Offline Reinforcement Learning Under Value and Density-Ratio Realizability: The Power of Gaps (Supplementary Material)

Jinglin Chen, Nan Jiang[1]

[1]Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, IL, USA

A PROOF OF MAIN RESULTS

In this section, we provide the complete proofs of our main results in Section 4. We start with some helper lemmas in Appendix A.1. Then we show the proof of Theorem 1 in Appendix A.2. Finally, we provide the proof of Theorem 2 in Appendix A.3.

A.1 HELPER LEMMAS

Lemma 1 (Concentration)**.** *With probability at least $1-\delta$, for any $f \in \mathcal{F}, w \in \mathcal{W}, h \in [H]$ we have,*

$$|\mathcal{L}_{\mathcal{D}}(f,w,h) - \mathbb{E}[\mathcal{L}_{\mathcal{D}}(f,w,h)]| \le 2CH\sqrt{\frac{\log(2|\mathcal{F}||\mathcal{W}|H/\delta)}{2n}} =: \varepsilon_{\mathrm{stat},n}.$$

Remark Here we apply Hoeffding's inequality to show the concentration result. Similar as Xie and Jiang (2020), we can also apply Bernstein's inequality, but the dominating rate would be the same.

Proof. Firstly, we fix $f \in \mathcal{F}, w \in \mathcal{W}, h \in [H]$. From the boundedness assumptions (Assumption 3 and Assumption 4), for any sample $(x_h^{(i)}, a_h^{(i)}, r_h^{(i)}, x_{h+1}^{(i)})$ in the dataset, we have

$$\left| w_h(x_h^{(i)}, a_h^{(i)})(f_h(x_h^{(i)}, a_h^{(i)}) - r_h^{(i)} - f_h(x_{h+1}^{(i)}, \pi_f(x_{h+1}^{(i)}))) \right| \le CH.$$

Then since our dataset is i.i.d., applying Hoeffding's inequality yields that with probability at least $1-\delta/(|\mathcal{F}||\mathcal{W}|H)$,

$$|\mathcal{L}_{\mathcal{D}}(f,w,h) - \mathbb{E}[\mathcal{L}_{\mathcal{D}}(f,w,h)]| \le 2CH\sqrt{\frac{\log(2|\mathcal{F}||\mathcal{W}|H/\delta)}{2n}}.$$

Finally, union bounding over $f \in \mathcal{F}, w \in \mathcal{W}, h \in [H]$ gives us that with probability at least $1-\delta$, for any $f \in \mathcal{F}, w \in \mathcal{W}, h \in [H]$,

$$|\mathcal{L}_{\mathcal{D}}(f,w,h) - \mathbb{E}[\mathcal{L}_{\mathcal{D}}(f,w,h)]| \le 2CH\sqrt{\frac{\log(2|\mathcal{F}||\mathcal{W}|H/\delta)}{2n}} := \varepsilon_{\mathrm{stat},n}.$$

This completes the proof. □

Lemma 2 (Population loss and average Bellman error)**.** *For any $f \in \mathcal{F}, w \in \mathcal{W}, h \in [H]$, we have*

$$\mathbb{E}[\mathcal{L}_{\mathcal{D}}(f,w,h)] = \mathbb{E}_{(x_h,a_h)\sim d_h^D}[w_h(x_h,a_h)(f_h(x_h,a_h) - (\mathcal{T}_h f_{h+1})(x_h,a_h)))]$$

and

$$\mathbb{E}[\mathcal{L}_{\mathcal{D}}(f,w^*,h)] = \mathcal{E}(f,\pi^*,h) = \mathbb{E}[f_h(x_h,a_h) - R_h(x_h,a_h) - f_{h+1}(x_{h+1},a_{h+1}) \mid a_{0:h}\sim\pi^*, a_{h+1}\sim\pi_f],$$

where $\mathcal{E}(\cdot)$ is the Q-type average Bellman error (Jin et al., 2021; Du et al., 2021)

$$\mathcal{E}(f,\pi,h) = \mathbb{E}[f_h(x_h,a_h) - R_h(x_h,a_h) - f_{h+1}(x_{h+1},a_{h+1}) \mid a_{0:h}\sim\pi, a_{h+1}\sim\pi_f].$$

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

Proof. These equations can be simply shown from the data generating process and the definition of population loss and empirical loss. For any $f \in \mathcal{F}, w \in \mathcal{W}, h \in [H]$, we have

$$
\begin{aligned}
&\mathbb{E}[\mathcal{L}_{\mathcal{D}}(f, w, h)] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}[w_h(x_h^{(i)}, a_h^{(i)})(f_h(x_h^{(i)}, a_h^{(i)}) - r_h^{(i)} - f_{h+1}(x_{h+1}^{(i)}, \pi_f(x_{h+1}^{(i)})))]\right] \\
&= \mathbb{E}_{(x_h,a_h)\sim d_h^D, x_{h+1}\sim P_h(\cdot|x_h,a_h)}[w_h(x_h, a_h)(f_h(x_h, a_h) - r_h - f_{h+1}(x_{h+1}, \pi_f(x_{h+1})))] \\
&= \mathbb{E}_{(x_h,a_h)\sim d_h^D}[w_h(x_h, a_h)(f_h(x_h, a_h) - R_h(x_h, a_h) - \mathbb{E}_{x_{h+1}\sim P_h(\cdot|x_h,a_h)}[f_{h+1}(x_{h+1}, \pi_f(x_{h+1}))])] \\
&= \mathbb{E}_{(x_h,a_h)\sim d_h^D}[w_h(x_h, a_h)(f_h(x_h, a_h) - (\mathcal{T}_h f_{h+1})(x_h, a_h)))].
\end{aligned}
$$

For any $f \in \mathcal{F}, h \in [H]$, we similarly have

$$
\begin{aligned}
\mathbb{E}[\mathcal{L}_{\mathcal{D}}(f, w^*, h)] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}[w_h^*(x_h^{(i)}, a_h^{(i)})(f_h(x_h^{(i)}, a_h^{(i)}) - r_h^{(i)} - f_{h+1}(x_{h+1}^{(i)}, \pi_f(x_{h+1}^{(i)})))]\right] \\
&= \mathbb{E}_{(x_h,a_h)\sim d_h^D, x_{h+1}\sim P_h(\cdot|x_h,a_h)}[w_h^*(x_h, a_h)(f_h(x_h, a_h) - r_h - f_{h+1}(x_{h+1}, \pi_f(x_{h+1})))] \\
&= \mathbb{E}_{(x_h,a_h)\sim d_h^*, x_{h+1}\sim P_h(\cdot|x_h,a_h)}[f_h(x_h, a_h) - r_h - f_{h+1}(x_{h+1}, \pi_f(x_{h+1}))] \\
&= \mathbb{E}[f_h(x_h, a_h) - R_h(x_h, a_h) - f_{h+1}(x_{h+1}, a_{h+1}) \mid a_{0:h} \sim \pi^*, a_{h+1} \sim \pi_f].
\end{aligned}
$$

This completes the proof. □

A.2 PROOF OF THEOREM 1

Theorem (Sample complexity of identifying v^*, restatement of Theorem 1)**.** *Suppose Assumption 1, Assumption 2, Assumption 3, Assumption 4 hold and the total number of samples nH satisfies*

$$
nH \geq \frac{8C^2 H^5 \log(2|\mathcal{F}||\mathcal{W}|H/\delta)}{\varepsilon^2}.
$$

Then with probability at least $1 - \delta$, running Algorithm 1 with $C_{\text{gap}} = 0$ and $\alpha = \varepsilon/(2H)$ guarantees

$$
|V_{\hat{f}}(x_0) - v^*| \leq \varepsilon.
$$

Proof. From our choice of n and Lemma 1, with probability at least $1 - \delta$, for any $f \in \mathcal{F}, w \in \mathcal{W}, h \in [H]$, we have

$$
|\mathcal{L}_{\mathcal{D}}(f, w, h) - \mathbb{E}[\mathcal{L}_{\mathcal{D}}(f, w, h)]| \leq \varepsilon_{\text{stat},n} \leq \varepsilon/(2H).
$$

Throughout the proof, we condition on this high probability event.

From Lemma 2, for any $w \in \mathcal{W}, h \in [H]$, we have

$$
\begin{aligned}
\mathbb{E}[\mathcal{L}_{\mathcal{D}}(Q^*, w, h)] &= \mathbb{E}_{(x_h,a_h)\sim d_h^D}[w_h(x_h, a_h)(Q_h^*(x_h, a_h) - \mathcal{T}_h Q_{h+1}^*(x_h, a_h)] \\
&= \mathbb{E}_{(x_h,a_h)\sim d_h^D}[w_h(x_h, a_h) \cdot 0] \\
&= 0.
\end{aligned}
$$

Therefore, we further have

$$
\mathcal{L}_{\mathcal{D}}(Q^*, w, h) \leq \mathbb{E}[\mathcal{L}_{\mathcal{D}}(Q^*, w, h)] + \varepsilon_{\text{stat},n} \leq \varepsilon/(2H) = \alpha,
$$

which means Q^* satisfies all the constraints.

Then we show that any value function satisfying all constraints (though it may have large average Bellman errors under some distributions) can not be much more pessimistic than Q^*.

From Lemma 1 and Lemma 2, we know that for any $f \in \mathcal{F}, h \in [H]$,

$$
|\mathcal{E}(f, \pi^*, h)|
$$

$$
\begin{aligned}
&= |\mathbb{E}[f_h(x_h, a_h) - R_h(x_h, a_h) - f_{h+1}(x_{h+1}, a_{h+1}) \mid a_{0:h} \sim \pi^*, a_{h+1} \sim \pi_f]| \\
&= |\mathbb{E}[\mathcal{L}_{\mathcal{D}}(f, w^*, h)]| \\
&\leq \mathcal{L}_{\mathcal{D}}(f, w^*, h) + \varepsilon_{\text{stat},n} \\
&\leq \alpha + \varepsilon_{\text{stat},n} \leq \varepsilon/H.
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
V_f(x_0) = \;& f_0(x_0, \pi_f(x_0)) \\
\geq \;& f_0(x_0, \pi^*(x_0)) \\
\geq \;& \mathbb{E}[R_0(x_0, a_0) + f_1(x_1, a_1) \mid a_0 \sim \pi^*, a_1 \sim \pi_f] - \varepsilon/H && (|\mathcal{E}(f, \pi^*, 0)| \leq \varepsilon/H) \\
\geq \;& \mathbb{E}[R_0(x_0, a_0) \mid a_0 \sim \pi^*] + \mathbb{E}[f_1(x_1, a_1) \mid a_{0:1} \sim \pi^*] - \varepsilon/H \\
\geq \;& \mathbb{E}[R_0(x_0, a_0) \mid a_0 \sim \pi^*] + \mathbb{E}[R_1(x_1, a_1) + f_2(x_2, a_2) \mid a_{0:1} \sim \pi^*, a_2 \sim \pi_f] - 2\varepsilon/H && (|\mathcal{E}(f, \pi^*, 1)| \leq \varepsilon/H) \\
\geq \;& \ldots \\
\geq \;& \mathbb{E}\left[\sum_{h=0}^{H-1} R_h(x_h, a_h) \mid a_{0:H-1} \sim \pi^*\right] - H \times \varepsilon/H = V_0^*(x_0) - \varepsilon.
\end{aligned}
$$

Combining the two arguments above, we know that the pessimistic value function $\hat{f}$ found by the algorithm satisfies

$$
v^* - \varepsilon = V_0^*(x_0) - \varepsilon \leq V_{\hat{f}}(x_0) \leq V_0^*(x_0) = v^*,
$$

where the second inequality is due to pessimism. This completes the proof. □

A.3 PROOF OF THEOREM 2

Theorem (Sample complexity of learning a near-optimal policy, restatement of Theorem 2)**.** *Suppose Assumption 1, Assumption 2, Assumption 3, Assumption 4, Assumption 5 hold and the total number of samples nH satisfies*

$$
nH \geq \frac{8C^2 H^7 \log(2|\mathcal{F}||\mathcal{W}|H/\delta)}{\varepsilon^2 \text{gap}(Q^*)^2}.
$$

Then with probability at least $1-\delta$, running Algorithm 1 with $\alpha = \varepsilon \text{gap}(Q^*)/(2H^2)$ and $C_{\text{gap}} = \text{gap}(Q^*)$ guarantees

$$
v^{\pi_{\hat{f}}} \geq v^* - \varepsilon.
$$

Proof. From our choice of n and Lemma 1, we know that with probability at least $1-\delta$, for any $f \in \mathcal{F}, w \in \mathcal{W}, h \in [H]$, we have

$$
|\mathcal{L}_{\mathcal{D}}(f, w, h) - \mathbb{E}[\mathcal{L}_{\mathcal{D}}(f, w, h)]| \leq \varepsilon_{\text{stat},n} \leq \varepsilon \text{gap}(Q^*)/(2H^2).
$$

Throughout the proof, we condition on this high probability event.

