Rethinking Offline Reinforcement Learning via Implicit State-State Planning (Appendixs)

Shuo Cao, Xuesong Wang, *Member, IEEE*, and Yuhu Cheng, *Member, IEEE*

APPENDIX A
PROOF FOR THEOREM 1

Proof: Consider two distinct SAV-functions, $\hat{Q}(s,a)$ and $\hat{Q}'(s,a)$. According to the definition of $\hat{\mathcal{B}}^{\pi_g}$, its infinite norm operation satisfies

$$\begin{aligned}
&\left\|(\hat{\mathcal{B}}^{\pi_g}\hat{Q})(s,a)-(\hat{\mathcal{B}}^{\pi_g}\hat{Q}')(s,a)\right\|_\infty \\
=&\max_{s,a}\left|\left(r+\gamma\mathbb{E}_{s',a'_{IS}\sim I(s',\chi_g(s'))}\left[\hat{Q}(s',a'_{IS})\right]\right)-\left(r+\gamma\mathbb{E}_{s',a'_{IS}\sim I(s',\chi_g(s'))}\left[\hat{Q}'(s',a'_{IS})\right]\right)\right| \\
\leq&\max_{s'}\left|\gamma\mathbb{E}_{s',a'_{IS}\sim I(s',\chi_g(s'))}\left[\hat{Q}(s',a'_{IS})-\hat{Q}'(s',a'_{IS})\right]\right| \\
\leq&\gamma\max_{s',a'_{IS}\sim I(s',\chi_g(s'))}\left|\hat{Q}(s',a'_{IS})-\hat{Q}'(s',a'_{IS})\right| \\
\leq&\gamma\max_{s,a}\left|\hat{Q}(s,a)-\hat{Q}'(s,a)\right| \\
=&\gamma\left\|\hat{Q}(s,a)-\hat{Q}'(s,a)\right\|_\infty.
\end{aligned} \tag{34}$$

According to Eq. (34), $\hat{\mathcal{B}}^{\pi_g}$ is a γ-contraction mapping, which satisfies $\left\|(\hat{\mathcal{B}}^{\pi_g}\hat{Q})-(\hat{\mathcal{B}}^{\pi_g}\hat{Q}')\right\|_\infty \leq \gamma\left\|\hat{Q}-\hat{Q}'\right\|_\infty$. Setting $\hat{Q}^*$ as the unique fixed point of the SAV-function $\hat{Q}$, then repeatedly executing $\hat{\mathcal{B}}^{\pi_g}$ yields

$$\left\|\hat{Q}_{k+1}-\hat{Q}^*\right\|_\infty=\left\|(\hat{\mathcal{B}}^{\pi_g}\hat{Q}_k)-(\hat{\mathcal{B}}^{\pi_g}\hat{Q}^*)\right\|_\infty\leq\gamma\left\|\hat{Q}_k-\hat{Q}^*\right\|_\infty\leq\cdots\leq\gamma^{k+1}\left\|\hat{Q}_0-\hat{Q}^*\right\|_\infty. \tag{35}$$

Eq. (35) indicates that for $\gamma<1$, as the number of iterations k increases, $\hat{Q}_{k+1}$ converges to $\hat{Q}^*$, i.e., $\lim_{k\to\infty}\hat{Q}_{k+1}=\hat{Q}^*$. Thus, Theorem 1 is proofed. ■

APPENDIX B
PROOF FOR THEOREM 2

Proof: Consider an MDP with the deterministic FDM and IDM. Solving this MDP through explicit state-stitching is equivalent to solving a sub-MDP via explicit action-stitching, wherein the sub-MDP includes only the $I(s,s')$ mappings that return a set of actions for each state s (i.e., the proposed implicit state-stitching). Specifically, for each state s, the following holds:

$$Q(s,a)\geq\max_{s'}\tilde{Q}(s,s')=\max_{s'}\hat{Q}(s,a=I(s,s'))\quad\text{for least one } a. \tag{36}$$

The actions obtained by $I(s,s')$ may constitute only a subset of the full MDP action space. However, under the assumption that s' is the optimal choice, the return associated with any MDP action must be less than or equal to the return achieved by $I(s,s')$. Specifically, for each state s, the following holds:

$$Q(s,a)\leq\max_{s'}\hat{Q}(s,I(s,s'))\quad\text{for all } a. \tag{37}$$

Combining Eqs. (36) and (37), we conclude that for a deterministic MDP, there exist

$$\max_a Q(s,a)=\max_{s'}\tilde{Q}(s,s')=\max_{s'}\hat{Q}(s,a=I(s,s')). \tag{38}$$

