Asymptotically Efficient Off-Policy Evaluation for Tabular Reinforcement Learning

Ming Yin
ming_yin@uscb.edu

Yu-Xiang Wang
yuxiangw@cs.ucsb.edu

Abstract

We consider the problem of off-policy evaluation for reinforcement learning, where the goal is to estimate the expected reward of a target policy π using offline data collected by running a logging policy μ. Standard importance-sampling based approaches for this problem suffer from a variance that scales exponentially with time horizon H, which motivates a splurge of recent interest in alternatives that break the "Curse of Horizon" (Liu et al., 2018a; Xie et al., 2019). In particular, it was shown that a marginalized importance sampling (MIS) approach can be used to achieve an estimation error of order $O(H^3/n)$ in mean square error (MSE) under an episodic Markov Decision Process model with finite states and potentially infinite actions. The MSE bound however is still a factor of H away from a Cramer-Rao lower bound of order $\Omega(H^2/n)$. In this paper, we prove that with a simple modification to the MIS estimator, we can asymptotically attain the Cramer-Rao lower bound, provided that the action space is finite. We also provide a general method for constructing MIS estimators with high-probability error bounds.

1 Introduction

Off-policy evaluation (OPE), which predicts the performance of a policy with data only sampled by a logging/behavior policy (Sutton and Barto, 2018), plays a key role for using reinforcement learning (RL) algorithms responsibly in many real-world decision-making problems such as marketing, finance, robotics, and healthcare. Deploying a policy without having an accurate evaluate of its performance could be costly, illegal, and can even break down the machine learning system. There is a large body of literature that studied the off-policy evaluation problem in both theoretical and application-oriented aspects. From the theoretical perspective, OPE problem is extensively studied in contextual bandits (Li et al., 2011; Dudík et al., 2011; Swaminathan et al., 2017; Wang et al., 2017) and reinforcement learning (RL) (Li et al., 2015; Jiang and Li, 2016; Thomas and Brunskill, 2016; Farajtabar et al., 2018; Xie et al., 2019) and the results of OPE studies have been applied to real-world applications including marketing (Theocharous et al., 2015; Thomas et al., 2017) and education (Mandel et al., 2014).

Problem setup. In the reinforcement learning (RL) problem the agent interacts with an underlying unknown dynamics which is modeled as a Markov decision process (MDP). An MDP is defined by a tuple $M = (\mathcal{S}, \mathcal{A}, r, P, d_1, H)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $P_t : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ is the transition kernel with $P_t(s'|s,a)$ representing the probability of seeing state s' after taking action a at state s, $r_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the mean reward function with $r_t(s,a)$ being the average immediate goodness of (s,a) at time t. Also, d_1 is denoted as the initial state distribution and H is the time horizon. The subscript t in P_t means the transition dynamics are non-stationary and could be different at each time t. A (non-stationary) policy $\pi : \mathcal{S} \to \mathbb{P}_{\mathcal{A}}^H$[1] assigns each state $s_t \in \mathcal{S}$ a distribution over actions at each time t, *i.e.* $\pi_t(\cdot|s_t)$ is a probability simplex with dimension $|\mathcal{S}|$.

Given a target policy of interest π, then the distribution of one H-step trajectory $\tau = (s_1, a_1, r_1, ..., s_H, a_H, r_H, s_{H+1})$ is specified by $\pi := (d_1, \pi)$[2] as follows: $s_1 \sim d_1^\pi$, for $t = 1, ..., H$, $a_t \sim \pi_t(\cdot|s_t)$ and random reward r_t has mean $r_t(s_t, a_t)$.

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

[1] Here $\mathbb{P}_{\mathcal{A}}^H = \mathbb{P}_{\mathcal{A}} \times \mathbb{P}_{\mathcal{A}} \times \mathbb{P}_{\mathcal{A}} \times ... \times \mathbb{P}_{\mathcal{A}}$, where "$\times$" represents Cartesian product and the product is performed for H times.

[2] For brevity, $\forall \pi$ we use π to denote the pair (d_1, π). This can be understood as: $\forall \pi$, $d_1^\pi = d_1$.