From the definition of $\text{gap}(Q^*)$, we know that prescreening will not eliminate Q^*, i.e., $Q^* \in \mathcal{F}(\text{gap}(Q^*))$. Then similar as the proof of Theorem 1, we have

$$
\mathcal{L}_{\mathcal{D}}(Q^*, w, h) \leq \mathbb{E}[\mathcal{L}_{\mathcal{D}}(Q^*, w, h)] + \varepsilon_{\text{stat},n} = \varepsilon_{\text{stat},n} \leq \varepsilon \text{gap}(Q^*)/(2H^2) = \alpha,
$$

which means that Q^* satisfies all the constraints.

For any $f \in \mathcal{F}(\text{gap}(Q^*))$ that satisfies all the constraints and any $h \in [H]$, we have

$$
\begin{aligned}
&\mathcal{E}(f, \pi^*, h) \\
&= |\mathbb{E}[f_h(x_h, a_h) - R_h(x_h, a_h) - f_{h+1}(x_{h+1}, a_{h+1}) \mid a_{0:h} \sim \pi^*, a_{h+1} \sim \pi_f]| \\
&= |\mathbb{E}[\mathcal{L}_{\mathcal{D}}(f, w^*, h)]| \\
&\leq \mathcal{L}_{\mathcal{D}}(f, w^*, h) + \varepsilon_{\text{stat},n} \\
&\leq \alpha + \varepsilon_{\text{stat},n}
\end{aligned}
$$

$$\leq \varepsilon \text{gap}(Q^*)/H^2.$$

Now we have the following stronger result compared with the proof of Theorem 1

$$\begin{aligned}
& V_f(x_0) \\
&= f_0(x_0, \pi_f(x_0)) \\
&\geq f_0(x_0, \pi^*(x_0)) + \text{gap}(Q^*)\mathbf{1}\{\pi_f(x_0) \neq \pi^*(x_0)\} \\
&\geq \mathbb{E}[R_0(x_0, a_0) + f_1(x_1, a_1) \mid a_0 \sim \pi^*, a_1 \sim \pi_f] \\
&\quad + \text{gap}(Q^*)\mathbf{1}\{\pi_f(x_0) \neq \pi^*(x_0)\} - \varepsilon \text{gap}(Q^*)/H^2 && (|\mathcal{E}(f, \pi^*, 0)| \leq \varepsilon \text{gap}(Q^*)/H^2) \\
&\geq \mathbb{E}[R_0(x_0, a_0) \mid a_0 \sim \pi^*] + \mathbb{E}[f_1(x_1, \pi^*(x_1)) + \text{gap}(Q^*)\mathbf{1}\{\pi_f(x_1) \neq \pi^*(x_1)\} \mid a_0 \sim \pi^*] \\
&\quad + \text{gap}(Q^*)\mathbf{1}\{\pi_f(x_0) \neq \pi^*(x_0)\} - \varepsilon \text{gap}(Q^*)/H^2 \\
&= \mathbb{E}[R_0(x_0, a_0) \mid a_0 \sim \pi^*] + \mathbb{E}[f_1(x_1, a_1) \mid a_{0:1} \sim \pi^*] + \text{gap}(Q^*)\mathbb{E}[\mathbf{1}\{\pi_f(x_1) \neq \pi^*(x_1)\} \mid a_0 \sim \pi^*] \\
&\quad + \text{gap}(Q^*)\mathbf{1}\{\pi_f(x_0) \neq \pi^*(x_0)\} - \varepsilon \text{gap}(Q^*)/H^2 \\
&\geq \mathbb{E}[R_0(x_0, a_0) \mid a_0 \sim \pi^*] + \mathbb{E}[R_1(x_1, a_1) + f_2(x_2, a_2) \mid a_{0:1} \sim \pi^*, a_2 \sim \pi_f] \\
&\quad + \text{gap}(Q^*)[\mathbf{1}\{\pi_f(x_0) \neq \pi^*(x_0)\} + \mathbb{E}[\mathbf{1}\{\pi_f(x_1) \neq \pi^*(x_1)\} \mid a_0 \sim \pi^*]] \\
&\quad - 2\varepsilon \text{gap}(Q^*)/H^2 && (|\mathcal{E}(f, \pi^*, 1)| \leq \varepsilon \text{gap}(Q^*)/H^2) \\
&\geq \dots \\
&\geq \mathbb{E}\left[\sum_{h=0}^{H-1} R_h(x_h, a_h) \mid a_{0:H-1} \sim \pi^*\right] + \text{gap}(Q^*)\mathbb{E}\left[\sum_{h=0}^{H-1} \mathbf{1}\{\pi_f(x_h) \neq \pi^*(x_h)\} \mid a_{0:H-1} \sim \pi^*\right] \\
&\quad - H \times \varepsilon \text{gap}(Q^*)/H^2 \\
&= V_0^*(x_0) + \text{gap}(Q^*)\mathbb{E}\left[\sum_{h=0}^{H-1} \mathbf{1}\{\pi_f(x_h) \neq \pi^*(x_h)\} \mid a_{0:H-1} \sim \pi^*\right] - \varepsilon \text{gap}(Q^*)/H.
\end{aligned}$$

This implies the pessimistic value function $\hat{f}$ found by the Algorithm 1 satisfies

$$V_0^*(x_0) \geq V_{\hat{f}}(x_0) \geq V_0^*(x_0) + \text{gap}(Q^*)\mathbb{E}\left[\sum_{h=0}^{H-1} \mathbf{1}\{\pi_{\hat{f}}(x_h) \neq \pi^*(x_h)\} \mid a_{0:H-1} \sim \pi^*\right] - \varepsilon \text{gap}(Q^*)/H$$

and thus

$$\mathbb{E}\left[\sum_{h=0}^{H-1} \mathbf{1}\{\pi_{\hat{f}}(x_h) \neq \pi^*(x_h)\} \mid a_{0:H-1} \sim \pi^*\right] \leq \varepsilon/H. \tag{1}$$

On the other hand, define each trajectory τ as $(x_0, a_0, r_0, \dots, x_{H-1}, a_{H-1}, r_{H-1}, x_H)$, the return of τ as $\text{Return}(\tau) = r_0 + \dots + r_{H-1}$, and the probability of τ under policy π (i.e., $a_h = \pi(x_h), \forall h \in [H]$) as $\Pr_\pi(\tau)$. For any $f \in \mathcal{F}$, we can decompose the entire trajectory space into three disjoint sets $\mathcal{C}_1 = \{\tau = (x_0, a_0, r_0, \dots, x_{H-1}, a_{H-1}, r_{H-1}, x_H) : \forall h \in [H], a_h = \pi^*(x_h) = \pi_f(x_h)\}$, $\mathcal{C}_2 = \{\tau = (x_0, a_0, r_0, \dots, x_{H-1}, a_{H-1}, r_{H-1}, x_H) : \forall h \in [H], a_h = \pi^*(x_h), \exists h \in [H], \pi_f(x_h) \neq \pi^*(x_h)\}$, $\mathcal{C}_3 = (\mathcal{C}_1 \bigcup \mathcal{C}_2)^\complement$.

Then we calculate V^{π^*} and V^{π_f} with the definition of these three sets

$$\begin{aligned}
V_0^{\pi^*}(x_0) &= \sum_{\tau \in \mathcal{C}_1 \bigcup \mathcal{C}_2 \bigcup \mathcal{C}_3} \Pr_{\pi^*}(\tau)\text{Return}(\tau) \\
&= \sum_{\tau \in \mathcal{C}_1} \Pr_{\pi^*}(\tau)\text{Return}(\tau) + \sum_{\tau \in \mathcal{C}_2} \Pr_{\pi^*}(\tau)\text{Return}(\tau) \\
& && \text{(Because } \pi^* \text{ is greedy policy, any trajectory } \tau \in \mathcal{C}_3 \text{ has 0 probability)} \\
&= \sum_{\tau \in \mathcal{C}_1} \Pr_{\pi_f}(\tau)\text{Return}(\tau) + \sum_{\tau \in \mathcal{C}_2} \Pr_{\pi^*}(\tau)\text{Return}(\tau) && \text{(Definition of } \mathcal{C}_1) \\
&\leq \sum_{\tau \in \mathcal{C}_1} \Pr_{\pi_f}(\tau)\text{Return}(\tau) + \sum_{\tau \in \mathcal{C}_2} \Pr_{\pi^*}(\tau)H && (\text{Return}(\tau) \leq H)
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{\tau \in \mathcal{C}_1 \bigcup \mathcal{C}_2 \bigcup \mathcal{C}_3} \Pr_{\pi_f}(\tau)\text{Return}(\tau) + \sum_{\tau \in \mathcal{C}_2} \Pr_{\pi^*}(\tau) H && (\text{Return}(\tau) \geq 0) \\
&= V_0^{\pi_f}(x_0) + \sum_{\tau \in \mathcal{C}_2} \Pr_{\pi^*}(\tau) H.
\end{aligned}
$$

