Pessimistic Model Selection for Offline Deep Reinforcement Learning (Supplementary Material)

Submitted to the 39[th] *Conference on Uncertainty in Artificial Intelligence* (UAI 2023). **To be used for reviewing only**.

A COMMENTS ON ASYMPTOTIC RESULTS

We remark here that all theoretical justification in this paper is based on asymptotics. It might be possible to investigate finite sample regimes when one has an exact confidence interval or a non-asymptotic bound. However, having an exact confidence interval might require some model specification of the value function, and using non-asymptotic bounds might require additional tuning steps (e.g., constants in many concentration inequalities), which is beyond the scope of this paper. In addition, as seen from our empirical evaluations below, with a relatively large sample size, the proposed model selection approach performs well.

B TECHNICAL PROOFS

Notations: The notation $\xi(N) \lesssim \theta(N)$ (resp. $\xi(N) \gtrsim \theta(N)$) means that there exists a sufficiently large (resp. small) constant $c_1 > 0$ (resp. $c_2 > 0$) such that $\xi(N) \leq c_1\theta(N)$ (resp. $\xi(N) \geq c_2\theta(N)$) for some sequences $\theta(N)$ and $\xi(N)$ related to N. In the following proofs, N often refers to some quantity related to n and T.

Lemma 1 and its proof : Let J denotes some index of our batch data $\mathcal{D}_n$. Define

$$\phi(J, Q^\pi, \omega^{\pi,\nu}, \pi) = \frac{1}{|J|}\sum_{(i,t)\in J} \omega^{\pi,\nu}(S_{i,t}, A_t)\left(R_{i,t} + \gamma\sum_{a'\in\mathcal{A}}\pi(a'|S_{i,t+1})Q^\pi(S_{i,t+1}, a') - Q^\pi(S_{i,t}, A_{i,t})\right),$$

where $|J|$ is the cardinality of the index set J, e.g., $|J_o| = \frac{nT}{O}$ for every $1 \leq o \leq O$. Then we have the following Lemma 1 as an intermediate result to Theorem .

Lemma 1 *Under Assumptions , for every $1 \leq l \leq L$ and $1 \leq o \leq O-1$, the following asymptotic equivalence holds.*

$$\sqrt{\frac{nT}{O}}\left\{\hat{\mathcal{V}}_{\mathcal{D}_{o+1}}(\hat{\pi}_l^{(o)}) - \mathcal{V}(\hat{\pi}_l^{(o)})\right\} = \sqrt{\frac{nT}{O}}\phi(J, Q^{\hat{\pi}^{*(o)}}, \omega^{\hat{\pi}_l^{(o)},\nu}, \hat{\pi}_l^{(o)}) + o_p(1), \tag{1}$$

where $o_p(1)$ refers to a quantity that converges to 0 as n or T goes to infinity.

The proof is similar to that of Theorem 7 in Kallus and Uehara [2019]. First, notice that

$$\begin{aligned}
&\sqrt{\frac{nT}{O}}\left\{\hat{\mathcal{V}}_{\mathcal{D}_{o+1}}(\hat{\pi}_l^{(o)}) - \mathcal{V}(\hat{\pi}_l^{(o)})\right\}\\
=&\sqrt{\frac{nT}{O}}\Big\{\phi(J, \widehat{Q}^{\hat{\pi}^{*(o)}}, \widehat{\omega}^{\hat{\pi}_l^{(o)},\nu}, \hat{\pi}_l^{(o)}) - \phi(J, Q^{\hat{\pi}^{*(o)}}, \omega^{\hat{\pi}_l^{(o)},\nu}, \hat{\pi}_l^{(o)})\\
&+ (1-\gamma)\mathbb{E}_{S_0\sim\nu}[\sum_{a\in\mathcal{A}}\hat{\pi}_l^{(o)}(a|S_0)Q^{\hat{\pi}_l^{(o)}}(S_0, a)] - (1-\gamma)\mathbb{E}_{S_0\sim\nu}[\sum_{a\in\mathcal{A}}\hat{\pi}_l^{(o)}(a|S_0)Q^{\hat{\pi}_l^{(o)}}(S_0, a)]\Big\}\\
&+\sqrt{\frac{nT}{O}}\phi(J, Q^{\hat{\pi}^{*(o)}}, \omega^{\hat{\pi}_l^{(o)},\nu}, \hat{\pi}_l^{(o)}).
\end{aligned}$$

Then it suffices to show the term in the first bracket converges to 0 faster than $\sqrt{nT}$. Notice that

$$\begin{aligned}
&\Big\{\phi(J, \widehat{Q}^{\hat{\pi}^{*(o)}}, \widehat{\omega}^{\hat{\pi}_l^{(o)},\nu}, \hat{\pi}_l^{(o)}) - \phi(J, Q^{\hat{\pi}^{*(o)}}, \omega^{\hat{\pi}_l^{(o)},\nu}, \hat{\pi}_l^{(o)})\\
&+ (1-\gamma)\mathbb{E}_{S_0\sim\nu}[\sum_{a\in\mathcal{A}}\hat{\pi}_l^{(o)}(a|S_0)Q^{\hat{\pi}_l^{(o)}}(S_0, a)] - (1-\gamma)\mathbb{E}_{S_0\sim\nu}[\sum_{a\in\mathcal{A}}\hat{\pi}_l^{(o)}(a|S_0)Q^{\hat{\pi}_l^{(o)}}(S_0, a)]\Big\}\\
=&E_1 + E_2 + E_3,
\end{aligned}$$

where

$$\begin{aligned}
E_1 =&\frac{O}{nT}\sum_{(i,t)\in J_{o+1}}(\widehat{\omega}^{\hat{\pi}_l^{(o)},\nu}(S_{i,t}, A_{i,t}) - \omega^{\hat{\pi}_l^{(o)},\nu}(S_{i,t}, A_{i,t}))(R_{i,t} - Q^{\hat{\pi}_l^{(o)}}(S_{i,t}, A_{i,t})\\
&+\gamma\sum_{a\in\mathcal{A}}\hat{\pi}_l^{(o)}(a|S_{i,t+1})Q^{\hat{\pi}_l^{(o)}}(S_{i,t+1}, a)),
\end{aligned}$$

