A Proof of Proposition 1

Let $\mathcal{F}$ denote the function class $\{f_r : f_r(\tau^0, \tau^1) = P_r(o = 1|\tau^0, \tau^1), r \in \mathcal{G}_r\}$. Let $\mathcal{I}_{\mathcal{F}}(\epsilon)$ denote the ϵ-bracket number with respect to ℓ_∞-norm, i.e., the minimum integer M such that there exist M functions $\{f^i\}_{i=1}^M$ such that for each $f_r \in \mathcal{F}$, we have $\sup_{\tau^0,\tau^1} |f_r(\tau^0, \tau^1) - f^i(\tau^0, \tau^1)| \le \epsilon$ for some $i \in [M]$. Then we know there exists a set of function $\overline{\mathcal{F}}$ with $|\overline{\mathcal{F}}| = \mathcal{I}_{\mathcal{F}}(\epsilon/4)$ such that for each $f_r \in \mathcal{F}$, there exists $\overline{f} \in \overline{\mathcal{F}}$ satisfying

$$\sup_{\tau^0,\tau^1} |f_r(\tau^0, \tau^1) - \overline{f}(\tau^0, \tau^1)| \le \epsilon/4.$$

Now we construct a bracket $(g^1_{\overline{f}}, g^2_{\overline{f}})$ defined as follows:

$$\begin{aligned}
&g^1_{\overline{f}}(o = 1|\tau^0, \tau^1) = \overline{f}(\tau^0, \tau^1) - \epsilon/4, g^1_{\overline{f}}(o = 0|\tau^0, \tau^1) = 1 - \overline{f}(\tau^0, \tau^1) - \epsilon/4,\\
&g^2_{\overline{f}}(o = 1|\tau^0, \tau^1) = \overline{f}(\tau^0, \tau^1) + \epsilon/4, g^2_{\overline{f}}(o = 0|\tau^0, \tau^1) = 1 - \overline{f}(\tau^0, \tau^1) + \epsilon/4.
\end{aligned}$$

Then clearly we have $g^1_{\overline{f}}(\cdot|\tau^0, \tau^1) \le P_r(\cdot|\tau^0, \tau^1) \le g^2_{\overline{f}}(\cdot|\tau^0, \tau^1)$ and $\|g^1_{\overline{f}}(\cdot|\tau^0, \tau^1) - g^2_{\overline{f}}(\cdot|\tau^0, \tau^1)\|_1 \le \epsilon$. This implies that $\mathcal{N}_{\mathcal{G}_r}(\epsilon) \le \mathcal{I}_{\mathcal{F}}(\epsilon/4)$.

Now we only need to bound $\mathcal{I}_{\mathcal{F}}(\epsilon/4)$. Consider θ and θ' with $\|\theta - \theta'\|_2 \le \epsilon_1$ and let r (r') denote the reward $\langle \phi, \theta \rangle$ ($\langle \phi, \theta' \rangle$). Then we know for all τ,

$$|r(\tau) - r'(\tau)| \le R\epsilon_1.$$

Fix the trajectory pair (τ^0, τ^1). Without loss of generality, we assume $\exp(r(\tau^0)) + \exp(r(\tau^1)) \le \exp(r'(\tau^0)) + \exp(r'(\tau^1))$. Then we have

$$\exp(r(\tau^0)) + \exp(r(\tau^1)) \le \exp(r'(\tau^0)) + \exp(r'(\tau^1)) \le \exp(R\epsilon_1)\Big(\exp(r(\tau^0)) + \exp(r(\tau^1))\Big).$$

On the other hand, we have

$$\begin{aligned}
&|f_r(\tau^0, \tau^1) - f_{r'}(\tau^0, \tau^1)|\\
&\quad= \frac{\Big|\exp(r(\tau^1))\Big(\exp(r'(\tau^0)) + \exp(r'(\tau^1))\Big) - \exp(r'(\tau^1))\Big(\exp(r(\tau^0)) + \exp(r(\tau^1))\Big)\Big|}{\Big(\exp(r'(\tau^0)) + \exp(r'(\tau^1))\Big)\Big(\exp(r(\tau^0)) + \exp(r(\tau^1))\Big)}.
\end{aligned}$$

Therefore, if $\exp(r(\tau^1))\Big(\exp(r'(\tau^0)) + \exp(r'(\tau^1))\Big) - \exp(r'(\tau^1))\Big(\exp(r(\tau^0)) + \exp(r(\tau^1))\Big) \ge 0$, then we have

$$\begin{aligned}
&\Big|\exp(r(\tau^1))\Big(\exp(r'(\tau^0)) + \exp(r'(\tau^1))\Big) - \exp(r'(\tau^1))\Big(\exp(r(\tau^0)) + \exp(r(\tau^1))\Big)\Big|\\
\le &\exp(R\epsilon_1)\exp(r(\tau^1))\Big(\exp(r(\tau^0)) + \exp(r(\tau^1))\Big) - \exp(-R\epsilon_1)\exp(r(\tau^1))\Big(\exp(r(\tau^0)) + \exp(r(\tau^1))\Big)\\
= &(\exp(R\epsilon_1) - \exp(-R\epsilon_1))\exp(r(\tau^1))\Big(\exp(r(\tau^0)) + \exp(r(\tau^1))\Big).
\end{aligned}$$

Otherwise, we have

$$\begin{aligned}
&\Big|\exp(r(\tau^1))\Big(\exp(r'(\tau^0)) + \exp(r'(\tau^1))\Big) - \exp(r'(\tau^1))\Big(\exp(r(\tau^0)) + \exp(r(\tau^1))\Big)\Big|\\
\le &\exp(R\epsilon_1)\exp(r(\tau^1))\Big(\exp(r(\tau^0)) + \exp(r(\tau^1))\Big) - \exp(r(\tau^1))\Big(\exp(r(\tau^0)) + \exp(r(\tau^1))\Big)\\
= &(\exp(R\epsilon_1) - 1)\exp(r(\tau^1))\Big(\exp(r(\tau^0)) + \exp(r(\tau^1))\Big).
\end{aligned}$$

Therefore we have

$$|f_r(\tau^0, \tau^1) - f_{r'}(\tau^0, \tau^1)|$$

$$\leq \frac{(\exp(R\epsilon_1) - \exp(-R\epsilon_1))\exp(r(\tau^1))\Big(\exp(r(\tau^0)) + \exp(r(\tau^1))\Big)}{\Big(\exp(r'(\tau^0)) + \exp(r'(\tau^1))\Big)\Big(\exp(r(\tau^0)) + \exp(r(\tau^1))\Big)} \leq \exp(2R\epsilon_1) - 1.$$

This implies that for any $\epsilon \leq 1$,

$$\log \mathcal{I}_{\mathcal{F}}(\epsilon/4) \leq \log \mathcal{I}_{d,B}\Big(\frac{2\ln 2}{R}\epsilon\Big) \leq \mathcal{O}\Big(d \log \frac{BR}{\epsilon}\Big),$$

where $\mathcal{I}_{d,B}(\cdot)$ is the covering number of a d-dimensional ball centered at the origin with radius B with respect to ℓ_2-norm and the last step is from Wainwright (2019). This concludes our proof.

B Feasible Implementation of FREEHAND

In this section we show how to implement the robust optimization step (Line 4) of FREEHAND in practice. Our idea is inspired by standard offline RL (Rigter et al., 2022) where the authors rely on Lagrangian formulation to make the theoretical algorithm CPPO (Uehara and Sun, 2021) practical enough to achieve good performance on the D4RL datasets. We believe the empirical insights provided in (Rigter et al., 2022) can be applied here as well.

First for the Lagrangian relaxation, the original inner minimization problem in Line 4 of FREEHAND is

$$\min_{r \in \mathcal{R}(\mathcal{D})} J(\pi; r, P^*) - \mathbb{E}_{\tau \sim \mu_{\mathrm{ref}}}[r(\tau)].$$

Note that the only constraint is $r \in \mathcal{R}(\mathcal{D})$. Then by introducing a Lagrangian multiplier β, we can convert such constrained minimization problem into an unconstrained regularized minimization problem:

$$\min_r J(\pi; r, P^*) - \mathbb{E}_{\tau \sim \mu_{\mathrm{ref}}}[r(\tau)] - \beta \sum_{n=1}^{N} \log P_r(o = o^n | \tau^{n,0}, \tau^{n,1}).$$

Consequently, Line 4 in FREEHAND can be converted to the following unconstrained regularized max-min problem:

$$\max_\pi \min_r \mathcal{L}(\pi, r) := J(\pi; r, P^*) - \mathbb{E}_{\tau \sim \mu_{\mathrm{ref}}}[r(\tau)] - \beta \sum_{n=1}^{N} \log P_r(o = o^n | \tau^{n,0}, \tau^{n,1}).$$

Since now we are facing an unregularized problem, the most common way to solve $\mathcal{L}(\pi, r)$ in practice is gradient ascent-descent. Suppose π and r are parametrized by θ and λ (usaully neural networks). Then gradient ascent-descent requires us to compute an unbiased stochastic gradient with respect to θ and λ respectively. Fortunately, this can be easy to achieve in practice. On the one hand, for the gradient of θ, we only need to compute $\nabla_\theta J(\pi_\theta; r, P^*)$. This task has been thoroughly discussed in the literature of policy gradient and one example is REINFORCE, which samples a trajectory τ by executing π_θ in P^* and then the estimated graidient can be expressed as

$$r(\tau) \sum_{h=1}^{H} \nabla_\theta \pi_{\theta,h}(a_h | s_h),$$

where (s_h, a_h) is the h-step of τ.

