PESSIMISTIC MODEL SELECTION FOR OFFLINE DEEP REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep Reinforcement Learning (DRL) has demonstrated great potentials in solving sequential decision making problems in many applications. Despite its promising performance, practical gaps exist when deploying DRL in real-world scenarios. One main barrier is the over-fitting issue that leads to poor generalizability of the policy learned by DRL. In particular, for offline DRL with observational data, model selection is a challenging task as there is no ground truth available for performance demonstration, in contrast with the online setting with simulated environments. In this work, we propose a pessimistic model selection (PMS) approach for offline DRL with a theoretical guarantee, which features a tuning-free framework for finding the best policy among a set of candidate models. Two refined approaches are also proposed to address the potential bias of DRL model in identifying the optimal policy. Numerical studies demonstrated the superior performance of our approach over existing methods.

1 INTRODUCTION

The success of deep reinforcement learning (Mnih et al., 2013; Henderson et al., 2018) (DRL) often leverages upon executive training data with considerable efforts to select effective neural architectures. Deploying online simulation to learn useful representations for DRL is not always realistic and possible especially in some high-stake environments, such as automatic navigation (Kahn et al., 2018; Hase et al., 2020), dialogue learning (Jaques et al., 2020), and clinical applications (Tang et al., 2020a). *Offline reinforcement learning* (Singh & Sutton, 1996; Levine et al., 2020; Agarwal et al., 2020) (ORL) has prompted freshened interests (Paine et al., 2020; Kidambi et al., 2020) to empower DRL toward training tasks associated with severe potential cost and risks. The idea of ORL is to train DRL models with only logged data and recorded trajectories. However, with given observational data, designing a successful neural architecture in ORL is often expensive (Levine et al., 2020), requiring intensive experiments, time, and computing resources. Meanwhile, automated machine learning (Hutter et al., 2019) (AutoML) shows many advantages to accelerate the development cycle for given data and desired tasks. AutoML has been used to provide simpler solutions, faster creation of those solutions, and specific models and hyperparameters that perform equally or better than hand-designed architectures from domain experts. AutoML demonstrates many competitive solutions, such as neural architectures search (Zoph & Le, 2016), hyperparameter optimization (Bergstra et al., 2011; Maclaurin et al., 2015), and model selection (Farahmand & Szepesvári, 2011; Fard & Pineau, 2010; Jiang et al., 2015; Xie & Jiang, 2021), for images classification, language modeling, and RL respectively. Unlike most aforementioned applications with online interaction, *Offline* tasks for reinforcement learning often suffer the challenges from insufficient observational data from offline collection to construct a universal approximated model for fully capturing the temporal dynamics. Therefore, relatively few attempts in the literature have been presented for provide a pipeline to automate the development process for model selection and neural architecture search in ORL settings.

In this work, we propose a novel model selection approach to automate ORL development process, which provides an evaluation mechanism to identify a well-performed DRL model given offline data. Our method utilizes statistical inference to provide uncertainty quantification on value functions trained by different DRL models, based on which a pessimistic idea is incorporated to select the best model/policy. In addition, two refined approaches are further proposed to address the possible biases of DRL models in identifying the optimal policy. In this work, we mainly focus on deep Q-network (Mnih et al., 2013; 2015) (DQN) based architectures for ORL applications, while our proposed methods can be extended to other settings. Figure 1 demonstrates the superior performance of the proposed pessimistic model selection (PMS) method in identifying the best model among 70 DRL models of different algorithms on one navigation task (See Appendix B for details), compared with the model selection method by (Tang &

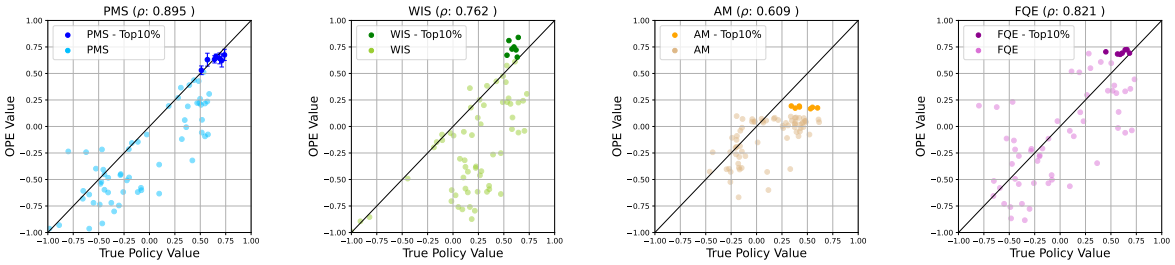


Figure 1: Model selection algorithms for offline DQN learning: (a) proposed pessimistic model selection (PMS); (b) weighted importance sampling (WIS) (Gottesman et al., 2018); (c) approximate model (AM) (Voloshin et al., 2019); (d) fitted Q evaluation (FQE) (Le et al., 2019). In this figure, the algorithms are trained and evaluated in a navigation task ($\mathbf{E}_2$) discussed in Section 7.

Wiens, 2021) which uses three OPE estimates for validation. Specifically, in this application, based on the derived confidence interval of the OPE value for each candidate model, the final selected model by our PMS method is the one that has the largest lower confidence limit, which exactly has the largest true OPE value among all candidate models. In contrast, none of three OPE estimates used for model selection by Tang & Wiens (2021) can identify the best model due to the inevitable overfitting issue during the validation procedure.

To close this section, we summarize the contributions of this work as follows:

- We propose a novel PMS framework, which targets finding the best policy from given candidate models (e.g., neural architecture, hyperparameters, etc) with offline data for DQN learning. Unlike many existing methods, our approach essentially does not involve additional hyperparameter tuning.
- Leveraging asymptotic analysis in statistical inference, we provide uncertainty quantification on each candidate model, based on which our method can guarantee that the worst performance of finally selected model is the best among all candidate models. See Corollary 1 for more details.
- Furthermore, to address potential biases of candidate models in identifying the optimal policy, two refined approaches are also proposed, one of which can be shown to have regret bounded by the smallest error bound among all candidate models. See Corollary 2. To the best of our knowledge, this is the first model-selection method in offline DRL with such guarantee.
- The numerical results demonstrate that the proposed PMS shows superior performance in different DQN benchmark environments.

2 Related Work

Model Selection for Reinforcement Learning: Model selection has been studied in online decision-making environments (Fard & Pineau, 2010; Lee & Taylor, 2014). Searching nearly optimal online model is a critical topic for online bandits problems with limited information feed-backs. For linear contextual bandits, Abbasi-Yadkori et al. (2011); Chu et al. (2011) are aiming to find the best worst-case bound when the optimal model class is given. For model-based reinforcement learning, Pacchiano et al. (2020) introduces advantages of using noise augmented Markov Decision Processes (MDP) to archive a competitive regret bound to select an individual model with constraints for ensemble training. Recently, Lee et al. (2021) utilized an online algorithm to select a low-complexity model based on a statistical test. However, most of the previous model selection approaches are focused on the online reinforcement learning setting. Very few works including Farahmand & Szepesvári (2011); Paine et al. (2020); Su et al. (2020); Tang & Wiens (2021); Xie & Jiang (2021) are focused on the offline setting. In this work, we provide one of the first model selection approaches based on statistical inference for reinforcement learning tasks with offline data collection.