It remains to show that $\Pr_{\pi^*}(\mathcal{C}_2) = \sum_{\tau \in \mathcal{C}_2} \Pr_{\pi^*}(\tau)$ is small. From the definition, any trajectory $\tau = (x_0, a_0, r_0, \ldots, x_{H-1}, a_{H-1}, r_{H-1}, x_H) \in \mathcal{C}_2$ satisfies that $\forall h \in [H], a_h = \pi^*(x_h)$ and $\exists h \in [H], a_h \neq \pi_f(x_h)$. Then for any $\tau \in \mathcal{C}_2$, we can find a unique index $h' \in [H]$ such that $a_0 = \pi^*(x_0) = \pi_f(x_0), \ldots, a_{h'-1} = \pi^*(x_{h'-1}) = \pi_f(x_{h'-1})$, $a_{h'} = \pi^*(x_{h'}) \neq \pi_f(x_{h'})$ (i.e., h' is the smallest index that π_f differs from π^* in trajectory τ). This implies that $\mathcal{C}_2 \subseteq \bigcup_{h'=0}^{H-1} \mathcal{C}_2^{h'}$, where $\mathcal{C}_2^{h'} = \{\tau = (x_0, a_0, r_0, \ldots, x_{H-1}, a_{H-1}, r_{H-1}, x_H) : a_0 = \pi^*(x_0) = \pi_f(x_0), \ldots, a_{h'-1} = \pi^*(x_{h'-1}) = \pi_f(x_{h'-1}), a_{h'} = \pi^*(x_{h'}) \neq \pi_f(x_{h'})\}$. Since $\mathbb{E}[\mathbf{1}\{\pi_f(x_{h'}) \neq \pi^*(x_{h'}) \mid a_{0:h'-1} \sim \pi^*\}] = \Pr_{\pi^*}(\mathcal{C}_2^{h'})$, we have

$$
\begin{aligned}
\sum_{\tau \in \mathcal{C}_2} \Pr_{\pi^*}(\tau) \leq \sum_{h'=0}^{H-1} \sum_{\tau \in \mathcal{C}_2^{h'}} \Pr_{\pi^*}(\tau) &= \mathbb{E}\left[\sum_{h=0}^{H-1} \mathbf{1}\{\pi_{\hat{f}}(x_h) \neq \pi^*(x_h)\} \mid a_{0:h-1} \sim \pi^*\right] \\
&= \mathbb{E}\left[\sum_{h=0}^{H-1} \mathbf{1}\{\pi_{\hat{f}}(x_h) \neq \pi^*(x_h)\} \mid a_{0:H-1} \sim \pi^*\right].
\end{aligned}
$$

Finally, combining all the results above gives us

$$
\begin{aligned}
V_0^{\pi_{\hat{f}}}(x_0) &\geq V_0^*(x_0) - \sum_{\tau \in \mathcal{C}_2} \Pr_{\pi^*}(\tau) H \\
&\geq V_0^*(x_0) - H\mathbb{E}\left[\sum_{h=0}^{H-1} \mathbf{1}\{\pi_{\hat{f}}(x_h) \neq \pi^*(x_h)\} \mid a_{0:H-1} \sim \pi^*\right] \\
&\geq v^* - H \times \varepsilon / H = v^* - \varepsilon. \qquad (2)
\end{aligned}
$$

This completes the proof. □

Remark We notice that Eq. (1) is the error of supervised learning (SL) with 0/1 loss. Therefore, we can directly use the RL to SL reduction in imitation learning literature (e.g., Theorem 2.1 in Ross and Bagnell (2010)) to translate it to the final performance difference. It gives us the same as our result in Eq. (2). This second part of the proof is different from the one in Ross and Bagnell (2010) and is potentially easier to understand. We believe that it is also of its independent interest.

B PROOF OF ROBUSTNESS RESULTS

In this section, we provide the complete proof of misspecificed cases in Section 5. We start with some helper lemmas in Appendix B.1. Then we show the proof of Theorem 3 in Appendix B.2 and the proof of Theorem 4 in Appendix B.3.

B.1 HELPER LEMMAS

Lemma 3 (Population loss bound for approximately realizable $\mathcal{W}$). *Recall that the definitions of $\varepsilon_{\mathcal{W}}$ and $\tilde{w}^*$ are*

$$
\varepsilon_{\mathcal{W}} = \min_{w \in \mathcal{W}} \max_{f \in \mathcal{F}} \max_{h \in [H]} \left| \mathbb{E}_{d_h^D}[w_h \cdot (f_h - \mathcal{T}_h f_{h+1})] - \mathbb{E}_{d_h^*}[f_h - \mathcal{T}_h f_{h+1}] \right|
$$

and

$$
\tilde{w}^* = \operatorname*{argmin}_{w \in \mathcal{W}} \max_{f \in \mathcal{F}} \max_{h \in [H]} \left| \mathbb{E}_{d_h^D}[w_h \cdot (f_h - \mathcal{T}_h f_{h+1})] - \mathbb{E}_{d_h^*}[f_h - \mathcal{T}_h f_{h+1}] \right|.
$$

For any $f \in \mathcal{F}, h \in [H]$, we have

$$
|\mathcal{E}(f, \pi^*, h)| \leq |\mathbb{E}[\mathcal{L}_{\mathcal{D}}(f, \tilde{w}^*, h)]| + \varepsilon_{\mathcal{W}},
$$

where $\mathcal{E}(\cdot)$ is the Q-type average Bellman error (Jin et al., 2021; Du et al., 2021)

$$
\mathcal{E}(f, \pi, h) = \mathbb{E}[f_h(x_h, a_h) - R_h(x_h, a_h) - f_{h+1}(x_{h+1}, a_{h+1}) \mid a_{0:h} \sim \pi, a_{h+1} \sim \pi_f].
$$

Proof. For any $f \in \mathcal{F}, h \in [H]$, we have

$$
\begin{aligned}
&|\mathcal{E}(f,\pi^*,h)| \\
=& \mathbb{E}[f_h(x_h,a_h) - R_h(x_h,a_h) - f_{h+1}(x_{h+1},a_{h+1}) \mid a_{0:h} \sim \pi^*, a_{h+1} \sim \pi_f]. \\
=& \left|\mathbb{E}_{(x_h,a_h)\sim d_h^*, x_{h+1}\sim P_h(\cdot|x_h,a_h)}[f_h(x_h,a_h) - R_h - f_{h+1}(x_{h+1},\pi_f(x_{h+1}))]\right| \\
=& \left|\mathbb{E}_{(x_h,a_h)\sim d_h^*}[f_h(x_h,a_h) - (\mathcal{T}_h f_{h+1})(x_h,a_h)]\right| \\
=& \left|\mathbb{E}_{d_h^*}[f_h - \mathcal{T}_h f_{h+1}]\right| \\
\le& \left|\mathbb{E}_{d_h^D}[\tilde{w}_h^*(f_h - \mathcal{T}_h f_{h+1}]\right| + \left|\mathbb{E}_{d_h^D}[\tilde{w}_h^* \cdot (f_h - \mathcal{T}_h f_{h+1})] - \mathbb{E}_{d_h^*}[f_h - \mathcal{T}_h f_{h+1}]\right| \\
\le& |\mathbb{E}[\mathcal{L}_{\mathcal{D}}(f,\tilde{w}^*,h)]| + \varepsilon_{\mathcal{W}},
\end{aligned}
$$

which completes the proof. □

Lemma 4 ($\varepsilon_{\mathcal{F}}$ is weaker than ℓ_∞ approximation error)**.** *Recall that the definitions of $\varepsilon_{\mathcal{F}}$ and $\tilde{Q}^*_{\mathcal{F}}$ are*

$$\varepsilon_{\mathcal{F}} = \min_{f\in\mathcal{F}} \max_{w\in\mathcal{W}} \max_{h\in[H]} \left(\left|\mathbb{E}_{d_h^D}[w_h \cdot (f_h - \mathcal{T}_h f_{h+1})]\right| + |f_0(x_0,\pi_f(x_0)) - Q_0^*(x_0,\pi^*(x_0))|\right)$$

and

$$\tilde{Q}^*_{\mathcal{F}} = \operatorname*{argmin}_{f\in\mathcal{F}} \max_{w\in\mathcal{W}} \max_{h\in[H]} \left(\left|\mathbb{E}_{d_h^D}[w_h \cdot (f_h - \mathcal{T}_h f_{h+1})]\right| + |f_0(x_0,\pi_f(x_0)) - Q_0^*(x_0,\pi^*(x_0))|\right).$$

Suppose additionally we have mild regularity assumptions on $\mathcal{W}$, i.e., for any $w \in \mathcal{W}, h \in [H]$, $\mathbb{E}_{d_h^D}[w_h] = 1$ and $w_h \in (\mathcal{X}\times\mathcal{A} \to [0,\infty))$. Then we have

$$\varepsilon_{\mathcal{F}} \le 3 \min_{f\in\mathcal{F}} \max_{h\in[H]} \|f_h - Q_h^*\|_\infty.$$

Proof. For any $f \in \mathcal{F}, w \in \mathcal{W}, h \in [H]$, we have the following

$$
\begin{aligned}
&\left|\mathbb{E}_{d_h^D}[w_h \cdot (f_h - \mathcal{T}_h f_{h+1})]\right| \\
\le& \left|\mathbb{E}_{d_h^D}[w_h \cdot (f_h - Q_h^* - \mathcal{T}_h f_{h+1} + \mathcal{T}_h Q_{h+1}^*)]\right| + \left|\mathbb{E}_{d_h^D}[w_h \cdot (Q_h^* - \mathcal{T}_h Q_{h+1}^*)]\right| \\
\le& \left|\mathbb{E}_{d_h^D}[w_h \cdot (f_h - Q_h^*)]\right| + \left|\mathbb{E}_{d_h^D}[w_h \cdot (\mathcal{T}_h f_{h+1} - \mathcal{T}_h Q_{h+1}^*)]\right| + 0 \\
\le& \mathbb{E}_{d_h^D}[w_h \cdot \|f_h - Q_h^*\|_\infty] + \left|\mathbb{E}_{(x_h,a_h)\sim d_h^D, x_{h+1}\sim P_h(\cdot|x_h,a_h)}[w_h \cdot (f_{h+1}(x_{h+1},\pi_f(x_{h+1}) - Q^*(x_{h+1},\pi^*(x_{h+1}))))]\right| \\
\le& \|f_h - Q_h^*\|_\infty + \mathbb{E}_{(x_h,a_h)\sim d_h^D\cdot w_h, x_{h+1}\sim P_h(\cdot|x_h,a_h)}[|f(x_{h+1},\pi_f(x_{h+1}) - Q_{h+1}^*(x_{h+1},\pi^*(x_{h+1}))|],
\end{aligned} \tag{3}
$$

where the last inequality is due to the $\mathbb{E}_{d_h^D}[w_h] = 1$ and $w_h \ge 0$.