On the other hand, for the gradient of λ, we only need to sample independent trajecotories τ' by executing π_θ in P^* and τ'' from μ_{ref} and an index $i \in [N]$. Then the unbiased estimated gradient can be directly written as

$$\nabla_\lambda r_\lambda(\tau') - \nabla_\lambda r_\lambda(\tau'') - \beta \nabla_\lambda \log P_{r_\lambda}(o = o^i | \tau^{i,0}, \tau^{i,1}).$$

Therefore, with the above estimated gradients, we can then run graident ascent-descent happily to solve $\max_\pi \min_r \mathcal{L}(\pi, r)$ in practice.

C Proof of Theorem 1

The proof of Theorem 1 consists of two steps, deriving the guarantee of MLE and analyzing the performance of pessimistic offline RL.

Step 1: MLE guarantee. We first need to show that the confidence set $\mathcal{R}(\mathcal{D})$ contains the true reward $r^\star$ with high probability. This can be proved via the following lemma which characterizes the guarantee of MLE:

Lemma 1 (Performance of MLE)**.** *Fix any* $\delta \in (0, 1]$. *Then with probability at least* $1 - \delta/2$ *we have that for all reward function* $r \in \mathcal{G}_r$,

$$\sum_{n=1}^{N} \log\left(\frac{P_r(o^n|\tau^{n,0},\tau^{n,1})}{P_{r^\star}(o^n|\tau^{n,0},\tau^{n,1})}\right) \leq c_{\mathrm{MLE}} \log(\mathcal{N}_{\mathcal{G}_r}(1/N)/\delta),$$

where $c_{\mathrm{MLE}} > 0$ *is a universal constant.*

We defer the proof to Appendix C.1. Denote the event in Lemma 1 by $\mathcal{E}_1$, then we know $\mathbb{P}(\mathcal{E}_1) \geq 1 - \delta/2$. Under the event $\mathcal{E}_1$, we have

$$\sum_{n=1}^{N} \log P_{r^\star}(o^n|\tau^{n,0},\tau^{n,1}) \geq \sum_{n=1}^{N} \log P_{\widehat{r}}(o^n|\tau^{n,0},\tau^{n,1}) - c_{\mathrm{MLE}} \log(\mathcal{N}_{\mathcal{G}_r}(1/N)/\delta),$$

which implies that $r^\star \in \mathcal{R}(\mathcal{D})$ since we know $r^\star \in \mathcal{G}_r$ from Assumption 2.

Nevertheless, the confidence set $\mathcal{R}(\mathcal{D})$ is constructed via loglikelihood and we indeed prefer a bound on the total variation (TV) distance between P_r and $P_{r^\star}$ where $r \in \mathcal{R}(\mathcal{D})$ to facilitate our subsequent analysis. We can obtain such a bound as shown in the following lemma from the literature (Liu et al. (2022)[Proposition 14],Zhan et al. (2022b)[Lemma 9]):

Lemma 2. *With probability at least* $1 - \delta/2$, *we have for all reward function* $r \in \mathcal{G}_r$ *that*

$$\mathbb{E}_{\tau^0\sim\mu_0,\tau^1\sim\mu_1}\left[\left\|P_r(\cdot|\tau^0,\tau^1) - P_{r^\star}(\cdot|\tau^0,\tau^1)\right\|_1^2\right] \leq \frac{c_{\mathrm{TV}}}{N}\left(\sum_{n=1}^{N}\log\left(\frac{P_{r^\star}(o^n|\tau^{n,0},\tau^{n,1})}{P_r(o^n|\tau^{n,0},\tau^{n,1})}\right) + \log(\mathcal{N}_{\mathcal{G}_r}(1/N)/\delta)\right),$$

where $c_{\mathrm{TV}} > 0$ *is a universal constant.*

Denote the event in Lemma 2 by $\mathcal{E}_2$ and then we know $\mathbb{P}(\mathcal{E}_2) \geq 1 - \delta/2$. Then from Lemma 1 and Lemma 2 we know that under event $\mathcal{E}_1 \cap \mathcal{E}_2$, we have for all $r \in \mathcal{R}(\mathcal{D})$:

$$\mathbb{E}_{\tau^0\sim\mu_0,\tau^1\sim\mu_1}\left[\left\|P_r(\cdot|\tau^0,\tau^1) - P_{r^\star}(\cdot|\tau^0,\tau^1)\right\|_1^2\right] \leq \frac{c\log(\mathcal{N}_{\mathcal{G}_r}(1/N)/\delta)}{N}, \tag{4}$$

where $c > 0$ is a universal constant.

Then under Assumption 3, we can apply the mean value theorem between $r^\star(\tau_1) - r^\star(\tau_0)$ and $r(\tau_1) - r(\tau_0)$ to (4) and ensure for all $r \in \mathcal{R}(\mathcal{D})$ that

$$\mathbb{E}_{\tau^0\sim\mu_0,\tau^1\sim\mu_1}[|(r^\star(\tau_1) - r^\star(\tau_0)) - (r(\tau_1) - r(\tau_0))|^2] \leq \frac{c\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/N)/\delta)}{N}, \tag{5}$$

where $\kappa := \frac{1}{\inf_{x\in[-r_{\max},r_{\max}]}\Phi'(x)}$ measures the non-linearity of the link function Φ.

Step 2: Pessimistic offline RL. Let $r_\pi^{\inf}$ denote $\mathrm{argmin}_{r\in\mathcal{R}(\mathcal{D})} J(\pi; r, P^\star) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r(\tau)]$. Then we can bound the suboptimality of $\widehat{\pi}$ as follows:

$$\begin{aligned}
&J(\pi_{\mathrm{tar}}; r^\star, P^\star) - J(\widehat{\pi}; r^\star, P^\star)\\
=&\big(J(\pi_{\mathrm{tar}}; r^\star, P^\star) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r^\star(\tau)]\big) - \big(J(\widehat{\pi}; r^\star, P^\star) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r^\star(\tau)]\big)\\
\leq&\Big(\big(J(\pi_{\mathrm{tar}}; r^\star, P^\star) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r^\star(\tau)]\big) - \big(J(\pi_{\mathrm{tar}}; r_{\pi_{\mathrm{tar}}}^{\inf}, P^\star) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r_{\pi_{\mathrm{tar}}}^{\inf}(\tau)]\big)\Big)\\
&- \Big(\big(J(\widehat{\pi}; r^\star, P^\star) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r^\star(\tau)]\big) - \big(J(\widehat{\pi}; r_{\widehat{\pi}}^{\inf}, P^\star) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r_{\widehat{\pi}}^{\inf}(\tau)]\big)\Big)
\end{aligned}$$

$$
\begin{aligned}
&\le \big(J(\pi_{\rm tar}; r^\star, P^\star) - \mathbb{E}_{\tau\sim\mu_{\rm ref}}[r^\star(\tau)]\big) - \big(J(\pi_{\rm tar}; r^{\rm inf}_{\pi_{\rm tar}}, P^\star) - \mathbb{E}_{\tau\sim\mu_{\rm ref}}[r^{\rm inf}_{\pi_{\rm tar}}(\tau)]\big)\\
&= \mathbb{E}_{\tau^0\sim\pi_{\rm tar},\tau^1\sim\mu_{\rm ref}}[(r^\star(\tau^0) - r^\star(\tau^1)) - (r^{\rm inf}_{\pi_{\rm tar}}(\tau^0) - r^{\rm inf}_{\pi_{\rm tar}}(\tau^1))]\\
&\le C_r(\mathcal{G}_r, \pi_{\rm tar}, \mu_{\rm ref})\sqrt{\mathbb{E}_{\tau_0\sim\mu_0,\tau_1\sim\mu_1}[|r^\star(\tau^0) - r^\star(\tau^1) - r^{\rm inf}_{\pi_{\rm tar}}(\tau^0) + r^{\rm inf}_{\pi_{\rm tar}}(\tau^1)|^2]}\\
&\le \sqrt{\frac{cC_r^2(\mathcal{G}_r, \pi_{\rm tar}, \mu_{\rm ref})\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/N)/\delta)}{N}},
\end{aligned}
$$

where the second step is due to $\widehat{\pi} = \operatorname{argmax}_{\pi\in\Pi_{\rm his}}\min_{r\in\mathcal{R}(\mathcal{D})} J(\pi; r, P^\star) - \mathbb{E}_{\tau\sim\mu_{\rm ref}}[r(\tau)]$, the third step is due to $r^{\rm inf}_{\widehat{\pi}} = \operatorname{argmin}_{r\in\mathcal{R}(\mathcal{D})} J(\widehat{\pi}; r, P^\star) - \mathbb{E}_{\tau\sim\mu_{\rm ref}}[r(\tau)]$, the fifth step comes from the definition of $C_r(\mathcal{G}_r, \pi_{\rm tar}, \mu_{\rm ref})$ (Definition 2) and the last step leverages (5). This concludes our proof.