Offline-Policy Learning: Training a DRL agent with offline data collection often relies on batch-wise optimization. Batch-Constrained deep Q-learning (Fujimoto et al., 2019) (BCQ) is considered one ORL benchmark that uses a generative model to minimize the distance of selected actions to the batch-wise data with a perturbation model to maximize its value function. Other popular ORL approaches, such as behavior regularized actor-critic (BRAC) (Wu et al., 2019), and random ensemble mixture (Agarwal et al., 2020) (REM) (as an optimistic perspective on large dataset), have also been studied in RL Unplugged (RLU) (Gulcehre et al., 2020) benchmark together with behavior cloning (Bain & Sammut, 1995; Ross & Bagnell, 2010) (BC), DQN, and DQN with quantile regression (Dabney

et al., 2018) (QR-DQN). RLU suggests a naive approach based on human experience for offline policy selection, which requires independent modification with shared domain expertise (e.g., Atari environments) for tuning each baseline. Meanwhile, how to design a model selection algorithm for ORL remains an open question. Motivated by the benefits and the challenges as mentioned earlier of the model selection for *offline* DRL, we aim to develop a tuning-free approach for model selection in offline DRL with theoretical guarantee.

3 Background and Notations

Consider a time-homogeneous Markov decision process (MDP) characterized by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, p is the transition kernel, i.e., $p(s'|s,a)$ is the probability mass (density) of transiting to s' given current state-action (s,a), r is the reward function, i.e., $\mathbb{E}(R_t|S_t = s, A_t = a) = r(s,a)$ for $t \geq 0$, and $0 \leq \gamma < 1$ is a discount factor. For simplicity, we assume $\mathcal{A}$ is finite. But our method can also be applied in continuous case. Under this MDP setting, it is sufficient to consider stationary Markovian policies for optimizing discounted sum of rewards (Puterman, 1994). Denote π as a stationary Markovian policy mapping from the state space $\mathcal{S}$ into a probability distribution over the action space. For example, $\pi(a|s)$ denotes the probability of choosing action a given the state value s. One essential goal of RL is to learn an optimal policy that maximizes the value function. Define $V^\pi(s) = \sum_{t=0}^{+\infty} \gamma^t \mathbb{E}^\pi[R_t|S_0 = s]$ and then the optimal policy is defined as

$$\pi^* \in \text{argmax}_\pi \{\mathcal{V}(\pi) \triangleq (1-\gamma) \sum_{s \in \mathcal{S}} V^\pi(s)\nu(s)\}, \tag{1}$$

where ν denotes some reference distribution function over $\mathcal{S}$. In addition, we denote Q-function as $Q^\pi(s,a) = \sum_{t=0}^{+\infty} \gamma^t \mathbb{E}^\pi(R_t|A_0 = a, S_0 = s)$ for $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

In this work, we consider the offline RL setting. The observed data consist of N trajectories, corresponding to N independent and identically distributed copies of $\{(S_t, A_t, R_t)\}_{t \geq 0}$. For any $i \in \{1, \cdots, n\}$, data collected from the ith trajectory can be summarized by $\{(S_{i,t}, A_{i,t}, R_{i,t}, S_{i,t+1})\}_{0 \leq t < T}$, where T denotes the termination time. We assume that the data are generated by some fixed stationary policy denoted by b.

Among many RL algorithms, we focus on Q-learning type of methods. The foundation is the optimal Bellman equation given below.

$$Q^*(s,a) = \mathbb{E}\left[R_t + \gamma \max_{a' \in \mathcal{A}} Q^*(S_{t+1}, a') \mid S_t = s, A_t = a\right], \tag{2}$$

where Q^* is called optimal Q-function, i.e., Q-function under π^*. Among others, fitted q-iteration (FQI) is one of the most popular RL algorithms (Ernst et al., 2005). FQI leverages supervised learning techniques to iteratively solve the optimal Bellman equation (2) and shows competitive performance in ORL.

To facilitate our model-selection algorithm, we introduce the discounted visitation probability, motivated by the marginal importance sampling estimator in (Liu et al., 2018). For any $t \geq 0$, let $p_t^\pi(s,a)$ denote the t-step visitation probability $\Pr^\pi(S_t = s, A_t = a)$ assuming the actions are selected according to π at time $1, \cdots, t$. We define the discounted visitation probability function as $d^\pi(s,a) = (1-\gamma)\sum_{t \geq 0} \gamma^t p_t^\pi(s,a)$. To adjust the distribution from behavior policy to any target policy π, we use the discounted probability ratio function defined as

$$\omega^{\pi,\nu}(s,a) = \frac{d^{\pi,\nu}(s)\pi(a|s)}{\frac{1}{T}\sum_{t=0}^{T-1} p_t^b(s,a)}, \tag{3}$$

where $p_t^b(s,a)$ is the t-step visitation probability under the behavior policy b, i.e., $\Pr^b(S_t = s, A_t = a)$. The ratio function $\omega^{\pi,\nu}(s,a)$ is always assumed well defined. The estimation of ratio function is motivated by the observation that for every measurable function f defined over $\mathcal{S} \times \mathcal{A}$,

$$\mathbb{E}[\frac{1}{T}\sum_{t=0}^{T-1} \omega^{\pi,\nu}(S_t, A_t)(f(S_t, A_t) - \gamma \sum_{a' \in \mathcal{A}} \pi(a' \mid S_{t+1}) f(S_{t+1}, a'))] = (1-\gamma)\mathbb{E}_{S_0 \sim \nu}[\sum_{a \in \mathcal{A}} \pi(a \mid S_0) f(a, S_0)], \tag{4}$$

based on which several min-max estimation methods has been proposed such as (Liu et al., 2018; Nachum et al., 2019; Uehara & Jiang, 2019).

Finally, because our proposed model selection algorithm relies on an efficient evaluation of any target policy using batch data, we introduce three types of offline policy evaluation estimators in the existing RL literature. The first type is called direct method via estimating Q-function, based on the relationship that $\mathcal{V}(\pi) =$

$(1-\gamma)\sum_{s\in\mathcal{S},a\in\mathcal{A}}\pi(a|s)Q(s,a)\nu(s)$. The second type is motivated by the importance sampling (Precup, 2000). Based on the definition of ratio function, we can see $\mathcal{V}(\pi) = \mathbb{E}[\frac{1}{T}\sum_{t=0}^{T-1}\omega^{\pi,\nu}(S_t,A_t)R_t]$, from which a plugin estimator can be constructed. The last type of OPE methods combines the first two types of methods and construct a so-called doubly robust estimator (Kallus & Uehara, 2019; Tang et al., 2020b). These estimators are mainly motivated by *efficient influence function* (EIF) in semiparametric statistics. Specifically, the EIF of $\mathcal{V}(\pi)$ is given by (Kallus & Uehara, 2019),

$$\frac{1}{T}\sum_{t=0}^{T-1}\omega^{\pi,\nu}(S_t,A_t)(R_t+\gamma\sum_{a\in\mathcal{A}}\pi(a|S_{t+1})Q^{\pi}(S_{t+1},a)-Q^{\pi}(S_t,A_t))+(1-\gamma)\mathbb{E}_{S_0\sim\nu}[\sum_{a\in\mathcal{A}}\pi(a|S_0)Q^{\pi}(S_0,a)]-\mathcal{V}(\pi). \tag{5}$$

A nice property of doubly robust estimators is that as long as either the Q-function or the ratio function can be estimated consistently, the final estimator is consistent (Robins et al., 1994; Jiang & Li, 2015; Kallus & Uehara, 2019; Tang et al., 2020b). Furthermore, doubly robust estimators can achieve semiparametric efficiency under general conditions, even if nuisance parameters are estimated via black box models such as deep neural networks. Our proposed algorithm will rely on this doubly robust type of OPE estimator.

4 Pessimistic Model Selection (PMS) for Best Policy

In this section, we discuss our pessimistic model selection approach. For the ease of presentation, we focus on the framework of (deep) Q-learning, where policy optimization is performed via estimating the optimal Q-function. While this covers a wide range of state-of-the-art RL algorithms such as FQI (Ernst et al., 2005), DQN (Mnih et al., 2013) and QR-DQN (Dabney et al., 2018), we remark that our method is not restricted to this class of algorithms.