Then value function under policy π is defined as:

$$v^{\pi} = \mathbb{E}_{\pi}\left[\sum_{t=1}^{H} r_t\right].$$

The OPE problem aims at estimating v^{π} while given that n episodic data[3] $\mathcal{D} = \left\{(s_t^{(i)}, a_t^{(i)}, r_t^{(i)})\right\}_{i\in[n]}^{t\in[H]}$ are actually coming from a different logging policy μ.

Existing methods. The classical way to tackle the problem of OPE relies on incorporating importance sampling weights (IS), which corrects the mismatch in the distributions under the behavior policy and target policy. Specifically, define the t-step importance ratio as $\rho_t := \pi_t(a_t|s_t)/\mu_t(a_t|s_t)$, then it uses the cumulative importance ratio $\rho_{1:t} := \prod_{t'=1}^{t} \rho_{t'}$ to create IS based estimators:

$$\widehat{V}_{\text{IS}} := \frac{1}{n}\sum_{i=1}^{n} \widehat{V}_{\text{IS}}^{(i)}, \quad \widehat{V}_{\text{IS}}^{(i)} := \rho_{1:H}^{(i)} \cdot \sum_{t=1}^{H} r_t^{(i)};$$

$$\widehat{V}_{\text{step-IS}} := \frac{1}{n}\sum_{i=1}^{n} \widehat{V}_{\text{step-IS}}^{(i)}, \quad \widehat{V}_{\text{step-IS}}^{(i)} := \sum_{t=1}^{H} \rho_{1:t}^{(i)} r_t^{(i)},$$

where $\rho_{1:t}^{(i)} = \prod_{t'=1}^{t} \pi_{t'}(a_{t'}^{(i)}|s_{t'}^{(i)})/\mu_{t'}(a_{t'}^{(i)}|s_{t'}^{(i)})$. There are different versions of IS estimators including weighted IS estimators and doubly robust estimators (Murphy et al., 2001; Hirano et al., 2003; Dudík et al., 2011; Jiang and Li, 2016).

Even though IS-based off-policy evaluation methods possess a lot of advantages (*e.g.* unbiasedness), the variance of the cumulative importance ratios $\rho_{1:t}$ may grow exponentially as the horizon goes long. Attempts to break the barriers of horizon have been tried using model-based approaches (Liu et al., 2018b; Gottesman et al., 2019), which builds the whole MDP using either parametric or nonparametric models for estimating the value of target policy. (Liu et al., 2018a) considers breaking the curse of horizon of time-invariant MDPs by deploying importance sampling on the average visitation distribution of state-action pairs, (Hallak and Mannor, 2017) considers leveraging the stationary ratio of state-action pairs to replace the trajectory weights in an online fashion and (Gelada and Bellemare, 2019) further applies the same idea in the deep reinforcement learning regime. Recently, (Kallus and Uehara, 2019a,b) propose double reinforcement learning (DRL), which is based on doubly robust estimator with cross-fold estimation of q-functions and marginalized density ratios. It was shown that DRL is asymptotically efficient when both components are estimated at fourth-root rates, however no finite sample error bounds are given.

[3] To distinguish the data from different episodes, we use superscript to denote which episode they belong to throughout the rest of the paper.

Our goal. In this paper, our goal is to obtain the optimality of IS-based methods through marginalized importance sampling (MIS). As an earlier attempt, Xie et al. (2019) constructs MIS estimator by aggregating all trajectories that share the same state transition patterns to directly estimate the state distribution shifts after the change of policies from the behavioral to the target. However, as pointed by Remark 4 in Xie et al. (2019), the MSE upper bound of MIS estimator is asymptotically inefficient by a multiplicative factor of H. Xie et al. (2019) conjectures that the lower bound is not achievable in their infinite action setting. To bridge the gap and ultimately achieve the optimality, we consider the Tabular MDPs, where both the state space and action space are finite (*i.e.* $S = |\mathcal{S}| < \infty, A = |\mathcal{A}| < \infty$) and each state-action pair can be visited frequently as long as the logging policy μ does sufficient exploration (which corresponds to Assumption 2.2). Under the Tabular MDP setting, we can show the MSE upper bound of MIS estimator matches the Cramer-Rao lower bound provided by Jiang and Li (2016). To distinguish the difference, throughout the rest of paper we call the modified MIS estimator Tabular-MIS (TMIS) and the MIS estimator in Xie et al. (2019) State-MIS (SMIS).