C.1 PROOF OF LEMMA 1

The proof largely follows Zhan et al. (2022b). Suppose $\overline{\mathcal{F}}$ is a $1/N$-bracket of $\mathcal{G}_r$ with $|\overline{\mathcal{F}}| = \mathcal{N}_{\mathcal{G}_r}(1/N)$ and we denote the set of all right brackets in $\overline{\mathcal{F}}$ by $\widetilde{\mathcal{F}}$, i.e., $\widetilde{\mathcal{F}} := \{f : \exists f', \text{ such that } [f', f] \in \overline{\mathcal{F}}\}$. Then fix any $f \in \widetilde{\mathcal{F}}$, we have:

$$
\begin{aligned}
\mathbb{E}\bigg[\exp\bigg(\sum_{n=1}^N \log\bigg(\frac{f(o^n|\tau^{n,0},\tau^{n,1})}{P_{r^\star}(o^n|\tau^{n,0},\tau^{n,1})}\bigg)\bigg)\bigg] &= \prod_{n=1}^N \mathbb{E}\bigg[\exp\bigg(\log\bigg(\frac{f(o^n|\tau^{n,0},\tau^{n,1})}{P_{r^\star}(o^n|\tau^{n,0},\tau^{n,1})}\bigg)\bigg)\bigg]\\
= \prod_{n=1}^N \mathbb{E}\bigg[\frac{f(o^n|\tau^{n,0},\tau^{n,1})}{P_{r^\star}(o^n|\tau^{n,0},\tau^{n,1})}\bigg] &= \prod_{n=1}^N \mathbb{E}\bigg[\sum_o f(o|\tau^{n,0},\tau^{n,1})\bigg] \le \bigg(1 + \frac{1}{N}\bigg)^N \le e,
\end{aligned}
$$

where the first step is due to each sample in $\mathcal{D}$ is i.i.d., the third step uses Tower property and the fourth step is from the fact that $\overline{\mathcal{F}}$ is a minimum $1/N$-bracket.

Then by Markov's inequality we have for any $\delta \in (0, 1]$,

$$
\begin{aligned}
&\mathbb{P}\bigg(\sum_{n=1}^N \log\bigg(\frac{f(o^n|\tau^{n,0},\tau^{n,1})}{P_{r^\star}(o^n|\tau^{n,0},\tau^{n,1})}\bigg) > \log(1/\delta)\bigg)\\
&\qquad\le \mathbb{E}\bigg[\exp\bigg(\sum_{n=1}^N \log\bigg(\frac{f(o^n|\tau^{n,0},\tau^{n,1})}{P_{r^\star}(o^n|\tau^{n,0},\tau^{n,1})}\bigg)\bigg)\bigg] \cdot \exp[-\log(1/\delta)] \le e\delta.
\end{aligned}
$$

By union bound, we have for all $f \in \widetilde{\mathcal{F}}$,

$$
\mathbb{P}\bigg(\sum_{n=1}^N \log\bigg(\frac{f(o^n|\tau^{n,0},\tau^{n,1})}{P_{r^\star}(o^n|\tau^{n,0},\tau^{n,1})}\bigg) > c_{\rm MLE}\log(\mathcal{N}_{\mathcal{G}_r}(1/N)/\delta)\bigg) \le \delta/2,
$$

where $c_{\rm MLE} > 0$ is a universal constant.

Therefore from the definition of $1/N$-bracket net, we know for all $r \in \mathcal{G}_r$, there exists $f \in \widetilde{\mathcal{F}}$ such that $P_r(\cdot|\tau^0,\tau^1) \le f(\cdot|\tau^0,\tau^1)$ for any trajectories (τ^0, τ^1). This implies that for all $r \in \mathcal{G}_r$,

$$
\mathbb{P}\bigg(\sum_{n=1}^N \log\bigg(\frac{P_r(o^n|\tau^{n,0},\tau^{n,1})}{P_{r^\star}(o^n|\tau^{n,0},\tau^{n,1})}\bigg) > c_{\rm MLE}\log(\mathcal{N}_{\mathcal{G}_r}(1/N)/\delta)\bigg) \le \delta/2,
$$

This concludes our proof.

D COMPARISON WITH ZHU ET AL. (2023)

Zhu et al. (2023) considers the linear reward setting under BTL model and they can achieve the following sample complexity:

$$
N = \mathcal{O}\bigg(\frac{C_{\rm lin}^2\exp(4BR)d\log(1/\delta)}{\epsilon^2}\bigg),
$$

where R and B are the norm bounds on the feature vectors ϕ and parameter θ (defined in Proposition 1).The concentrability coefficient C_{lin} is defined as

$$C_{\text{lin}} := \|\mathbb{E}_{\tau^0\sim\pi_{\text{tar}},\tau^1\sim\mu_{\text{ref}}}[\phi(\tau^0)-\phi(\tau^1)]\|_{\Sigma_{\mathcal{D}}^{-1}},$$

and $\Sigma_{\mathcal{D}}$ is the empirical covariance matrix of the dataset $\frac{1}{N}\sum_{n=1}^{N}(\phi(\tau^{n,0})-\phi(\tau^{n,1}))(\phi(\tau^{n,0})-\phi(\tau^{n,1}))^{\top}$.

Note that all the analysis and proofs in this paper still hold when we define the concentrability coefficient as

$$C'_r(\mathcal{G}_r,\pi_{\text{tar}},\mu_{\text{ref}}) := \max\left\{0, \sup_{r\in\mathcal{G}_r}\frac{\mathbb{E}_{\tau^0\sim\pi_{\text{tar}},\tau^1\sim\mu_{\text{ref}}}[r^\star(\tau^0)-r^\star(\tau^1)-r(\tau^0)+r(\tau^1)]}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}|r^\star(\tau^{n,0})-r^\star(\tau^{n,1})-r(\tau^{n,0})+r(\tau^{n,1})|^2}}\right\}.$$

Then when specializing the result in Theorem 1 to the linear reward setting under BTL model with this version of concentrability coefficient, the sample complexity is

$$N = \widetilde{\mathcal{O}}\left(\frac{(C'_r(\mathcal{G}_r,\pi_{\text{tar}},\mu_{\text{ref}}))^2\exp(2r_{\max})d\log(BR/\delta)}{\epsilon^2}\right).$$

We know that $BR \geq r_{\max}$. In addition, note that in this case, we have $C_{\text{lin}} \geq 0$ and for any $r \in \mathcal{G}_r$,

$$\begin{aligned}
&\left|\mathbb{E}_{\tau^0\sim\pi_{\text{tar}},\tau^1\sim\mu_{\text{ref}}}[r^\star(\tau^0)-r^\star(\tau^1)-r(\tau^0)+r(\tau^1)]\right| \\
&\quad= \left|\langle\mathbb{E}_{\tau^0\sim\pi_{\text{tar}},\tau^1\sim\mu_{\text{ref}}}[\phi(\tau^0)-\phi(\tau^1)],\theta^\star-\theta\rangle\right| \\
&\quad\leq \|\mathbb{E}_{\tau^0\sim\pi_{\text{tar}},\tau^1\sim\mu_{\text{ref}}}[\phi(\tau^0)-\phi(\tau^1)]\|_{\Sigma_{\mathcal{D}}^{-1}}\cdot\|\theta^\star-\theta\|_{\Sigma_{\mathcal{D}}} \\
&\quad= \|\mathbb{E}_{\tau^0\sim\pi_{\text{tar}},\tau^1\sim\mu_{\text{ref}}}[\phi(\tau^0)-\phi(\tau^1)]\|_{\Sigma_{\mathcal{D}}^{-1}}\cdot\sqrt{\frac{1}{N}\sum_{n=1}^{N}|r^\star(\tau^{n,0})-r^\star(\tau^{n,1})-r(\tau^{n,0})+r(\tau^{n,1})|^2},
\end{aligned}$$

where we suppose $r^\star(\tau)=\langle\phi(\tau),\theta^\star\rangle$ and $r(\tau)=\langle\phi(\tau),\theta\rangle$. Therefore we have

$$C'_r(\mathcal{G}_r,\pi_{\text{tar}},\mu_{\text{ref}}) \leq C_{\text{lin}}.$$

This implies that Theorem 1 can recover the sample complexity for linear reward setting under BTL model in Zhu et al. (2023) with only some additional log factors.

E Omitted Details

In this section we supplement the definition of bracket number for the transition class and advantage function class.