Suppose we have total number of L candidate models for policy optimization, where each candidate model will output an estimated policy, say $\hat{\pi}_l$ for $1 \le l \le L$. Our goal is to select the best policy among L policies during our training procedure. Note that these L models can be different deep neural network architectures, hyper-parameters, and various classes of functions for approximating the optimal Q-function or policy class, etc.

4.1 Difficulties and Challenges

Given a candidate model l among L models, we can apply for example FQI using the whole batch data $\mathcal{D}_n$ to learn an estimate of Q^* as $\widehat{Q}_l$ and an estimated optimal policy $\hat{\pi}_l$ defined as $\hat{\pi}_l(a|s) \in \text{argmax}_{a\in\mathcal{A}}\widehat{Q}_l(s,a)$, for every $s \in \mathcal{S}$. In order to select the final policy, one may use a naive greedy approach to choose some $\tilde{l}$ such that $\tilde{l} \in \text{argmax}_l \mathbb{E}_{S_0\sim\nu}[\sum_{a\in\mathcal{A}}\hat{\pi}_l(a|s)\widehat{Q}_l(S_0,a)]$, as our goal is to maximize $\mathcal{V}(\pi)$. However, using this criterion will lead to some over-fitting issue. Specifically, due to the distributional mismatch between the behavior policy and target policies, which is regarded as a fundamental challenge in offline RL (Levine et al., 2020), we may easily overestimate Q-function, especially when some state-action pairs are not sufficiently visited in the batch data. This issue becomes more serious when we apply max-operator during our policy optimization procedure. Such observations have already been noticed in the literature. See recent work such as (Kumar et al., 2019; 2020; Paine et al., 2020; Yu et al., 2020; Tang & Wiens, 2021; Jin et al., 2021) for more details. Therefore it may not be appropriate to use this criterion for selecting the best policy among L models.

One may also use cross-validation procedure to address the issue of over-fitting or overestimating Q-function for model selection. For example, one can use OPE approaches on the validate dataset to evaluate the performance of estimated policies from the training data set (see Tang & Wiens (2021) for more details). However, since there is no ground truth for the value function of any policies, the OPE procedure on the validation dataset cannot avoid involving additional tuning on hyperparameters. Therefore such procedure may still incur a large variability due to the over-fitting issue. In addition, arbitrarily splitting the dataset for cross-validation and ignoring the Markov dependent structure will cause additional errors, which should be seriously taken care of.

4.2 Sequential Model Selection

In the following, we propose a pessimistic model selection algorithm for finding an optimal policy among L candidate models. Our goal is to develop an approach to estimate the value function under each candidate model during

our policy optimization procedure with theoretical guarantee. The proposed algorithm is motivated by recent development in statistical inference of sequential decision making (Luedtke & Van Der Laan, 2016; Shi et al., 2020). The idea is to first estimate optimal Q-function Q^*, optimal policy π^* and the resulting ratio function based on a chunk of data, and evaluate the performance of the estimated policy on the next chunk of data using previously estimated nuisance functions. Then we combine the first two chunks of data, perform the same estimation procedure and evaluation on the next chunk of data. The framework of MDP provides a nature way of splitting the data.

Specifically, denote the index of our batch dataset $\mathcal{D}_n$ as $J_0 = \{(i,t) : 1 \le i \le n, 0 \le t < T\}$. We divide J_0 into O number of non-overlapping subsets, denoted by $J_1, \cdots, J_O$ and the corresponding data subsets are denoted by $\mathcal{D}_1, \cdots, \mathcal{D}_O$. Without loss of generality, we assume these data subsets have equal size. We require that for any $1 \le o_1 < o_2 \le O$, any $(i_1, t_1) \in J_{o_1}$ and $(i_2, t_2) \in J_{o_2}$, either $i_2 \ne i_1$ or $t_1 < t_2$. For $1 \le o \le O$, denote the aggregate chunks of data as $\bar{\mathcal{D}}_o = \left\{(S_{i,t}, A_{i,t}, R_{i,t}, S_{i,t+1}), (i,t) \in \bar{J}_o = J_1 \cup \cdots \cup J_o\right\}$.

We focus on FQI algorithm for illustrative purpose and it should be noticed that our algorithm can be applied to other RL algorithms. Starting from the first chunk of our batch data, at the o-th step ($o = 1, \cdots, O-1$), for each candidate model $l = 1, \cdots, L$, we apply FQI on $\bar{\mathcal{D}}_o$ to compute $\widehat{Q}_l^{(o)}$ as an estimate of optimal Q-function and obtain $\hat{\pi}_l^{(o)}$ correspondingly. Additionally, we compute an estimate of ratio function $\omega^{\hat{\pi}_l^{(o)},\nu}$ using $\bar{\mathcal{D}}_o$ by many existing algorithms such as Nachum et al. (2019). Denote the resulting estimator as $\widehat{\omega}^{\hat{\pi}_l^{(o)},\nu}$. The purpose of estimating this ratio function is to improve the efficiency and robustness of our value function estimation for each candidate model. Then we compute the estimated value function of $\hat{\pi}_l^{(o)}$ on $\mathcal{D}_{o+1}$ as

$$\hat{\mathcal{V}}_{\mathcal{D}_{o+1}}(\hat{\pi}_l^{(o)}) = (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[\sum_{a_0\in\mathcal{A}} \hat{\pi}_l^{(o)}(a_0|S_0)\widehat{Q}_l^{(o)}(S_0,a_0)\right] \tag{6}$$

$$+\mathbb{E}_{\mathcal{D}_{o+1}}\left[\widehat{\omega}^{\hat{\pi}_l^{(o)},\nu}(S,A)\left(R+\gamma\sum_{a'\in\mathcal{A}}\hat{\pi}_l^{(o)}(a'|S')\widehat{Q}_l^{(o)}(S',a') - \widehat{Q}_l^{(o)}(S,A)\right)\right], \tag{7}$$

where $\mathbb{E}_{\mathcal{D}_{o+1}}$ denotes the empirical average over the $(o+1)$ chunk of dataset and (S, A, R, S') is one transition tuple in $\mathcal{D}_{o+1}$. While one can aggregate $\hat{\mathcal{V}}_{\mathcal{D}_{o+1}}(\hat{\pi}_l^{(o)})$ for $1 \le o \le (O-1)$ to evaluate the performance of L models, the uncertainty of these estimates due to the finite sample estimation should not be ignored. Therefore, in the following, we derive an uncertainty quantification of our estimated value function for each candidate model, for performing model selection. Based on the EIF of the value function, (conditioning on $\bar{\mathcal{D}}_o$), the variance of $\hat{\mathcal{V}}_{\mathcal{D}_{o+1}}(\hat{\pi}_l^{(o)})$ is

$$\sigma^2(\hat{\pi}_l^{(o)}) = \mathbb{E}\left[\left\{\widehat{\omega}^{\hat{\pi}_l^{(o)},\nu}(S,A)\left(R+\gamma\sum_{a'\in\mathcal{A}}\hat{\pi}_l^{(o)}(a'|S')\widehat{Q}_l^{*(o)}(S',a') - \widehat{Q}_l^{*(o)}(S,A)\right)\right\}^2\right], \tag{8}$$

where (S, A, S') is a transition tuple with (S, A) follows some stationary distribution. See Assumption 1. Correspondingly we have an estimate defined as

$$\hat{\sigma}_{o+1}^2(\hat{\pi}_l^{(o)}) = \mathbb{E}_{\mathcal{D}_{o+1}}\left[\left\{\widehat{\omega}^{\hat{\pi}_l^{(o)},\nu}(S,A)\left(R+\gamma\sum_{a'\in\mathcal{A}}\hat{\pi}_l^{(o)}(a'|S')\widehat{Q}_l^{*(o)}(S',a') - \widehat{Q}_l^{*(o)}(S,A)\right)\right\}^2\right]. \tag{9}$$