$$E_2 = \frac{O}{nT} \sum_{(i,t)\in J_{o+1}} \omega^{\hat{\pi}_l^{(o)},\nu}(S_{i,t},A_{i,t})(\widehat{Q}^{\hat{\pi}_l^{(o)}}(S_{i,t},A_{i,t}) - Q^{\hat{\pi}_l^{(o)}}(S_{i,t},A_{i,t})$$
$$+\gamma \sum_{a\in\mathcal{A}} \hat{\pi}_l^{(o)}(a|S_{i,t+1})(\widehat{Q}^{\hat{\pi}_l^{(o)}}(S_{i,t+1},a) - Q^{\hat{\pi}_l^{(o)}}(S_{i,t+1},a))),$$

and

$$E_3 = \frac{O}{nT} \sum_{(i,t)\in J_{o+1}} (\widehat{\omega}^{\hat{\pi}_l^{(o)},\nu}(S_{i,t},A_{i,t}) - \omega^{\hat{\pi}_l^{(o)},\nu}(S_{i,t},A_{i,t}))(\widehat{Q}^{\hat{\pi}_l^{(o)}}(S_{i,t},A_{i,t}) - Q^{\hat{\pi}_l^{(o)}}(S_{i,t},A_{i,t})$$
$$+\gamma \sum_{a\in\mathcal{A}} \hat{\pi}_l^{(o)}(a|S_{i,t+1})(\widehat{Q}^{\hat{\pi}_l^{(o)}}(S_{i,t+1},a) - Q^{\hat{\pi}_l^{(o)}}(S_{i,t+1},a))).$$

Next, we bound each of the above three terms. For term E_1, it can be seen that

$$\mathbb{E}[E_1|\bar{J}_o] = 0.$$

In addition, by the previous Assumptions, we can show

$$\mathrm{Var}[E_1] = \mathbb{E}[\mathrm{Var}(E_1|\bar{J}_o)] \lesssim \frac{O}{nT}(nT/O)^{-2\kappa_2},$$

where the inequality is based on that each item in E_3 is uncorrelated with others. Then by Markov's inequality, we can show

$$|E_1| = O_p((\frac{O}{nT})^{-1/2-\kappa_2}).$$

Similarly, we can show

$$|E_2| = O_p((\frac{O}{nT})^{-1/2-\kappa_1}).$$

For term (E_3), by Cauchy Schwarz inequality and similar arguments as before, we can show

$$|E_3| = O_p((\frac{O}{nT})^{-(\kappa_2+\kappa_1)}).$$

Therefore, as long as $(\kappa_2+\kappa_1) > 1/2$, we have $E_1 + E_2 + E_3 = o(\sqrt{O/nT})$, which concludes our proof.

Proof of Theorem We aim to show that

$$\frac{\sqrt{nT(O-1)/O}\left(\hat{\mathcal{V}}(\hat{\pi}_l) - \mathcal{V}(\hat{\pi}_l)\right)}{\hat{\sigma}(l)} \Longrightarrow \mathcal{N}(0,1).$$

It can be seen that

$$\begin{aligned}\frac{\sqrt{nT(O-1)/O}\left(\hat{\mathcal{V}}(\hat{\pi}_l) - \mathcal{V}(\hat{\pi}_l)\right)}{\hat{\sigma}(l)} &= \sqrt{\frac{nT}{O(O-1)}}\left(\sum_{o=1}^{O-1} \frac{\hat{\mathcal{V}}_{\mathcal{D}_{o+1}}(\hat{\pi}_l^{(o)}) - \mathcal{V}(\hat{\pi}_l)}{\hat{\sigma}_{o+1}(\hat{\pi}_l^{(o)})}\right) \\ &= \sqrt{\frac{nT}{O(O-1)}}\left(\sum_{o=1}^{O-1} \frac{\hat{\mathcal{V}}_{\mathcal{D}_{o+1}}(\hat{\pi}_l^{(o)}) - \mathcal{V}(\hat{\pi}_l^{(o)})}{\hat{\sigma}_{o+1}(\hat{\pi}_l^{(o)})}\right) \\ &+ \sqrt{\frac{nT}{O(O-1)}}\left(\sum_{o=1}^{O-1} \frac{\mathcal{V}(\hat{\pi}_l^{(o)}) - \mathcal{V}(\hat{\pi}_l)}{\hat{\sigma}_{o+1}(\hat{\pi}_l^{(o)})}\right).\end{aligned}$$

Define

$$\phi(J, Q^\pi, w^\pi, \pi) = \frac{1}{|J|}\sum_{(i,t)\in J} w^{\pi,\nu}(S_{i,t},A_t)\left(R_{i,t} + \gamma\sum_{a'\in\mathcal{A}} \pi(a'|S_{i,t+1})Q^\pi(S_{9,t+1},a') - Q^\pi(S_{i,t},A_{i,t})\right),$$

where $|J|$ is the cardinality of the index set J, i.e., $|J| = \frac{nT}{O}$. Then by Lemma 1, we show that

$$\sqrt{\frac{nT}{O}} \frac{\hat{\mathcal{V}}_{\mathcal{D}_{o+1}}(\hat{\pi}_l^{(o)}) - \mathcal{V}(\hat{\pi}_l^{(o)})}{\hat{\sigma}_{o+1}(\hat{\pi}_l^{(o)})} = \sqrt{\frac{nT}{O}} \frac{\phi(J_{o+1}, Q^{\hat{\pi}_l^{(o)}}, w^{\hat{\pi}_l^{(o)}}, \hat{\pi}_l^{(o)})}{\hat{\sigma}_{o+1}(\hat{\pi}_l^{(o)})} + o_p(1). \tag{2}$$