Definition 5 (ϵ-bracket number of transition probability classes)**.** *Suppose f^1, f^2 is a function with $f^1(\cdot|s,a), f^2(\cdot|s,a) \in \mathbb{R}^{|\mathcal{S}|}$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. Then we say (f^1,f^2) is a ϵ-bracket if $f^1(\cdot|s,a)\leq f^2(\cdot|s,a)$ and $\|f^1(\cdot|s,a)-f^2(\cdot|s,a)\|_1\leq\epsilon$ for all (s,a). The ϵ-bracket number of a transition probability class $\mathcal{G}_{P_h}$ where $h\in[H-1]$ is the minimum integer N satisfying that there exist N ϵ-brackets $(f^{n,1},f^{n,2})_{n=1}^{N}$ such that for any function $P_h\in\mathcal{G}_{P_h}$ there is a bracket $(f^{i,1},f^{i,2})$ where $i\in[N]$ containing it, i.e., $f^{i,1}(\cdot|s,a)\leq P_h(\cdot|s,a)\leq f^{i,2}(\cdot|s,a)$ for all (s,a).*

Definition 6 (ϵ-bracket number of initial state distribution classes)**.** *Suppose $f^1,f^2\in\mathbb{R}^{|\mathcal{S}|}$. Then we say (f^1,f^2) is a ϵ-bracket if $f^1\leq f^2$ and $\|f^1-f^2\|_1\leq\epsilon$. The ϵ-bracket number of a initial state distribution class $\mathcal{G}_{P_0}$ is the minimum integer N satisfying that there exist N ϵ-brackets $(f^{n,1},f^{n,2})_{n=1}^{N}$ such that for any $P_0\in\mathcal{G}_{P_0}$ there is a bracket $(f^{i,1},f^{i,2})$ where $i\in[N]$ containing it, i.e., $f^{i,1}\leq P_0\leq f^{i,2}$.*

Definition 7 (ϵ-bracket number of advantage function classes)**.** *Suppose g^1,g^2 is a function with $g^1(\cdot|s,a^0,a^1), g^2(\cdot|s,a^0,a^1)\in\mathbb{R}^2$ for all $(s,a^0,a^1)\in\mathcal{S}\times\mathcal{A}\times\mathcal{A}$. Then we say (g^1,g^2) is a ϵ-bracket if $g^1(\cdot|s,a^0,a^1)\leq g^2(\cdot|s,a^0,a^1)$ and $\|g^1(\cdot|s,a^0,a^1)-g^2(\cdot|s,a^0,a^1)\|_1\leq\epsilon$ for all $(s,a^0,a^1)\in\mathcal{S}\times\mathcal{A}\times\mathcal{A}$. The ϵ-bracket number of a reward class $\mathcal{G}_{A_h}$ where $h\in[H]$ is the minimum integer N satisfying that there exist N ϵ-brackets $(g^{n,1},g^{n,2})_{n=1}^{N}$ such that for any function $A_h\in\mathcal{G}_{A_h}$ there is a bracket $(g^{i,1},g^{i,2})$ where $i\in[N]$ containing it, i.e., $g^{i,1}(\cdot|s,a^0,a^1)\leq P_{A_h}(\cdot|s,a^0,a^1)\leq g^{i,2}(\cdot|s,a^0,a^1)$ for all $(s,a^0,a^1)\in\mathcal{S}\times\mathcal{A}\times\mathcal{A}$.*

We use $\mathcal{N}_{\mathcal{G}_{P_h}}(\epsilon)$ and $\mathcal{N}_{\mathcal{G}_{A_h}}(\epsilon)$ to denote the ϵ-bracket number of $\mathcal{G}_{P_h}$ and $\mathcal{G}_{A_h}$. Similarly, when the transition probability or the advantage function possesses a low-dimension embedding, we can also bound the ϵ-bracket number efficiently.

F Proofs of Lower Bounds

F.1 Proof of Proposition 2

Given any S, A, H, consider a MDP with horizon H, state space $\mathcal{S} = \{s^1, s^2, \cdots, s^S\}$ and action space $\mathcal{A} = \{a^1, a^2, \cdots, a^A\}$. In the following discussion we consider the case $C \geq 2$ and $1 < C < 2$ respectively.

Case 1: $C \geq 2$**.** Consider the case where the state is fixed throughout an episode. We suppose the initial state distribution $P_0^\star$ is $P_0^\star(s^1) = \frac{1}{2}$ and $P_0^\star(s^i) = \frac{1}{2(S-1)}$ for all $2 \leq i \leq S$. Let $\pi_{\mathrm{tar},h}(a^1|s) = 1$ for all $h \in [H]$ and $s \in \mathcal{S}$. Then we can set the dataset distribution μ_0 as

$$\mu_0(\tau) = \begin{cases} \frac{1}{2C}, & \text{if the state is } s^1 \text{ and all actions in } \tau \text{ are } a^1 \text{ except } a_{H-1} = a^2, \\ \frac{1}{2} - \frac{1}{2C}, & \text{if the state is } s^1 \text{ and all actions in } \tau \text{ are } a^1 \text{ except } a_H = a^2, \\ \frac{1}{2(S-1)}, & \text{if the state is not } s^1 \text{ and all actions in } \tau \text{ are } a^1, \\ 0, & \text{otherwise,} \end{cases}$$

where a_h is the action at step h in τ. Then we know

$$\mu_{0,h}(s, a^1) = \frac{1}{2(S-1)}, \qquad \forall h \in [H], s \in \mathcal{S} \setminus \{s^1\},$$
$$\mu_{0,h}(s^1, a^1) = \frac{1}{2}, \quad \mu_{0,H-1}(s, a^1) = \frac{1}{2} - \frac{1}{2C}, \quad \mu_{0,H}(s, a^1) = \frac{1}{2C}.$$

It is obvious we have $C_{\mathrm{st}} \leq C$ in this setting. On the other hand, since the trajectory whose state is s^1 and all actions are a^1 is covered by π_{tar} but not by μ_0, we have $C_{\mathrm{tr}} = \infty$.

Case 2: $1 < C < 2$**.** Consider the case where the state is fixed throughout an episode. We suppose the initial state distribution of $P_0^\star$ is $P_0^\star(s^1) = \frac{C-1}{2}$, $P_0^\star(s^2) = \frac{2-C}{2}$ and $P_0^\star(s^i) = \frac{1}{2(S-2)}$ for all $3 \leq i \leq S$. Note that here we require $S \geq 3$. When $S = 2$, we can let $P_0^\star(s^1) = C - 1$ and $P_0^\star(s^2) = 2 - C$ and the following analysis will still hold. Therefore here we assume $S \geq 3$ without loss of generality. Let $\pi_{\mathrm{tar},h}(a^1|s) = 1$ for all $h \in [H]$ and $s \in \mathcal{S}$. Then we can set the dataset distribution μ_0 as

$$\mu_0(\tau) = \begin{cases} \frac{C-1}{2C}, & \text{if the state of } \tau \text{ is } s^1 \text{ and all actions in } \tau \text{ are } a^1 \text{ except } a_{H-1} = a^2, \\ \frac{C-1}{2C}, & \text{if the state of } \tau \text{ is } s^1 \text{ and all actions in } \tau \text{ are } a^1 \text{ except } a_H = a^2, \\ \frac{2-C}{2C}, & \text{if the state of } \tau \text{ is } s^2 \text{ and the actions are all } a^1, \\ \frac{1}{2(S-2)}, & \text{if the state of } \tau \text{ is not } s^1 \text{ or } s^2 \text{ and the actions are all } a^1, \\ 0, & \text{otherwise.} \end{cases}$$

Then we know

$$\mu_{0,h}(s, a^1) = \frac{1}{2(S-2)}, \qquad \forall h \in [H], s \in \mathcal{S} \setminus \{s^1, s^2\},$$
$$\mu_{0,h}(s^2, a^1) = \frac{2-C}{2C}, \qquad \forall h \in [H],$$
$$\mu_{0,h}(s^1, a^1) = \frac{2C-2}{2C}, \qquad \forall h \in [H-2],$$
$$\mu_{0,H-1}(s^1, a^1) = \mu_{0,H}(s^1, a^1) = \frac{C-1}{2C}.$$

It is obvious we have $C_{\mathrm{st}} \leq C$ in this setting. On the other hand, since the trajectory whose state is s^1 and all actions are a^1 is covered by π_{tar} but not by μ_0, we have $C_{\mathrm{tr}} = \infty$. This concludes our proof.

F.2 Proof of Theorem 2

We consider the case $C \geq 2$ and $1 < C < 2$ respectively.

Case 1: $C \geq 2$**.** Consider the case where there is only one state s and two actions a^1, a^2. Set the dataset distribution $\mu_0 = \mu_1$ where

$$\mu_0(\tau) = \begin{cases} \frac{1}{C}, & \text{if all actions in } \tau \text{ are } a^1 \text{ except } a_{H-1} = a^2, \\ 1 - \frac{1}{C}, & \text{if all actions in } \tau \text{ are } a^1 \text{ except } a_H = a^2, \\ 0, & \text{otherwise,} \end{cases}$$

where a_h is the action at step h in τ. In the following discussion we will use τ^1 to denote the trajectory where all actions are a^1 except $a_{H-1} = a^2$ and τ^2 to denote the trajectory where all actions are a^1 except $a_{H-1} = a^2$. Then we know

$$\mu_{0,h}(s, a^1) = 1, \qquad \forall h \in [H-2],$$
$$\mu_{0,H-1}(s, a^1) = 1 - \frac{1}{C}, \qquad \mu_{0,H-1}(s, a^2) = \frac{1}{C},$$
$$\mu_{0,H}(s, a^1) = \frac{1}{C}, \qquad \mu_{0,H}(s, a^2) = 1 - \frac{1}{C}.$$

We consider two different reward function r^1 and r^2:

$$r^1_h(s, a^1) = r^1_h(s, a^2) = r^2_h(s, a^1) = r^2_h(s, a^2) = 0, \qquad \forall h \in [H-2],$$
$$r^1_{H-1}(s, a^1) = r^2_{H-1}(s, a^2) = 1, \qquad r^1_{H-1}(s, a^2) = r^2_{H-1}(s, a^1) = 0,$$
$$r^1_H(s, a^1) = r^2_H(s, a^2) = 1, \qquad r^1_H(s, a^2) = r^2_H(s, a^1) = 0,$$

Then we have two MDPs, $\mathcal{M}_1$ and $\mathcal{M}_2$ whose reward functions are r^1 and r^2 respectively. It can be easily verified that $(\mathcal{M}_1, \mu_0) \in \overline{\Theta}_{\mathrm{st}}(C), (\mathcal{M}_2, \mu_0) \in \overline{\Theta}_{\mathrm{st}}(C)$.