1.1 Summary of results.

This work considers the problem of off-policy evaluation for a finite horizon, nonstationary, episodic MDP under tabular MDP setting. We propose and analyze Tabular-MIS estimator, which closes the gap between Cramer-Rao lower bound provided by Jiang and Li (2016) (on the variance of any unbiased estimator for a simplified setting of an nonstationary episodic MDP) and the MSE upper bound of State-MIS estimator (Xie et al., 2019). We also provide a high probability result by introducing a data-splitting type Tabular-MIS estimator, which retains the asymptotic efficiency while having an exponential tail. To the best of our knowledge, Split-TMIS is the first IS-based estimator in OPE that achieves asymptotic sample efficiency while having finite sample guarantees in high probability.

Moreover, the calculation of Tabular-MIS estimator and Split-TMIS does not explicitly incorporate the importance weights, which in turn implies our off-policy evaluation algorithm can be implemented without needing to know logging probabilities μ. Such logging-policy-free feature makes our Tabular-MIS estimator estimator more practical in the real-world applications.

Finally, we conduct a numerical simulation in Section 4 to empirically validate our theoretical results. We see that Tabular-MIS estimator improves over State-MIS estimator in MSE by a factor of H as expected.

1.2 Other related work

Markov Decision Processes have a long history of associated research (Puterman, 1994; Sutton and Barto, 1998), but many theoretical problems in the basic tabular setting remain an active area of research as of today. In particular, other than off-policy setting, there are two types of questions: *Regret bound and sample complexity in the online setting* and *Sample complexity with a generative model*. A detailed discussion can be found in Section B in appendix.

Our setting is different in two ways compared to those mentioned above. First, we consider a fixed pair of logging and target policy μ and π, so our bounds can depend explicitly on π and μ instead of S, A. Second, we do not have either online access to the environment (to change policies) or a generative model. Our high-probability bound with a direct union bound argument, implies a sample complexity of $\widetilde{O}(H^3S^2A/\epsilon^2)$ for identifying the optimal policy, which is suboptimal up to a factor of S, but notably has the optimal dependence in H. We remark that achieving the optimal dependence in the planning horizon H is generally tricky (see, e.g., the COLT open problem (Jiang and Agarwal, 2018) for more details). The current paper is among the few instances where we know how to obtain the optimal parameters.

Finally, the tabular RL setting is a basic abstraction that is relatively far away from real applications, which might have unobserved states, continuous state, non-zero Bellman error in the value function approximation. We leave generalization of the techniques in this paper to these more practical settings as future work.

2 Method

2.1 Problem description

In addition to the non-stationary, finite horizon tabular MDP $M = (\mathcal{S}, \mathcal{A}, r, T, d_1, H)$ (where $S := |\mathcal{S}| < \infty$ and $A := |\mathcal{A}| < \infty$), non-stationary logging policy μ and target policy π in Section 1, we denote $d_t^\mu(s_t, a_t)$ and $d_t^\pi(s_t, a_t)$ the induced joint state-action distribution at time t and the state distribution counterparts $d_t^\mu(s_t)$ and $d_t^\pi(s_t)$, satisfying $d_t^\pi(s_t, a_t) = d_t^\pi(s_t) \cdot \pi(a_t|s_t)$.[4] The initial distributions are identical $d_1^\mu = d_1^\pi = d_1$. Moreover, we use $P_{i,j}^\pi \in \mathbb{R}^{S\times S}$, $\forall j < i$ to represent the state transition probability from step j to step i under policy π, where $P_{t+1,t}^\pi(s'|s) = \sum_a P_{t+1,t}(s'|s,a)\pi_t(a|s)$. The marginal state distribution vector $d_t^\pi(\cdot)$ satisfies $d_t^\pi = P_{t,t-1}^\pi d_{t-1}^\pi$.

[4] For μ, $d_t^\mu(s_t, a_t) = d_t^\mu(s_t) \cdot \mu(a_t|s_t)$.