If we can show that

$$\max_{1\le o\le (O-1)} \left| \frac{\hat{\sigma}_{o+1}(\hat{\pi}_l^{(o)})}{\sigma_{o+1}(\hat{\pi}_l^{(o)})} - 1 \right| = o_p(1),$$

which will be shown later, then by Slutsky theorem, we can show that

$$\begin{aligned}
&\sqrt{\frac{nT}{O(O-1)}} \left(\sum_{o=1}^{O-1} \frac{\hat{\mathcal{V}}_{\mathcal{D}_{o+1}}(\hat{\pi}_l^{(o)}) - \mathcal{V}(\hat{\pi}_l^{(o)})}{\hat{\sigma}_{o+1}(\hat{\pi}_l^{(o)})} \right) \\
&= \underbrace{\sqrt{\frac{nT}{O(O-1)}} \left(\sum_{o=1}^{O-1} \frac{\phi(J_{o+1}, Q^{\hat{\pi}_l^{(o)}}, w^{\hat{\pi}_l^{(o)}}, \hat{\pi}_l^{(o)})}{\sigma_{o+1}(\hat{\pi}_l^{(o)})} \right)}_{(I)} + o_p(1).
\end{aligned}$$

For (I), we can see that

$$(I) = \sqrt{\frac{O}{nT(O-1)}} (\sum_{o=1}^{O-1} \sum_{(i,t)\in J_{o+1}} w^{\hat{\pi}_l^{(o)},\nu}(S_{i,t}, A_{i,t})(R_{i,t} \tag{3}$$

$$+ \gamma \sum_{a'\in\mathcal{A}} \hat{\pi}_l^{(o)}(a'|S_{i,t+1}) Q^{\hat{\pi}_l^{(o)}}(S_{i,t+1}, a') - Q^{\hat{\pi}_l^{(o)}}(S_{i,t}, A_{i,t}))/\sigma_{o+1}(\hat{\pi}_l^{(o)})). \tag{4}$$

By the sequential structure of our proposed algorithm, (I) forms a mean zero martingale. Then we use Corollary 2.8 of [McLeish, 1974] to show its asymptotic distribution. First of all, by the uniformly bounded assumption on Q-function, ratio function and the variance, we can show that

$$\begin{aligned}
\sqrt{\frac{O}{nT(O-1)}} \max_{1\le o\le (O-1)} \max_{(i,t)\in J_0} \Big| w^{\hat{\pi}_l^{(o)},\nu}(S_{i,t}, A_{i,t})(R_{i,t} + \gamma \sum_{a'\in\mathcal{A}} \hat{\pi}_l^{(o)}(a'|S_{i,t+1}) Q^{\hat{\pi}_l^{(o)}}(S_{i,t+1}, a') - \\
Q^{\hat{\pi}_l^{(o)}}(S_{i,t}, A_{i,t}))/\sigma_{o+1}(\hat{\pi}_l^{(o)}) \Big| = o_p(1).
\end{aligned}$$

Next, we aim to show that

$$\frac{O}{nT(O-1)} \Big| (\sum_{o=1}^{O-1} \sum_{(i,t)\in J_{o+1}} \{w^{\hat{\pi}_l^{(o)},\nu}(S_{i,t}, A_{i,t})(R_{i,t} \tag{5}$$

$$+\gamma \sum_{a'\in\mathcal{A}} \hat{\pi}_l^{(o)}(a'|S_{i,t+1}) Q^{\hat{\pi}_l^{(o)}}(S_{i,t+1}, a') - Q^{\hat{\pi}_l^{(o)}}(S_{i,t}, A_{i,t}))\}^2/\sigma_{o+1}^2(\hat{\pi}_l^{(o)})) - 1 \Big| = o_p(1).$$

Notice that the left hand side of the above is bounded above by

$$\frac{O}{nT} \max_{1\le o\le (O-1)} \Big| (\sum_{(i,t)\in J_{o+1}} \{w^{\hat{\pi}_l^{(o)},\nu}(S_{i,t}, A_{i,t})(R_{i,t} \tag{6}$$

$$+\gamma \sum_{a'\in\mathcal{A}} \hat{\pi}_l^{(o)}(a'|S_{i,t+1}) Q^{\hat{\pi}_l^{(o)}}(S_{i,t+1}, a') - Q^{\hat{\pi}_l^{(o)}}(S_{i,t}, A_{i,t}))\}^2/\sigma_{o+1}^2(\hat{\pi}_l^{(o)})) - 1 \Big|. \tag{7}$$

Because, for each $1 \le o \le (O-1)$,

$$\frac{O}{nT}\Bigg\{(\sum_{(i,t)\in J_{o+1}}\{w^{\hat{\pi}_l^{(o)},\nu}(S_{i,t},A_{i,t})(R_{i,t} \tag{8}$$

$$+\gamma\sum_{a'\in\mathcal{A}}\hat{\pi}_l^{(o)}(a'|S_{i,t+1})Q^{\hat{\pi}_l^{(o)}}(S_{i,t+1},a')-Q^{\hat{\pi}_l^{(o)}}(S_{i,t},A_{i,t}))\}^2-\mathbb{E}[\{w^{\hat{\pi}_l^{(o)},\nu}(S,A)(R \tag{9}$$

$$+\gamma\sum_{a'\in\mathcal{A}}\hat{\pi}_l^{(o)}(a'|S')Q^{\hat{\pi}_l^{(o)}}(S',a')-Q^{\hat{\pi}_l^{(o)}}(S,A))\}]/\sigma^2_{o+1}(\hat{\pi}_l^{(o)}))\Bigg\}, \tag{10}$$

forms a mean zero martingale, we apply Freedman's inequality in [Freedman, 1975] with Assumptions to show it is bounded by $O_p(\sqrt{\frac{O}{nT}})$. Applying union bound shows (5) is $o_p(1)$ and furthermore consistency of $\hat{\sigma}(\hat{\pi}_l)$ in (2) holds. Then we apply the martingale central limit theorem to show

$$\sqrt{\frac{nT}{O(O-1)}}\left(\sum_{o=1}^{O-1}\frac{\phi(J_{o+1},Q^{\hat{\pi}_l^{(o)}},w^{\hat{\pi}_l^{(o)}},\hat{\pi}_l^{(o)})}{\sigma_{o+1}(\hat{\pi}_l^{(o)})}\right)\Longrightarrow\mathcal{N}(0,1).$$