Further, let $L(\pi; \mathcal{M})$ denote the suboptimality of policy π in $\mathcal{M}$, then we have for all policies π,

$$L(\pi; \mathcal{M}_1) + L(\pi; \mathcal{M}_2) \geq 2.$$

Now we can apply Le Cam's method, which leads to the following inequality

$$\inf_{\widehat{\pi}} \sup_{\mathcal{M} \in \{\mathcal{M}_1, \mathcal{M}_2\}} \mathbb{E}_{\mathcal{D}}[L(\pi, \mathcal{M})] \geq \frac{1}{2} \exp(-N \mathbf{KL}\Big(\mu_0 \otimes \mu_1 \otimes P_{r^1} \| \mu_0 \otimes \mu_1 \otimes P_{r^2}\Big)).$$

It can be observed that $\mathbf{KL}\Big(\mu_0 \otimes \mu_1 \otimes P_{r^1} \| \mu_0 \otimes \mu_1 \otimes P_{r^2}\Big) = 0$ since $r^1(\tau^1) = r^1(\tau^2) = r^2(\tau^1) = r^2(\tau^2) = 1$. Therefore we have

$$\inf_{\widehat{\pi}} \sup_{\mathcal{M} \in \{\mathcal{M}_1, \mathcal{M}_2\}} \mathbb{E}_{\mathcal{D}}[L(\pi, \mathcal{M})] \geq \frac{1}{2}.$$

Case 2: $1 < C < 2$**.** Consider the case where there are two one states s^1, s^2 and two actions a^1, a^2. We suppose the initial state distribution of $P_0^\star$ is fixed as $P_0^\star(s^1) = C - 1$ and $P_0^\star(s^2) = 2 - C$. In addition, the state will stay the same throughout the whole episode. Then we can set the dataset distribution $\mu_0 = \mu_1$ where

$$\mu_0(\tau) = \begin{cases} \frac{C-1}{C}, & \text{if the state of } \tau \text{ is } s^1 \text{ and all actions in } \tau \text{ are } a^1 \text{ except } a_{H-1} = a^2, \\ \frac{C-1}{C}, & \text{if the state of } \tau \text{ is } s^1 \text{ and all actions in } \tau \text{ are } a^1 \text{ except } a_H = a^2, \\ \frac{2-C}{C}, & \text{if the state of } \tau \text{ is } s^2 \text{ and the actions are all } a^1, \\ 0, & \text{otherwise.} \end{cases}$$

In the following discussion we will use τ^3 to denote the trajectory where state is s^1 and all actions are a^1 except $a_{H-1} = a^2$; τ^4 to denote the trajectory where state is s^1 and all actions are a^1 except $a_{H-1} = a^2$; τ^5 to denote the trajectory where state is s^2 and all actions are a^1. Then we know

$$\mu_{0,h}(s^2, a^1) = \frac{2-C}{C}, \qquad \forall h \in [H],$$
$$\mu_{0,h}(s^1, a^1) = \frac{2C-2}{C}, \qquad \forall h \in [H-2],$$

$$
\begin{aligned}
\mu_{0,H-1}(s^1, a^1) = \mu_{0,H-1}(s^1, a^2) = \frac{C-1}{C}, \\
\mu_{0,H}(s^1, a^1) = \mu_{0,H}(s^1, a^2) = \frac{C-1}{C}.
\end{aligned}
$$

We consider two different reward function r^1 and r^2:

$$
\begin{aligned}
&r_h^1(s^1, a^1) = r_h^1(s^1, a^2) = r_h^2(s^1, a^1) = r_h^2(s^1, a^2) = 0, \qquad \forall h \in [H-2], \\
&r_{H-1}^1(s^1, a^1) = r_{H-1}^2(s^1, a^2) = 1, \qquad r_{H-1}^1(s^1, a^2) = r_{H-1}^2(s^1, a^1) = 0, \\
&r_H^1(s^1, a^1) = r_H^2(s^1, a^2) = 1, \qquad r_H^1(s^1, a^2) = r_H^2(s^1, a^1) = 0, \\
&r_h^1(s^2, a^1) = r_h^1(s^2, a^2) = r_h^2(s^2, a^1) = r_h^2(s^2, a^2) = 0, \qquad \forall h \in [H-1] \\
&r_H^1(s^2, a^1) = r_H^2(s^2, a^1) = 1, \qquad r_H^1(s^2, a^2) = r_H^2(s^2, a^2) = 0.
\end{aligned}
$$

Then we have two MDPs, $\mathcal{M}_1$ and $\mathcal{M}_2$ whose reward functions are r^1 and r^2 respectively. It can be easily verified that $(\mathcal{M}_1, \mu_0) \in \overline{\Theta}_{\rm st}(C), (\mathcal{M}_2, \mu_0) \in \Theta_{\rm st}(C)$.

In addition, we have for all policies π,

$$
L(\pi; \mathcal{M}_1) + L(\pi; \mathcal{M}_2) \geq 2(C-1).
$$

Therefore by Le Cam's method, we have

$$
\inf_{\widehat{\pi}} \sup_{\mathcal{M} \in \{\mathcal{M}_1, \mathcal{M}_2\}} \mathbb{E}_{\mathcal{D}}[L(\pi, \mathcal{M})] \geq \frac{(C-1)}{2} \exp\left(-N \cdot \mathbf{KL}\Big(\mu_0 \otimes \mu_1 \otimes P_{r^1} \| \mu_0 \otimes \mu_1 \otimes P_{r^2}\Big)\right),
$$

where the KL divergence is 0 since $r(\tau) = 1$ for all $r \in \{r^1, r^2\}$ and $\tau \in \{\tau^3, \tau^4, \tau^5\}$. Therefore, we have

$$
\inf_{\widehat{\pi}} \sup_{\mathcal{M} \in \{\mathcal{M}_1, \mathcal{M}_2\}} \mathbb{E}_{\mathcal{D}}[L(\pi, \mathcal{M})] \geq \frac{C-1}{2}.
$$

In conclusion, we have for any $C > 1$ and $H \geq 2$,

$$
\inf_{\widehat{\pi}} \sup_{(\mathcal{M}, \mu_0) \in \Theta_{\rm st}(C)} \mathbb{E}_{\mathcal{D}}[J(\pi^\star; r^\star, P^\star) - J(\widehat{\pi}; r^\star, P^\star)] \gtrsim \min\Big\{C-1, 1\Big\}.
$$

F.3 Proof of Theorem 3

The proof is inspired by the hard instances in Rashidinejad et al. (2021b). We consider the case $C \geq 2$ and $1 < C < 2$ respectively.

Case 1: $C \geq 2$. Consider the case where there is only one state s and two actions a^1, a^2. Set the dataset distribution $\mu_0 = \mu_1$ where

$$
\mu_0(\tau^\star) = \frac{1}{C}, \qquad \mu_0(\tau^\dagger) = 1 - \frac{1}{C},
$$

where $\tau^\star$ is the trajecotry where the actions are all a^1 and $\tau^\dagger$ is the trajecotry where the actions are all a^2.

We consider two different reward function r^1 and r^2:

$$
r^1(\tau) = \begin{cases} \frac{1}{2} + x, & \text{if all the actions in } \tau \text{ are } a^1, \\ \frac{1}{2}, & \text{otherwise.} \end{cases}
$$

$$
r^2(\tau) = \begin{cases} \frac{1}{2} - x, & \text{if all the actions in } \tau \text{ are } a^1, \\ \frac{1}{2}, & \text{otherwise.} \end{cases}
$$

Here $0 < x < \frac{1}{2}$ is a quantity we will specify later. Then we have two MDPs, $\mathcal{M}_1$ and $\mathcal{M}_2$ whose reward functions are r^1 and r^2 respectively. It can be easily verified that $(\mathcal{M}_1, \mu_0) \in \Theta_{\rm tr}(C), (\mathcal{M}_2, \mu_0) \in \Theta_{\rm tr}(C)$.