Historical data $\mathcal{D} = \left\{(s_t^{(i)}, a_t^{(i)}, r_t^{(i)})\right\}_{i\in[n]}^{t\in[H]}$ was obtained by logging policy μ and we can only use $\mathcal{D}$ to estimate the value of target policy π, *i.e.* v^π. Suppose we only assume knowledge about π and *do not observe* $r_t(s_t, a_t)$ for any actions other than the noisy immediate reward $r_t^{(i)}$ after observing $s_t^{(i)}, a_t^{(i)}$. The goal is to find an estimator to minimize the mean-square error (MSE):

$$\mathrm{MSE}(\pi, \mu, M) = \mathbb{E}_\mu[(\widehat{v}^\pi - v^\pi)^2].$$

Assumption 2.1 (Bounded rewards). *$\forall\, t = 1, ..., H$ and $i = 1, ..., n$, $0 \le r_t^{(i)} \le R_{\max}$.*

The bounded reward assumption can be relaxed to: $\exists R_{\max}, \sigma < +\infty$ such that $0 \le \mathbb{E}[r_t|s_t, a_t, s_{t+1}] \le R_{\max}$, $\mathrm{Var}[r_t|s_t, a_t, s_{t+1}] \le \sigma^2$ (as in Xie et al. (2019)), for achieving Cramer-Rao lower bound. However, the boundedness will become essential for applying concentrate inequalities in deriving high probability bounds.

Assumption 2.2 (Sufficient exploration). *Logging policy μ obeys that $d_m := \min_{t,s_t} d_t^\mu(s_t) > 0$.*

In fact this assumption can be relaxed to: require $d_t^\mu(s_t) > 0$ whenever $d_t^\pi(s_t) > 0$, and the corresponding $d_m := \min_{t,s_t}\{d_t^\mu(s_t) : d_t^\mu(s_t) > 0\}$. However, for the illustration purpose we stick to the above assumption. This assumption is always required for the consistency of off-policy evaluation estimator.

Assumption 2.3 (Bounded weights). *$\tau_s := \max_{t,s_t} \frac{d_t^\pi(s_t)}{d_t^\mu(s_t)} < +\infty$ and $\tau_a := \max_{t,s_t,a_t} \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)} < +\infty$.*

Assumption 2.3 is also necessary for discrete state and actions, as otherwise the second moments of the importance weight would be unbounded and the MSE of estimators will become intractable . The bound on τ_s is natural since $\tau_s \le \max_{t,s_t} \frac{1}{d_t^\mu(s_t)} = \frac{1}{\min_{t,s_t} d_t^\mu(s_t)} = \frac{1}{d_m}$ and it is finite by the Assumption 2.2; similarly, $\tau_a < \infty$ is also automatically satisfied if $\min_{t,s_t,a_t} \mu(a_t|s_t) > 0$. Finally, as we will see in the results, explicit dependence on τ_s, τ_a and d_m only appear in the low-order terms of the error bound.

2.2 Tabular-MIS estimator

To overcome the barrier caused by cumulative importance weights in IS type estimators, marginalized importance sampling directly estimates the marginalized state visitation distribution $\widehat{d}_t$ and defines the MIS estimator:

$$\widehat{v}_{MIS}^\pi = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{H} \frac{\widehat{d}_t^\pi(s_t^{(i)})}{\widehat{d}_t^\mu(s_t^{(i)})} \widehat{r}_t^\pi(s^{(i)}). \tag{1}$$

and $\widehat{d}_t^\mu(\cdot)$ is directly estimated using the empirical mean, *i.e.* $\widehat{d}_t^\mu(s_t) := \frac{1}{n}\sum_i \mathbf{1}(s_t^{(i)} = s_t) := \frac{n_{s_t}}{n}$ whenever

$n_{s_t} > 0$ and $\widehat{d}_t^{\pi}(s_t)/\widehat{d}_t^{\mu}(s_t) = 0$ when $n_{s_t} = 0$. Then the MIS estimator (1) becomes:

$$\widehat{v}_{MIS}^{\pi} = \sum_{t=1}^{H} \sum_{s_t} \widehat{d}_t^{\pi}(s_t)\widehat{r}_t^{\pi}(s_t) \tag{2}$$