The estimation procedure stops once we have used all our offline data and denote the final estimated policy as $\hat{\pi}_l$ for each $l = 1, \cdots, L$. Notice that $\hat{\pi}_l = \hat{\pi}_l^{(O)}$. Finally, we compute the weighted average of all the intermediate value functions as our final evaluation of the estimated policy $\hat{\pi}_l$, i.e.,

$$\hat{\mathcal{V}}(\hat{\pi}_l) = \left(\sum_{o=1}^{O-1}\frac{1}{\hat{\sigma}_{o+1}(\hat{\pi}_l^{(o)})}\right)^{-1}\left(\sum_{o=1}^{O-1}\frac{\hat{\mathcal{V}}_{\mathcal{D}_{o+1}}(\hat{\pi}_l^{(o)})}{\hat{\sigma}_{o+1}(\hat{\pi}_l^{(o)})}\right). \tag{10}$$

In Section 5, we show that under some technical conditions, the following asymptotic result holds, i.e.,

$$\frac{\sqrt{nT(O-1)/O}\left(\hat{\mathcal{V}}(\hat{\pi}_l) - \mathcal{V}(\hat{\pi}_l)\right)}{\hat{\sigma}(l)} \Longrightarrow \mathcal{N}(0,1), \tag{11}$$

where $\hat{\sigma}(l) = (O-1)\left(\sum_{o=1}^{O-1}\{\sigma_{o+1}(\hat{\pi}_l^{(o)})\}^{-1}\right)^{-1}$, $\Longrightarrow$ refers to weak convergence when either n or T goes to infinity, and $\mathcal{N}(0,1)$ refers to the standard normal distribution. Based on the asymptotic result in (11), we can construct a confidence interval for the value function of each policy $\hat{\pi}_l$. Given a confidence level α, for each l, we can compute $U(l) = \hat{\mathcal{V}}(\hat{\pi}_l) - z_{\alpha/2}\sqrt{O/nT(O-1)}\hat{\sigma}(l)$, where $z_{\alpha/2}$ is $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution. Our final selected one is $\hat{l} \in \text{argmax}_{1\le l\le L}\, U(l)$.

The use of $U(l)$ is motivated by the recent proposed pessimistic idea to address the overestimation issue of value (or Q) function in the offline RL setting. See Kumar et al. (2019; 2020); Jin et al. (2021); Xie et al. (2021); Uehara & Sun (2021); Zanette et al. (2021) for more details. The final output of our algorithm is $\hat{\pi}_{\hat{l}}$ and an outline of the proposed algorithm can be found in Algorithm 1. As we can see, our algorithm is essentially tuning free, which provides great flexibility in real-world applications. The only two adjustable parameters is O and α. The size of O balances the computational cost and the accuracy of evaluating each candidate model. In specific, large O will not only increase the accuracy of our procedure but also result in a large computational cost. As a sacrifice for the tuning free algorithm, we need to apply ORL algorithms O times for each candidate model. The parameter α determines how worst the performance of each policy we should use to evaluate each policy. See Corollary 1 for more insights.

Algorithm 1: Pessimistic Model Selection (PMS) for offline RL

Input: Dataset $\mathcal{D}_n$ and L candidate models for estimating optimal Q-function and policy; We divide $\mathcal{D}_n$ into non-overlapping subsets denoted by $\mathcal{D}_1, \cdots, \mathcal{D}_O$. We require that for any $1 \le o_1 < o_2 \le O$, any $(i_1, t_1) \in J_{o_1}$ and $(i_2, t_2) \in J_{o_2}$, either $i_2 \ne i_1$ or $t_1 \le t_2$.

for $l \in L$ **do**
  for $o = 1$ ***to*** $O-1$ **do**
    For $l \in L$ model (or algorithm), construct the optimal $\widehat{Q}_l^{(o)}$ and $\hat{\pi}_l^{(o)}$ using $\bar{\mathcal{D}}_o$ data subset.
    Compute $\widehat{\omega}^{\hat{\pi}_l^{(o)},\nu}$ using $\bar{\mathcal{D}}_o$ by Nachum et al. (2019), which solves a min-max procedure related to (4).
    Compute $\hat{\mathcal{V}}_{\mathcal{D}_{o+1}}(\hat{\pi}_l^{(o)})$ and $\hat{\sigma}^2_{o+1}(l)$ using $\mathcal{D}_{o+1}$ given in (6) and (9) respectively.
  end
  For l-th model, we compute $U(l) = \hat{\mathcal{V}}(\hat{\pi}_l) - z_{\alpha/2}\sqrt{nT(O-1)/O}\hat{\sigma}(l)$, where $\hat{\mathcal{V}}(\hat{\pi}_l)$ and $\hat{\sigma}(l)$ are given in (10) and (11) respectively.
end
Pick $\hat{l} = \arg\max_l U(l)$ as our selected models
Return $\hat{\pi}_{\hat{l}}$.

5 Theoretical Results

In this section, we justify our asymptotic result given in (11). We use O_p to denote the stochastic boundedness. Before that, we make several technical assumptions:

Assumption 1 *The stochastic process $\{A_t, S_t\}_{t\ge 0}$ is stationary with stationary distribution p_∞.*
Assumption 2 *For every $1 \le l \le L$ and $1 \le o \le O$, we have $\mathbb{E}|\mathcal{V}(\hat{\pi}_l^{(o)}) - \mathcal{V}(\pi^*)| \le C_0(nT/O)^{-\alpha}$, for some constant C_0 and $\alpha > 1/2$.*

Assumption 1 is standard in the existing literature such as (Kallus & Uehara, 2019). Assumption 2 is key to our developed asymptotic results developed. This assumption essentially states that all candidate models are good enough so that eventually their value functions will converge to that of the optimal one. This implies that there is no asymptotic bias in identifying the optimal policy. While this is reasonable thanks to the capability of deep neutral networks, which has demonstrated its empirical success in many RL applications, such an assumption could still be strong. In Section 6, we relax this assumption and provide two remedies for addressing possibly biased estimated policies. In addition, Assumption 1 also requires that the convergence rates of value functions under estimated policies are fast enough. This has been shown to hold under the margin condition on π^*, see e.g., (Hu et al., 2021) for more details.

Assumption 3 *For every $1 \le l \le L$ and $1 \le o \le O-1$, suppose $\mathbb{E}_{(A,S)\sim p_\infty}|\widehat{Q}_l^{(o)} - Q^{\hat{\pi}_l^{(o)}}(S,A)|^2 = O_p\{(nT/O)^{-2\kappa_1}\}$ for some constant $\kappa_1 \ge 0$. In addition, $\widehat{Q}_l^{(o)}$ is uniformly bounded almost surely.*

Assumption 4 *For every* $1 \leq l \leq L$ *and* $1 \leq o \leq O-1$*, suppose* $\mathbb{E}_{(A,S)\sim p_\infty}|\widehat{\omega}^{\hat{\pi}_l^{(o)},\nu}(S,A) - \omega^{\hat{\pi}_l^{(o)},\nu}(S,A)|^2 = O_p\{(nT/O)^{-2\kappa_2}\}$ *for some constant* $\kappa_2 \geq 0$*. In addition, both* $\omega^{\hat{\pi}_l^{(o)},\nu}$ *and* $\widehat{\omega}^{\hat{\pi}_l^{(o)},\nu}$ *are uniformly bounded above and below away from* 0 *almost surely.*

Assumption 5 *For every* $1 \leq l \leq L$ *and* $1 \leq o \leq O-1$*,* $\sigma^2(\hat{\pi}_l^{(o)})$ *and* $\hat{\sigma}^2_{o+1}(\hat{\pi}_l^{(o)})$ *are bounded above and below from* 0 *almost surely.*

Assumptions 3 and 4 impose high-level conditions on two nuisance functions. Our theoretical results only require $\kappa_1 + \kappa_2 > 1/2$, which is a mild assumption. For example, if considered parametric models for both Q-function and ratio function, then $\kappa_1 = \kappa_2 = 1/2$. If considered nonparametric models for these two nuisance functions such as deep neural networks, then $1/4 < \kappa_1, \kappa_2 < 1/2$ can be obtained under some regularity conditions. See Fan et al. (2020) and Liao et al. (2020); Uehara et al. (2021) for the convergence rates of Q-function and ratio function by nonparametric models respectively. In addition, Assumption 5 is a mild assumption, mainly for theoretical justification. Then we have the following main theorem as a foundation of our proposed algorithm.