Further, let $L(\pi;\mathcal{M})$ denote the suboptimality of policy π in $\mathcal{M}$, then we have for all policies π,

$$L(\pi;\mathcal{M}_1)+L(\pi;\mathcal{M}_2)\geq x.$$

Now we can apply Le Cam's method, which leads to the following inequality

$$\inf_{\widehat{\pi}}\sup_{\mathcal{M}\in\{\mathcal{M}_1,\mathcal{M}_2\}}\mathbb{E}_{\mathcal{D}}[L(\pi,\mathcal{M})]\geq\frac{x}{4}\exp\Bigg(-N\cdot\mathbf{KL}\Big(\mu_0\otimes\mu_1\otimes P_{r^1}\|\mu_0\otimes\mu_1\otimes P_{r^2}\Big)\Bigg).$$

Now we only need to bound $\mathbf{KL}\Big(\mu_0\otimes\mu_1\otimes P_{r^1}\|\mu_0\otimes\mu_1\otimes P_{r^2}\Big)$, which can be computed as follows:

$$\begin{aligned}&\mathbf{KL}\Big(\mu_0\otimes\mu_1\otimes P_{r^1}\|\mu_0\otimes\mu_1\otimes P_{r^2}\Big)\\=&2\sum_{\tau^0=\tau^\star,\tau^1=\tau^\dagger}\mu_0(\tau^0)\mu_1(\tau^1)\mathbf{KL}\big(\mathrm{Bern}(\sigma(x))\|\mathrm{Bern}(\sigma(-x))\big)\\\leq&\frac{2\exp(1/2)x^2}{C}.\end{aligned}$$

Then by letting $x=\min\left\{\frac{1}{2},\sqrt{\frac{C}{2\exp(1/2)N}}\right\}$, we have

$$\inf_{\widehat{\pi}}\sup_{\mathcal{M}\in\{\mathcal{M}_1,\mathcal{M}_2\}}\mathbb{E}_{\mathcal{D}}[L(\pi,\mathcal{M})]\geq\frac{\exp(-1)}{4}x=\frac{\exp(-1)}{4}\min\left\{\frac{1}{2},\sqrt{\frac{C}{2\exp(1/2)N}}\right\}.$$

Case 2: $1<C<2$**.** Consider the case where there are two one states s^1,s^2 and two actions a^1,a^2. We suppose the initial state distribution of $P_0^\star$ is fixed as $P_0^\star(s^1)=C-1$ and $P_0^\star(s^2)=2-C$. In addition, the state will stay the same throughout the whole episode. Then we can set the dataset distribution $\mu_0=\mu_1$ where

$$\mu_0(\tau)=\begin{cases}\frac{2(C-1)}{C}\cdot\frac{1}{2}, & \text{if the state of } \tau \text{ is } s^1 \text{ and the actions are all } a^1 \text{ or all } a^2,\\ \frac{2-C}{C}, & \text{if the state of } \tau \text{ is } s^2 \text{ and the actions are all } a^1,\\ 0, & \text{if the state of } \tau \text{ is } s^2 \text{ and the actions contain } a^2.\end{cases}$$

Let $\tau^\star$ be the trajectory where the state is s^1 and the actions are all a^1.

We further consider two different reward function r^1 and r^2:

$$r^1(\tau)=\begin{cases}\frac{1}{2}+x, & \text{if the state is } s^1 \text{ and all the actions in } \tau \text{ are } a^1,\\ \frac{1}{2}, & \text{otherwise.}\end{cases}$$

$$r^2(\tau)=\begin{cases}\frac{1}{2}-x, & \text{if the state is } s^1 \text{ and all the actions in } \tau \text{ are } a^1,\\ \frac{1}{2}, & \text{otherwise.}\end{cases}$$

Here $0<x<\frac{1}{2}$ is a quantity we will specify later. Then we have two MDPs, $\mathcal{M}_1$ and $\mathcal{M}_2$ whose reward functions are r^1 and r^2 respectively. It can be easily verified that $(\mathcal{M}_1,\mu_0)\in\Theta_{\mathrm{tr}}(C),(\mathcal{M}_2,\mu_0)\in\Theta_{\mathrm{tr}}(C)$.

In addition, we have for all policies π,

$$L(\pi;\mathcal{M}_1)+L(\pi;\mathcal{M}_2)\geq(C-1)x.$$

Therefore by Le Cam's method, we have

$$\inf_{\widehat{\pi}}\sup_{\mathcal{M}\in\{\mathcal{M}_1,\mathcal{M}_2\}}\mathbb{E}_{\mathcal{D}}[L(\pi,\mathcal{M})]\geq\frac{(C-1)x}{4}\exp\Bigg(-N\cdot\mathbf{KL}\Big(\mu_0\otimes\mu_1\otimes P_{r^1}\|\mu_0\otimes\mu_1\otimes P_{r^2}\Big)\Bigg),$$

where the KL divergence can be computed as follows:

$$\begin{aligned}&\mathbf{KL}\Big(\mu_0\otimes\mu_1\otimes P_{r^1}\|\mu_0\otimes\mu_1\otimes P_{r^2}\Big)\\=&2\sum_{\tau^0=\tau^\star,\tau^1\neq\tau^\star}\mu_0(\tau^0)\mu_1(\tau^1)\mathbf{KL}\big(\mathrm{Bern}(\sigma(x))\|\mathrm{Bern}(\sigma(-x))\big)\end{aligned}$$

$$\leq \frac{2(C-1)\exp(1/2)x^2}{C}.$$

Then by letting $x = \min\left\{\frac{1}{2}, \sqrt{\frac{C}{2\exp(1/2)(C-1)N}}\right\}$, we have

$$\inf_{\widehat{\pi}} \sup_{\mathcal{M}\in\{\mathcal{M}_1,\mathcal{M}_2\}} \mathbb{E}_{\mathcal{D}}[L(\pi,\mathcal{M})] \geq \frac{(C-1)\exp(-1)}{4}x = \frac{\exp(-1)}{4}\min\left\{\frac{C-1}{2}, \sqrt{\frac{(C-1)}{2\exp(1/2)N}}\right\}.$$

In conclusion, we have for any $C > 1$ and $H \geq 1$,

$$\inf_{\widehat{\pi}} \sup_{(\mathcal{M},\mu_0)\in\Theta_{\mathrm{st}}(C)} \mathbb{E}_{\mathcal{D}}[J(\pi^\star; r^\star, P^\star) - J(\widehat{\pi}; r^\star, P^\star)] \gtrsim \min\left\{C-1, \sqrt{\frac{C-1}{N}}\right\}.$$

G Proof of Theorem 4

The proof still consists of two steps, deriving the guarantee of MLE and analyzing the performance of pessimistic offline RL.

Step 1: MLE guarantee. Note that Lemma 1 and Lemma 2 still applies here. Let $\mathcal{E}_1$ and $\mathcal{E}_2$ denote the event in Lemma 1 and Lemma 2 respectively. Following almost the same arguments, we have the following guarantee for the estimation of the system dynamics:

Lemma 3. *Under Assumption 4, with probability at least* $1-\delta/2$, *the following event holds true:*

$$\begin{aligned}
&(1) P_h^\star \in \mathcal{P}_h(\mathcal{D}), P_0^\star \in \mathcal{P}_{\mathrm{ini}}(\mathcal{D}), \qquad \forall h \in [H-1],\\
&(2) \mathbb{E}_{(s_h,a_h)\sim\mu_{0,h}}\left[\left\|P_h(\cdot|s,a) - P_h^\star(\cdot|s,a)\right\|_1^2\right] + \mathbb{E}_{(s_h,a_h)\sim\mu_{1,h}}\left[\left\|P_h(\cdot|s,a) - P_h^\star(\cdot|s,a)\right\|_1^2\right]\\
&\qquad \leq \frac{c\log(H\mathcal{N}_{\mathcal{G}_{P_h}}(1/N)/\delta)}{N}, \qquad \forall h \in [H-1], P_h \in \mathcal{P}_h(\mathcal{D}),\\
&(3) \mathbb{E}_{s\sim\mu_{0,1}}\left[\left\|P_0(s) - P_0^\star(s)\right\|_1^2\right] + \mathbb{E}_{s\sim\mu_{1,1}}\left[\left\|P_0(s) - P_0^\star(s)\right\|_1^2\right]\\
&\qquad \leq \frac{c\log(H\mathcal{N}_{\mathcal{G}_{P_0}}(1/N)/\delta)}{N}, \qquad \forall P_0 \in \mathcal{P}_0(\mathcal{D}).
\end{aligned}$$

The proof is omitted here. Let $\mathcal{E}_3$ denote the event in Lemma 3.

Step 2: Pessimistic offline RL. We first introduce the following lemma which suggests that under event $\mathcal{E}_3$, we can evaluate the expected cumulative reward of π_{tar} with respect to any reward function $r \in \mathcal{G}_r$ via the system dynamics $P_h \in \mathcal{P}_h(\mathcal{D})$:

Lemma 4. *Suppose Asusmption 3 is true. Then under* $\mathcal{E}_3$, *we have for all reward function* $r \in \mathcal{G}_r$ *and* $P = (\{P_h\}_{h=0}^{H-1})$ *where* $P_h \in \mathcal{P}_h(\mathcal{D})$ *that*

$$J(\pi_{\mathrm{tar}}; r, P^\star) - J(\pi_{\mathrm{tar}}; r, P) \leq Hr_{\max}\sqrt{\frac{cC_P^2(\{\mathcal{G}_{P_h}\}, \pi_{\mathrm{tar}})\log(H\mathcal{N}_P(1/N)/\delta)}{N}},$$

where $\mathcal{N}_P = \max_{0\leq h\leq H-1}\{\mathcal{N}_{\mathcal{G}_{P_h}}\}$.

The proof is deferred to Appendix G.1.