Construction of State-MIS estimator. Based on the estimated marginal state transition $\widehat{d}_t^{\pi} = \widehat{P}_t^{\pi}\widehat{d}_{t-1}^{\pi}$, State-MIS estimator in Xie et al. (2019) directly estimates the state transition $P_t^{\pi}(s_t|s_{t-1})$ and state reward $r_t^{\pi}(s_t)$ as:

$$\widehat{P}_t^{\pi}(s_t|s_{t-1}) = \frac{1}{n_{s_{t-1}}} \sum_{i=1}^{n} \frac{\pi(a_{t-1}^{(i)}|s_{t-1})}{\mu(a_{t-1}^{(i)}|s_{t-1})} \tag{3}$$

$$\cdot \mathbf{1}((s_{t-1}^{(i)}, s_t^{(i)}, a_t^{(i)}) = (s_{t-1}, s_t, a_t)); \tag{4}$$

$$\widehat{r}_t^{\pi}(s_t) = \frac{1}{n_{s_t}} \sum_{i=1}^{n} \frac{\pi(a_t^{(i)}|s_t)}{\mu(a_t^{(i)}|s_t)} r_t^{(i)} \cdot \mathbf{1}(s_t^{(i)} = s_t). \tag{5}$$

State-MIS estimator directly constructs state transitions $\widehat{P}_t^{\pi}(s_t|s_{t-1})$ without explicitly modeling actions. Therefore, it is still valid when action space $\mathcal{A}$ is unbounded. However, importance weights must be explicitly utilized for compensating the discrepancy between μ and π and the knowledge of $\mu(a|s)$ at each state-action pair (s, a) is required.

Construction of Tabular-MIS estimator. Alternatively, we can go beyond importance weights and construct empirical estimates for $\widehat{P}_{t+1}(s_{t+1}|s_t, a_t)$ and $\widehat{r}_t(s_t, a_t)$ as:

$$\widehat{P}_{t+1}(s_{t+1}|s_t, a_t) = \frac{\sum_{i=1}^{n} \mathbf{1}[(s_{t+1}^{(i)}, a_t^{(i)}, s_t^{(i)}) = (s_{t+1}, s_t, a_t)]}{n_{s_t,a_t}}$$

$$\widehat{r}_t(s_t, a_t) = \frac{\sum_{i=1}^{n} r_t^{(i)} \mathbf{1}[(s_t^{(i)}, a_t^{(i)}) = (s_t, a_t)]}{n_{s_t,a_t}}, \tag{6}$$

where we set $\widehat{P}_{t+1}(s_{t+1}|s_t, a_t) = 0$ and $\widehat{r}_t(s_t, a_t) = 0$ if $n_{s_t,a_t} = 0$, with n_{s_t,a_t} the empirical visitation frequency to state-action (s_t, a_t) at time t. The corresponding estimation of $\widehat{P}_t^{\pi}(s_t|s_{t-1})$ and $\widehat{r}_t^{\pi}(s_t)$ are defined as:

$$\widehat{P}_t^{\pi}(s_t|s_{t-1}) = \sum_{a_{t-1}} \widehat{P}_t(s_t|s_{t-1}, a_{t-1})\pi(a_{t-1}|s_{t-1}),$$

$$\widehat{r}_t^{\pi}(s_t) = \sum_{a_t} \widehat{r}_t(s_t, a_t)\pi(a_t|s_t), \ \widehat{d}_t^{\pi} = \widehat{P}_t^{\pi}\widehat{d}_{t-1}^{\pi}. \tag{7}$$

In conclusion, by using the same estimator for $\widehat{d}_t^{\mu}$, $\widehat{v}_{\text{TMIS}}^{\pi}$ and $\widehat{v}_{\text{SMIS}}^{\pi}$ share the same form of (2). However, Tabular-MIS estimator constructs a different estimation of component $\widehat{d}_t^{\pi}$ though (6)-(7) by leveraging the fact that each state-action pair is visited frequently under Tabular setting.

The motivation of MIS-type estimators comes from the fact that we have a nonstationary MDP model and its underlying state marginal transition follows $d_t^{\pi} = P_t^{\pi} d_{t-1}^{\pi}$. The MIS estimators are then obtained by using corresponding plug-in estimators for each different components (*i.e.* $\widehat{d}_t^{\pi}$ for d_t^{π}, $\widehat{P}_t^{\pi}$ for P_t^{π}). On the other hand, IS-type estimators design the value function in a more straightforward way without needing to estimate the transition environment (Mahmood et al., 2014). Therefore in this sense MIS-type estimators are essentially model-based estimators with the model of interactive environment $M = (\mathcal{S}, \mathcal{A}, r, T, d_1, H)$.