Now, we bound the second term in Eq. (3). Using ε' to denote $\max_{h\in[H]} \|f_h - Q_h^*\|_\infty$, we have

$$
\begin{aligned}
&Q_{h+1}^*(x_{h+1},\pi^*(x_{h+1})) - \varepsilon' \le f_{h+1}(x_{h+1},\pi^*(x_{h+1})) \\
\le& f_{h+1}(x_{h+1},\pi_f(x_{h+1})) \le Q_{h+1}^*(x_{h+1},\pi_f(x_{h+1})) + \varepsilon' \le Q_{h+1}^*(x_{h+1},\pi^*(x_{h+1})) + \varepsilon'.
\end{aligned}
$$

This implies that

$$|f_{h+1}(x_{h+1},\pi_f(x_{h+1}) - Q_{h+1}^*(x_{h+1},\pi^*(x_{h+1}))| \le \varepsilon' = \max_{h\in[H]} \|f_h - Q_h^*\|_\infty.$$

Therefore, we have

$$\left|\mathbb{E}_{d_h^D}[w_h \cdot (f_h - \mathcal{T}_h f_{h+1})]\right| \le \|f_h - Q_h^*\|_\infty + \mathbb{E}_{(x_h,a_h)\sim d_h^D\cdot w_h, x_{h+1}\sim P_h(\cdot|x_h,a_h)}[\|f_{h+1} - Q_{h+1}^*\|_\infty].$$

Since $\mathbb{E}_{d_h^D}[w_h] = 1$, we know that $\mathbb{E}_{(x_h,a_h)\sim d_h^D\cdot w_h, x_{h+1}\sim P_h(\cdot|x_h,a_h)}[\cdot]$ is a probability distribution over x_{h+1}. This implies that

$$\left|\mathbb{E}_{d_h^D}[w_h \cdot (f_h - \mathcal{T}_h f_{h+1})]\right| \le 2 \max_{h\in[H]} \|f_h - Q_h^*\|_\infty.$$

Similarly, we have $|f_0(x_0, \pi_f(x_0)) - Q_0^*(x_0, \pi^*(x_0))| \le \max_{h\in[H]} \|f_h - Q_h^*\|_\infty$, thus

$$\left|\mathbb{E}_{d_h^D}[w_h \cdot (f_h - \mathcal{T}_h f_{h+1})]\right| + |f_0(x_0, \pi_f(x_0)) - Q_0^*(x_0, \pi^*(x_0))| \le 3 \max_{h\in[H]} \|f_h - Q_h^*\|_\infty.$$

Taking $\max$ over $h \in [h], w \in \mathcal{W}$ and then taking $\min$ over $f \in \mathcal{F}$ on both sides completes the proof. $\square$

B.2 PROOF OF THEOREM 3

Theorem (Robust version of Theorem 1, Restatement of Theorem 3)**.** *Suppose Assumption 3, Assumption 4 hold and the total number of samples nH satisfies*

$$nH \ge \frac{8C^2H^5 \log(2|\mathcal{F}||\mathcal{W}|H/\delta)}{\varepsilon^2}.$$

Then with probability $1-\delta$, running Algorithm 1 with $\alpha = \varepsilon/(2H) + \varepsilon_{\mathcal{F}}$ and $C_{\text{gap}} = 0$ guarantees

$$|V_{\hat{f}}(x_0) - v^*| \le \varepsilon + H\varepsilon_{\mathcal{F}} + H\varepsilon_{\mathcal{W}}.$$

Proof. From Lemma 1 and our choice $n \ge \frac{8C^2H^4 \log(2|\mathcal{F}||\mathcal{W}|H/\delta)}{\varepsilon^2}$, with probability at least $1-\delta$, for any $f \in \mathcal{F}, w \in \mathcal{W}, h \in [H]$, we have

$$|\mathcal{L}_{\mathcal{D}}(f, w, h) - \mathbb{E}[\mathcal{L}_{\mathcal{D}}(f, w, h)]| \le \varepsilon_{\text{stat},n} \le \varepsilon/(2H).$$

Throughout the proof, we will condition on this high probability event.

From Lemma 2, we have

$$\begin{aligned}
|\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\tilde{Q}_{\mathcal{F}}^*, w, h)]| = & \left|\mathbb{E}_{(x_h,a_h)\sim d_h^D}[w_h(x_h,a_h)(\tilde{Q}_{\mathcal{F},h}^*(x_h,a_h) - (\mathcal{T}_h\tilde{Q}_{\mathcal{F},h+1}^*)(x_h,a_h))]\right| \\
\le & \left|\mathbb{E}_{(x_h,a_h)\sim d_h^D}[w_h(x_h,a_h)(\tilde{Q}_{\mathcal{F},h}^*(x_h,a_h) - (\mathcal{T}_h\tilde{Q}_{\mathcal{F},h+1}^*)(x_h,a_h))]\right| \\
& + \left|\tilde{Q}_{\mathcal{F},0}^*(x_0, \pi_{\tilde{Q}_{\mathcal{F}}^*}(x_0)) - Q_0^*(x_0, \pi^*(x_0))\right| \\
\le & \ \varepsilon_{\mathcal{F}}.
\end{aligned}$$

When using the relaxed constraints by setting $\alpha = \varepsilon/(2H) + \varepsilon_{\mathcal{F}}$, we can incorporate the approximation errors. More specifically, we have

$$\left|\mathcal{L}_{\mathcal{D}}(\tilde{Q}_{\mathcal{F}}^*, w, h)\right| \le \left|\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\tilde{Q}_{\mathcal{F}}^*, w, h)]\right| + \varepsilon_{\text{stat},n} \le \varepsilon_{\mathcal{F}} + \varepsilon_{\text{stat},n} \le \varepsilon/(2H) + \varepsilon_{\mathcal{F}} = \alpha,$$

which implies that $\tilde{Q}_{\mathcal{F}}^*$ will satisfy all constraints.

In addition, for any $f \in \mathcal{F}$ that satisfies all constraints, we have that for any $w \in \mathcal{W}, h \in [H]$,

$$|\mathbb{E}[\mathcal{L}_{\mathcal{D}}(f, w, h)]| \le \mathcal{L}_{\mathcal{D}}(f, w, h) + \varepsilon_{\text{stat},n} \le \alpha + \varepsilon_{\text{stat},n} = \varepsilon/H + \varepsilon_{\mathcal{F}}.$$

From Lemma 3, we further have

$$|\mathcal{E}(f, \pi^*, h)| \le |\mathbb{E}[\mathcal{L}_{\mathcal{D}}(f, \tilde{w}^*, h)]| + \varepsilon_{\mathcal{W}}.$$

Since $\tilde{w}^* \in \mathcal{W}$, we get

$$|\mathcal{E}(f, \pi^*, h)| \le |\mathbb{E}[\mathcal{L}_{\mathcal{D}}(f, \tilde{w}^*, h)]| + \varepsilon_{\mathcal{W}} \le \varepsilon/H + \varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{W}} := \varepsilon'.$$

Following telescoping step in the proof of Theorem 1, for any $f \in \mathcal{F}, h \in [H]$ that satisfies all constraints, we have

$$V_f(x_0) = f_0(x_0, \pi_f(x_0)) \ge V_0^*(x_0) - H\varepsilon'.$$

Therefore, we have

$$V_0^*(x_0) + \varepsilon_{\mathcal{F}} = Q_0^*(x_0, \pi^*(x_0)) + \varepsilon_{\mathcal{F}} \ge \tilde{Q}_0^*(x_0, \pi_{\tilde{Q}^*}(x_0)) \ge \hat{f}_0(x_0, \pi_{\hat{f}}(x_0)) \ge V_0^*(x_0) - H\varepsilon',$$

where the first inequality is due to the definition of approximation error $\varepsilon_{\mathcal{F}}$ and the second inequality is due to pessimism. This gives us

$$|V_{\hat{f}}(x_0) - v^*| \le \max\{H\varepsilon', \varepsilon_{\mathcal{F}}\} \le \varepsilon + H\varepsilon_{\mathcal{F}} + H\varepsilon_{\mathcal{W}},$$

which completes the proof. $\square$

B.3 PROOF OF THEOREM 4

Theorem (Robust version of Theorem 2, restatement of Theorem 4)**.** *Suppose Assumption 3, Assumption 4 hold and the total number of samples nH satisfies*

$$nH \geq \frac{8C^2 H^7 \log(2|\mathcal{F}||\mathcal{W}|H/\delta)}{\varepsilon^2 C_{\text{gap}}^2}.$$

Then with probability $1-\delta$, running Algorithm 1 with a user-specified C_{gap} and $\alpha = \varepsilon C_{\text{gap}}/(2H^2) + \varepsilon_{\mathcal{F}(C_{\text{gap}})}$ guarantees

$$v^{\pi_{\hat{f}}} \geq v^* - \varepsilon - \frac{(H^2+H)\varepsilon_{\mathcal{F}(C_{\text{gap}})} + H^2 \varepsilon_{\mathcal{W}}}{C_{\text{gap}}}.$$

Proof. From Lemma 1 and our choice $n \geq \frac{8C^2 H^6 \log(2|\mathcal{F}||\mathcal{W}|H/\delta)}{\varepsilon^2 C_{\text{gap}}^2}$, with probability at least $1-\delta$, for any $f \in \mathcal{F}, w \in \mathcal{W}, h \in [H]$, we have

$$|\mathcal{L}_{\mathcal{D}}(f,w,h) - \mathbb{E}[\mathcal{L}_{\mathcal{D}}(f,w,h)]| \leq \varepsilon_{\text{stat},n} \leq \varepsilon C_{\text{gap}}/(2H^2).$$

Throughout the proof, we will condition on this high probability event.

From Lemma 2, we have

$$\begin{aligned}
|\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\tilde{Q}^*_{\mathcal{F}(C_{\text{gap}})}, w, h)]| = &\left|\mathbb{E}_{(x_h,a_h)\sim d_h^D}[w_h(x_h,a_h)(\tilde{Q}^*_{\mathcal{F}(C_{\text{gap}}),h}(x_h,a_h) - (\mathcal{T}_h \tilde{Q}^*_{\mathcal{F}(C_{\text{gap}}),h+1})(x_h,a_h))]\right| \\
\leq &\left|\mathbb{E}_{(x_h,a_h)\sim d_h^D}[w_h(x_h,a_h)(\tilde{Q}^*_{\mathcal{F}(C_{\text{gap}}),h}(x_h,a_h) - (\mathcal{T}_h \tilde{Q}^*_{\mathcal{F}(C_{\text{gap}}),h+1})(x_h,a_h))]\right| \\
&+ \left|\tilde{Q}^*_{\mathcal{F}(C_{\text{gap}}),0}(x_0, \pi_{\tilde{Q}^*_{\mathcal{F}(C_{\text{gap}})}}(x_0)) - Q_0^*(x_0,\pi^*(x_0))\right| \\
\leq &\ \varepsilon_{\mathcal{F}(C_{\text{gap}})}.
\end{aligned}$$

When using the relaxed constraints of $\alpha = \varepsilon C_{\text{gap}}/(2H^2) + \varepsilon_{\mathcal{F}(C_{\text{gap}})}$, we can incorporate the approximation errors. More specifically, we have

$$\begin{aligned}
\left|\mathcal{L}_{\mathcal{D}}(\tilde{Q}^*_{\mathcal{F}(C_{\text{gap}})}, w, h)\right| \leq &\left|\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\tilde{Q}^*_{\mathcal{F}(C_{\text{gap}})}, w, h)]\right| + \varepsilon_{\text{stat},n} \\
\leq &\ \varepsilon_{\mathcal{F}(C_{\text{gap}})} + \varepsilon_{\text{stat},n} \\
\leq &\ \varepsilon C_{\text{gap}}/(2H^2) + \varepsilon_{\mathcal{F}(C_{\text{gap}})} = \alpha,
\end{aligned}$$

which implies that $\tilde{Q}^*_{\mathcal{F}(C_{\text{gap}})}$ will satisfy all constraints.