Theorem 1 *Under Assumptions 1-5, we have*

$$\frac{\sqrt{nT(O-1)/O}\left(\hat{\mathcal{V}}(\hat{\pi}_l) - \mathcal{V}(\hat{\pi}_l)\right)}{\hat{\sigma}(l)} \Longrightarrow \mathcal{N}(0,1). \tag{12}$$

Theorem 1 provides an uncertainty quantification of each candidate model used in policy optimization. Such uncertainty quantification is essential in offline RL as data are often limited. A consequent result following Theorem 1 validates the proposed Algorithm 1:

Corollary 1 *Under Assumptions 1-5,* $\liminf_{nT\to\infty} \Pr(\mathcal{V}(\hat{\pi}_{\hat{l}}) \geq \max_{1\leq l\leq L} \mathcal{V}(\hat{\pi}_l) - 2z_{\alpha/2}\sqrt{nT(O-1)/O}\hat{\sigma}(l)) \geq 1 - L\alpha$.

As can be seen clearly from Corollary 1 and the proposed PMS, with a large probability (by letting α small), we consider the worst performance of each candidate model $\hat{\pi}_l$ in the sense of the lower confidence limit of the value function, and then select the best one among all models.

6 Two Refined Approaches

In this section, we relax Assumption 2 by allowing possibly non-negligible bias in estimating the optimal policy and introduce two refined approaches for addressing this issue. Instead of imposing Assumption 2, we make an alternative assumption below.

Assumption 6 *For every* $1 \leq l \leq L$*, there exists* $B(l)$ *such that* $\max_{1\leq o\leq (O-1)} |\mathcal{V}(\hat{\pi}_l^{(o)}) - \mathcal{V}(\pi^*)| \leq B(l)$ *almost surely.*

Assumption 6 is a very mild assumption. It essentially states that the biases for all our intermediate value function estimates are bounded by some constant, which is much weaker than Assumption 2. In this case, the asymptotic results in (12) may not hold in general. Correspondingly, we have the following result.

Theorem 2 *Under Assumptions 1, 3-6, for every* $1 \leq l \leq L$*, the following inequality holds.*

$$\liminf_{nT\to\infty} \Pr\left(|\mathcal{V}(\pi^*) - \hat{\mathcal{V}}(\hat{\pi}_l)| \leq z_{\alpha/2}\sqrt{O/nT(O-1)}\hat{\sigma}(l) + B(l)\right) \geq 1-\alpha. \tag{13}$$

Motivated by Lepski's principle (Lepski & Spokoiny, 1997) from nonparametric statistics and (Su et al., 2020) studying the model selection of OPE, we consider the following refined model-selection procedure to find the best policy. We first rank L candidate models in an non-increasing order based on the value of $\hat{\sigma}(l)$, i.e., for $1 \leq i < j \leq L$, $\hat{\sigma}(i) \geq \hat{\sigma}(j)$. Then for i-th model, we construct an interval as

$$I(i) = \left[\hat{\mathcal{V}}(\hat{\pi}_l) - 2z_{\alpha/(2L)}\sqrt{O/nT(O-1)}\hat{\sigma}(l), \hat{\mathcal{V}}(\hat{\pi}_l) + 2z_{\alpha/(2L)}\sqrt{O/nT(O-1)}\hat{\sigma}(l)\right].$$

Finally the optimal model/policy we choose is $\hat{\pi}_{\hat{i}}$ such that $\hat{i} = \max\{i : 1 \leq i \leq L, \cap_{1\leq j\leq i} I(j) \neq \emptyset\}$. To show this procedure is valid, we need to make one additional assumption.

Assumption 7 *There exists a $\zeta < 1$ such that for $1 \leq i \leq L$, $B(i) \leq B(i+1)$ and $\zeta\hat{\sigma}(i) \leq \hat{\sigma}(i+1) \leq \hat{\sigma}(i)$ almost surely.*

This assumption is borrowed from Su et al. (2020). It typically assumes that after sorting our model based on $\hat{\sigma}(l)$, the bias of estimated policy is monotonically increasing and the standard deviation is monotonically deceasing but not too quickly. This is commonly seen when all candidate estimators exhibit some bias-variance trade-off phenomena. Define the following event

$$\mathcal{E} = |\hat{\mathcal{V}}(\hat{\pi}_{\hat{i}}) - \mathcal{V}(\pi^*)| \leq 6(1+\zeta^{-1}) \min_{1\leq i\leq L} \{B(i) + z_{\alpha/(2L)}\sqrt{O/nT(O-1)}\hat{\sigma}(i)\}. \tag{14}$$

Then we have the following theoretical guarantee for our refined procedure.

Corollary 2 *Under Assumptions 1, 3-7, we have $\liminf_{nT\to\infty} \Pr(\mathcal{E}) \geq 1-\alpha$. If we further assume that for any $\delta > 0$, with probability at least $1-\delta$, for every $1 \leq i \leq L$, $|\mathcal{V}(\hat{\pi}_i) - \hat{\mathcal{V}}(\hat{\pi}_i)| \leq c(\delta)\log(L)\hat{\sigma}(i)/\sqrt{NT}$ for some constant $c(\delta)$, then $\liminf_{nT\to\infty} \Pr(\bar{\mathcal{E}}) \geq 1-\alpha-\delta$, where*

$$\bar{\mathcal{E}} = |\mathcal{V}(\hat{\pi}_{\hat{i}}) - \mathcal{V}(\pi^*)| \leq 3(1+\zeta^{-1}) \min_{1\leq i\leq L} \{B(i) + (c(\delta)\log(L) + z_{\alpha/(2L)})\sqrt{O/nT(O-1)}\hat{\sigma}(i)\}. \tag{15}$$

The additional assumption (i.e., the high probability bound) in Corollary 2 can be shown to hold by the empirical process theory under some technical conditions (Van de Geer, 2000). Hence Corollary 2 provides a strong guarantee that the regret of the final selected policy is bounded by the smallest error bound among all L candidate policies.

