard assumptions in batch reinforcement learning (Chen and Jiang, 2019). However, in assumption A.3, we only require the data coverage of stochastic states which is the fundamental difference.

Assumption A.4. *We assume that there exists a set $\mathcal{B}$ of typical pseudo-stochastic states such that the distributions $\beta_k(x), k = 1, \ldots, K$ used for augmenting virtual samples satisfy $\beta_k(x) \geq \sigma_1, \forall x \in \mathcal{B}$. We also assume that the marginal distribution over using a reasonable policy π, that is, $d^\pi_{\eta_0}(s, x) := (1 - \gamma)\sum_{t=0}^{\infty} \gamma^{t-1} \Pr(s_t = s, x_t = x|(s_0, x_0) \sim \eta_0, \pi)$ satisfies $\sum_{s\in\mathcal{S}} \sum_{x\in\mathcal{X},x\notin\mathcal{B}} d^\pi_{\eta_0}(s, x) \leq \sigma_2$, where η_0 is the initial distribution. Furthermore, if we have for a distribution η of the states satisfying $\sum_{s\in\mathcal{S}} \sum_{x\in\mathcal{X},x\notin\mathcal{B}} \eta(s, x) \leq \sigma_2$, then under any reasonable policy π, the marginal distribution $\eta^\pi_h(s, x) := \Pr(s_h = s, x_h = x|(s_0, x_0) \sim \eta, \pi)$ satisfies $\sum_{s\in\mathcal{S}} \sum_{x\in\mathcal{X},x\notin\mathcal{B}} \eta^\pi_h(s, x) \leq c_0\sigma_2, \forall h > 0$, for some constant $c_0 \geq 1$. In particular, all the policies at each iteration, the optimal policy and their joint policies are assumed to be reasonable policies.*

We remark that in a queuing network, under any stable policy, the queue distribution has an exponential tail; in other words, large queue lengths occur with a small probability. In such a case, we can use a uniform distribution for pseudo-stochastic states in set $\mathcal{B}$ to guarantee that $\sigma_1 = \Theta\left(\frac{1}{\log(1/\sigma_2)})\right)$. Therefore, if we choose $\sigma_2 = \frac{1}{n}$, then $\sigma_1 = \frac{1}{\log n}$.

Auxiliary Lemmas

In the following lemma, we will show that when all admissible distributions are not far away from the data distribution μ over stochastic state $\mathcal{S}$ and action $\mathcal{A}$, we can have a good coverage of $\mathcal{S} \times \mathcal{X} \times \mathcal{A}$ by generating virtual samples.

For a given dataset D of size $|D| = n$ and data distribution μ, let $\bar{\mu}_\beta$ denote the expected distribution of the the state action pair (s, x, a) in the combined datase after using Algorithm 2 with a virtual sample distribution $\beta(x)$.

Lemma A.1. *Given a virtual sample generating distribution $\beta(x)$ of the pseudo-stochastic state, if $\beta(x) \geq \sigma_1, \forall x \in \mathcal{B}$. Then given any admissible distribution ν, then under Assumptions A.3, we have for any $(s, x, a) \in \mathcal{S} \times \mathcal{X} \times \mathcal{A}, x \in \mathcal{B}$,*

$$\frac{\nu(s,x,a)}{\bar{\mu}_\beta(s,x,a)} \leq \frac{(m+1)C}{m\sigma_1},$$

where

$$\bar{\mu}_\beta(s,x,a) = \mu(s,x,a)\frac{n}{nm+n} + \sum_{\hat{x}\in\mathcal{X}} \mu(s,\hat{x},a)\left(\beta(x)\frac{nm}{nm+n}\right) \tag{8}$$

Proof. Given the data distribution μ, we know that the real samples are drawn according to $\mu(s, x, a)$. Then

$$\begin{aligned}
\frac{\nu(s,x,a)}{\bar{\mu}_\beta(s,x,a)} \leq& \frac{\sum_{\hat{x}\in\mathcal{X}} \nu(s,\hat{x},a)}{\bar{\mu}_\beta(s,x,a)} = \frac{\sum_{\hat{x}\in\mathcal{X}} \nu(s,\hat{x},a)}{\sum_{\hat{x}\in\mathcal{X}} \mu(s,\hat{x},a)} \times \frac{\sum_{\hat{x}\in\mathcal{X}} \mu(s,\hat{x},a)}{\bar{\mu}_\beta(s,x,a)} \\
=& \frac{\nu(s,a)}{\mu(s,a)} \times \frac{\sum_{\hat{x}\in\mathcal{X}} \mu(s,\hat{x},a)}{\bar{\mu}_\beta(s,x,a)} \\
\leq_{(1)}& C \times \frac{\mu(s,a)}{\bar{\mu}_\beta(s,x,a)} \\
=_{(2)}& C\left(\frac{\mu(s,a)}{\frac{\mu(s,x,a)}{m+1} + \frac{m}{m+1}\cdot\sum_{\hat{x}\in\mathcal{X}} \mu(s,\hat{x},a)\beta(x)}\right) \\
\leq& C\left(\frac{(m+1)\mu(s,a)}{m\sum_{\hat{x}\in\mathcal{X}} \mu(s,\hat{x},a)\beta(x)}\right) \\
\leq& C\left(\frac{(m+1)\mu(s,a)}{m\sum_{\hat{x}\in\mathcal{B}} \mu(s,\hat{x},a)\beta(x)}\right) \\
\leq& C\cdot\frac{m+1}{m}\cdot\frac{1}{\sigma_1},
\end{aligned}$$

where the inequality (1) holds because of Assumption A.3, the equality (2) holds by substituting equation (8) and the last inequality is true because the fact that $\beta(x) \geq \sigma_1, \forall x \in \mathcal{B}$. $\square$

The next lemma transforms the norm in terms of distribution ν to distribution $\bar{\mu}_\beta$ (Eq. (8)).