3 Main Results

We now show that our Tabular-MIS estimator achieves the asymptotic Cramer-Rao lower bound for DAG-MDP (Jiang and Li, 2016) and therefore is asymptotically sample efficient. To formalize our statement, we pre-specify the following boundary conditions: $r_0(s_0) \equiv 0$, $\sigma_0(s_0, a_0) \equiv 0$, $\frac{d_0^{\pi}(s_0)}{d_0^{\mu}(s_0)} \equiv 1$, $\frac{\pi(a_0|s_0)}{\mu(a_0|s_0)} \equiv 1$, $V_{H+1}^{\pi} \equiv 0$, and, as a reminder, $\tau_a := \max_{t,s_t,a_t} \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}$ and $\tau_s := \max_{t,s_t} \frac{d_t^{\pi}(s_t)}{d_t^{\mu}(s_t)}$.

Theorem 3.1. *Suppose the n episodic historical data $\mathcal{D} = \left\{(s_t^{(i)}, a_t^{(i)}, r_t^{(i)})\right\}_{i=1,\dots,n}^{t=1,\dots,H}$ is obtained by running a logging policy μ and π is the new target policy which we want to test. If the number of episodes n satisfies*

$$n > \max\left[\frac{16\log n}{\min_{t,s_t,a_t} d_t^{\mu}(s_t, a_t)}, \frac{4H\tau_a\tau_s}{\min_{t,s_t} \max\{d_t^{\pi}(s_t), d_t^{\mu}(s_t)\}}\right]$$

then under Assumption 2.1-2.3 our Tabular-MIS estimator $\widehat{v}_{\text{TMIS}}^{\pi}$ has the following Mean-Square-Error upper bound:

$$\mathbb{E}[(\widehat{v}_{\text{TMIS}}^{\pi} - v^{\pi})^2] \le \frac{1}{n} \sum_{h=0}^{H} \sum_{s_h,a_h} \frac{d_h^{\pi}(s_h)^2}{d_h^{\mu}(s_h)} \frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)}$$

$$\cdot \text{Var}\left[(V_{h+1}^{\pi}(s_{h+1}^{(1)}) + r_h^{(1)})\middle| s_h^{(1)} = s_h, a_h^{(1)} = a_h\right] \tag{8}$$

$$\cdot \left(1 + \sqrt{\frac{16\log n}{n \min_{t,s_t} d_t^{\mu}(s_t)}}\right) + O(\frac{\tau_a^2\tau_s H^3}{n^2 \cdot d_m}),$$

where the value function is defined as: $V_h^{\pi}(s_h) := \mathbb{E}_{\pi}\left[\sum_{t=h}^{H} r_t^{(1)}\middle| s_h^{(1)} = s_h\right]$, $\forall h \in \{1, 2, ..., H\}$.

The proof of this theorem, and all the other technical results we present in this section, are deferred to the appendix. We summarize the novel ingredients in the proof in Section 3.1. Before that, we make a few remarks about this interesting result.

Remark 3.2 (Asymptotic efficiency and local minimaxity)**.** *The error bound implies that* $\lim_{n\to\infty} n \cdot \mathbb{E}[(\widehat{v}_{\text{TMIS}}^{\pi} - v^{\pi})^2]$

$$\sum_{t=0}^{H} \mathbb{E}_{\mu}\left[\frac{d^{\pi}(s_t^{(1)}, a_t^{(1)})^2}{d^{\mu}(s_t^{(1)}, a_t^{(1)})^2} \text{Var}\left[V_{t+1}^{\pi}(s_{t+1}^{(1)}) + r_t^{(1)}\middle| s_t^{(1)}, a_t^{(1)}\right]\right].$$