Thus, Theorem 2 is proofed. ■

APPENDIX C
PROOF FOR THEOREM 3

Proof: Recalling Eq. (27) and applying the triangle inequality, we obtain

$$\begin{aligned}
&\left\|\hat{Q}^{\pi_g}\left(s,\pi_g(s)\right)-Q^{\pi_\beta}\left(s,\pi_\beta(s)\right)\right\| \\
=&\left\|\hat{Q}^{\pi_g}\left(s,I(s,\chi_g(s))\right)-Q^{\pi_\beta}\left(s,\pi_\beta(s)\right)\right\| \\
=&\Big\|\left(\hat{Q}^{\pi_g}\left(s,I(s,\chi_g(s))\right)-\hat{Q}^{\pi_g}\left(s,\pi_\beta(s)\right)\right)+\left(\hat{Q}^{\pi_g}\left(s,\pi_\beta(s)\right)-\hat{Q}^{\pi_\beta}\left(s,\pi_\beta(s)\right)\right) \\
&\quad+\left(\hat{Q}^{\pi_\beta}\left(s,\pi_\beta(s)\right)-\bar{Q}^{\pi_\beta}\left(s,\pi_\beta(s)\right)\right)+\left(\bar{Q}^{\pi_\beta}\left(s,\pi_\beta(s)\right)-Q^{\pi_\beta}\left(s,\pi_\beta(s)\right)\right)\Big\| \\
\leq&\left\|\hat{Q}^{\pi_g}\left(s,I(s,\chi_g(s))\right)-\hat{Q}^{\pi_g}\left(s,\pi_\beta(s)\right)\right\|+\left\|\hat{Q}^{\pi_g}\left(s,\pi_\beta(s)\right)-\hat{Q}^{\pi_\beta}\left(s,\pi_\beta(s)\right)\right\| \\
&+\left\|\hat{Q}^{\pi_\beta}\left(s,\pi_\beta(s)\right)-\bar{Q}^{\pi_\beta}\left(s,\pi_\beta(s)\right)\right\|+\left\|\bar{Q}^{\pi_\beta}\left(s,\pi_\beta(s)\right)-Q^{\pi_\beta}\left(s,\pi_\beta(s)\right)\right\|. \qquad (39)
\end{aligned}$$

Considering $\left\|\hat{Q}^{\pi_g}\left(s,I(s,\chi_g(s))\right)-\hat{Q}^{\pi_g}\left(s,\pi_\beta(s)\right)\right\|$, and given the K_Q-Lipschitz as stated in Assumption 1, we can derive

$$\left\|\hat{Q}^{\pi_g}\left(s,I(s,\chi_g(s))\right)-\hat{Q}^{\pi_g}\left(s,\pi_\beta(s)\right)\right\| \leq K_Q\left\|I(s,\chi_g(s))-\pi_\beta(s)\right\|. \qquad (40)$$

Based on $\pi_\beta(s)=I(s,s')$ and the K_I-Lipschitz as stated in Assumption 1, $\|I(s,\chi_g(s))-\pi_\beta(s)\|=\|I(s,\chi_g(s))-I(s,s')\|\leq K_I\|\chi_g(s)-s'\|$ holds. Recalling the state constraint threshold $\|\chi_g(s)-s'\|\leq\epsilon_g$ from Eq. (14), there exist

$$\|I(s,\chi_g(s))-\pi_\beta(s)\|\leq K_I\epsilon_g. \qquad (41)$$

Combining Eqs. (40) and (41), we obtain

$$\left\|\hat{Q}^{\pi_g}\left(s,I(s,\chi_g(s))\right)-\hat{Q}^{\pi_g}\left(s,\pi_\beta(s)\right)\right\|\leq K_QK_I\epsilon_g. \qquad (42)$$

Next, considering $\left\|\hat{Q}^{\pi_g}\left(s,\pi_\beta(s)\right)-\hat{Q}^{\pi_\beta}\left(s,\pi_\beta(s)\right)\right\|$, and based on the value gaps under different policies as presented in Lemma 1, it follows that

$$\left\|\hat{Q}^{\pi_g}\left(s,\pi_\beta(s)\right)-\hat{Q}^{\pi_\beta}\left(s,\pi_\beta(s)\right)\right\|\leq\left\|\hat{Q}^{\pi_g}\left(s,a\right)-\hat{Q}^{\pi_\beta}\left(s,a\right)\right\|\leq R_{\max}\sqrt{2K_I\epsilon_g}/(1-\gamma)^2. \qquad (43)$$

Due to the estimation deviation between the $\hat{Q}^{\pi_\beta}$ estimated by off-policy expectile regression and the $\bar{Q}^{\pi_\beta}$ estimated by MSE, we define the upper bound of the bias between them as δ_τ:

$$\left\|\hat{Q}^{\pi_\beta}\left(s,\pi_\beta(s)\right)-\bar{Q}^{\pi_\beta}\left(s,\pi_\beta(s)\right)\right\|\leq\left\|\hat{Q}^{\pi_\beta}\left(s,a\right)-\bar{Q}^{\pi_\beta}\left(s,a\right)\right\|\leq\delta_\tau. \qquad (44)$$