The remaining is to show

$$\sqrt{\frac{nT}{O(O-1)}}\left(\sum_{o=1}^{O-1}\frac{\mathcal{V}(\hat{\pi}_l^{(o)})-\mathcal{V}(\hat{\pi}_l)}{\hat{\sigma}_{o+1}(\hat{\pi}_l^{(o)})}\right)$$

is asymptotically negligible. Consider

$$\mathbb{E}\left|\mathcal{V}(\hat{\pi}_l^{(o)})-\mathcal{V}(\hat{\pi}_l)\right| \tag{11}$$

$$\le\mathbb{E}\left|\mathcal{V}(\hat{\pi}_l^{(o)})-\mathcal{V}(\pi_l^*)\right|+\mathbb{E}\left|\mathcal{V}(\hat{\pi}_l)-\mathcal{V}(\pi_l^*)\right| \tag{12}$$

$$\le\mathbb{E}\left|\mathcal{V}(\hat{\pi}_l^{(o)})-\mathcal{V}(\pi_l^*)\right|+\mathbb{E}\left|\mathcal{V}(\hat{\pi}_l)-\mathcal{V}(\pi_l^*)\right| \tag{13}$$

$$\le(nTo)^{-\kappa}O^{\kappa}+(nT)^{-\kappa}, \tag{14}$$

where we use Assumption for the last inequality. Summarizing together, we can show that

$$\begin{aligned}
&\sqrt{\frac{nT}{O(O-1)}}\mathbb{E}\left|\sum_{o=1}^{O-1}\mathcal{V}(\hat{\pi}_l^{(o)})-\mathcal{V}(\hat{\pi}_l)\right|\\
\le&\sqrt{\frac{nT}{O(O-1)}}\sum_{o=1}^{O-1}(nTo)^{-\kappa}O^{\kappa}+\sqrt{\frac{nT(O-1)}{O}}(nT)^{-\kappa}\\
\le&\sqrt{\frac{nTO^2}{O(O-1)}}\sum_{o=1}^{O-1}(nT)^{-\kappa}+\sqrt{\frac{nT(O-1)}{O}}(nT)^{-\kappa}\\
=&o(1),
\end{aligned}$$

where we obtain the second inequality by that $\sum_{o=1}^{O-1}o^{-\kappa}\le 1+\int_1^O o^{-\kappa}do\lesssim O^{1-\kappa}$. In the last inequality, we use $\kappa>1$ in Assumption. Then Markov inequality gives that

$$\sqrt{\frac{nT}{O(O-1)}}\left(\sum_{o=1}^{O-1}\mathcal{V}(\hat{\pi}_l^{(o)})-\mathcal{V}(\hat{\pi}_l)\right)=o_p(1).$$

Moreover, by Assumption that $\inf_{1\le o\le O-1}\hat{\sigma}_{o+1}(\hat{\pi}_l^{(o)})\ge c$ for some constant $c>0$, we can further show that

$$\sqrt{\frac{nT}{O(O-1)}}\left(\sum_{o=1}^{O-1}\frac{\mathcal{V}(\hat{\pi}_l^{(o)})-\mathcal{V}(\hat{\pi}_l)}{\hat{\sigma}_{o+1}(\hat{\pi}_l^{(o)})}\right)=o_p(1),$$

which completes our proof.

Proof of Corollary Denote the sets $E_l = \{|\mathcal{V}(\hat{\pi}_l) - \hat{\mathcal{V}}(\hat{\pi}_l)| \leq \hat{u}(l)\}$, $l = 1, \ldots, L$, where $\hat{u}(l) = z_{\alpha/2}\sqrt{nT(O-1)/O}\hat{\sigma}(l)$. Note that $\liminf_{nT\to\infty} \Pr(\cap_{j=1}^L E_j) \geq 1 - L\alpha$ and

$$\begin{aligned}
&\Pr(\mathcal{V}(\hat{\pi}_{\hat{l}}) \geq \max_{1\leq l\leq L} \mathcal{V}(\hat{\pi}_l) - 2\hat{u}(l)) \\
=&\Pr(\mathcal{V}(\hat{\pi}_{\hat{l}}) - \hat{\mathcal{V}}(\hat{\pi}_{\hat{l}}) + \hat{\mathcal{V}}(\hat{\pi}_{\hat{l}}) \geq \max_{1\leq l\leq L} \mathcal{V}(\hat{\pi}_l) - \hat{\mathcal{V}}(\hat{\pi}_l) - 2\hat{u}(l) + \hat{\mathcal{V}}(\hat{\pi}_l)) \\
\geq&\Pr(\mathcal{V}(\hat{\pi}_{\hat{l}}) - \hat{\mathcal{V}}(\hat{\pi}_{\hat{l}}) + \hat{\mathcal{V}}(\hat{\pi}_{\hat{l}}) \geq \max_{1\leq l\leq L} \mathcal{V}(\hat{\pi}_l) - \hat{\mathcal{V}}(\hat{\pi}_l) - 2\hat{u}(l) + \hat{\mathcal{V}}(\hat{\pi}_l) | \cap_{j=1}^L E_j) \Pr(\cap_{j=1}^L E_j) \\
\geq&\Pr(\hat{\mathcal{V}}(\hat{\pi}_{\hat{l}}) - \hat{u}(\hat{l}) \geq \max_{1\leq l\leq L} \hat{\mathcal{V}}(\hat{\pi}_l) - \hat{u}(l) | \cap_{j=1}^L E_j) \Pr(\cap_{j=1}^L E_j) \\
=&\Pr(\cap_{j=1}^L E_j),
\end{aligned}$$

where the last inequality holds because given the event $\cap_{j=1}^L E_j$, one has $-\hat{u}(\hat{l}) \leq \mathcal{V}(\hat{\pi}_{\hat{l}}) - \hat{\mathcal{V}}(\hat{\pi}_{\hat{l}})$ and $\mathcal{V}(\hat{\pi}_l) - \hat{\mathcal{V}}(\hat{\pi}_l) \leq \hat{u}(l)$ for any l. This completes the proof by taking $\liminf$ on both sides.