Let $(r_\pi^{\inf}, P_\pi^{\inf})$ denote $\mathrm{argmin}_{r\in\mathcal{R}(\mathcal{D}),P\in\mathcal{P}_{\mathrm{ini}}(\mathcal{D})\times\prod_{h=1}^{H-1}\mathcal{P}_h(\mathcal{D})} J(\pi; r, P) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r(\tau)]$. Then under the event $\mathcal{E}_3$, we can bound the suboptimality of $\widehat{\pi}$ as follows:

$$\begin{aligned}
&J(\pi_{\mathrm{tar}}; r^\star, P^\star) - J(\widehat{\pi}; r^\star, P^\star)\\
=&\big(J(\pi_{\mathrm{tar}}; r^\star, P^\star) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r^\star(\tau)]\big) - \big(J(\widehat{\pi}; r^\star, P^\star) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r^\star(\tau)]\big)\\
=&\Big(\big(J(\pi_{\mathrm{tar}}; r^\star, P^\star) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r^\star(\tau)]\big) - \big(J(\pi_{\mathrm{tar}}; r_{\pi_{\mathrm{tar}}}^{\inf}, P^\star) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r_{\pi_{\mathrm{tar}}}^{\inf}(\tau)]\big)\Big)
\end{aligned}$$

$$
\begin{aligned}
&+\Big(\big(J(\pi_{\mathrm{tar}}; r^{\mathrm{inf}}_{\pi_{\mathrm{tar}}}, P^\star) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r^{\mathrm{inf}}_{\pi_{\mathrm{tar}}}(\tau)]\big) - \big(J(\pi_{\mathrm{tar}}; r^{\mathrm{inf}}_{\pi_{\mathrm{tar}}}, P^{\mathrm{inf}}_{\pi_{\mathrm{tar}}}) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r^{\mathrm{inf}}_{\pi_{\mathrm{tar}}}(\tau)]\big)\Big)\\
&+\Big(\big(J(\pi_{\mathrm{tar}}; r^{\mathrm{inf}}_{\pi_{\mathrm{tar}}}, P^{\mathrm{inf}}_{\pi_{\mathrm{tar}}}) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r^{\mathrm{inf}}_{\pi_{\mathrm{tar}}}(\tau)]\big) - \big(J(\widehat{\pi}; r^{\mathrm{inf}}_{\widehat{\pi}}, P^{\mathrm{inf}}_{\widehat{\pi}}) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r^{\mathrm{inf}}_{\widehat{\pi}}(\tau)]\big)\Big)\\
&+\Big(\big(J(\widehat{\pi}; r^{\mathrm{inf}}_{\widehat{\pi}}, P^{\mathrm{inf}}_{\widehat{\pi}}) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r^{\mathrm{inf}}_{\widehat{\pi}}(\tau)]\big) - \big(J(\widehat{\pi}; r^\star, P^\star) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r^\star(\tau)]\big)\Big)\\
\leq&\Big(\big(J(\pi_{\mathrm{tar}}; r^\star, P^\star) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r^\star(\tau)]\big) - \big(J(\pi_{\mathrm{tar}}; r^{\mathrm{inf}}_{\pi_{\mathrm{tar}}}, P^\star) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r^{\mathrm{inf}}_{\pi_{\mathrm{tar}}}(\tau)]\big)\Big)\\
&+\Big(\big(J(\pi_{\mathrm{tar}}; r^{\mathrm{inf}}_{\pi_{\mathrm{tar}}}, P^\star) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r^{\mathrm{inf}}_{\pi_{\mathrm{tar}}}(\tau)]\big) - \big(J(\pi_{\mathrm{tar}}; r^{\mathrm{inf}}_{\pi_{\mathrm{tar}}}, P^{\mathrm{inf}}_{\pi_{\mathrm{tar}}}) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r^{\mathrm{inf}}_{\pi_{\mathrm{tar}}}(\tau)]\big)\Big)\\
&+\Big(\big(J(\widehat{\pi}; r^{\mathrm{inf}}_{\widehat{\pi}}, P^{\mathrm{inf}}_{\widehat{\pi}}) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r^{\mathrm{inf}}_{\widehat{\pi}}(\tau)]\big) - \big(J(\widehat{\pi}; r^\star, P^\star) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r^\star(\tau)]\big)\Big)\\
\leq&\Big(\big(J(\pi_{\mathrm{tar}}; r^\star, P^\star) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r^\star(\tau)]\big) - \big(J(\pi_{\mathrm{tar}}; r^{\mathrm{inf}}_{\pi_{\mathrm{tar}}}, P^\star) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r^{\mathrm{inf}}_{\pi_{\mathrm{tar}}}(\tau)]\big)\Big)\\
&+\Big(\big(J(\pi_{\mathrm{tar}}; r^{\mathrm{inf}}_{\pi_{\mathrm{tar}}}, P^\star) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r^{\mathrm{inf}}_{\pi_{\mathrm{tar}}}(\tau)]\big) - \big(J(\pi_{\mathrm{tar}}; r^{\mathrm{inf}}_{\pi_{\mathrm{tar}}}, P^{\mathrm{inf}}_{\pi_{\mathrm{tar}}}) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[r^{\mathrm{inf}}_{\pi_{\mathrm{tar}}}(\tau)]\big)\Big)\\
\leq&\sqrt{\frac{cC_r^2(\mathcal{G}_r, \pi_{\mathrm{tar}}, \mu_{\mathrm{ref}})\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/N)/\delta)}{N}} + Hr_{\max}\sqrt{\frac{cC_P^2(\{\mathcal{G}_{P_h}\}, \pi_{\mathrm{tar}})\log(H\mathcal{N}_P(1/N)/\delta)}{N}},
\end{aligned}
$$

where the third and fourth step are due to the definition of $\widehat{\pi}$, $(r^{\mathrm{inf}}_{\widehat{\pi}}, P^{\mathrm{inf}}_{\widehat{\pi}})$ and (1) in Lemma 3. The last step comes from Lemma 4 and the proof of Theorem 1. This concludes our proof.

G.1 Proof of Lemma 4

Let P^h be the system dynamics $(P_0^\star, \{P_t^\star\}_{t=1}^h, \{P_t\}_{t=h+1}^{H-1})$ for all $0 \leq h \leq H-1$. Then we have

$$
J(\pi_{\mathrm{tar}}; r, P^\star) - J(\pi_{\mathrm{tar}}; r, P) = \sum_{h=1}^{H-1}(J(\pi_{\mathrm{tar}}; r, P^h) - J(\pi_{\mathrm{tar}}; r, P^{h-1})) + (J(\pi_{\mathrm{tar}}; r, P^0) - J(\pi_{\mathrm{tar}}; r, P)).
$$

For any $h \in [H-1]$, we have

$$
\begin{aligned}
&J(\pi_{\mathrm{tar}}; r, P^h) - J(\pi_{\mathrm{tar}}; r, P^{h-1})\\
=&\mathbb{E}_{(s_1,a_1,\cdots,s_h,a_h)\sim(\pi_{\mathrm{tar}},P^\star)}\Big[\sum_{s_{h+1}} P_h^\star(s_{h+1}|s_h,a_h)\mathbb{E}_{(\pi_{\mathrm{tar}},P)}\big[r(\tau)|s_1,a_1,\cdots,s_{h+1}\big]\\
&\qquad\qquad\qquad\qquad\qquad - \sum_{s_{h+1}} P_h(s_{h+1}|s_h,a_h)\mathbb{E}_{(\pi_{\mathrm{tar}},P)}\big[r(\tau)|s_1,a_1,\cdots,s_{h+1}\big]\Big]\\
=&\mathbb{E}_{(s_1,a_1,\cdots,s_h,a_h)\sim(\pi_{\mathrm{tar}},P^\star)}\Big[\sum_{s_{h+1}} (P_h^\star(s_{h+1}|s_h,a_h) - P_h(s_{h+1}|s_h,a_h))\mathbb{E}_{(\pi_{\mathrm{tar}},P)}\big[r(\tau)|s_1,a_1,\cdots,s_{h+1}\big]\Big]\\
\leq& r_{\max}\mathbb{E}_{(s_h,a_h)\sim(\pi_{\mathrm{tar}},P^\star)}\Big[\big\|P_h^\star(\cdot|s_h,a_h) - P_h(\cdot|s_h,a_h)\big\|_1\Big]\\
\leq& r_{\max}\sqrt{\frac{cC_P^2(\pi_{\mathrm{tar}})\log(H\mathcal{N}_{\mathcal{G}_{P_h}}(1/N)/\delta)}{N}},
\end{aligned}
$$

where $\mathbb{E}_{(\pi_{\mathrm{tar}},P)}\big[\cdot|s_1,a_1,\cdots,s_{h+1}\big]$ is the distribution of the trajectory τ when executing policy π_{tar} under the transition probability $\{P_t\}_{t=h+1}^{H-1}$ while fixing the history to be $s_1, a_1, \cdots, s_{h+1}$. Here the first step utilizes the Tower property, the third and fourth step uses Cuachy-Schwartz inequality and the last step comes from Lemma 3.