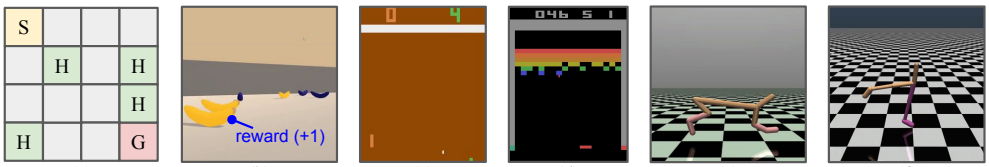
Another refined approach: Notice that the above refined approach indeed focuses on OPE estimates to select the best policy with regret warranty. The rough motivation behind is to find a policy that has the smallest estimation error to the optimal one. However, such procedure may not directly match the goal of maximizing the value function in ORL. To relieve this issue , we can alternatively choose the final policy as $\hat{\pi}_{\hat{\hat{i}}}$ such that

$$\hat{\hat{i}} = \text{argmax}_{1\leq i\leq \hat{i}} \hat{\mathcal{V}}(\hat{\pi}_i) - 2z_{\alpha/2}\sqrt{nT(O-1)/O}\hat{\sigma}(i),$$

where the argmax is taken over $\hat{i}$ models. This approach can be viewed as a combination of PMS and the above refined approach. By adopting this approach, candidate models with large biases are removed by the truncation on $\hat{i}$ in the first step. In the second step, we use the idea of PMS to select the best model having the best worst performance among remaining candidate models. Unfortunately, we do not have theoretical guarantee for this combined approach.

7 Experimental Results

We select six DQN environments ($\mathbf{E}_1$ to $\mathbf{E}_6$) from open-source benchmarks (Brockman et al., 2016; Juliani et al., 2018) to conduct numerical experiments, as shown in Fig. 2. These tasks of deployed environments cover different domains that include tabular learning (Fig 2(a)); automatic navigation in a geometry environment with a physical ray-tracker (Fig 2(b)); Atari digital gaming (Fig 2(c) and (d)), and continuous control (Fig 2(e) and (f)). We provide detailed task description and targeted reward for each environment in Appendix B.



(a) (b) (c) (d) (e) (f)

Figure 2: DQN environments deployed in our studies: (a) $\mathbf{E}_1$: *FrozenLake-v0*; (b) $\mathbf{E}_2$: *Banana Collectors* from (Juliani et al., 2018) as a 3D geometrical navigation task; (c) $\mathbf{E}_3$: *Pong-v0*; (d) $\mathbf{E}_4$: *Breakout-v0*; (e) $\mathbf{E}_5$: *Halfcheetah-v1*; (f) $\mathbf{E}_6$: *Walker2d-v1*.

Experiment setups. To evaluate the performance of PMS with DQN models in the offline setting RL, we choose different neural network architectures under five competitive DRL algorithms including DQN by (Mnih et al., 2013; 2015), BCQ by (Fujimoto et al., 2019), BC by (Bain & Sammut, 1995; Ross & Bagnell, 2010), BRAC by (Wu et al., 2019) from RLU benchmarks, and REM by (Agarwal et al., 2020). Within each architecture, 70 candidate models

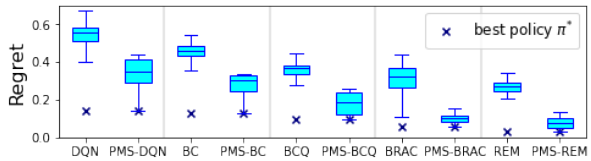


Figure 3: Box plots of model selection performance from offline Q-networks in each DRL algorithm.

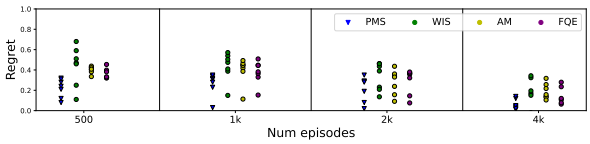


Figure 4: Sensitivity analysis for different training data size. PMS attains the best performance and has the least sensitivity.

are created by assigning different hyperparameters and training setups. See Appendix B.1 for details. We then conduct performance evaluation of different ORL model selection methods on these generated candidate models.

Evaluation procedure. We utilize validation scores from OPE for each model selection algorithm, which picks the best (or a good) policy from the candidate set based on its own criterion. Regret is used as the evaluation metric for the candidate set of L models. The regret for model l is defined as $\mathcal{V}(\pi_{l^*}) - \mathcal{V}(\pi_l)$, where $l^* = \arg\max_{l'=1\ldots L} \mathcal{V}(\pi_{l'})$ corresponds to the candidate policy with the best OPE validation performance. In our implementation, we treat π_{l^*} as π^*, the oracle but unknown best possible policy. A small regret is desirable after model selection.

Performance comparison. As highlighted in Fig. 1 in the introduction, we report estimated OPE values by different model selection approaches, i.e. PMS and three methods by (Tang & Wiens, 2021), versus the true OPE values. In this experiment, we consider 70 DQN models under the above mentioned five DRL algorithms, i.e., 14 models are considered for each architecture. We use fewer models for each DRL algorithm mainly for clear presentation. By using the confidence interval constructed by our PMS procedure, our method is able to correctly select the top models, while the other three methods fail. To further investigate the performance of PMS, we implement model selection among 70 models within each DRL algorithm separately. Fig. 3 shows the box plots of averaged regret over six environments after OPE per neural network architecture. Each subfigure contains results from one particular DRL algorithm with different hyperparameters or training setups. The left box plot refers to the regrets of all 70 models and the right one represents the regrets of top 10% models selected by the proposed PMS method. Note that the right box plot is a subset of the left one. The results show that our proposed PMS successfully helps to select models with the best policies and improve the average regret by a significant margin. In particular, PMS-REM-based models attain the lowest regrets, due to the benefit from its ensemble process. Detailed results for each environment is given in Appendix B.

Sensitivity analysis. Fig. 4 compares different selection algorithms with varying training data size. PMS outperforms the others across all scales, and larger number of episodes gives smaller variation and lower sensitivity.

PMS algorithm with refinements. We replicate our experiments on in the offline navigation task in $\mathbf{E}_2$ (*Banana Collector*) for 30 times and report regrets of top 10% models selected by PMS and two refinements in Fig. 5. As we can see, while the overall performances of the proposed three model selection methods are similar, two refined approaches have better regrets than PMS in terms of median, demonstrating their promising performances in identifying the best model. OPE results have been also evaluated also in DRL tasks with $\mathbf{E}_1$ and $\mathbf{E}_3$ to $\mathbf{E}_6$, where the refinement algorithms (PMS R1/R2) have only a relative $\pm$ 0.423 % performance compared to its original PMS setups.

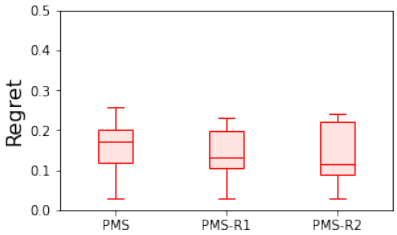


Figure 5: PMS and its refinements (R1/R2).

8 Conclusion

We propose a new theory-driven model selection framework (PMS) for offline deep reinforcement learning based on statistical inference. The proposed pessimistic mechanism is tuning-free and warrants that the worst performance of the selected model is the best among all candidate models. Two refined approaches are further proposed to address the biases of DRL models. Extensive experimental results on six DQN environments with various network architectures and training hyperparameters demonstrate that our proposed PMS method consistently yields improved model selection performance over existing baselines. The results suggest the effectiveness of PMS as a powerful tool for automating model selection in offline DRL. We will open source our implements for future studies.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pp. 104–114. PMLR, 2020.

Michael Bain and Claude Sammut. A framework for behavioural cloning. In *Machine Intelligence 15*, pp. 103–129, 1995.

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.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.

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

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.

Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Damien Ernst, Pierre Geurts, Louis Wehenkel, and L. Littman. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.

Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pp. 486–489. PMLR, 2020.

Amir-massoud Farahmand and Csaba Szepesvári. Model selection in reinforcement learning. *Machine learning*, 85 (3):299–332, 2011.

M Fard and Joelle Pineau. Pac-bayesian model selection for reinforcement learning. *Advances in Neural Information Processing Systems*, 23:1624–1632, 2010.

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

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pp. 2052–2062. PMLR, 2019.