Proof of Theorem on Bias To show the results in Theorem, it can be seen that

$$\begin{aligned}
\left|\frac{\sqrt{nT(O-1)/O}\left(\hat{\mathcal{V}}(\hat{\pi}_l) - \mathcal{V}(\pi^*)\right)}{\hat{\sigma}(l)}\right| \leq& \left|\sqrt{\frac{nT}{O(O-1)}}\left(\sum_{o=1}^{O-1}\frac{\hat{\mathcal{V}}_{\mathcal{D}_{o+1}}(\hat{\pi}_l^{(o)}) - \mathcal{V}(\hat{\pi}_l^{(o)})}{\hat{\sigma}_{o+1}(\hat{\pi}_l^{(o)})}\right)\right| \\
&+ \sqrt{\frac{nT}{O(O-1)}}\left(\sum_{o=1}^{O-1}\frac{\mathcal{V}(\hat{\pi}_l^{(o)}) - \mathcal{V}(\pi^*)}{\hat{\sigma}_{o+1}(\hat{\pi}_l^{(o)})}\right) \\
\leq& \underbrace{\left|\sqrt{\frac{nT}{O(O-1)}}\left(\sum_{o=1}^{O-1}\frac{\hat{\mathcal{V}}_{\mathcal{D}_{o+1}}(\hat{\pi}_l^{(o)}) - \mathcal{V}(\hat{\pi}_l^{(o)})}{\hat{\sigma}_{o+1}(\hat{\pi}_l^{(o)})}\right)\right|}_{(I)} \\
&+ B(l)\sqrt{\frac{nT}{O(O-1)}}\left(\sum_{o=1}^{O-1}\frac{1}{\hat{\sigma}_{o+1}(\hat{\pi}_l^{(o)})}\right).
\end{aligned}$$

Then by results in the proof of Theorem, we can show that

$$\lim_{nT\to\infty} \Pr((I) > z_{\alpha/2}) = \alpha. \tag{15}$$

This implies that

$$\liminf_{nT\to\infty} \Pr\left(|\mathcal{V}(\pi^*) - \hat{\mathcal{V}}(\hat{\pi}_l)| \leq z_{\alpha/2}\sqrt{O/nT(O-1)}\hat{\sigma}(l) + B(l)\right) \tag{16}$$

$$\geq \lim_{nT\to\infty} \Pr((I) \leq z_{\alpha/2}) = 1 - \alpha, \tag{17}$$

which concludes our proof.

Proof of Corollary: We mainly show the proof of the second claim in the corollary, based on which the first claim can be readily seen. Define an event E such that $1 \leq l \leq L$, $|\mathcal{V}(\hat{\pi}_l) - \hat{\mathcal{V}}(\hat{\pi}_l)| \leq c(\delta)\log(L)\hat{\sigma}(i)/\sqrt{NT}$ and $|\mathcal{V}(\pi^*) - \hat{\mathcal{V}}(\hat{\pi}_l)| \leq z_{\alpha/(2L)}\sqrt{O/nT(O-1)}\hat{\sigma}(l) + B(l)$. Based on the assumption given in Corollary and Theorem , we have $\liminf_{nT\to\infty} P(E) \geq 1 - \delta - \alpha$. In the following, we suppose event E holds.

Inspired by the proofs of Corollary 1 in [Mathé, 2006] and Theorem 3 of [Su et al., 2020], we define $\tilde{l} = \max\{l : B(l) \leq u_1(l) + u_2(l)\}$, where $u_1(l) = z_{\alpha/(2L)}\sqrt{O/nT(O-1)}\hat{\sigma}(l)$. Let $u_2(l) = c(\delta)\log(L)\hat{\sigma}(i)/\sqrt{NT}$. By Assumption , for $l \leq \tilde{l}$,

$$B(l) \leq B(\tilde{l}) \leq u_1(\tilde{l}) \leq u_1(l),$$

which further implies that for any $l \leq \tilde{l}$,

$$|\hat{\mathcal{V}}(\hat{\pi}_l) - \mathcal{V}(\pi^*)| \leq B(l) + u_1(l) \leq 2u_1(l).$$

Then $\mathcal{V}(\pi^*) \in I(l)$ based on the construction of $I(l)$ for all $l \leq \tilde{l}$. In addition, we have for $l \leq \tilde{l}$

$$|\mathcal{V}(\hat{\pi}_l) - \mathcal{V}(\pi^*)| \leq 2u_1(l) + u_2(l), \tag{18}$$

by triangle inequality and event E. Since $I(l)$ share at least one common element for $1 \leq l \leq \tilde{l}$, we have $\hat{i} \geq \tilde{l}$. Moreover, there must exist an element x such that $x \in I(\tilde{l}) \cap I(\hat{i})$, where $|\hat{\mathcal{V}}(\hat{\pi}_{\tilde{l}}) - x| \leq u_1(\tilde{l})$ and $|\hat{\mathcal{V}}(\hat{\pi}_{\hat{i}}) - x| \leq u_1(\hat{i})$. This indicates that

$$
\begin{aligned}
|\hat{\mathcal{V}}(\hat{\pi}_{\hat{i}} - \mathcal{V}(\pi^*)| &\leq |\hat{\mathcal{V}}(\hat{\pi}_{\hat{i}}) - x| + |\hat{\mathcal{V}}(\hat{\pi}_{\tilde{l}}) - x| + |\hat{\mathcal{V}}(\hat{\pi}_{\tilde{l}}) - \mathcal{V}(\pi^*)| && (19)\\
&\leq u_1(\hat{i}) + 2u_1(\tilde{l}) \leq 3u_1(\tilde{l}), && (20)
\end{aligned}
$$

by again triangle inequality and Assumption , and

$$
|\mathcal{V}(\hat{\pi}_{\hat{i}}) - \mathcal{V}(\pi^*)| \leq u_2(\hat{i}) + 3u_1(\tilde{l}) \leq u_2(\tilde{l}) + 3u_1(\tilde{l}), \tag{21}
$$

by event E and Assumption . Define $l^* = \min\{l : B(l) + u_1(l) + u_2(l)\}$. Then following the similar proof of [Su et al., 2020], we consider two cases:

Case 1: If $l^* \leq \tilde{l}$, then we have

$$
u_2(\tilde{l}) + B(\tilde{l}) + u_1(\tilde{l}) \leq 2u_1(l^*) + u_2(l^*) \leq 2u_1(l^*) + 2B(l^*) + u_2(l^*),
$$

where we use Assumption .

Case 2: If $l^* > \tilde{l}$, then we have

$$
\zeta(u_2(\tilde{l}) + u_1(\tilde{l})) \leq (u_2(\tilde{l} + 1) + u_1(\tilde{l} + 1)) \leq B(\tilde{l} + 1) \leq B(l^*),
$$

where we use Assumption . This implies that

$$
u_2(\tilde{l}) + u_1(\tilde{l}) + B(\tilde{l}) \leq (1 + 1/\zeta)B(l^*).
$$

Combining two cases, we can show that

$$
u_2(\tilde{l}) + u_1(\tilde{l}) + B(\tilde{l}) \leq (1 + 1/\zeta)(B(l^*) + u_1(l^*) + u_2(l^*)),
$$

as $\zeta < 1$. Together with (19), we have

$$
|\mathcal{V}(\hat{\pi}_{\hat{i}}) - \mathcal{V}(\pi^*)| \leq u_2(\hat{i}) + 3u_1(\tilde{l}) \leq 3(1 + 1/\zeta)(B(l^*) + u_1(l^*) + u_2(l^*)), \tag{22}
$$

which concludes our proof.

C MORE DETAILS ON DQN ENVIRONMENTS

We introduce our deployed DQN environments in this section, which included four environments with discrete action ($\mathbf{E}_1$ to $\mathbf{E}_4$) and two environments ($\mathbf{E}_5$ to $\mathbf{E}_6$) with continuous action. These environments cover wide applications, including tabular learning ($\mathbf{E}_1$), navigation to a target object in a geometrical space ($\mathbf{E}_2$), digital gaming ($\mathbf{E}_3$ to $\mathbf{E}_4$), and continuous control ($\mathbf{E}_5$ to $\mathbf{E}_6$).

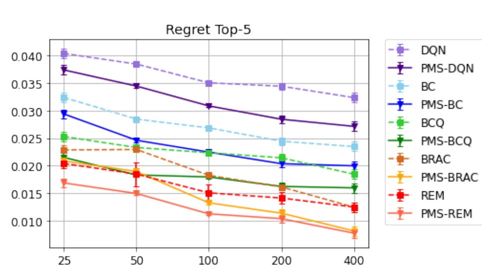

Figure 1: Policy selection using top-k ranking regret score in $\mathbf{E}_1$ (Frozen Lake).

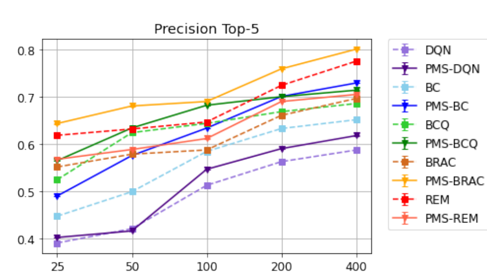

Figure 2: Policy selection using top-k ranking precision in $\mathbf{E}_1$ (Frozen Lake).

E_1: Frozen Lake: The Frozen Lake is a maze environment that manipulates an agent to walk from a starting point (S) to a goal point without failing into the hole (H). We use *FrozenLake-v0* from OpenAI Gym [Brockman et al., 2016]. We provide top-5 regret and precision results shown in Figure and 2.

E_2: Banana Collector: The Banana collector is one popular 3D-graphical navigation environment that compresses discrete actions and states as an open source DQN benchmark from Unity [1] ML-Agents v0.3.[Juliani et al., 2018]. The DRL agent controls an automatic vehicle with 37 dimensions of state observations including velocity and a ray-based perceptional information from objects around the agent. The targeted reward is 12.0 points by accessing correct yellow bananas $(+1)$ and avoiding purple bananas (-1) in first-person point of view as shown in Fig(b). We provide the related top-5 regret and precision results shown in Figure 3 and 4.

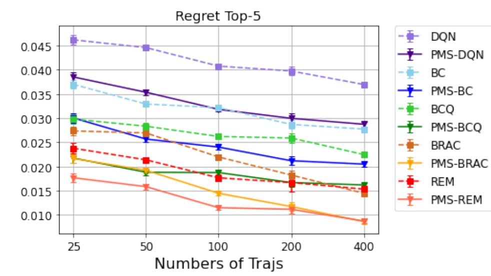

Figure 3: Policy selection using top-k ranking regret score in $\mathbf{E}_2$ (Banana Collector).

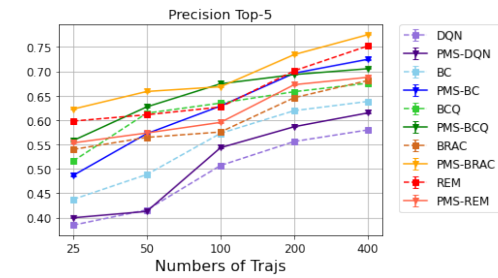

Figure 4: Policy selection using top-k ranking precision in $\mathbf{E}_2$ (Banana Collector).