In addition, for any $f \in \mathcal{F}(C_{\text{gap}})$ that satisfies all constraints, we have that for any $w \in \mathcal{W}, h \in [H]$,

$$|\mathbb{E}[\mathcal{L}_{\mathcal{D}}(f,w,h)]| \leq \mathcal{L}_{\mathcal{D}}(f,w,h) + \varepsilon_{\text{stat},n} \leq \alpha + \varepsilon_{\text{stat},n} = \varepsilon C_{\text{gap}}/H^2 + \varepsilon_{\mathcal{F}(C_{\text{gap}})}.$$

From Lemma 3, we further have

$$|\mathcal{E}(f,\pi^*,h)| \leq |\mathbb{E}[\mathcal{L}_{\mathcal{D}}(f,\tilde{w}^*,h)]| + \varepsilon_{\mathcal{W}}.$$

Since $\tilde{w}^* \in \mathcal{W}$, we get

$$|\mathcal{E}(f,\pi^*,h)| \leq |\mathbb{E}[\mathcal{L}_{\mathcal{D}}(f,\tilde{w}^*,h)]| + \varepsilon_{\mathcal{W}} \leq \varepsilon C_{\text{gap}}/H^2 + \varepsilon_{\mathcal{F}(C_{\text{gap}})} + \varepsilon_{\mathcal{W}} := \varepsilon'.$$

Since we run the algorithm on $\mathcal{F}(C_{\text{gap}})$, the gap parameter will be C_{gap} instead of $\text{gap}(Q^*)$ in Theorem 2. Following the proof of Theorem 2, for any $f \in \mathcal{F}(C_{\text{gap}}), h \in [H]$ that satisfies all constraints, we have

$$V_f(x_0) = f_0(x_0,\pi_f(x_0)) \geq Q_0^*(x_0,\pi^*(x_0)) + C_{\text{gap}}\mathbb{E}\left[\sum_{h=0}^{H-1} \mathbf{1}\{\pi_f(x_h) \neq \pi^*(x_h)\} \mid a_{0:H-1} \sim \pi^*\right] - H\varepsilon'.$$

Therefore, we have

$$Q_0^*(x_0,\pi^*(x_0)) + \varepsilon_{\mathcal{F}(C_{\text{gap}})}$$

$$
\begin{aligned}
&\geq \tilde{Q}^*_{\mathcal{F}(C_{\text{gap}}),0}(x_0, \pi_{Q^*_{\mathcal{F}(C_{\text{gap}})}}(x_0)) && \text{(Definition of approximation error } \varepsilon_{\mathcal{F}(C_{\text{gap}})}) \\
&\geq \hat{f}_0(x_0, \pi_{\hat{f}}(x_0)) && \text{(Pessimism)} \\
&\geq Q^*_0(x_0, \pi^*(x_0)) + C_{\text{gap}} \mathbb{E}\left[\sum_{h=0}^{H-1} \mathbf{1}\{\pi_f(x_h) \neq \pi^*(x_h)\} \mid a_{0:H-1} \sim \pi^*\right] - H\varepsilon',
\end{aligned}
$$

which yields

$$
\mathbb{E}\left[\sum_{h=0}^{H-1} \mathbf{1}\{\pi_f(x_h) \neq \pi^*(x_h)\} \mid a_{0:H-1} \sim \pi^*\right] \leq \left(H\varepsilon' + \varepsilon_{\mathcal{F}(C_{\text{gap}})}\right)/C_{\text{gap}}.
$$

This translates to the performance difference bound of

$$
V_0^{\pi_{\hat{f}}}(x_0) \geq v^* - H\left(H\varepsilon' + \varepsilon_{\mathcal{F}(C_{\text{gap}})}\right)/C_{\text{gap}} \geq v^* - \varepsilon - \frac{(H^2+H)\varepsilon_{\mathcal{F}(C_{\text{gap}})} + H^2\varepsilon_{\mathcal{W}}}{C_{\text{gap}}},
$$

which completes the proof. □

B.4 COROLLARY FROM THEOREM 4

Theorem 4 gives us a convenient way to set the gap parameter C_{gap}. We show that it can easily handle the case that ℓ_∞ approximation error of $\mathcal{F}$ and $\text{gap}(Q^*)$ are known. We formally define ℓ_∞ approximation error and the corresponding best approximator w.r.t. $\mathcal{F}$ as

$$
\varepsilon_{\mathcal{F},\infty} = \min_{f \in \mathcal{F}} \max_{h \in [H]} \|f_h - Q^*_h\|_\infty, \quad \tilde{Q}^*_{\mathcal{F},\infty} = \operatorname*{argmin}_{f \in \mathcal{F}} \max_{h \in [H]} \|f_h - Q^*_h\|_\infty.
$$

Similarly, we can define the version for $\mathcal{F}(\text{gap}(Q^*))$.

Then we have the following corollary.

Corollary 5 (Corollary from Theorem 4)**.** *Suppose Assumption 3, Assumption 4 hold, the weight function class satisfies the additional mild regularity assumptions stated in Lemma 4. Assume we are given $\varepsilon_{\mathcal{F},\infty}$, $\text{gap}(Q^*)$ and $2\varepsilon_{\mathcal{F},\infty} < \text{gap}(Q^*)$. If the total number of samples nH satisfies*

$$
nH \geq \frac{8C^2H^7 \log(2|\mathcal{F}||\mathcal{W}|H/\delta)}{\varepsilon^2(\text{gap}(Q^*) - 2\varepsilon_{\mathcal{F},\infty})^2},
$$

then with probability $1-\delta$, running Algorithm 1 with $C_{\text{gap}} = \text{gap}(Q^*) - 2\varepsilon_{\mathcal{F},\infty}$ and $\alpha = \varepsilon(\text{gap}(Q^*) - 2\varepsilon_{\mathcal{F},\infty})/(2H^2) + 2\varepsilon_{\mathcal{F},\infty}$ guarantees

$$
v^{\pi_{\hat{f}}} \geq v^* - \varepsilon - \frac{(2H^2+H)\varepsilon_{\mathcal{F},\infty} + H^2\varepsilon_{\mathcal{W}}}{\text{gap}(Q^*) - 2\varepsilon_{\mathcal{F},\infty}}.
$$

Proof. From the definition of $\text{gap}(Q^*)$, $\varepsilon_{\mathcal{F},\infty}$ and $\tilde{Q}^*_{\mathcal{F},\infty}$, we know that

$$
\text{gap}(\tilde{Q}^*_{\mathcal{F},\infty}) \geq \text{gap}(Q^*) - 2\varepsilon_{\mathcal{F},\infty} > 0.
$$

Therefore, we have $\tilde{Q}^*_{\mathcal{F},\infty} \in \mathcal{F}(\text{gap}(Q^*) - 2\varepsilon_{\mathcal{F},\infty})$. Together with the definition that $\tilde{Q}^*_{\mathcal{F},\infty}$ is the best approximator of Q^* within $\mathcal{F}$ (under ℓ_∞ norm), we know that $\tilde{Q}^*_{\mathcal{F},\infty}$ is also the best approximator within $\mathcal{F}(\text{gap}(Q^*) - 2\varepsilon_{\mathcal{F},\infty})$ (under ℓ_∞ norm). This implies that

$$
\varepsilon_{\mathcal{F}(\text{gap}(Q^*) - 2\varepsilon_{\mathcal{F},\infty}),\infty} = \varepsilon_{\mathcal{F},\infty}.
$$

In addition, under the mild regularity assumptions stated in Lemma 4, applying Lemma 4 tells us

$$
\varepsilon_{\mathcal{F}(\text{gap}(Q^*) - 2\varepsilon_{\mathcal{F},\infty})} \leq 3 \min_{f \in \mathcal{F}(\text{gap}(Q^*) - 2\varepsilon_{\mathcal{F},\infty})} \max_{h \in [H]} \|f_h - Q^*_h\|_\infty = 3\varepsilon_{\mathcal{F}(\text{gap}(Q^*) - 2\varepsilon_{\mathcal{F},\infty}),\infty} = 3\varepsilon_{\mathcal{F},\infty}.
$$

The remaining part of the proof follows a similar approach as the proof of Theorem 4. Firstly, we have the $1-\delta$ high probability event that for any $f \in \mathcal{F}, w \in \mathcal{W}, h \in [H]$

$$
|\mathcal{L}_{\mathcal{D}}(f, w, h) - \mathbb{E}[\mathcal{L}_{\mathcal{D}}(f, w, h)]| \leq \varepsilon_{\text{stat},n} \leq \varepsilon(\text{gap}(Q^*) - 2\varepsilon_{\mathcal{F},\infty})/(2H^2).
$$

Then following the proof Lemma 4, we have

$$\begin{aligned}|\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\tilde{Q}^*_{\mathcal{F},\infty}, w, h)]| = & \left|\mathbb{E}_{d_h^D}[w_h \cdot (\tilde{Q}^*_{\mathcal{F},\infty,h} - \mathcal{T}_h \tilde{Q}^*_{\mathcal{F},\infty,h+1})]\right| \\ \leq & \left|\mathbb{E}_{d_h^D}[w_h \cdot (\tilde{Q}^*_{\mathcal{F},\infty,h} - Q^*_h)]\right| + \left|\mathbb{E}_{d_h^D}[w_h \cdot (\mathcal{T}_h \tilde{Q}^*_{\mathcal{F},\infty,h+1} - \mathcal{T}_h Q^*_{h+1})]\right| + 0 \\ \leq & \ 2 \max_{h\in[H]} \|\tilde{Q}^*_{\mathcal{F},\infty,h} - Q^*_h\|_\infty = 2\varepsilon_{\mathcal{F},\infty}.\end{aligned}$$

The empirical loss of $\tilde{Q}^*_{\mathcal{F},\infty}$ satisfies

$$\begin{aligned}\left|\mathcal{L}_{\mathcal{D}}(\tilde{Q}^*_{\mathcal{F},\infty}, w, h)\right| \leq & \left|\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\tilde{Q}^*_{\mathcal{F},\infty}, w, h)]\right| + \varepsilon_{\text{stat},n} \\ \leq & \ \varepsilon(\text{gap}(Q^*) - 2\varepsilon_{\mathcal{F},\infty})/(2H^2) + 2\varepsilon_{\mathcal{F},\infty} = \alpha,\end{aligned}$$

which implies that $\tilde{Q}^*_{\mathcal{F},\infty}$ will satisfy all constraints.

In addition, for any $f \in \mathcal{F}(\text{gap}(Q^*) - 2\varepsilon_{\mathcal{F},\infty})$ that satisfies all constraints, we have that for any $w \in \mathcal{W}, h \in [H]$,

$$|\mathbb{E}[\mathcal{L}_{\mathcal{D}}(f, w, h)]| \leq \mathcal{L}_{\mathcal{D}}(f, w, h) + \varepsilon_{\text{stat},n} \leq \alpha + \varepsilon_{\text{stat},n} = \varepsilon(\text{gap}(Q^*) - 2\varepsilon_{\mathcal{F},\infty})/H^2 + 2\varepsilon_{\mathcal{F},\infty}.$$

Similarly, we further have

$$|\mathcal{E}(f, \pi^*, h)| \leq |\mathbb{E}[\mathcal{L}_{\mathcal{D}}(f, \tilde{w}^*, h)]| + \varepsilon_{\mathcal{W}} \leq \varepsilon(\text{gap}(Q^*) - 2\varepsilon_{\mathcal{F},\infty})/H^2 + 2\varepsilon_{\mathcal{F},\infty} + \varepsilon_{\mathcal{W}} := \varepsilon'.$$

The final performance difference bound is

$$V_0^{\pi_{\hat{f}}}(x_0) \geq v^* - H\left(H\varepsilon' + \varepsilon_{\mathcal{F},\infty}\right)/(\text{gap}(Q^*) - 2\varepsilon_{\mathcal{F},\infty}) \geq v^* - \varepsilon - \frac{(2H^2 + H)\varepsilon_{\mathcal{F},\infty} + H^2\varepsilon_{\mathcal{W}}}{\text{gap}(Q^*) - 2\varepsilon_{\mathcal{F},\infty}},$$

where the difference compared with the derivation in the proof of Theorem 4 is that we use ℓ_∞ bound to get

$$Q_0^*(x_0, \pi^*(x_0)) + \varepsilon_{\mathcal{F},\infty} \geq \tilde{Q}^*_{\mathcal{F},\infty,0}(x_0, \pi_{Q^*_{\mathcal{F},\infty}}(x_0)).$$

This completes the proof. $\square$

C PROOF OF THE UNKNOWN GAP PARAMETER SETTING

In this section, we present the formal proof of Theorem 5. We start with a standard helper lemma in Appendix C.1, which shows the concentration result of Monte Carlo estimate. Then we show the proof of Theorem 5 in Appendix C.2.