Furthermore, based on the sampling error result from Lemma 2, we can conclude that

$$\left\|\bar{Q}^{\pi_\beta}\left(s,\pi_\beta(s)\right)-Q^{\pi_\beta}\left(s,\pi_\beta(s)\right)\right\|\leq\left\|\bar{Q}^{\pi_\beta}\left(s,a\right)-Q^{\pi_\beta}\left(s,a\right)\right\|\leq\delta_{\mathcal{D}}. \qquad (45)$$

Finally, substituting Eqs. (42)-(45) into Eq. (39), we can derive

$$\left\|\hat{Q}^{\pi_g}\left(s,I(s,\chi_g(s))\right)-Q^{\pi_\beta}\left(s,\pi_\beta(s)\right)\right\|\leq\delta_{\mathcal{D}}+\delta_\tau+K_QK_I\epsilon_g+R_{\max}\sqrt{2K_I\epsilon_g}/(1-\gamma)^2. \qquad (46)$$

Thus, Theorem 3 is proofed. ■

APPENDIX D
PROOF FOR THEOREM 4

Proof: In ORL, once the training dataset is fixed, the performance gap $\varepsilon_{\mathcal{D}}$ between the actual global optimal policy π^* and the local optimal behavior policy π^*_β derived from the finite dataset, becomes fixed:

$$\varepsilon_{\mathcal{D}}=J(\pi^*)-J(\pi^*_\beta). \qquad (47)$$

While the performance gap between the executed-policy π_e and the optimal policy π^* is $J(\pi^*)-J(\pi_e)$, by applying Eq. (31) to extend $|J(\pi^*)-J(\pi_e)|$ and using the triangle inequality, we obtain

$$\begin{aligned}
|J(\pi^*)-J(\pi_e)|&=\left|\left(J(\pi^*)-J(\pi_g)\right)+\left(J(\pi_g)-J(\pi^*_\beta)\right)+\left(J(\pi^*_\beta)-J(\pi_e)\right)\right| \\
&\leq|J(\pi^*)-J(\pi_g)|+\left|J(\pi_g)-J(\pi^*_\beta)\right|+\left|J(\pi^*_\beta)-J(\pi_e)\right|. \qquad (48)
\end{aligned}$$

Considering $|J(\pi^*) - J(\pi_g)|$, based on the definition of $J(\pi)$ in Section III-A, Lemma 3, and Eq. (47), we can derive

$$\begin{aligned}
|J(\pi^*) - J(\pi_g)| =& \left| \frac{1}{1-\gamma} \mathbb{E}_{s\sim d^{\pi^*}(s)} \left[r(s) \right] - \frac{1}{1-\gamma} \mathbb{E}_{s\sim d^{\pi_g}(s)} \left[r(s) \right] \right| \\
\leq& \frac{1}{1-\gamma} \left| \int_s \left(d^{\pi^*}(s) - d^{\pi_g}(s) \right) r(s) \mathrm{d}s \right| \\
\leq& \frac{R_{\max}}{1-\gamma} \int_s \left| d^{\pi^*}(s) - d^{\pi_g}(s) \right| \mathrm{d}s \\
\leq& \frac{R_{\max}}{1-\gamma} C K_T K_I \max_s \| \chi^*(s) - \chi_g(s) \| \\
=& \frac{R_{\max}}{1-\gamma} C K_T K_I \max_s \| \chi^*(s) - s' + s' - \chi_g(s) \| \\
\leq& \frac{R_{\max}}{1-\gamma} C K_T K_I \max_s \left(\| \chi^*(s) - s' \| + \| s' - \chi_g(s) \| \right) \\
\leq& \frac{R_{\max}}{1-\gamma} C K_T K_I (\varepsilon_{\mathcal{D}} + \epsilon_g).
\end{aligned} \tag{49}$$

By considering $\left|J(\pi_g) - J(\pi^*_\beta)\right|$ and applying Lemma 3, we obtain

$$\begin{aligned}
\left|J(\pi_g) - J(\pi^*_\beta)\right| =& \left| \frac{1}{1-\gamma} \mathbb{E}_{s\sim d^{\pi_g}(s)} \left[r(s) \right] - \frac{1}{1-\gamma} \mathbb{E}_{s\sim d^{\pi^*_\beta}(s)} \left[r(s) \right] \right| \\
\leq& \frac{1}{1-\gamma} \left| \int_s \left(d^{\pi_g}(s) - d^{\pi^*_\beta}(s) \right) r(s) \mathrm{d}s \right| \\
\leq& \frac{R_{\max}}{1-\gamma} \int_s \left| d^{\pi_g}(s) - d^{\pi^*_\beta}(s) \right| \mathrm{d}s \\
\leq& \frac{R_{\max}}{1-\gamma} C K_T K_I \max_s \| \chi_g(s) - s' \| \\
\leq& \frac{R_{\max}}{1-\gamma} C K_T K_I \epsilon_g.
\end{aligned} \tag{50}$$