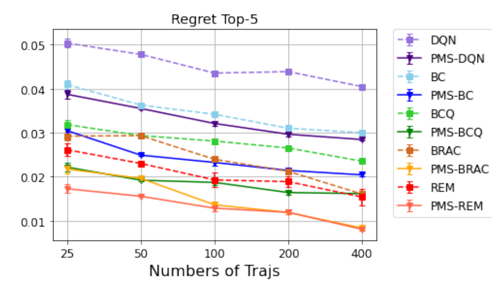

Figure 5: Policy selection using top-k ranking regret score in $\mathbf{E}_3$ (Pong).

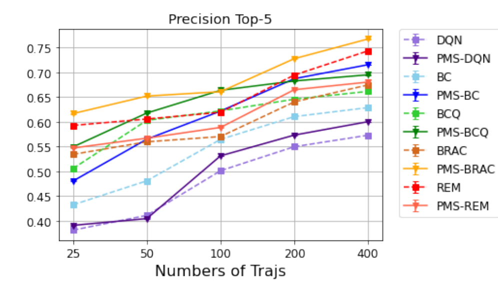

Figure 6: Policy selection using top-k ranking precision in $\mathbf{E}_3$ (Pong).

E_3: Pong: Pong is one Atari game environment from OpenAI Gym [Brockman et al., 2016] as shown in (c). We provide its top-5 regret and precision results shown in Figure 5 and 6.

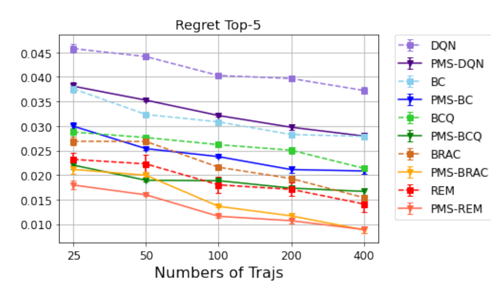

Figure 7: Policy selection using top-k ranking regret score in $\mathbf{E}_4$ (Breakout).

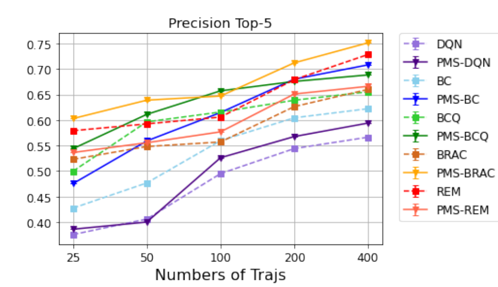

Figure 8: Policy selection using top-k ranking precision in $\mathbf{E}_4$ (HalfCheetah-v1).

[1] `https://www.youtube.com/watch?v=heVMs3t9qSk`

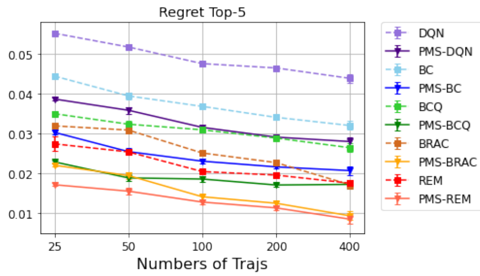

Figure 9: Policy selection using top-k ranking regret score in $\mathbf{E}_5$ (HalfCheetah-v1).

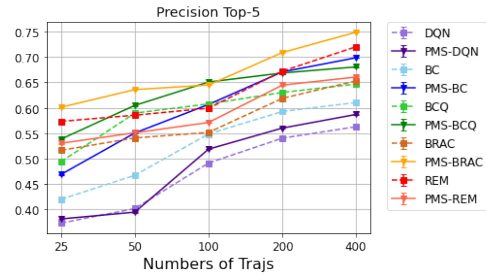

Figure 10: Policy selection using top-k ranking precision in $\mathbf{E}_5$ (HalfCheetah-v1).

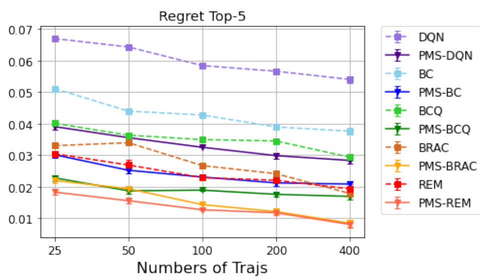

Figure 11: Policy selection using top-k ranking regret score in $\mathbf{E}_6$ (Walker2d-v1).

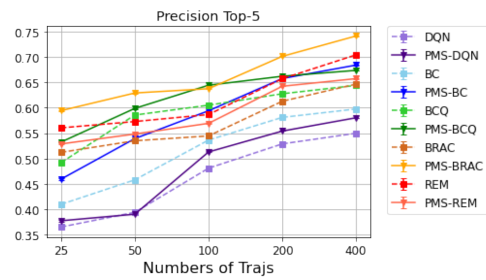

Figure 12: Policy selection using top-k ranking precision in $\mathbf{E}_6$ (Walker2d-v1).

$\mathbf{E}_4$: Breakout: Breakout is one Atari game environment from OpenAI Gym [Brockman et al., 2016] as shown in Fig 7(d). We provide the related top-5 regret and precision results shown in Figure 7 and 8.

$\mathbf{E}_5$: HalfCheetah-v1: Halfcheetah is a continuous action and state environment to control agent with monuments made by MuJoCo simulators as shown in (e). We provide the related top-5 regret and precision results shown in Figure 9 and 10.

$\mathbf{E}_6$: Walker2d-v1: Walker2d-v1 is a continuous action and state environment to control agent with monuments made by MuJoCo simulators as shown in (f). We provide the related top-5 regret and precision results shown in Figure 11 and 12.