Lemma A.2. *Let ν be any admissible distribution, $\bar{\mu}_\beta$ denote the new data distribution defined in Eq.* (8) *after generating virtual samples with $\beta(x)$. If $\beta(x) \geq \sigma_1, \forall x \in \mathcal{B}$, then under Assumption A.3, for any function $f : \mathcal{S} \times \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, we have $\|f\|_{\underline{2},\nu} \leq \sqrt{\frac{m+1}{m}\frac{C}{\sigma_1}}\|f\|_{\underline{2},\bar{\mu}_\beta}$, where*

$$\|f\|_{\underline{2},\nu} = \left(\sum_{(s,x,a)\in\mathcal{S}\times\mathcal{X}\times\mathcal{A}, x\in\mathcal{B}} |f(s,x,a)|^2\nu(s,x,a)\right)^{1/2}.$$

Proof. For any function f, we have

$$\|f\|_{\underline{2},\nu} = \left(\sum_{(s,x,a)\in\mathcal{S}\times\mathcal{X}\times\mathcal{A}, x\in\mathcal{B}} |f(s,x,a)|^2\nu(s,x,a)\right)^{1/2}$$

$$\le \left(\sum_{(s,x,a)\in\mathcal{S}\times\mathcal{X}\times\mathcal{A}, x\in\mathcal{B}} |f(x,x,a)|^2 \bar{\mu}_\beta(s,x,a) \frac{(m+1)C}{m\sigma_1} \right)^{1/2}$$
$$\le \sqrt{\frac{(m+1)C}{m\sigma_1}} \|f\|_{\underline{2},\bar{\mu}_\beta},$$

where the first inequality is a result of Lemma A.1. □

Lemma A.3. *Consider two functions $f, f' : \mathcal{S}\times\mathcal{X}\times\mathcal{A} \to \mathbb{R}$ and define a policy $\pi_{f,f'}(s,x) := \arg\min_{a\in\mathcal{A}} \min\{f(s,x,a), f'(s,x,a)\}$. Then we have $\forall \nu \in \Delta(\mathcal{S}\times\mathcal{X}\times\mathcal{A})$,*

$$\|V_f - V_{f'}\|_{2,P(\nu)} = \|f - f'\|_{2,P(\nu)\times\pi_{f,f'}}. \tag{9}$$

Proof.

$$\begin{aligned}
&\|V_f - V_{f'}\|^2_{2,P(\nu)} \\
=& \sum_{(s,x,a)\in\mathcal{S}\times\mathcal{X}\times\mathcal{A}} \sum_{s'\in\mathcal{S}} P(s'|s,a) \left(\min_{a'\in\mathcal{A}} f(s', g(s,x,a,s'), a') - \min_{a'\in\mathcal{A}} f'(s', g(s,x,a,s'), a') \right)^2 \\
\le& \sum_{(s,x,a)\in\mathcal{S}\times\mathcal{X}\times\mathcal{A}} \sum_{s'\in\mathcal{S}} P(s'|s,a) \left(f(s', g(s,x,a,s'), \pi_{f,f'}(s', g(s,x,a,s'))) \right. \\
&\quad \left. - f'(s', g(s,x,a,s'), \pi_{f,f'}(s', g(s,x,a,s'))) \right)^2 \\
=& \|f - f'\|^2_{2,P(v)\times\pi_{f,f'}}.
\end{aligned}$$

□

Lemma A.4. *Under Assumptions A.3 and A.4, for any admissible distribution $\nu \in \Delta(\mathcal{S}\times\mathcal{X}\times\mathcal{A})$, and a data distribution $\bar{\mu}_\beta$ associated with a virtual sample distribution $\beta(x)$, define $P(\nu)$ as a distribution generated as $s'_i \sim P(\nu)$, then for any policy π, and $f, f' : \mathcal{S}\times\mathcal{X}\times\mathcal{A} \to \mathbb{R}$, we have*

$$\|f - Q^*\|_{\underline{2},\nu} \le \sqrt{\frac{(m+1)C}{m\sigma}} \|f - \mathcal{T}f'\|_{\underline{2},\bar{\mu}_\beta} + \gamma \|f' - Q^*\|_{2,P(\nu)\times\pi_{f',Q^*}} \tag{10}$$

Proof.

$$\begin{aligned}
\|f - Q^*\|_{\underline{2},\nu} &= \|f - \mathcal{T}f' + \mathcal{T}f' - Q^*\|_{\underline{2},\nu} \\
&\le_{(1)} \|f - \mathcal{T}f'\|_{\underline{2},\nu} + \|\mathcal{T}f' - Q^*\|_{\underline{2},\nu} \\
&\le_{(2)} \sqrt{\frac{(m+1)C}{m\sigma}} \|f - \mathcal{T}f'\|_{\underline{2},\bar{\mu}_\beta} + \|\mathcal{T}f' - Q^*\|_{\underline{2},\nu} \\
&\le_{(3)} \sqrt{\frac{(m+1)C}{m\sigma}} \|f - \mathcal{T}f'\|_{\underline{2},\bar{\mu}_\beta} + \gamma \|V_{f'} - V^*\|_{2,P(\nu)} \\
&\le \sqrt{\frac{(m+1)C}{m\sigma}} \|f - \mathcal{T}f'\|_{\underline{2},\bar{\mu}_\beta} + \gamma \|f' - Q^*\|_{2,P(\nu)\times\pi_{f',Q^*}},
\end{aligned}$$

where inequality (1) holds because of triangle inequality, inequality (2) comes from lemma A.2, inequality (3) holds because

$$\begin{aligned}
\|\mathcal{T}f' - Q^*\|^2_{2,\nu} = \|\mathcal{T}^* f' - \mathcal{T}Q^*\|^2_{2,\nu} &= \mathbb{E}_{(s,x,a)\sim\nu}\left[((\mathcal{T}f')(s,x,a) - (\mathcal{T}Q^*)(s,x,a))^2 \right] \\
&= \mathbb{E}_{(s,x,a)\sim\nu}\left[\left(\gamma \mathbb{E}_{s'\sim P(\cdot|s,a)} \left[V_{f'}(s', g(s,x,a,s')) - V^*(s', g(s,x,a,s')) \right] \right)^2 \right] \\
&\le \gamma^2 \mathbb{E}_{(s,x,a)\sim\nu, s'\sim P(\cdot|s,a)} \left[(V_{f'}(s', g(s,x,a,s')) - V^*(s', g(s,x,a,s')))^2 \right] \\
&= \gamma^2 \mathbb{E}_{s'\sim P(\nu)} \left[(V_{f'}(s', g(s,x,a,s')) - V^*(s', g(s,x,a,s')))^2 \right] \\
&= \gamma^2 \|V_{f'} - V^*\|^2_{2,P(\nu)},
\end{aligned}$$

and the last inequality holds due to Lemma A.3.