Omer Gottesman, Fredrik Johansson, Joshua Meier, Jack Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xuefeng Peng, et al. Evaluating reinforcement learning algorithms in observational health settings. *arXiv preprint arXiv:1805.12298*, 2018.

Caglar Gulcehre, Ziyu Wang, Alexander Novikov, Tom Le Paine, Sergio Gomez Colmenarejo, Konrad Zolna, Rishabh Agarwal, Josh Merel, Daniel Mankowitz, Cosmin Paduraru, et al. Rl unplugged: Benchmarks for offline reinforcement learning. *arXiv e-prints*, pp. arXiv–2006, 2020.

Hannes Hase, Mohammad Farid Azampour, Maria Tirindelli, Magdalini Paschali, Walter Simson, Emad Fatemizadeh, and Nassir Navab. Ultrasound-guided robotic navigation with deep reinforcement learning. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5534–5541. IEEE, 2020.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Yichun Hu, Nathan Kallus, and Masatoshi Uehara. Fast rates for the regret of offline reinforcement learning. *arXiv preprint arXiv:2102.00479*, 2021.

Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. *Automated machine learning: methods, systems, challenges*. Springer Nature, 2019.

Natasha Jaques, Judy Hanwen Shen, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Human-centric dialog training via offline reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3985–4003, 2020.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. *arXiv preprint arXiv:1511.03722*, 2015.

Nan Jiang, Alex Kulesza, and Satinder Singh. Abstraction selection in model-based reinforcement learning. In *International Conference on Machine Learning*, pp. 179–188. PMLR, 2015.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021.

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.

Gregory Kahn, Adam Villaflor, Bosen Ding, Pieter Abbeel, and Sergey Levine. Self-supervised deep reinforcement learning with generalized computation graphs for robot navigation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5129–5136. IEEE, 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.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. In *NeurIPS*, 2020.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, pp. 11784–11794, 2019.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*, 2020.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pp. 3703–3712, 2019.

Jason D Lee and Jonathan E Taylor. Exact post model selection inference for marginal screening. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 1*, pp. 136–144, 2014.

Jonathan Lee, Aldo Pacchiano, Vidya Muthukumar, Weihao Kong, and Emma Brunskill. Online model selection for reinforcement learning with function approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3340–3348. PMLR, 2021.

Oleg V Lepski and Vladimir G Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, pp. 2512–2546, 1997.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Peng Liao, Zhengling Qi, and Susan Murphy. Batch policy learning in average reward markov decision processes. *arXiv preprint arXiv:2007.11771*, 2020.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pp. 5356–5366, 2018.

Alexander R Luedtke and Mark J Van Der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of statistics*, 44(2):713, 2016.

Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, pp. 2113–2122. PMLR, 2015.

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.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, pp. 2315–2325, 2019.

Aldo Pacchiano, Philip Ball, Jack Parker-Holder, Krzysztof Choromanski, and Stephen Roberts. On optimism in model-based reinforcement learning. *arXiv preprint arXiv:2006.11911*, 2020.

Tom Le Paine, Cosmin Paduraru, Andrea Michi, Caglar Gulcehre, Konrad Zolna, Alexander Novikov, Ziyu Wang, and Nando de Freitas. Hyperparameter selection for offline reinforcement learning. *arXiv preprint arXiv:2007.09055*, 2020.

Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, pp. 80, 2000.

Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. ISSN 01621459. URL `http://www.jstor.org/stable/2290910`.

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 661–668. JMLR Workshop and Conference Proceedings, 2010.

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

Chengchun Shi, Sheng Zhang, Wenbin Lu, and Rui Song. Statistical inference of the value function for reinforcement learning in infinite horizon settings. *arXiv preprint arXiv:2001.04515*, 2020.

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.

Satinder P Singh and Richard S Sutton. Reinforcement learning with replacing eligibility traces. *Machine learning*, 22(1):123–158, 1996.

Yi Su, Pavithra Srinath, and Akshay Krishnamurthy. Adaptive estimator selection for off-policy evaluation. In *International Conference on Machine Learning*, pp. 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.

Shengpu Tang, Aditya Modi, Michael Sjoding, and Jenna Wiens. Clinician-in-the-loop decision making: Reinforcement learning with near-optimal set-valued policies. In *International Conference on Machine Learning*, pp. 9387–9396. PMLR, 2020a.

Ziyang Tang, Yihao Feng, Lihong Li, Dengyong Zhou, and Qiang Liu. Doubly robust bias reduction in infinite horizon off-policy estimation. In *International Conference on Learning Representations*, 2020b.

Masatoshi Uehara and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. *arXiv preprint arXiv:1910.12809*, 2019.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline rl: Pac bounds and posterior sampling under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Masatoshi Uehara, Masaaki Imaizumi, Nan Jiang, Nathan Kallus, Wen Sun, and Tengyang Xie. Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*, 2021.

Sara Van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

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.

Cameron Voloshin, Hoang M Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. *arXiv preprint arXiv:1911.06854*, 2019.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pp. 11404–11413. PMLR, 2021.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *arXiv preprint arXiv:2106.06926*, 2021.

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.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*, 2020.

Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. *arXiv preprint arXiv:2108.08812*, 2021.

Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.

A 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 $\xi(N) \geq c_1\theta(N)$ (resp. $\xi(N) \leq c_2\theta(N)$) 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 1.

Lemma 1 *Under Assumptions 1 and 3-5, for every $1 \leq l \leq L$ and $1 \leq o \leq O-1$, the following asymptotic equivalence holds.*

$$\sqrt{\frac{nT}{O}} \left\{ \hat{\mathcal{V}}_{\mathcal{D}_{o+1}}(\hat{\pi}_l^{(o)}) - \mathcal{V}(\hat{\pi}_l^{(o)}) \right\} = \sqrt{\frac{nT}{O}} \phi(J, Q^{\hat{\pi}^{*(o)}}, \omega^{\hat{\pi}_l^{(o)},\nu}, \hat{\pi}_l^{(o)}) + o_p(1), \tag{16}$$

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 & 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

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

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

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 Assumptions 3 and 4, 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 1 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

$$\sqrt{\frac{nT}{O}}\frac{\hat{\mathcal{V}}_{\mathcal{D}_{o+1}}(\hat{\pi}_l^{(o)}) - \mathcal{V}(\hat{\pi}_l^{(o)})}{\hat{\sigma}_{o+1}(\hat{\pi}_l^{(o)})} = \sqrt{\frac{nT}{O}}\frac{\phi(J_{o+1}, Q^{\hat{\pi}_l^{(o)}}, w^{\hat{\pi}_l^{(o)}}, \hat{\pi}_l^{(o)})}{\hat{\sigma}_{o+1}(\hat{\pi}_l^{(o)})} + o_p(1). \tag{17}$$

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

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

For (I), we can see that

$$(I)=\sqrt{\frac{O}{nT(O-1)}}(\sum_{o=1}^{O-1}\sum_{(i,t)\in J_{o+1}}w^{\hat{\pi}_l^{(o)},\nu}(S_{i,t},A_{i,t})(R_{i,t} \tag{18}$$
$$+\gamma\sum_{a'\in\mathcal{A}}\hat{\pi}_l^{(o)}(a'|S_{i,t+1})Q^{\hat{\pi}_l^{(o)}}(S_{i,t+1},a')-Q^{\hat{\pi}_l^{(o)}}(S_{i,t},A_{i,t}))/\sigma_{o+1}(\hat{\pi}_l^{(o)})). \tag{19}$$