C.1 A HELPER LEMMA

Lemma 6 (Concentration for Monte Carlo estimate)**.** *Assume we run policy π and collect m trajectories $\left\{x_0^{(i)}, a_0^{(i)}, r_0^{(i)}, \ldots, x_{H-1}^{(i)}, a_{H-1}^{(i)}, r_{H-1}^{(i)}\right\}_{i=1}^m$ and our Monte Carlo estimate is defined as*

$$\hat{v}^\pi := \frac{1}{m}\sum_{i=1}^m \sum_{h=0}^{H-1} r_h^{(i)}.$$

Then we have

$$|\hat{v}^\pi - v^\pi| \leq 2H\sqrt{\frac{\log(2/\delta)}{2m}}.$$

Proof. Define random variable $Y_i := \sum_{h=0}^{H-1} r_h^{(i)}$. From the definition, we know that Y_i are i.i.d. samples with mean v^π. Applying Hoeffding's inequality and noticing that $|Y_i| \leq H$ gives us with probability $1-\delta$,

$$\left|\frac{1}{m}\sum_{i=1}^m Y_i - v^\pi\right| \leq 2H\sqrt{\frac{\log(2/\delta)}{2m}}.$$

This completes the proof. $\square$

C.2 PROOF OF THEOREM 5

Theorem (Sample complexity of finding a near-optimal policy with unknown $\text{gap}(Q^*)$, restatement of Theorem 5)**.** *Suppose Assumption 1, Assumption 2, Assumption 3, Assumption 4, Assumption 5 hold but $\text{gap}(Q^*)$ is unknown. Assume we have a dataset $\mathcal{D}$ with size n for each $\mathcal{D}_h$ and additional online access to collect*

$$(\log(2H/\text{gap}(Q^*)))^2 \cdot \frac{n\log(24/\delta)}{C^2H} = \tilde{O}\left(\frac{n\log(1/\delta)}{C^2H}\right)$$

samples. Then with probability at least $1-\delta$, the output policy $\hat{\pi}$ from Algorithm 2 satisfies

$$v^{\hat{\pi}} \geq v^* - 5\sqrt{\frac{32C^2H^6\iota(\log(2H/\text{gap}(Q^*)))}{n\text{gap}(Q^*)^2}},$$

where $\iota(t) = \log(24|\mathcal{F}||\mathcal{W}|H \cdot 2^t/\delta)$.

Proof. For Theorem 1, Theorem 2 and Monte Carlo roll out estimate at iteration t, we set their high probability event parameter as $\delta_t' := \delta/(6\times 2^t)$. Then union bounding over all of them gives us $1-\delta$ high probability event. Our following analysis is conditioned on these high probability events.

Firstly, we show that Algorithm 2 will terminate once our guess $\text{gap}_t^{\text{guess}}$ drops below the true $\text{gap}(Q^*)$. From Theorem 1, we know that $|\hat{v}_t^* - v^*| \leq \varepsilon_t$. Further, when $\text{gap}_t^{\text{guess}} \leq \text{gap}(Q^*)$, we can guarantee that $Q^* \in \mathcal{F}(\text{gap}_t^{\text{guess}})$. Therefore, Theorem 2 tells us $v^{\hat{\pi}_t} \geq v^* - \varepsilon_t$. Finally, for Monte Carlo estimate $\hat{v}^{\hat{\pi}_t}$, we have $|\hat{v}^{\hat{\pi}_t} - v^{\hat{\pi}_t}| \leq \varepsilon_t$. Combining them together yields

$$\hat{v}^{\hat{\pi}_t} \geq v^{\hat{\pi}_t} - \varepsilon_t \geq v^* - \varepsilon_t - \varepsilon_t \geq \hat{v}_t^* - \varepsilon_t - \varepsilon_t - \varepsilon_t = \hat{v}_t^* - 3\varepsilon_t,$$

which means our algorithm will stop in this iteration.

So if we assume the algorithm terminates at iteration T, then T satisfies $H/2^T \geq \text{gap}(Q^*)/2$, thus

$$T \leq \log(2H/\text{gap}(Q^*)).$$

Then we prove that the output policy $\hat{\pi}_T$ satisfies $v^{\hat{\pi}_T} \geq v^* - 5\varepsilon_t$. This can be seen from

$$v^{\hat{\pi}_T} \geq \hat{v}^{\hat{\pi}_T} - \varepsilon_T \geq \hat{v}_T^* - 3\varepsilon_T - \varepsilon_T \geq v^* - \varepsilon_T - 3\varepsilon_T - \varepsilon_T = v^* - 5\varepsilon_T.$$

Notice that ε_t will increase as t increases. Therefore, if our algorithm terminates before $\text{gap}_t^{\text{guess}}$ drops below $\text{gap}(Q^*)$, we will have a better performance guarantee. More specifically, we have

$$\varepsilon_T \leq \varepsilon_{\log(2H/\text{gap}(Q^*))} = \sqrt{\frac{32C^2H^6\iota(\log(2H/\text{gap}(Q^*)))}{n\text{gap}(Q^*)^2}}.$$

Therefore, $\hat{\pi}_T$ satisfies

$$v^{\hat{\pi}_T} \geq v^* - 5\sqrt{\frac{32C^2H^6\iota(\log(2H/\text{gap}(Q^*)))}{n\text{gap}(Q^*)^2}},$$

which has the same order of the accuracy as running Algorithm 1 with known $\text{gap}(Q^*)$ in Theorem 2 up to polylog terms.

Finally we calculate the required number of online samples. For iteration t, applying Lemma 6, we require

$$H \cdot \frac{2H^2\log(12\times 2^t/\delta)}{\varepsilon_t^2} \leq \frac{2H^3\log(12\times 2^T/\delta)}{\varepsilon_t^2} = \frac{n\log(12\times 2^T/\delta)}{4C^2H\iota(t)2^{2t}} \leq \frac{n\log(12\times 2^T/\delta)}{C^2H} \leq \frac{nT\log(12\times 2/\delta)}{C^2H}$$

samples. Then since we have at most $\log(2H/\text{gap}(Q^*))$ iterations, the required number of online samples is at most

$$\log(2H/\text{gap}(Q^*)) \cdot \frac{nT\log(12\times 2/\delta)}{C^2H} \leq (\log(2H/\text{gap}(Q^*)))^2 \cdot \frac{n\log(24/\delta)}{C^2H}.$$

This completes the proof. $\square$

D LAGRANGIAN FORM ALGORITHM AND RESULTS

In this section, we introduce the Lagrangian form variant of PABC (Algorithm 1) and its sample complexity guarantees. We start with showing its variant PABC-L (Algorithm 1) in Appendix D.1. Then we provide the main results of PABC-L in Appendix D.2 and its robustness results in Appendix D.3.

D.1 ALGORITHM

In this part, we introduce the PABC-L (PABC with Lagrangian form) algorithm as shown in Algorithm 1. Compared with PABC (Algorithm 1), PABC-L does not take the threshold α as input. In addition, it moves the constraints (Eq. (2)) to the objective (Eq. (5)). Furthermore, to estimate v^*, it returns $\hat{f}_0(x_0, \pi_{\hat{f}}(x_0)) + H \cdot \max_{w\in\mathcal{W}, h\in[H]} |\mathcal{L}_{\mathcal{D}}(\hat{f}, w, h)|$ instead of $\hat{f}_0(x_0, \pi_{\hat{f}}(x_0))$.

Algorithm 1 PABC-L (PABC with Lagrangian form)

Input: gap factor C_{gap}, function class $\mathcal{F}$, weight function class $\mathcal{W}$, and dataset $\mathcal{D}$.

1: Perform prescreening according to input C_{gap}:

$$\mathcal{F}(C_{\text{gap}}) := \{f \in \mathcal{F} : \text{gap}(f) \geq C_{\text{gap}}\}. \tag{4}$$

2: Find the pessimism value function in $\mathcal{F}(C_{\text{gap}})$ with the Lagrangian form objective

$$\hat{f} = \underset{f\in\mathcal{F}(C_{\text{gap}})}{\text{argmin}} \left(f_0(x_0, \pi_f(x_0)) + H \cdot \max_{w\in\mathcal{W}, h\in[H]} |\mathcal{L}_{\mathcal{D}}(f, w, h)| \right) \tag{5}$$

where the empirical loss $\mathcal{L}_{\mathcal{D}}(f, w, h)$ is defined as

$$\mathcal{L}_{\mathcal{D}}(f, w, h) = \frac{1}{n}\sum_{i=1}^{n}[w_h(x_h^{(i)}, a_h^{(i)})(f_h(x_h^{(i)}, a_h^{(i)}) - r_h^{(i)} - f_{h+1}(x_{h+1}^{(i)}, \pi_f(x_{h+1}^{(i)})))]. \tag{6}$$

Output: policy $\pi_{\hat{f}}$ and return estimation $\hat{f}_0(x_0, \pi_{\hat{f}}(x_0)) + H \cdot \max_{w\in\mathcal{W}, h\in[H]} |\mathcal{L}_{\mathcal{D}}(\hat{f}, w, h)|$.

Remark In the objective (Eq. (5)), we can also use

$$\hat{f} = \underset{f\in\mathcal{F}(C_{\text{gap}})}{\text{argmin}} \left(f_0(x_0, \pi_f(x_0)) + \sum_{h=0}^{H-1} \max_{w\in\mathcal{W}} |\mathcal{L}_{\mathcal{D}}(f, w, h)| \right). \tag{7}$$

From the detailed proofs in the subsequent parts, it is easy to see that the theoretical results hold under this objective (Eq. (7)).

D.2 MAIN GUARANTEES

In this part, we present the main sample complexity results of PABC-L (Algorithm 1). In parallel with Section 4, we show that PABC-L can identify v^* without the gap assumption in Appendix D.2.1 and show that PABC-L with the gap assumption learns a near-optimal policy in Appendix D.2.2.

D.2.1 ESTIMATING OPTIMAL EXPECTED RETURN

We show the sample complexity bound and the proof for PABC-L to identify v^*. The bound is the same as that of PABC (Theorem 1).

Theorem 7 (Sample complexity of identifying v^*, Lagrangian version)**.** *Suppose Assumption 1, Assumption 2, Assumption 3, Assumption 4 hold and the total number of samples nH satisfies*

$$nH \geq \frac{8C^2H^5 \log(2|\mathcal{F}||\mathcal{W}|H/\delta)}{\varepsilon^2}.$$

Then with probability at least $1-\delta$, running Algorithm 1 with $C_{\text{gap}} = 0$ guarantees

$$|V_{\hat{f}}(x_0) - v^*| \leq \varepsilon.$$

Proof. The proof mostly follows the proof of Theorem 1, and we only show the different and crucial steps here. We still condition on the high probability event from concentration (Lemma 1).