□

Lemma A.5. *For a given data sample (s, x, a, r, s', a') generated from a data distribution μ, such that $(s, x, a) \sim \mu, s' \sim P(\cdot|s, a), x' = g(s, x, a, s')$, for any $f, f' \in \mathcal{F}$, define $V_f(s, x) = \min_{a'} f(s, x, a')$, then*

$$\begin{aligned}
&\mathbb{E}\left[(f(s,x,a) - r - \gamma V_{f'}(s',x'))^2\right] \\
=&\|f - \mathcal{T}f'\|_{2,\mu}^2 + \gamma^2 \mathbb{E}_{(s,x,a)\sim\mu}[Var(V_{f'}(s',x')|s,x,a)]
\end{aligned} \tag{11}$$

Proof.

$$\begin{aligned}
&\mathbb{E}\left[(f(s,x,a) - r - \gamma V_{f'}(s',x'))^2\right] \\
=&\mathbb{E}\left[(f(s,x,a) - (\mathcal{T}f')(s,x,a) + (\mathcal{T}f')(s,x,a) - (r + \gamma V_{f'}(s',x')))^2\right] \\
=&\mathbb{E}\left[(f(s,x,a) - (\mathcal{T}f')(s,x,a))^2\right] + \mathbb{E}\left[((\mathcal{T}f')(s,x,a) - (r + \gamma V_{f'}(s',x')))^2\right] \\
&+ \underbrace{2\mathbb{E}\left[(f(s,x,a) - (\mathcal{T}f')(s,x,a))\left((\mathcal{T}f')(s,x,a) - (r + \gamma V_{f'}(s',x'))\right)\right]}_{(1)=0} \\
=&\mathbb{E}\left[(f(s,x,a) - (\mathcal{T}f')(s,x,a))^2\right] + \gamma^2 \mathbb{E}_{(s,x,a)\sim\mu}[\mathrm{Var}(V_{f'}(s',x')|s,x,a)] \\
=&\|f - \mathcal{T}f'\|_{2,\mu}^2 + \gamma^2 \mathbb{E}_{(s,x,a)\sim\mu}[\mathrm{Var}(V_{f'}(s',x')|s,x,a)],
\end{aligned}$$

where the equation $(1) = 0$ because that condition on (s, x, a), we have f and $V_{f'}$ are independent. ☐

Lemma A.6. *Under Algorithm 2, at iteration k, we have*

$$\mathcal{L}_{\hat{\mu}_{\beta_k}}(f; f') - \mathcal{L}_{\hat{\mu}_{\beta_k}}(\mathcal{T}f; f') = \|f - \mathcal{T}f'\|_{2,\bar{\mu}_{\beta_k}}^2, \tag{12}$$

where $\mathcal{L}_{\hat{\mu}_{\beta_k}}(f; f') = \mathbb{E}[\mathcal{L}_{\hat{D}_k}(f; f')]$.

Proof. Recall that $\hat{D}_k = D \cup D_k$ and $|\hat{D}_k| = nm + n$. The expectation is w.r.t. the random draw of the dataset D and the random generation of dataset D_k with virtual sample distribution β_k. We know that

$$\begin{aligned}
\mathcal{L}_{\hat{D}_k}(f; f') =& \frac{1}{|\hat{D}_k|} \sum_{(s,x,a,r,s',x')\in D} (f(s,x,a) - r - \gamma V_{f'}(s',x'))^2 \\
&+ \frac{1}{|\hat{D}_k|} \sum_{(s,x,a,r,s',x')\in D_k} (f(s,x,a) - r - \gamma V_{f'}(s',x'))^2
\end{aligned}$$

Let $\mathcal{M}_k^{(s,x,a,s',x')}$ denote the set of virtual samples that are associated with the real sample (s, x, a, s', x') at iteration k. Then

$$\begin{aligned}
&\mathcal{L}_{\hat{\mu}_{\beta_k}}(f; f') := \mathbb{E}[\mathcal{L}_{\hat{D}_k}(f; f')] \\
=&\frac{n}{nm+n}\left(\|f - \mathcal{T}f'\|_{2,\mu}^2 + \gamma^2 \mathbb{E}_{(s,x,a)\sim\mu}[\mathrm{Var}(V_{f'}(s',x')|s,x,a)]\right) \qquad \text{(by using Lemma } A.5\text{)} \\
&+ \frac{1}{nm+n}\mathbb{E}\left[\sum_{(s,x,a,r,s',x')\in D} \sum_{(s,\bar{x},a,\bar{r},\bar{s}',\bar{x}')\in\mathcal{M}_k^{(s,x,a,r,s',x')}} (f(s,\bar{x},a) - \bar{r} - \gamma V_{f'}(\bar{s}',\bar{x}'))^2\right] \\
=&\frac{n}{nm+n}\left(\|f - \mathcal{T}f'\|_{2,\mu}^2 + \gamma^2 \mathbb{E}_{(s,x,a)\sim\mu}[\mathrm{Var}(V_{f'}(s',x')|s,x,a)]\right) \\
&+ \frac{1}{nm+n}\mathbb{E}\left[\sum_{(s,x,a,r,s',x')\in D} \mathbb{E}\left[\sum_{(s,\bar{x},a,\bar{r},\bar{s}',\bar{x}')\in\mathcal{M}_k^{(s,x,a,r,s',x')}} (f(s,\bar{x},a) - \bar{r} - \gamma V_{f'}(\bar{s}',\bar{x}'))^2 \,\middle|\, s,x,a,r,s',x'\right]\right] \\
=&\frac{n}{nm+n}\left(\|f - \mathcal{T}f'\|_{2,\mu}^2 + \gamma^2 \mathbb{E}_{(s,x,a)\sim\mu}[\mathrm{Var}(V_{f'}(s',x')|s,x,a)]\right)
\end{aligned}$$