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

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

Next, we aim to show that

$$\frac{O}{nT(O-1)}\Bigg|(\sum_{o=1}^{O-1}\sum_{(i,t)\in J_{o+1}}\{w^{\hat{\pi}_l^{(o)},\nu}(S_{i,t},A_{i,t})(R_{i,t} \tag{20}$$
$$+\gamma\sum_{a'\in\mathcal{A}}\hat{\pi}_l^{(o)}(a'|S_{i,t+1})Q^{\hat{\pi}_l^{(o)}}(S_{i,t+1},a')-Q^{\hat{\pi}_l^{(o)}}(S_{i,t},A_{i,t}))\}^2/\sigma_{o+1}^2(\hat{\pi}_l^{(o)}))-1\Bigg|=o_p(1).$$

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

$$\frac{O}{nT}\max_{1\le o\le(O-1)}\Bigg|(\sum_{(i,t)\in J_{o+1}}\{w^{\hat{\pi}_l^{(o)},\nu}(S_{i,t},A_{i,t})(R_{i,t} \tag{21}$$
$$+\gamma\sum_{a'\in\mathcal{A}}\hat{\pi}_l^{(o)}(a'|S_{i,t+1})Q^{\hat{\pi}_l^{(o)}}(S_{i,t+1},a')-Q^{\hat{\pi}_l^{(o)}}(S_{i,t},A_{i,t}))\}^2/\sigma_{o+1}^2(\hat{\pi}_l^{(o)}))-1\Bigg|. \tag{22}$$

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

$$\frac{O}{nT}\Bigg\{(\sum_{(i,t)\in J_{o+1}}\{w^{\hat{\pi}_l^{(o)},\nu}(S_{i,t},A_{i,t})(R_{i,t} \tag{23}$$
$$+\gamma\sum_{a'\in\mathcal{A}}\hat{\pi}_l^{(o)}(a'|S_{i,t+1})Q^{\hat{\pi}_l^{(o)}}(S_{i,t+1},a')-Q^{\hat{\pi}_l^{(o)}}(S_{i,t},A_{i,t}))\}^2-\mathbb{E}[\{w^{\hat{\pi}_l^{(o)},\nu}(S,A)(R \tag{24}$$
$$+\gamma\sum_{a'\in\mathcal{A}}\hat{\pi}_l^{(o)}(a'|S')Q^{\hat{\pi}_l^{(o)}}(S',a')-Q^{\hat{\pi}_l^{(o)}}(S,A))\}]/\sigma_{o+1}^2(\hat{\pi}_l^{(o)}))\Bigg\}, \tag{25}$$

forms a mean 0 martingale, we apply Freedman's inequality in (Freedman, 1975) with Assumptions 3-5 to show it is bounded by $O_p(\sqrt{\frac{O}{nT}})$. Applying union bound shows (20) is $o_p(1)$ and furthermore consistency of $\hat{\sigma}(\hat{\pi}_l)$ in (17) holds. Then we apply 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

$$\begin{aligned}
&\mathbb{E}\left|\mathcal{V}(\hat{\pi}_l^{(o)}) - \mathcal{V}(\hat{\pi}_l)\right| && (26)\\
\leq&\mathbb{E}\left|\mathcal{V}(\hat{\pi}_l^{(o)}) - \mathcal{V}(\pi_l^*)\right| + \mathbb{E}\left|\mathcal{V}(\hat{\pi}_l) - \mathcal{V}(\pi_l^*)\right| && (27)\\
\leq&\mathbb{E}\left|\mathcal{V}(\hat{\pi}_l^{(o)}) - \mathcal{V}(\pi_l^*)\right| + \mathbb{E}\left|\mathcal{V}(\hat{\pi}_l) - \mathcal{V}(\pi_l^*)\right| && (28)\\
\leq&(nTo)^{-\alpha}O^{\alpha} + (nT)^{-\alpha}, && (29)
\end{aligned}$$

where we use Assumption 2 for the last inequality. Summarizing together, we can show that

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

where we obtain the second inequality by that $\sum_{o=1}^{O-1} o^{-\alpha} \leq 1 + \int_1^O o^{-\alpha}do \lesssim O^{1-\alpha}$. In the last inequality, we use $\alpha > 1$ in Assumption 2. 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 5 that $\inf_{1\leq o\leq O-1}\hat{\sigma}_{o+1}(\hat{\pi}_l^{(o)}) \geq c$ 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 1 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

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

which completes the proof by taking $\liminf$ on both sides.

Proof of Theorem 2 To show the results in Theorem 2, it can be seen that

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

Then by results in the proof of Theorem 1, we can show that

$$\lim_{nT\to\infty}\Pr((I) > z_{\alpha/2}) = \alpha. \tag{30}$$

This implies that

$$\liminf_{nT\to\infty}\Pr\left(|\mathcal{V}(\pi^*)-\hat{\mathcal{V}}(\hat{\pi}_l)| \leq z_{\alpha/2}\sqrt{O/nT(O-1)}\hat{\sigma}(l)+B(l)\right) \tag{31}$$

$$\geq \lim_{nT\to\infty}\Pr((I) \leq z_{\alpha/2}) = 1-\alpha, \tag{32}$$

which concludes our proof.

Proof of Corollary 2: 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 2 and Theorem 2, 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 7, 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 $l \leq \tilde{l}$

$$|\mathcal{V}(\hat{\pi}_l) - \mathcal{V}(\pi^*)| \leq 2u_1(l) + u_2(l), \tag{33}$$

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

$$|\hat{\mathcal{V}}(\hat{\pi}_{\hat{i}} - \mathcal{V}(\pi^*)| \leq |\hat{\mathcal{V}}(\hat{\pi}_{\hat{i}}) - x| + |\hat{\mathcal{V}}(\hat{\pi}_{\tilde{l}}) - x| + |\hat{\mathcal{V}}(\hat{\pi}_{\tilde{l}}) - \mathcal{V}(\pi^*)| \tag{34}$$

$$\leq u_1(\hat{i}) + 2u_1(\tilde{l}) \leq 3u_1(\tilde{l}), \tag{35}$$

by again triangle inequality and Assumption 7, and

$$|\mathcal{V}(\hat{\pi}_{\hat{i}}) - \mathcal{V}(\pi^*)| \leq u_2(\hat{i}) + 3u_1(\tilde{l}) \leq u_2(\tilde{l}) + 3u_1(\tilde{l}), \tag{36}$$

by event E and Assumption 7. 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 7.

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 7. 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 (36), we have

$$|\mathcal{V}(\hat{\pi}_{\hat{i}}) - \mathcal{V}(\pi^*)| \leq u_2(\hat{i}) + 3u_1(\tilde{l}) \leq 3(1 + 1/\zeta)(B(l^*) + u_1(l^*) + u_2(l^*)), \tag{37}$$

which concludes our proof.

B 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 covers wide applications includes 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 $\mathbf{E}_6$).

$\mathbf{E}_1$: 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) as shown in Fig 2(a).

$\mathbf{E}_2$: 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 2(b).

$\mathbf{E}_3$: Pong: Pong is one Atari game environment from OpenAI Gym (Brockman et al., 2016) as shown in Fig 2(c).

$\mathbf{E}_4$: Breakout: Breakout is one Atari game environment from OpenAI Gym (Brockman et al., 2016) as shown in Fig 2(d).

$\mathbf{E}_5$: HalfCheetah-v1: Halfcheetah is a continuous action and state environment to control agent with monuments made by MuJoCo simulators as shown in Fig 2(e).

$\mathbf{E}_6$: Walker2d-v1: Walker2d-v1 is a continuous action and state environment to control agent with monuments made by MuJoCo simulators as shown in Fig 2(f).

B.1 HYPER-PARAMETERS INFORMATION

We select a total of 70 DQN based models for each environments. We will open source the model and implementation for future studies. Table 1, Table 2, and Table 3 summarize their hyper-parameter and setups.

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

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 reply (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	$\{1000\}$

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 careful control in clinical settings (Tang & Wiens, 2021) or resilience-oriented (Yang et al., 2021) reinforcement learning.