From the concentration result and the choice of n, we have the bound for Q^*:

$$V_0^*(x_0) + H \cdot \max_{w\in\mathcal{W}, h\in[H]} |\mathcal{L}_{\mathcal{D}}(Q^*, w, h)| \leq V_0^*(x_0) + H\varepsilon_{\text{stat},n},$$

where $\varepsilon_{\text{stat},n} \leq \varepsilon/H$.

From pessimism and the objective in Algorithm 1, we have

$$V_0^*(x_0) + H \cdot \max_{w\in\mathcal{W}, h\in[H]} |\mathcal{L}_{\mathcal{D}}(Q^*, w, h)| \geq \ V_{\hat{f}}(x_0) + H \cdot \max_{w\in\mathcal{W}, h\in[H]} |\mathcal{L}_{\mathcal{D}}(\hat{f}, w, h)|.$$

Therefore, we get

$$V_0^*(x_0) + H\varepsilon_{\text{stat},n} \geq V_{\hat{f}}(x_0) + H \cdot \max_{w\in\mathcal{W}, h\in[H]} |\mathcal{L}_{\mathcal{D}}(\hat{f}, w, h)|. \tag{8}$$

For any $f \in \mathcal{F}$, following the telescoping step in the proof of Theorem 1, we know that

$$\begin{aligned}
V_f(x_0) =&\ f_0(x_0, \pi_f(x_0)) \\
\geq&\ f_0(x_0, \pi^*(x_0)) \\
=&\ \mathbb{E}[R_0(x_0, a_0) + f_1(x_1, a_1) \mid a_0 \sim \pi^*, a_1 \sim \pi_f] + \mathcal{E}(f, \pi^*, 0) \\
\geq&\ \mathbb{E}[R_0(x_0, a_0) \mid a_0 \sim \pi^*] + \mathbb{E}[f_1(x_1, a_1) \mid a_{0:1} \sim \pi^*] + \mathcal{E}(f, \pi^*, 0) \\
\geq&\ \mathbb{E}[R_0(x_0, a_0) \mid a_0 \sim \pi^*] + \mathbb{E}[R_1(x_1, a_1) + f_2(x_2, a_2) \mid a_{0:1} \sim \pi^*, a_2 \sim \pi_f] + \mathcal{E}(f, \pi^*, 1) + \mathcal{E}(f, \pi^*, 0) \\
\geq&\ \ldots \\
\geq&\ \mathbb{E}\left[\sum_{h=0}^{H-1} R_h(x_h, a_h) \mid a_{0:H-1} \sim \pi^*\right] + \sum_{h=0}^{H-1} \mathcal{E}(f, \pi^*, h) \\
\geq&\ V_0^*(x_0) - \sum_{h=0}^{H-1} |\mathcal{E}(f, \pi^*, h)|.
\end{aligned}$$

Therefore, we get

$$\begin{aligned}
& V_{\hat{f}}(x_0) + H \cdot \max_{w\in\mathcal{W}, h\in[H]} |\mathcal{L}_{\mathcal{D}}(\hat{f}, w, h)| \\
\geq&\ V_0^*(x_0) - \sum_{h=0}^{H-1} |\mathcal{E}(\hat{f}, \pi^*, h)| + H \cdot \max_{w\in\mathcal{W}, h\in[H]} |\mathcal{L}_{\mathcal{D}}(\hat{f}, w, h)| \\
\geq&\ V_0^*(x_0) - \sum_{h=0}^{H-1} |\mathcal{E}(\hat{f}, \pi^*, h)| + H \cdot \max_{w\in\mathcal{W}, h\in[H]} |\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\hat{f}, w, h)]| - H\varepsilon_{\text{stat},n} \\
\geq&\ V_0^*(x_0) - \sum_{h=0}^{H-1} |\mathcal{E}(\hat{f}, \pi^*, h)| + \sum_{h=0}^{H-1} |\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\hat{f}, w^*, h)]| - H\varepsilon_{\text{stat},n} \\
=&\ V_0^*(x_0) - \sum_{h=0}^{H-1} |\mathcal{E}(\hat{f}, \pi^*, h)| + \sum_{h=0}^{H-1} |\mathcal{E}(\hat{f}, \pi^*, h)| - H\varepsilon_{\text{stat},n}
\end{aligned}$$

$$= V_0^*(x_0) - H\varepsilon_{\text{stat},n}. \tag{9}$$

Combining Eq. (8) and Eq. (9) yields

$$|V_{\hat{f}}(x_0) + H \cdot \max_{w\in\mathcal{W},h\in[H]} |\mathcal{L}_{\mathcal{D}}(\hat{f}, w, h)| - v^*| = |V_{\hat{f}}(x_0) + H \cdot \max_{w\in\mathcal{W},h\in[H]} |\mathcal{L}_{\mathcal{D}}(\hat{f}, w, h)| - V_0^*(x_0)| \le H\varepsilon_{\text{stat},n} \le \varepsilon,$$

which completes the proof. □

D.2.2 LEARNING A NEAR-OPTIMAL POLICY

Here we present the result for learning a near optimal policy. Compared with its counterpart (Theorem 2), the sample complexity only differs in the constant.

Theorem 8 (Sample complexity of learning a near-optimal policy, Lagrangian version)**.** *Suppose Assumption 1, Assumption 2, Assumption 3, Assumption 4, Assumption 5 hold and the total number of samples nH satisfies*

$$nH \ge \frac{32C^2H^7\log(2|\mathcal{F}||\mathcal{W}|H/\delta)}{\varepsilon^2 \text{gap}(Q^*)^2}.$$

Then with probability at least $1-\delta$, running Algorithm 1 with $C_{\text{gap}} = \text{gap}(Q^*)$ guarantees

$$v^{\pi_{\hat{f}}} \ge v^* - \varepsilon.$$

Proof. The proof mostly follows the proof of Theorem 2 and Theorem 7, and we only show the different and crucial steps here. We still condition on the high probability event from concentration (Lemma 1).

Similar as the proof of Theorem 7, from pessimism, we have

$$V_0^*(x_0) + H\varepsilon_{\text{stat},n} \ge V_{\hat{f}}(x_0) + H \cdot \max_{w\in\mathcal{W},h\in[H]} |\mathcal{L}_{\mathcal{D}}(\hat{f}, w, h)|, \tag{10}$$

where $\varepsilon_{\text{stat},n} \le \varepsilon\text{gap}(Q^*)/(2H^2)$.

On the other hand, following the proof of Theorem 2 and Theorem 7, we have

$$\begin{aligned}
&V_f(x_0) + H \cdot \max_{w\in\mathcal{W},h\in[H]} |\mathcal{L}_{\mathcal{D}}(\hat{f}, w, h)| \\
&\ge V_0^*(x_0) + \text{gap}(Q^*)\mathbb{E}\left[\sum_{h=0}^{H-1} \mathbf{1}\{\pi_f(x_h) \ne \pi^*(x_h)\} \mid a_{0:H-1} \sim \pi^*\right] - \sum_{h=0}^{H-1} |\mathcal{E}(\hat{f}, w^*, h)| + H \cdot \max_{w\in\mathcal{W},h\in[H]} |\mathcal{L}_{\mathcal{D}}(\hat{f}, w, h)| \\
&\ge V_0^*(x_0) + \text{gap}(Q^*)\mathbb{E}\left[\sum_{h=0}^{H-1} \mathbf{1}\{\pi_f(x_h) \ne \pi^*(x_h)\} \mid a_{0:H-1} \sim \pi^*\right] - \sum_{h=0}^{H-1} |\mathcal{E}(\hat{f}, w^*, h)| + \sum_{h=0}^{H-1} |\mathcal{E}(\hat{f}, w^*, h)| - H\varepsilon_{\text{stat},n} \\
&\ge V_0^*(x_0) + \text{gap}(Q^*)\mathbb{E}\left[\sum_{h=0}^{H-1} \mathbf{1}\{\pi_f(x_h) \ne \pi^*(x_h)\} \mid a_{0:H-1} \sim \pi^*\right] - H\varepsilon_{\text{stat},n}.
\end{aligned} \tag{11}$$

Combining Eq. (10) and Eq. (11) yields

$$\mathbb{E}\left[\sum_{h=0}^{H-1} \mathbf{1}\{\pi_f(x_h) \ne \pi^*(x_h)\} \mid a_{0:H-1} \sim \pi^*\right] \le 2H\varepsilon_{\text{stat},n}/\text{gap}(Q^*) \le \varepsilon.$$

The remaining steps are followed from the proof of Theorem 2. □

D.3 ROBUSTNESS TO MISSPECIFICATION

In this part, we present the sample complexity results of PABC-L (Algorithm 1) under misspecification. In parallel with Section 5, we show that PABC-L can identify v^* in Appendix D.3.1 and show its results for learning a near-optimal policy in Appendix D.3.2. The major advantage of PABC-L is that it does not take α as the input, therefore, we no longer require the knowledge of approximation errors.

D.3.1 ESTIMATING OPTIMAL EXPECTED RETURN

We present the result for identifying v^*. The sample complexity of PABC-L is the same as its counterpart (Theorem 3).

Theorem 9 (Robust version of Theorem 7)**.** *Suppose Assumption 3, Assumption 4 hold and the total number of samples nH satisfies*

$$nH \geq \frac{8C^2H^5\log(2|\mathcal{F}||\mathcal{W}|H/\delta)}{\varepsilon^2}.$$

Then with probability $1-\delta$, running Algorithm 1 with $C_{\text{gap}}=0$ guarantees

$$|V_{\hat{f}}(x_0) - v^*| \leq \varepsilon + H\varepsilon_{\mathcal{F}} + H\varepsilon_{\mathcal{W}}.$$

Proof. The proof mostly follows the proof of Theorem 3 and Theorem 7, and we only show the different and crucial steps here. We still condition on the high probability event from concentration (Lemma 1).

For $\tilde{Q}^*_{\mathcal{F}}$, from the concentration result and the definition of $\varepsilon_{\mathcal{F}}$, we get

$$\tilde{Q}^*_{\mathcal{F},0}(x_0, \pi_{Q^*_{\mathcal{F}}(x_0)}) + H \cdot \max_{w\in\mathcal{W}, h\in[H]} |\mathcal{L}_{\mathcal{D}}(\tilde{Q}^*_{\mathcal{F}}, w, h)| \leq V^*_0(x_0) + H\varepsilon_{\mathcal{F}} + H\varepsilon_{\text{stat},n},$$

where $\varepsilon_{\text{stat},n} \leq \varepsilon/H$.