$$
\begin{aligned}
&+\frac{m}{nm+n}\mathbb{E}\left[\sum_{(s,x,a,r,s',x')\in D}\sum_{\bar{x}\in\mathcal{X}}\beta_k(\bar{x})\left(f(s,\bar{x},a)-\bar{r}-\gamma V_{f'}(\bar{s}',\bar{x}')\right)^2\right]\\
=&\frac{n}{nm+n}\left(\|f-\mathcal{T}f'\|_{2,\mu}^2+\gamma^2\mathbb{E}_{(s,x,a)\sim\mu}[\mathrm{Var}(V_{f'}(s',x')|s,x,a)]\right)\\
&+\frac{mn}{nm+n}\sum_{(s,x,a)\in\mathcal{S}\times\mathcal{X}\times\mathcal{A}}\mu(s,x,a)\sum_{\bar{x}\in\mathcal{X}}\beta_k(\bar{x})\left(f(s,\bar{x},a)-\bar{r}-\gamma V_{f'}(\bar{s}',\bar{x}')\right)^2\\
=&\frac{n}{nm+n}\left(\|f-\mathcal{T}f'\|_{2,\mu}^2+\gamma^2\mathbb{E}_{(s,x,a)\sim\mu}[\mathrm{Var}(V_{f'}(s',x')|s,x,a)]\right)\\
&+\frac{mn}{nm+n}\sum_{(s,\bar{x},a)\in\mathcal{S}\times\mathcal{X}\times\mathcal{A}}\mu(s,a)\beta_k(\bar{x})\left(f(s,\bar{x},a)-\bar{r}-\gamma V_{f'}(\bar{s}',\bar{x}')\right)^2\\
=&\frac{n}{nm+n}\left(\|f-\mathcal{T}f'\|_{2,\mu}^2+\gamma^2\mathbb{E}_{(s,x,a)\sim\mu}[\mathrm{Var}(V_{f'}(s',x')|s,x,a)]\right)\\
&+\frac{nm}{nm+n}\left(\|f-\mathcal{T}f'\|_{2,\mu_k}^2+\gamma^2\mathbb{E}_{(s,x,a)\sim\mu_k}[\mathrm{Var}(V_{f'}(s',x')|s,x,a)]\right), \qquad \text{(by using Lemma } A.5)
\end{aligned}
$$

where $\bar{r} = R(s,\bar{x},a), \mu_k(s,x,a) = \sum_{x'\in\mathcal{X}}\mu(s,x',a)\beta_k(x) = \mu(s,a)\beta_k(x)$. Since we have $\bar{\mu}_{\beta_k}(s,x,a) = \frac{1}{m+1}\mu(s,x,a) + \frac{m}{m+1}\mu(s,a)\beta_k(x)$.

Therefore,

$$
\mathcal{L}_{\hat{\mu}_{\beta_k}}(f;f') - \mathcal{L}_{\hat{\mu}_{\beta_k}}(\mathcal{T}f';f') = \|f-\mathcal{T}f'\|_{2,\bar{\mu}_{\beta_k}}^2.
$$

□

The next lemma shows an upper bound on $\|f_{k+1}-\mathcal{T}f_k\|_{2,\bar{\mu}_{\beta_k}}^2$.

Lemma A.7. *Given the MDP $M = (\mathcal{S},\mathcal{X},P,R,\gamma)$, we assume that the Q−function classes $\mathcal{F}$ satisfies $\forall f\in\mathcal{F},\mathcal{T}f\in\mathcal{F}$. The dataset D is generated as: $(s,x,a)\sim\mu, r = R(s,x,a), s'\sim P(\cdot|s,a), x' = g(s,x,a,s')$, and the new dataset $\hat{D}_k = D\cup D_k$ is generated by following Alg. 1 with virtual sample generating distribution $\beta_k(x)$ at kth iteration. Then with probability at least $1-\delta,\forall f\in\mathcal{F}$, and $k = 0,\ldots,K$ we hvae*

$$
\|f_{k+1}-\mathcal{T}f_k\|_{2,\bar{\mu}_{\beta_k}}^2 \le 5\left(\frac{1}{n}+\frac{1}{m}\right)V_{\max}^2\log(nK|\mathcal{F}|^2/\delta) + \frac{3\delta V_{\max}^2}{n} \tag{13}
$$

Proof. Using Lemma A.6 we know that

$$
\|f-\mathcal{T}f'\|_{2,\bar{\mu}_{\beta_k}}^2 = \mathcal{L}_{\hat{\mu}_{\beta_k}}(f;f') - \mathcal{L}_{\hat{\mu}_{\beta_k}}(\mathcal{T}f;f').
$$

Then it is sufficient to bound $\|f-\mathcal{T}f'\|_{2,\bar{\mu}_{\beta_k}}^2$ by bounding

$$
\mathcal{L}_{\hat{\mu}_{\beta_k}}(f;f') - \mathcal{L}_{\hat{\mu}_{\beta_k}}(\mathcal{T}f;f') = \mathbb{E}[\mathcal{L}_{\hat{D}_k}(f;f') - \mathcal{L}_{\hat{D}_k}(\mathcal{T}f;f')].
$$

For any f,f', recall that

$$
\begin{aligned}
\mathcal{L}_{\hat{D}_k}(f;f') = &\underbrace{\frac{1}{|\hat{D}_k|}\sum_{(s,x,a,r,s',x')\in D}\left(f(s,x,a)-r-\gamma V_{f'}(s',x')\right)^2}_{\mathcal{L}_D(f;f')}\\
&+\underbrace{\frac{1}{|\hat{D}_k|}\sum_{(s,x,a,r,s',x')\in D_k}\left(f(s,x,a)-r-\gamma V_{f'}(s',x')\right)^2}_{\mathcal{L}_{D_k}(f;f')}.
\end{aligned}
$$

For any f,f' define

$$
Y(f;f') := (f(s,x,a)-r-\gamma V_{f'}(s',x'))^2 - (\mathcal{T}f'(s,x,a)-r-\gamma V_{f'}(s',x'))^2
$$

Then for each $(s,x,a,s',x')\in D$, we get i.i.d. variables $Y_1(f;f'),\ldots,Y_n(f;f')$.