Similarly, by considering $\left|J(\pi_g) - J(\pi^*_\beta)\right|$ and utilizing Lemma 3, we obtain

$$\begin{aligned}
\left|J(\pi^*_\beta) - J(\pi_e)\right| =& \left| \frac{1}{1-\gamma} \mathbb{E}_{s\sim d^{\pi^*_\beta}(s)} \left[r(s) \right] - \frac{1}{1-\gamma} \mathbb{E}_{s\sim d^{\pi_e}(s)} \left[r(s) \right] \right| \\
\leq& \frac{1}{1-\gamma} \left| \int_s \left(d^{\pi^*_\beta}(s) - d^{\pi_e}(s) \right) r(s) \mathrm{d}s \right| \\
\leq& \frac{R_{\max}}{1-\gamma} \int_s \left| d^{\pi^*_\beta}(s) - d^{\pi_e}(s) \right| \mathrm{d}s \\
\leq& \frac{R_{\max}}{1-\gamma} C K_T K_I \max_s \| s' - \chi_e(s) \| \\
\leq& \frac{R_{\max}}{1-\gamma} C K_T K_I \epsilon_e.
\end{aligned} \tag{51}$$

Substituting Eqs. (49)-(51) into Eq. (48), we obtain

$$|J(\pi^*) - J(\pi_e)| \leq \frac{R_{\max}}{1-\gamma} C K_T K_I (\varepsilon_{\mathcal{D}} + 2\epsilon_g + \epsilon_e). \tag{52}$$

Furthermore, the performance lower bound of executed-policy π_e is

$$J(\pi_e) \geq J(\pi^*) - K(\varepsilon_{\mathcal{D}} + 2\epsilon_g + \epsilon_e) := \chi_e. \tag{53}$$

where $K = R_{\max} C K_T K_I / (1-\gamma)$.

Similar to Eqs. (47)-(53), if π_e and χ_e are replaced with π_g and χ_g, then the performance lower bound of guided-policy π_g satisfies

$$J(\pi_g) \geq J(\pi^*) - K(\varepsilon_{\mathcal{D}} + \epsilon_g + 2\epsilon_e) := \chi_g. \tag{54}$$

Recalling $\epsilon_g \leq \epsilon_e$ given in Eq. (18), it follows that $\chi_e \geq \chi_g$. Thus, Theorem 4 is proofed. ■

APPENDIX E
TASK GOALS AND DATASET TYPES FOR GYM-MUJOCO AND ANTMAZE IN D4RL

TABLE SI
DESCRIPTION AND REFERENCE MIN-MAX RETURN OF THREE GYM-MUJOCO AND THREE ANTMAZE ENVIRONMENTS

Control tasks	**Learning goal**	**Reward**	**Ref. min**	**Ref. max**
HalfCheetah	Control a cheetah-like robot to run as quickly forward as possible	Dense	-280.18	12135.0
Hopper	Control a monopod-like robot to jump as far forward as possible	Dense	-20.27	3234.3
Walker2d	Control a biped-like robot to walk as quickly forward as possible	Dense	1.63	4592.3
Antmaze-umaze	Control an 8-DOF ant-like robot to find the shortest path in a simple-maze layout	Sparse	0.0	1.0
Antmaze-medium	Control an 8-DOF ant-like robot to find the shortest path in a medium-maze layout	Sparse	0.0	1.0
Antmaze-large	Control an 8-DOF ant-like robot to find the shortest path in a difficult-maze layout	Sparse	0.0	1.0

TABLE SII
FIVE DATASET TYPES FOR EACH GYM-MUJOCO ENVIRONMENT AND THREE DATASET TYPES FORTHE ANTMAZE ENVIRONMENT

Dataset type	**Description of collected data quality/policy**	**Policy num.**	**Policy mix**	**Experiences**
Random	Random-level policy	Single	100%	10^6
Medium	Medium-level policy that achieves one-third of expert policy performance	Single	100%	10^6
Medium-replay	Medium dataset and replay buffer during medium-level agent training	Multiple	50% : 50%	$\approx 2 \times 10^6$
Medium-expert	Medium-level policy and expert-level policy	Multiple	50% : 50%	2×10^6
Expert	Expert-level agent policy or demonstration	Single	100%	10^6
Fixed	Goal-reaching policy from a fixed location to a specific goal	Single	100%	10^6
Diverse	Goal-reaching policy from a random location to a random goal	Single	100%	10^6
Play	Starting from hand-picked location set to specific hand-picked goal set	Single	100%	10^6