From pessimism and the objective in Algorithm 1, we have

$$\tilde{Q}^*_{\mathcal{F},0}(x_0, \pi_{Q^*_{\mathcal{F}}(x_0)}) + H \cdot \max_{w\in\mathcal{W}, h\in[H]} |\mathcal{L}_{\mathcal{D}}(\tilde{Q}^*_{\mathcal{F}}, w, h)| \geq \ V_{\hat{f}}(x_0) + H \cdot \max_{w\in\mathcal{W}, h\in[H]} |\mathcal{L}_{\mathcal{D}}(\hat{f}, w, h)|.$$

Therefore, we get

$$V^*_0(x_0) + H\varepsilon_{\mathcal{F}} + H\varepsilon_{\text{stat},n} \geq V_{\hat{f}}(x_0) + H \cdot \max_{w\in\mathcal{W}, h\in[H]} |\mathcal{L}_{\mathcal{D}}(\hat{f}, w, h)|. \tag{12}$$

For any $f\in\mathcal{F}$, following the telescoping step in the proof of Theorem 7, we know that

$$V_f(x_0) \geq V^*_0(x_0) - \sum_{h=0}^{H-1} |\mathcal{E}(f, \pi^*, h)|.$$

Therefore, similar as the proof of Theorem 7 and applying Lemma 3, we get

$$\begin{aligned}
&V_{\hat{f}}(x_0) + H \cdot \max_{w\in\mathcal{W}, h\in[H]} |\mathcal{L}_{\mathcal{D}}(\hat{f}, w, h)| \\
&\geq V^*_0(x_0) - \sum_{h=0}^{H-1} |\mathcal{E}(\hat{f}, \pi^*, h)| + H \cdot \max_{w\in\mathcal{W}, h\in[H]} |\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\hat{f}, w, h)]| - H\varepsilon_{\text{stat},n} \\
&\geq V^*_0(x_0) - \sum_{h=0}^{H-1} |\mathcal{E}(\hat{f}, \pi^*, h)| + \sum_{h=0}^{H-1} |\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\hat{f}, \tilde{w}^*, h)]| - H\varepsilon_{\text{stat},n} \\
&\geq V^*_0(x_0) - \sum_{h=0}^{H-1} |\mathcal{E}(\hat{f}, \pi^*, h)| + \sum_{h=0}^{H-1} |\mathcal{E}(\hat{f}, \pi^*, h)| - H\varepsilon_{\mathcal{W}} - H\varepsilon_{\text{stat},n} \\
&= V^*_0(x_0) - H\varepsilon_{\mathcal{W}} - H\varepsilon_{\text{stat},n}.
\end{aligned} \tag{13}$$

Combining Eq. (12) and Eq. (13) yields

$$\begin{aligned}
|V_{\hat{f}}(x_0) + H \cdot \max_{w\in\mathcal{W}, h\in[H]} |\mathcal{L}_{\mathcal{D}}(\hat{f}, w, h)| - v^*| =&\ |V_{\hat{f}}(x_0) + H \cdot \max_{w\in\mathcal{W}, h\in[H]} |\mathcal{L}_{\mathcal{D}}(\hat{f}, w, h)| - V^*_0(x_0)| \\
=&\ H(\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{W}} + \varepsilon_{\text{stat},n}) \\
\leq&\ \varepsilon + H(\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{W}}),
\end{aligned}$$

which completes the proof. $\square$

D.3.2 LEARNING A NEAR-OPTIMAL POLICY

In this part, we show the results for learning a near-optimal policy. Compared with the ones for PABC (Theorem 4 and Corollary 5), the differences are only the constants.

Theorem 10 (Robust version of Theorem 8)**.** *Suppose Assumption 3, Assumption 4 hold and the total number of samples nH satisfies*

$$nH \geq \frac{32C^2H^7 \log(2|\mathcal{F}||\mathcal{W}|H/\delta)}{\varepsilon^2 C_{\text{gap}}^2}.$$

Then with probability $1-\delta$, running Algorithm 1 with a user-specified C_{gap} guarantees

$$v^{\pi_{\hat{f}}} \geq v^* - \varepsilon - \frac{H^2 \varepsilon_{\mathcal{F}(C_{\text{gap}})} + H^2 \varepsilon_{\mathcal{W}}}{C_{\text{gap}}}.$$

Proof. The proof mostly follows the proof of Theorem 8 and Theorem 9, and we only show the different and crucial steps here. We still condition on the high probability event from concentration (Lemma 1).

Similar as the proof of Theorem 9, we have

$$V_0^*(x_0) + H\varepsilon_{\mathcal{F}(C_{\text{gap}})} + H\varepsilon_{\text{stat},n} \geq V_{\hat{f}}(x_0) + H \cdot \max_{w \in \mathcal{W}, h \in [H]} |\mathcal{L}_{\mathcal{D}}(\hat{f}, w, h)|, \tag{14}$$

where $\varepsilon_{\text{stat},n} \leq \varepsilon C_{\text{gap}}/(2H^2)$.

On the other hand, following the proof of Theorem 8 and Theorem 9, we have

$$\begin{aligned}
&V_{\hat{f}}(x_0) + H \cdot \max_{w \in \mathcal{W}, h \in [H]} |\mathcal{L}_{\mathcal{D}}(\hat{f}, w, h)| \\
&\geq V_0^*(x_0) + C_{\text{gap}} \mathbb{E}\left[\sum_{h=0}^{H-1} \mathbf{1}\{\pi_f(x_h) \neq \pi^*(x_h)\} \mid a_{0:H-1} \sim \pi^*\right] - \sum_{h=0}^{H-1} |\mathcal{E}(\hat{f}, w^*, h)| \\
&\quad + H \cdot \max_{w \in \mathcal{W}, h \in [H]} |\mathcal{L}_{\mathcal{D}}(\hat{f}, w, h)| \\
&\geq V_0^*(x_0) + C_{\text{gap}} \mathbb{E}\left[\sum_{h=0}^{H-1} \mathbf{1}\{\pi_f(x_h) \neq \pi^*(x_h)\} \mid a_{0:H-1} \sim \pi^*\right] - H\varepsilon_{\mathcal{W}} - H\varepsilon_{\text{stat},n}.
\end{aligned} \tag{15}$$

Combining Eq. (14) and Eq. (15) yields

$$\mathbb{E}\left[\sum_{h=0}^{H-1} \mathbf{1}\{\pi_f(x_h) \neq \pi^*(x_h)\} \mid a_{0:H-1} \sim \pi^*\right] \leq H(2\varepsilon_{\text{stat},n} + \varepsilon_{\mathcal{W}} + \varepsilon_{\mathcal{F}(C_{\text{gap}})})/C_{\text{gap}}.$$

The remaining steps can be followed from the proof of Theorem 2. ☐

Corollary 11 (Corollary from Theorem 10)**.** *Suppose Assumption 3, Assumption 4 hold, the weight function class satisfies the additional mild regularity assumptions stated in Lemma 4. Assume we are given $\varepsilon_{\mathcal{F},\infty}$, $\text{gap}(Q^*)$ and $2\varepsilon_{\mathcal{F},\infty} < \text{gap}(Q^*)$. If the total number of samples nH satisfies*

$$nH \geq \frac{8C^2H^7 \log(2|\mathcal{F}||\mathcal{W}|H/\delta)}{\varepsilon^2(\text{gap}(Q^*) - 2\varepsilon_{\mathcal{F},\infty})^2},$$

then with probability $1-\delta$, running Algorithm 1 with $C_{\text{gap}} = \text{gap}(Q^*) - 2\varepsilon_{\mathcal{F},\infty}$ guarantees

$$v^{\pi_{\hat{f}}} \geq v^* - \varepsilon - \frac{2H^2\varepsilon_{\mathcal{F},\infty} + H^2\varepsilon_{\mathcal{W}}}{\text{gap}(Q^*) - 2\varepsilon_{\mathcal{F},\infty}}.$$

Proof. The proof mostly follows the proof of Corollary 5 and Theorem 10, and we only show the different and crucial steps here. We still condition on the high probability event from concentration (Lemma 1).

Similar as the proof of Corollary 5 and Theorem 10, we have

$$V_0^*(x_0) + 2H\varepsilon_{\mathcal{F},\infty} + H\varepsilon_{\text{stat},n} \geq V_{\hat{f}}(x_0) + H \cdot \max_{w \in \mathcal{W}, h \in [H]} |\mathcal{L}_{\mathcal{D}}(\hat{f}, w, h)|. \tag{16}$$

On the other hand, following the proof of Theorem 10, we have

$$\begin{aligned} & V_{\hat{f}}(x_0) + H \cdot \max_{w \in \mathcal{W}, h \in [H]} |\mathcal{L}_{\mathcal{D}}(\hat{f}, w, h)| \\ \geq\ & V_0^*(x_0) + (\text{gap}(Q^*) - 2\varepsilon_{\mathcal{F},\infty})\mathbb{E}\left[\sum_{h=0}^{H-1} \mathbf{1}\{\pi_f(x_h) \neq \pi^*(x_h)\} \mid a_{0:H-1} \sim \pi^*\right] - H\varepsilon_{\mathcal{W}} - H\varepsilon_{\text{stat},n}. \end{aligned} \tag{17}$$

Combining Eq. (16) and Eq. (17) yields

$$\mathbb{E}\left[\sum_{h=0}^{H-1} \mathbf{1}\{\pi_f(x_h) \neq \pi^*(x_h)\} \mid a_{0:H-1} \sim \pi^*\right] \leq H(2\varepsilon_{\text{stat},n} + \varepsilon_{\mathcal{W}} + 2\varepsilon_{\mathcal{F},\infty})/(\text{gap}(Q^*) - 2\varepsilon_{\mathcal{F},\infty}).$$

The remaining steps can be followed from the proof of Theorem 2. ☐

E DISCUSSION ON THE DATA COVERAGE ASSUMPTION

In this section, we provide an example that shows our data coverage assumption is more relaxed than the π^*-concentrability assumption in Zhan et al. (2022) (their Assumption 1) based on raw density ratios. Notice that their assumption translates into $d_h^*(x_h, a_h)/d_h^D(x_h, a_h) \leq C, \forall h \in [H], x_h \in \mathcal{X}_h, a_h \in \mathcal{A}$ in our finite-horizon episodic setting. We will show an instance where there exists some $h, (x_h, a_h)$ such that $d_h^*(x_h, a_h)/d_h^D(x_h, a_h) = \infty$ and w^* does not even exist (thus $w^* \notin \mathcal{W}$), but we still have $\varepsilon_{\mathcal{W}} = 0$. Therefore, our robust version of sample complexity results can give us meaningful guarantees, however, we cannot apply the (robustness) results in Zhan et al. (2022).

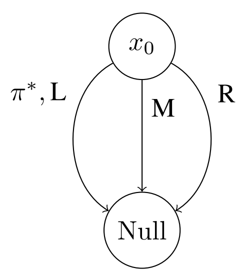

Figure 1: Example for comparison with π^*-concentrability assumption (Zhan et al., 2022).

	(x_0, L)	(x_0, M)	(x_0, R)
R	0.8	0.6	0.3
Q^*	0.8	0.6	0.3
f	0.7	0.3	0.8
d^*	1	0	0
d^D	0	0.5	0.5
w	0	1	1

Table 1: Example for comparison with π^*-concentrability assumption (Zhan et al., 2022).

As shown in Figure 1, circles denote states and arrows denote actions with deterministic transitions. In this MDP, the length of horizon is $H = 1$ and taking any action L, M, or R at the initial state x_0 transits to the Null terminal state. Since $H = 1$, in the following discussion we drop the subscript h for simplicity. In Table 1, we show the reward function, the optimal value function Q^*, the bad function f, the density-ratio function of the optimal policy d^*, the data distribution d^D, and the weight function w. We construct a singleton weight function class $\mathcal{W} = \{w\}$ and a realizable function class $\mathcal{F} = \{Q^*, f\}$. One can easily verify that $d^*(x_0, \text{L})/d^D(x_0, \text{L}) = \infty$, w^* does not exist, and the approximation error $\varepsilon_{\mathcal{W}}$ as defined in Eq. (4) is 0.

References

Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems*, 34, 2021.

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010.

Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, pages 550–559. PMLR, 2020.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason D Lee. Offline reinforcement learning with realizability and single-policy concentrability. *arXiv preprint arXiv:2202.04634*, 2022.