We also define

$$X_i(f;f') := (f(s_i,\hat{x}_i,a_i) - \hat{r}_i - \gamma V_{f'}({s_i}',\hat{x_i}'))^2 - (\mathcal{T}f'(s_i,\hat{x}_i,a_i) - \hat{r}_i - \gamma V_{f'}({s_i}',\hat{x_i}'))^2,$$

where $(s_i,\hat{x}_i,a_i,\hat{r}_i,s_i',\hat{x}_i')$ is an augmented sample based on the ith real sample (s_i,x_i,a_i,r_i,s_i'). Denote the m i.i.d virtual samples by $X_{i_1}(f;f'),\ldots,X_{i_m}(f;f')$. Therefore

$$\mathcal{L}_{\hat{D}_k}(f;f') - \mathcal{L}_{\hat{D}_k}(\mathcal{T}f';f') = \frac{n}{nm+n} \times \frac{1}{n}\sum_{i=1}^{n} Y_i(f;f') + \frac{nm}{nm+n} \times \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} X_{i_j}(f;f'). \tag{14}$$

Taking the expectations on both sides, we obtain for any $f,f'\in\mathcal{F}$,

$$\mathcal{L}_{\hat{\mu}_{\beta_k}}(f;f') - \mathcal{L}_{\hat{\mu}_{\beta_k}}(\mathcal{T}f';f') = \frac{n}{nm+n}\mathbb{E}\left[Y(f;f')\right] + \frac{nm}{nm+n} \times \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i(f;f')\right]$$

We need to introduce $\frac{1}{n}\sum_{i=1}^{n} Y_i(f;f')$ and $\frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m} X_{i_j}(f;f')$ to bound the above terms. For the first term, we know that the variance of Y can be bounded by:

$$\begin{aligned}
\mathrm{Var}(Y(f;f')) \leq & \mathbb{E}[Y(f;f')^2] \\
= & \mathbb{E}[\left((f(s,x,a) - r - \gamma V_{f'}(s',x'))^2 - (\mathcal{T}f'(s,x,a) - r - \gamma V_f(s',x'))^2\right)^2] \\
= & \mathbb{E}\left[\left(f(s,x,a) - \mathcal{T}f'(s,x,a)\right)^2\left(f(s,x,a) + \mathcal{T}f'(s,x,a) - 2r - 2\gamma V_{f'}(s',x')\right)^2\right] \\
\leq & 4V_{\max}^2\mathbb{E}\left[\left(f(s,x,a) - \mathcal{T}f'(s,x,a)\right)^2\right] \\
= & 4V_{\max}^2\|f - \mathcal{T}f'\|_{2,\mu}^2 \\
= & 4V_{\max}^2\mathbb{E}[Y(f;f')],
\end{aligned} \tag{15}$$

where the last equality is true because

$$\begin{aligned}
& \mathbb{E}[Y(f;f')] = \mathbb{E}[\mathcal{L}_D(f;f')] - \mathbb{E}[\mathcal{L}_D(\mathcal{T}f';f')] \\
= & \|f - \mathcal{T}f'\|_{2,\mu}^2 + \gamma^2\mathbb{E}_{(s,x,a)\sim\mu}[\mathrm{Var}(V_{f'}(s',x')|s,x,a] \\
& - \|\mathcal{T}f' - \mathcal{T}f'\|_{2,\mu}^2 - \gamma^2\mathbb{E}_{(s,x,a)\sim\mu}[\mathrm{Var}(V_{f'}(s',x')|s,x,a] \qquad \text{(using Lemma A.5)} \\
= & \|f - \mathcal{T}f'\|_{2,\mu}^2.
\end{aligned}$$

Then by applying Bernstein's inequality, together with a union bound over all $f,f'\in\mathcal{F}$, we obtain with probability $1-\delta$ we have

$$\begin{aligned}
\mathbb{E}\left[Y(f;f')\right] - \frac{1}{n}\sum_{i=1}^{n} Y_i(f;f') \leq & \sqrt{\frac{2\mathrm{Var}(Y(f;f'))\log(|\mathcal{F}|^2/\delta)}{n}} + \frac{4V_{\max}^2\log(|\mathcal{F}|^2/\delta)}{3n} \\
\leq & \sqrt{\frac{8V_{\max}^2\mathbb{E}[Y(f;f')]\log(|\mathcal{F}|^2/\delta)}{n}} + \frac{4V_{\max}^2\log(|\mathcal{F}|^2/\delta)}{3n}
\end{aligned} \tag{16}$$

For the second term, note that for any given ith sample (s_i,x_i,a_i,s_i',x_i') all the variables $\{X_{i_j}\}$ are i.i.d. Then following a similar argument, then for all $f,f'\in\mathcal{F}$, we have with probability at least $1-\delta/n$,

$$\begin{aligned}
& \mathbb{E}[X_i(f;f')|s_i,x_i,a_i] - \frac{1}{m}\sum_{j=1}^{m} X_{i_j}(f;f') \\
\leq & \sqrt{\frac{8V_{\max}^2\mathbb{E}[X_i(f;f')|s_i,x_i,a_i]\log(n|\mathcal{F}|/\delta)}{m}} + \frac{4V_{\max}^2\log(n|\mathcal{F}|^2/\delta)}{3m}
\end{aligned}$$

Then it is easy to obtain that we have for all $f,f'\in\mathcal{F}$,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i(f;f') - \frac{1}{m}\sum_{j=1}^{m} X_{i_j}(f;f')\right]$$

$$\leq \frac{1}{n}\sum_{i=1}^{n}\left(\sqrt{\frac{8V_{\max}^2\mathbb{E}[X_i(f;f')]\log(n|\mathcal{F}|/\delta)}{m}}+\frac{4V_{\max}^2\log(n|\mathcal{F}|^2/\delta)}{3m}\right)+\frac{\delta V_{\max}^2}{n} \tag{17}$$

Combining Eq.(17) and Eq.(16) we can obtain with probability at least $1-\delta$, for all $f, f' \in \mathcal{F}$,

$$\begin{aligned}
&\frac{n}{n+nm}\times\mathbb{E}\left[Y(f;f')\right]-\frac{n}{n+nm}\times\frac{1}{n}\sum_{i=1}^{n}Y_i(f;f')+\frac{nm}{nm+n}\times\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i(f;f')-\frac{1}{m}\sum_{j=1}^{m}X_{i_j}(f;f')\right]\\
\leq&\frac{1}{1+m}\times\left(\sqrt{\frac{8V_{\max}^2\mathbb{E}[Y(f;f')]\log(|\mathcal{F}|^2/\delta)}{n}}+\frac{4V_{\max}^2\log(|\mathcal{F}|^2/\delta)}{3n}\right)\\
&+\frac{nm}{nm+n}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\sqrt{\frac{8V_{\max}^2\mathbb{E}[X_i(f;f')]\log(n|\mathcal{F}|/\delta)}{m}}+\frac{4V_{\max}^2\log(n|\mathcal{F}|^2/\delta)}{3m}\right)+\frac{\delta V_{\max}^2}{n}\right)
\end{aligned} \tag{18}$$

Let $f=f_{k+1}, f'=f_k$, then according to Algorithm 2 we know that

$$f_{k+1}=\hat{\mathcal{T}}_{k,\mathcal{F}}f_k := \arg\min_{f\in\mathcal{F}}\mathcal{L}_{\hat{\mathcal{D}}_k}(f;f_k).$$