For $J(\pi_{\mathrm{tar}}; r, P^0) - J(\pi_{\mathrm{tar}}; r, P)$, similarly we have

$$
J(\pi_{\mathrm{tar}}; r, P^0) - J(\pi_{\mathrm{tar}}; r, P) \leq r_{\max}\sqrt{\frac{cC_P^2(\pi_{\mathrm{tar}})\log(H\mathcal{N}_{\mathcal{G}_{P_0}}(1/N)/\delta)}{N}}.
$$

Therefore we conclude that

$$
J(\pi_{\mathrm{tar}}; r, P^\star) - J(\pi_{\mathrm{tar}}; r, P) \leq Hr_{\max}\sqrt{\frac{cC_P^2(\pi_{\mathrm{tar}})\log(H\mathcal{N}_P(1/N)/\delta)}{N}}.
$$

H Proof of Theorem 5

We first derive the guarantee of MLE for estimating $A^\star$. Similar to Lemma 1 and Lemma 2, we have the following lemma in the action-based comparison setting:

Lemma 5. *Under Assumption 7, with probability at least $1-\delta$, the following event holds true:*

$$\mathbb{E}_{s\sim\mu_h, a^0\sim\mu_{0,h}(\cdot|s), a^1\sim\mu_{1,h}(\cdot|s)}\left[\left\|P_{\widehat{A}_h}(\cdot|s,a^0,a^1)-P_{A_h^\star}(\cdot|s,a^0,a^1)\right\|_1^2\right]\leq\frac{c\log(H\mathcal{N}_{\mathcal{G}_{A_h}}(1/N)/\delta)}{N}, \forall h\in[H].$$

The proof is omitted here. Let $\mathcal{E}_4$ denote the event in Lemma 5. Then under Assumption 8, we can apply the mean value theorem and obtain that under $\mathcal{E}_4$, we have for all $h\in[H]$ that

$$\begin{aligned}&\mathbb{E}_{s\sim\mu_h, a^0\sim\mu_{0,h}(\cdot|s), a^1\sim\mu_{1,h}(\cdot|s)}\left[|A_h^\star(s,a^0)-A_h^\star(s,a^1)-\widehat{A}_h(s,a^0)+\widehat{A}_h(s,a^1)|^2\right]\\&\leq\frac{c\kappa^2\log(H\mathcal{N}_{\mathcal{G}_{A_h}}(1/N)/\delta)}{N}, \forall h\in[H].\end{aligned}\tag{6}$$

Recall that $\kappa=\frac{1}{\inf_{x\in[-r_{\max},r_{\max}]}\Phi'(x)}$.

On the other hand, note that we have the following performance lemma:

Lemma 6. *For any deterministic Markovian policies π and π', we have*

$$J(\pi;r^\star,P^\star)-J(\pi';r^\star,P^\star)=\sum_{h=1}^H\mathbb{E}_{s\sim d_h^{\pi'}}\left[Q_h^\pi(s,\pi(s))-Q_h^\pi(s,\pi'(s))\right]$$

The proof is deferred to Appendix H.1.

The rest of the proof largely follows Uehara et al. (2023). Under the event $\mathcal{E}_4$, we can bound the suboptimality of $\widehat{\pi}$ as follows:

$$J(\pi^\star;r^\star,P^\star)-J(\widehat{\pi};r^\star,P^\star)\leq r_{\max}\sum_{h=1}^H\mathbb{E}_{s\sim d_h^{\pi^\star}}\left[\mathbb{1}(\pi_h^\star(s)\neq\widehat{\pi}_h(s))\cdot\mathbb{1}(Q_h^\star(s,\widehat{\pi}_h(s))<Q_h^\star(s,\pi_h^\star(s)))\right]$$

$$\leq r_{\max}\sum_{h=1}^H\mathbb{E}_{s\sim d_h^{\pi^\star}}\left[\sum_{a\in\mathcal{A}}\mathbb{1}\Big(\widehat{A}_h(s,a)\geq\widehat{A}_h(s,\pi_h^\star(s))\Big)\cdot\mathbb{1}\Big(Q_h^\star(s,a)<Q_h^\star(s,\pi_h^\star(s))\Big)\right],$$

where the first step comes from Lemma 6 and the second step is due to the definition of $\widehat{\pi}$. Then for any $\alpha>0$, we have

$$\begin{aligned}&\mathbb{E}_{s\sim d_h^{\pi^\star}}\left[\sum_{a\in\mathcal{A}}\mathbb{1}\Big(\widehat{A}_h(s,a)\geq\widehat{A}_h(s,\pi_h^\star(s))\Big)\cdot\mathbb{1}\Big(Q_h^\star(s,a)<Q_h^\star(s,\pi_h^\star(s))\Big)\right]\\\leq&\mathbb{E}_{s\sim d_h^{\pi^\star}}\left[\sum_{a\in\mathcal{A}}\mathbb{1}\Big(Q_h^\star(s,\pi_h^\star(s))>Q_h^\star(s,a)\geq Q_h^\star(s,\pi_h^\star(s))-\alpha\Big)\right]\\&+\mathbb{E}_{s\sim d_h^{\pi^\star}}\left[\sum_{a\in\mathcal{A}}\mathbb{1}\Big(Q_h^\star(s,\pi_h^\star(s))-Q_h^\star(s,a)-\widehat{A}_h(s,\pi_h^\star(s))+\widehat{A}_h(s,a)\geq\alpha\Big)\right].\end{aligned}$$

By Assumption 6, we have

$$\mathbb{E}_{s\sim d_h^{\pi^\star}}\left[\sum_{a\in\mathcal{A}}\mathbb{1}\Big(Q_h^\star(s,\pi_h^\star(s))>Q_h^\star(s,a)\geq Q_h^\star(s,\pi_h^\star(s))-\alpha\Big)\right]\leq|\mathcal{A}|(\alpha/\alpha_0)^\beta.$$

For the second term, we have

$$\mathbb{E}_{s\sim d_h^{\pi^\star}}\left[\sum_{a\in\mathcal{A}}\mathbb{1}\Big(Q_h^\star(s,\pi_h^\star(s))-Q_h^\star(s,a)-\widehat{A}_h(s,\pi_h^\star(s))+\widehat{A}_h(s,a)\geq\alpha\Big)\right]$$

$$
\begin{aligned}
&= \frac{1}{\alpha^2}\mathbb{E}_{s\sim d_h^{\pi^\star}}\left[\sum_{a\in\mathcal{A}} \alpha^2 \mathbb{1}\Big(A_h^\star(s,\pi_h^\star(s)) - A_h^\star(s,a) - \widehat{A}_h(s,\pi_h^\star(s)) + \widehat{A}_h(s,a) \geq \alpha\Big)\right] \\
&\leq \frac{1}{\alpha^2}\mathbb{E}_{s\sim d_h^{\pi^\star}}\left[\sum_{a\in\mathcal{A}} \Big|A_h^\star(s,\pi_h^\star(s)) - A_h^\star(s,a) - \widehat{A}_h(s,\pi_h^\star(s)) + \widehat{A}_h(s,a)\Big|^2\right] \\
&\leq \frac{c|\mathcal{A}|C_{\mathrm{act}}\kappa^2\log(H\mathcal{N}_{\mathcal{G}_{A_h}}(1/N)/\delta)}{\alpha^2 N},
\end{aligned}
$$

where the last step comes from the definition of C_{act} and (6).

Therefore by picking appropriate α, we have with probability at least $1-\delta$ that

$$
J(\pi^\star; r^\star, P^\star) - J(\widehat{\pi}; r^\star, P^\star) \leq cH|\mathcal{A}|\left(\frac{2}{\beta}\right)^{\frac{\beta-2}{\beta+2}}\left(\frac{1}{\alpha_0}\right)^{\frac{2\beta}{\beta+2}}\left(\frac{\kappa^2 C_{\mathrm{act}}\log(H\mathcal{N}_{\mathcal{G}_A}(1/N)/\delta)}{N}\right)^{\frac{\beta}{\beta+2}}.
$$

H.1 Proof of Lemma 6

For any two policies π and π', we have that

$$
\begin{aligned}
&J(\pi'; r^\star, P^\star) - J(\pi; r^\star, P^\star) \\
=&\mathbb{E}_{\pi'}\left[r_1^\star(s_1,a_1) + V_2^{\pi'}(s_2)\right] - \mathbb{E}_{\pi'}\left[V_1^\pi(s_1)\right] \\
=&\mathbb{E}_{\pi'}\left[V_2^{\pi'}(s_2) - (V_1^\pi(s_1) - r_1^\star(s_1,a_1))\right] \\
=&\mathbb{E}_{\pi'}\left[V_2^{\pi'}(s_2) - V_2^\pi(s_2)\right] + \mathbb{E}_{\pi'}\left[Q_1^\pi(s_1,a_1) - V_1^{r,\pi}(s_1)\right] \\
=&\mathbb{E}_{\pi'}\left[V_2^{\pi'}(s_2) - V_2^\pi(s_2)\right] + \mathbb{E}_{\pi'}\left[\langle Q_1^\pi(s_1,\cdot), \pi_1'(\cdot|s_1) - \pi_1(\cdot|s_1)\rangle\right] \\
=&\cdots = \sum_{h=1}^H \mathbb{E}_{\pi'}\left[\langle Q_h^\pi(s_h,\cdot), \pi_h'(\cdot|s) - \pi_h(\cdot|s)\rangle\right].
\end{aligned}
$$

This concludes our proof.