D HYPER-PARAMETERS INFORMATION

We select a total of 70 DQN based models for each environment. We will open source the model and implementation for future studies. Table 1, Table 2, and Table 3 summarize their hyper-parameter and setups. In addition, Figure 13 and Figure 14 provide ablation studies on different scales of α and O selection in PMS experiments for the deployed DRL navigation task ($\mathbf{E}_2$). From the experimental results, a more pessimistic α (e.g., 0.001) is associated with a slightly better attained top-5 regret. Meanwhile, the selection of O does not produce much different performance on selected policies but slightly affects the range of the selected policies.

Table 1: Hyper-parameters information for for DQN models used in $\mathbf{E}_1$ to $\mathbf{E}_2$

Hyper-parameters	Values
Hidden layers	$\{1,\ 2\}$
Hidden units	$\{16,\ 32,\ 64,\ 128\}$
Learning rate	$\{1 \times \mathrm{e}^{-3},\ 5 \times \mathrm{e}^{-4}\}$
DQN training iterations	$\{100,\ 500,\ 1k,\ 2k\}$
Batch size	$\{64\}$

Table 2: Hyper-parameters information for for DQN models used in $\mathbf{E}_3$ to $\mathbf{E}_4$

Hyper-parameters	Values
Convolutional layers	$\{2, 3\}$
Convolutional units	$\{16,\ 32\}$
Hidden layers	$\{2, 3\}$
Hidden units	$\{64,\ 256,\ 512\}$
Learning rate	$\{1 \times \mathrm{e}^{-3},\ 5 \times \mathrm{e}^{-4}\}$
DQN training iterations	$\{4M,\ 4.5M,\ 5M\}$
Batch size	$\{64\}$

Table 3: Hyper-parameters information for double DQN (DDQN) models [Van Hasselt et al., 2016] with a prioritized replay [Schaul et al., 2015] used in $\mathbf{E}_5$ to $\mathbf{E}_6$.

Hyper-parameters	Values
Hidden layers	$\{4,\ 5,\ 6\}$
Hidden units	$\{64,\ 128,\ 256,\ 512\}$
Learning rate	$\{1 \times \mathrm{e}^{-3},\ 5 \times \mathrm{e}^{-4}\}$
DDQN training frames	$\{40M,\ 45M,\ 50M\}$
Batch size	$\{256\}$
Buffer size	$\{10^6\}$
Updated target	$\{1000\}$

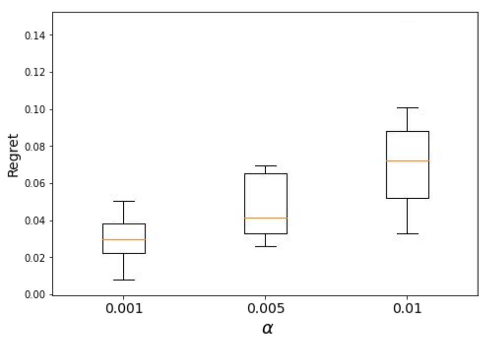

Figure 13: Different α for PMS selection.

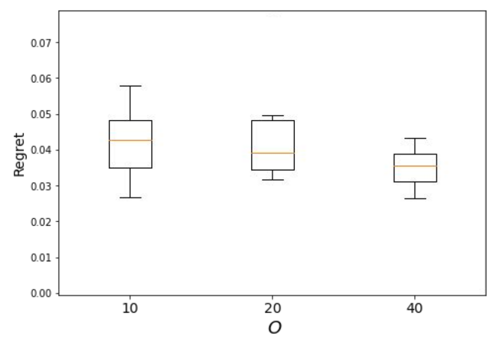

Figure 14: Different O for PMS selection.

E BROADER IMPACT

There are also some limitations of the proposed PMS as one of the preliminary attempts on model selection for offline reinforcement learning. When the benchmarks environments (excluded Atari games) are based on simulated environments to collect the true policy [Barth-Maron et al., 2018, Siegel et al., 2019], more real-world-based environments could be customized and studied in future works. For example, one experimental setup needs to be carefully controlled in clinical settings [Tang and Wiens, 2021] or resilience-oriented [Yang et al., 2021] reinforcement learning.

References

Gabriel Barth-Maron, Matthew W Hoffman, David Budden, Will Dabney, Dan Horgan, TB Dhruva, Alistair Muldal, Nicolas Heess, and Timothy Lillicrap. Distributed distributional deterministic policy gradients. In *International Conference on Learning Representations*, 2018.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

Arthur Juliani, Vincent-Pierre Berges, Ervin Teng, Andrew Cohen, Jonathan Harper, Chris Elion, Chris Goy, Yuan Gao, Hunter Henry, Marwan Mattar, et al. Unity: A general platform for intelligent agents. *arXiv preprint arXiv:1809.02627*, 2018.

Nathan Kallus and Masatoshi Uehara. Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes. *arXiv preprint arXiv:1909.05850*, 2019.

Peter Mathé. The lepskii principle revisited. *Inverse problems*, 22(3):L11, 2006.

Donald L McLeish. Dependent central limit theorems and invariance principles. *the Annals of Probability*, 2(4):620–628, 1974.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

Noah Siegel, Jost Tobias Springenberg, Felix Berkenkamp, Abbas Abdolmaleki, Michael Neunert, Thomas Lampe, Roland Hafner, Nicolas Heess, and Martin Riedmiller. Keep doing what worked: Behavior modelling priors for offline reinforcement learning. In *International Conference on Learning Representations*, 2019.

Yi Su, Pavithra Srinath, and Akshay Krishnamurthy. Adaptive estimator selection for off-policy evaluation. In *International Conference on Machine Learning*, pages 9196–9205. PMLR, 2020.

Shengpu Tang and Jenna Wiens. Model selection for offline reinforcement learning: Practical considerations for healthcare settings. *arXiv preprint arXiv:2107.11003*, 2021.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Chao-Han Huck Yang, I Hung, Te Danny, Yi Ouyang, and Pin-Yu Chen. Causal inference q-network: Toward resilient reinforcement learning. *arXiv preprint arXiv:2102.09677*, 2021.