According to Eq. (14), we have

$$\mathcal{L}_{\hat{D}_k}(f;f_k)-\mathcal{L}_{\hat{D}_k}(\mathcal{T}f_k;f_k)=\frac{n}{nm+n}\times\frac{1}{n}\sum_{i=1}^{n}Y_i(f;f_k)+\frac{nm}{nm+n}\times\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}X_{i_j}(f;f_k).$$

Then it is easy to observe that $\hat{\mathcal{T}}_{k,\mathcal{F}}f_k$ minimizes $\mathcal{L}_{\hat{D}_k}(\cdot;f_k)$, it also minimizes

$$\frac{n}{nm+n}\times\frac{1}{n}\sum_{i=1}^{n}Y_i(\cdot;f_k)+\frac{nm}{nm+n}\times\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}X_{i_j}(\cdot;f_k)$$

because the two objectives only differ by a constant $\mathcal{L}_{\hat{D}_k}(\mathcal{T}f_k;f_k)$. Therefore under assumption A.2 we know that $\mathcal{T}f_k\in\mathcal{F}$, we are able to obtain that

$$\begin{aligned}
&\frac{1}{nm+n}\sum_{i=1}^{n}Y_i(\hat{\mathcal{T}}_{k,\mathcal{F}}f_k;f_k)+\frac{1}{mn+n}\sum_{i=1}^{n}\sum_{j=1}^{m}X_{i_j}(\hat{\mathcal{T}}_{k,\mathcal{F}}f_k;f_k)\\
\leq&\frac{1}{nm+n}\sum_{i=1}^{n}Y_i(\mathcal{T}f_k;f_k)+\frac{1}{nm+n}\sum_{i=1}^{n}\sum_{j=1}^{m}X_{i_j}(\mathcal{T}f_k;f_k)=0,
\end{aligned} \tag{19}$$

where the last equality holds due to the definitions of Y_i and X_{i_j}. Therefore plugging the result from Eq.(19) into Eq.(18), we can obtain

$$\begin{aligned}
&\frac{1}{m+1}\mathbb{E}[Y(f_{k+1};f_k)]+\frac{m}{mn+n}\sum_{i=1}^{n}\mathbb{E}[X_i(f_{k+1};f_k)]\\
\leq&\frac{1}{1+m}\left(\sqrt{\frac{8V_{\max}^2\mathbb{E}[Y(f_{k+1};f_k)]\log(|\mathcal{F}|^2/\delta)}{n}}+\frac{4V_{\max}^2\log(|\mathcal{F}|^2/\delta)}{3n}\right)\\
&+\frac{nm}{nm+n}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\sqrt{\frac{8V_{\max}^2\mathbb{E}[X_i(f_{k+1};f_k)]\log(n|\mathcal{F}|^2/\delta)}{m}}+\frac{4V_{\max}^2\log(n|\mathcal{F}|^2/\delta)}{3m}\right)+\frac{\delta V_{\max}^2}{n}\right)
\end{aligned}$$

By solving the quadratic formula, we get

$$\begin{aligned}
&\mathcal{L}_{\hat{\mu}_{\beta_k}}(f_{k+1};f_k)-\mathcal{L}_{\hat{\mu}_{\beta_k}}(\mathcal{T}f_k;f_k)=\|f_{k+1}-\mathcal{T}f_k\|_{2,\bar{\mu}_{\beta_k}}^2\\
=&\frac{1}{m+1}\mathbb{E}[Y(f_{k+1};f_k)]+\frac{m}{nm+n}\sum_{j=1}^{n}\mathbb{E}[X_i(f_{k+1};f_k)]\\
\leq&5\left(\frac{1}{n}+\frac{1}{m}\right)V_{\max}^2\log(n|\mathcal{F}|^2/\delta)+\frac{3\delta V_{\max}^2}{n}
\end{aligned}$$

Finally, apply a union bound over all $t=0\dots K$, we conclude the proof.

$\square$

A.1 Proof of Theorem 1

Now we are ready to show the main theorem. Given a dataset D. After generating virtual samples D_k we get a new combined dataset $\hat{D}_k = D \cup D_k$ at each iteration k with virtual sample generating distribution $\beta_k(x)$. We first have

$$\begin{aligned}
v^{\pi_{f_k}} - v^* =& \frac{1}{1-\gamma}\mathbb{E}_{(s,x)\sim d_{\eta_0}^{\pi_{f_k}}(s,x)}\left[Q^*(s,x,\pi_{f_k}) - V^*(s,x)\right] \\
\leq& \frac{1}{1-\gamma}\mathbb{E}_{(s,x)\sim d_{\eta_0}^{\pi_{f_k}}(s,x)}\left[Q^*(s,x,\pi_{f_k}) - f_k(s,x,\pi_{f_k}) + f_k(s,x,\pi^*) - V^*(s,x)\right] \\
\leq& \frac{1}{1-\gamma}\left(\|Q^* - f_k\|_{1,d_{\eta_0}^{\pi_{f_k}}(s,x)\times\pi_{f_k}} + \|Q^* - f_k\|_{1,d_{\eta_0}^{\pi_{f_k}}(s,x)\times\pi^*}\right) \\
\leq& \frac{1}{1-\gamma}\left(\|Q^* - f_k\|_{2,d_{\eta_0}^{\pi_{f_k}}(s,x)\times\pi_{f_k}} + \|Q^* - f_k\|_{2,d_{\eta_0}^{\pi_{f_k}}(s,x)\times\pi^*}\right),
\end{aligned} \tag{20}$$

where the first equality follows from the performance difference lemma (Kakade and Langford, 2002), the first inequality holds because $\pi_{f_k} \in \arg\min_a f_k(s,x,a)$ and the last inequality is true by using the fact that for any vector $a = (a_1,\ldots,a_n)$ and a valid distribution $d = (d_1,\ldots,d_n)$, $\sum_i d_i = 1$

$$\begin{aligned}
\|a\|_{1,d} =& \sum_i |a_i| d_i = \sum_i |a_i|\sqrt{d_i}\sqrt{d_i} \qquad \text{(Cauchy–Schwarz inequality)} \\
\leq& \sqrt{\sum_i d_i} \times \sqrt{\sum_i |a_i|^2 d_i} = \|a\|_{2,d}.
\end{aligned}$$

According to Assumption A.4, we know that $\sum_{s\in\mathcal{S}}\sum_{x\in\mathcal{X},x\notin\mathcal{B}} d_{\eta_0}^{\pi_{f_k}}(s,x) \leq \sigma_2$, which implies that $\sum_{(s,x,a)\in\mathcal{S}\times\mathcal{X}\times\mathcal{A},x\notin\mathcal{B}}\{d_{\eta_0}^{\pi_{f_k}}(s,x) \times \pi^*\}(s,x,a) \leq \sigma_2$, and $\sum_{(s,x,a)\in\mathcal{S}\times\mathcal{X}\times\mathcal{A},x\notin\mathcal{B}}\{d_{\eta_0}^{\pi_{f_k}}(s,x) \times \pi_{f_k}\}(s,x,a) \leq \sigma_2$. Define $\xi = \{d_{\eta_0}^{\pi_{f_k}}(s,x) \times \pi_{f_k}\}(s,x,a)$, then we have

$$\begin{aligned}
\|f_k - Q^*\|_{2,\xi} =& \left(\sum_{(s,x,a)\in\mathcal{S}\times\mathcal{X}\times\mathcal{A}} |f_k(s,x,a) - Q^*(s,x,a)|^2 \xi(s,x,a)\right)^{1/2} \\
\leq& \left(\sum_{(s,x,a)\in\mathcal{S}\times\mathcal{X}\times\mathcal{A},x\in\mathcal{B}} |f_k(s,x,a) - Q^*(s,x,a)|^2 \xi(s,x,a)\right)^{1/2} \\
&+ \left(\sum_{(s,x,a)\in\mathcal{S}\times\mathcal{X}\times\mathcal{A},x\notin\mathcal{B}} |f_k(s,x,a) - Q^*(s,x,a)|^2 \xi(s,x,a)\right)^{1/2} \\
\leq& \|f_k - Q^*\|_{\underline{2},\xi} + \sqrt{\sigma_2}V_{\max} \\
\leq& \sqrt{\frac{(m+1)C}{m\sigma_1}}\|f_k - \mathcal{T}f_{k-1}\|_{\underline{2},\bar{\mu}_{\beta_k}} + \gamma\|f_{k-1} - Q^*\|_{2,P(\xi)\times\pi_{f_{k-1},Q^*}} + \sqrt{\sigma_2}V_{\max},
\end{aligned} \tag{21}$$

where the first inequality holds because $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ ($a \geq 0, b \geq 0$) and the last inequality comes from Lemma A.4.

By using LemmaA.7 we have with at least probability $1-\delta$

$$\|f_k - \mathcal{T}f_{k-1}\|^2_{\underline{2},\bar{\mu}_{\beta_k}} \leq \|f_k - \mathcal{T}f_{k-1}\|^2_{\underline{2},\bar{\mu}_{\beta_k}} \leq \epsilon_1 \tag{22}$$

where $\epsilon_1 = 5\left(\frac{1}{n} + \frac{1}{m}\right)V_{\max}^2 \log(nK|\mathcal{F}|^2/\delta) + \frac{3\delta V_{\max}^2}{n}$. Therefore, we obtain

$$\|f_k - Q^*\|_{2,\xi} \leq \gamma\|f_{k-1} - Q^*\|_{2,P(\xi)\times\pi_{f_{k-1},Q^*}} + \sqrt{\frac{(m+1)C\epsilon_1}{m\sigma_1}} + \sqrt{\sigma_2}V_{\max}. \tag{23}$$

Now define $\xi' = P(\xi) \times \pi_{f_{k-1},Q^*}$. Then based on Assumption A.4, it is easy to obtain $\sum_{(s,x,a)\in\mathcal{S}\times\mathcal{X}\times\mathcal{A},x\notin\mathcal{B}} \xi'(s,x,a) \leq c_0\sigma_2$. Note that the distribution $\xi'' = P(\xi') \times \pi_{f_{k-2},Q^*}$ still satisfies $\sum_{(s,x,a)\in\mathcal{S}\times\mathcal{X}\times\mathcal{A},x\notin\mathcal{B}} \xi''(s,x,a) \leq c_0\sigma_2$ according to Assumption A.4, because $\pi_{f_{k-1}}, \pi_{f_{k-2}}$ and Q^* are all assumed to be reasonable policies.

Repeating the expansion above for k times, we have

$$\|f_k - Q^*\|_{2,\xi} \leq \frac{1-\gamma^k}{1-\gamma}\left(\sqrt{\frac{(m+1)C\epsilon_1}{m\sigma_1}} + \sqrt{c_0\sigma_2}V_{\max}\right) + \gamma^k V_{\max}.$$

All the above analyses are still applied to the case when $\xi = \{d_{\eta_0}^{\pi_{f_k}}(s,x) \times \pi^*\}$. Therefore, it is straightforward to obtain

$$v^* - v^{\pi_{f_k}} \leq \frac{2}{(1-\gamma)^2}\left(\sqrt{\frac{(m+1)C\epsilon_1}{m\sigma_1}} + \sqrt{c_0\sigma_2}V_{\max} + \gamma^k(1-\gamma)V_{\max}\right). \tag{24}$$

Substituting ϵ_1 completes the proof.

A.2 Extension of Theorem 1

We repeat the assumption for extending our main results to the case when $|\mathcal{X}|$ can be infinite such that $f(s,x,a)$ may not be bounded by $V_{\max}$.

Repeat of Assumption 1: For the typical set $\mathcal{B}$ of pseudo-stochastic states (defined in Assumption A.4), for any $s, a \in \mathcal{S} \times \mathcal{A}$, $f \in \mathcal{F}$, if $x \in \mathcal{B}$, then $f(s,x,a) \leq V_{\max}$ otherwise if $x \notin \mathcal{B}$, we have $|f(s,x,a) - Q^*(s,x,a)| \leq V_{\max}$. Furthermore, for any given $f \in \mathcal{F}$, (s,x), $x \in \mathcal{B}$, we have $|V_f(s',x') - V_f(s'',x'')| \leq V_{\max}$, where $x' = g(s,x,\pi_f,s')$, $x'' = g(s,x,\pi_f,s'')$.

There are two places we need to pay attention to: (1) : a bound on $\|f_0 - Q^*\|_{2,P(\xi)\times\pi_{f_0,Q^*}}$, (2) : a bound on the variance of Y as shown in Eq. (15). For the first case, it automatically holds due to assumption 1. For the second term, we first have

$$\begin{aligned}
&\mathrm{Var}(Y(f;f')) \leq \mathbb{E}[Y(f;f')^2]\\
=&\mathbb{E}[\left((f(s,x,a) - r - \gamma V_{f'}(s',x'))^2 - (\mathcal{T}f'(s,x,a) - r - \gamma V_f(s',x'))^2\right)^2]\\
=&\mathbb{E}\left[(f(s,x,a) - \mathcal{T}f'(s,x,a))^2\left(f(s,x,a) + \mathcal{T}f'(s,x,a) - 2r - 2\gamma V_{f'}(s',x')\right)^2\right]\\
=&\mathbb{E}\left[(f(s,x,a) - \mathcal{T}f'(s,x,a))^2\left(f(s,x,a) - \mathcal{T}f'(s,x,a) + 2\mathcal{T}f'(s,x,a) - 2r - 2\gamma V_{f'}(s',x')\right)^2\right].
\end{aligned}$$

We also know that

$$\begin{aligned}
&(f(s,x,a) - \mathcal{T}f'(s,x,a) + 2\mathcal{T}f'(s,x,a) - 2r - 2\gamma V_{f'}(s',x'))^2\\
\leq&2(f(s,x,a) - \mathcal{T}f'(s,x,a))^2 + 2(2\mathcal{T}f'(s,x,a) - 2r - 2\gamma V_{f'}(s',x'))^2\\
=&2(f(s,x,a) + Q^*(s,x,a) - Q^*(s,x,a) - \mathcal{T}f'(s,x,a))^2 + 8\gamma^2(\mathbb{E}[V'_{f'}(\hat{s},\hat{x})|s,x,a] - V_{f'}(s',x'))^2\\
\leq&16V_{\max}^2.
\end{aligned}$$

Therefore, we have $\mathrm{Var}(Y(f;f')) \leq 16V_{\max}^2\mathbb{E}[Y(f;f')]$.

Then we can obtain a similar result of the same order, which only differs for some constant $\tilde{c}$ such that

$$v^* - v^{\pi_{f_k}} \leq \frac{2\tilde{c}}{(1-\gamma)^2}\left(\sqrt{\frac{(m+1)C\epsilon_1}{m\sigma_1}} + \sqrt{c_0\sigma_2}V_{\max} + \gamma^k(1-\gamma)V_{\max}\right). \tag{25}$$

B Additional Simulations

B.1 Combining PSG with Policy Gradient-type algorithms

In this section, we investigate the possibilities of using ASG in policy gradient-type algorithms. In particular, we use ASG in the phase of policy evaluation. An algorithm (SAC-ASG) that incorporates ASG into SAC is presented in Alg. 3. We also compare our algorithm SAC-ASG with state-of-art Dyna-type model-based approaches, i.e., MBPO (Janner et al., 2019) on the phases criss-cross network environment (Fig. 2b). The simulation results are shown in Fig.4. We can observe that

Algorithm 3: SAC-ASG

Input: Critic Networks: $Q_{\theta_1}, Q_{\theta_2}$, Target Critic Networks: $Q_{\theta'_1}, Q_{\theta'_2}$;
Actor-Network: A_ϕ, Empty sample reply buffer: $\mathcal{D}$, Learning rate: λ ;
for *each iteration* **do**
  for *each environment interaction* **do**
    Take action $a_t \sim A_\phi(a_t|s_t, x_t)$, observe next state (s_{t+1}, x_{t+1}), and reward r_t ;
    Store the transition into replay buffer: $\mathcal{D} \leftarrow \mathcal{D} \cup \{s_t, x_t, a_t, r_t, s_{t+1}, x_{t+1}\}$;
  for *each training step* **do**
    Sample mini-batch d of n transitions from replay buffer $\mathcal{D}$;
    for *Each virtual training loop* **do**
      Obtain virtual dataset $d' :=$ **ASG(d,m)** ;
      Combine training dataset $d \cup d' := \{s, x, a, r, s', x'\}$;
      $\tilde{a} \leftarrow A_\phi(s', x'), \ \ y \leftarrow r + \gamma(\min_{i=1,2} Q_{\theta'_i}(s', x', \tilde{a}) - \alpha \log(A_\phi(\tilde{a}|s', x'))$;
      $J_Q(\theta_i) = (nm + n)^{-1} \sum (y - Q_{\theta_i}(s, x, a))^2$ for $i \in \{1, 2\}$;
      $\theta_i \leftarrow \theta_i - \lambda \nabla_{\theta_i} J_Q(\theta_i)$ for $i \in \{1, 2\}$ `// Update Critic networks`
      $J_\pi(\phi) = (nm + n)^{-1} \sum (\alpha \log A_\phi(a|s) - \min_{i=1,2} Q_{\theta_i}(s, a)))$;
      $\phi \leftarrow \phi - \lambda \nabla_\phi J_\pi(\phi)$ `// Update Actor network`
      $\theta'_i \leftarrow \tau\theta_i + (1 - \tau)\theta'_i$ `// Update target network weights`
Output: Actor Network A_ϕ ;

the performance of our approach is significantly better than the baselines'. We also would like to emphasize that the training time of our approach is much less than that of MBPO (4 hours v.s. 3 days).

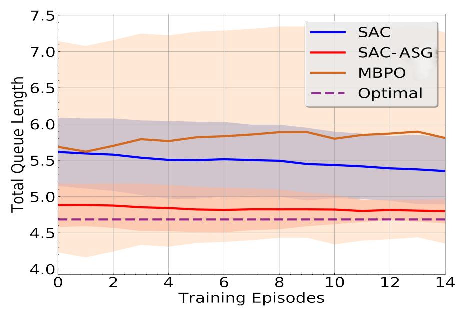

Figure 4: Performance on the Two Phases Criss-Cross Network

B.2 Details of the Environment in Section 4.4

In this section, we summarize the detailed parameters used in section 4.4 in Table 3.

Setting	Arrival Rates	Service Rates	Job Size Range
(a)	$\{0.6, 0.6\}$	$\{2, 1.5, 1.5\}$	2
(a)	$\{0.6, 0.6\}$	$\{7, 3.5, 7\}$	5
(c)	$\{0.6, 0.6\}$	$\{2.5, 4.5, 2.5\}$	5

Table 3: Detailed Environment Parameters

B.3 Experimental settings

For all the simulations, We used a single NVIDIA GeForce RTX 2080 Super with AMD Ryzen 7 3700 8-Core Processor.