This exactly matches the CR-lower bound in Jiang and

Li (2016, Proposition 3) for DAG-MDP[5]*. In contrast, the State-MIS estimator in (Xie et al., 2019) achieves an asymptotic MSE of*

$$\sum_{t=0}^{H} \mathbb{E}_{\mu}\left[\frac{d^{\pi}(s_t^{(1)})^2}{d^{\mu}(s_t^{(1)})^2}\mathrm{Var}\Big[\frac{\pi(a_t^{(1)}|s_t^{(1)})}{\mu(a_t^{(1)}|s_t^{(1)})}(V_{t+1}^{\pi}(s_{t+1}^{(1)})+r_t^{(1)})\Big|s_t^{(1)}\Big]\right]. \tag{9}$$

We note that while in classical literature CR-lower bound is often used to lower bound the variance of *unbiased* estimators, the modern theory of estimation establishes that it is also the correct asymptotic minimax lower bound for the MSE of *all* estimators in every local neighborhood of the parameter space (see, e.g., Van der Vaart, 2000, Chapter 8). In other words, our results imply that Tabular-MIS estimator is asymptotically, locally, uniformly minimax optimal, namely, optimal for every problem instance separately.

While asymptotically efficient estimators for this problem in related settings have been proposed in independent recent work (Kallus and Uehara, 2019a,b), our estimator is the first that comes with finite sample guarantees with an explicit expression on the low-order terms. Moreover, our estimator demonstrates that doubly robust estimation techniques is not essential for achieving asymptotic efficiency.

Remark 3.3 (Simplified finite sample error bound)**.** *The theory implies that there is universal constants C_1, C_2 such that for all $n \geq C_1 H \frac{\tau_a}{d_m}$, i.e., when we have a just visited every state-action pair for $\Omega(H)$ times, $\mathbb{E}[(\widehat{v}^{\pi}_{\mathrm{TMIS}} - v^{\pi})^2] = C_2 H^2 \tau_a \tau_s R_{\max}^2/n$.*

In deriving the above remark, we used the somewhat surprising observation that

$$\sum_{t=1}^{H} \mathbb{E}_{\pi}\left[\mathrm{Var}\Big[V_{t+1}^{\pi}(s_{t+1}^{(1)}) + r_t^{(1)}\Big|s_t^{(1)}, a_t^{(1)}\Big]\right] \leq H^2 R_{\max}^2.$$

Note that we are summing H quantities that are potentially on the order of $H^2 R_{\max}^2$, yet no additional factors of H shows up. This observation is folklore and has been used in deriving tight results for tabular RL in (e.g., Azar et al., 2017). It can be proven using the following decomposition of the variance of the empirical mean estimator and the fact it is bounded by $H^2 R_{\max}^2$.

Lemma 3.4. *For any policy π and any MDP.*

$$\mathrm{Var}_{\pi}\left[\sum_{t=1}^{H} r_t^{(1)}\right] = \sum_{t=1}^{H}\Big(\mathbb{E}_{\pi}\left[\mathrm{Var}\left[r_t^{(1)} + V_{t+1}^{\pi}(s_{t+1}^{(1)})\Big|s_t^{(1)}, a_t^{(1)}\right]\right] + \mathbb{E}_{\pi}\left[\mathrm{Var}\left[\mathbb{E}[r_t^{(1)} + V_{t+1}^{\pi}(s_{t+1}^{(1)})|s_t^{(1)}, a_t^{(1)}]\Big|s_t^{(1)}\right]\right]\Big).$$

[5] Jiang and Li (2016) focused on the special case with deterministic reward only at $t = H$. It is straightforward to show that the above expression is the CR-lower bound in the general tabular setting.

The proof, which applies the law-of-total-variance recursively, is deferred to the appendix.

Remark 3.5 (When $\pi = \mu$)**.** *One surprising observation is that Tabular-MIS estimator improves the efficiency even for the on-policy evaluation problem when $\pi = \mu$. In other word, the natural Monte Carlo estimator of the reward in the on-policy evaluation problem is in fact asymptotically inefficient.*

3.1 Building blocks of the analysis

At a high level, the techniques we used, including the idea of fictitious estimator and peeling the variance (expectation) of fictitious estimator $\widetilde{v}^{\pi}$ from behind by applying total law of variances (expectations) repeatedly, are consistent with Xie et al. (2019).

In addition to the above techniques, we leverage the fact of frequent state-action visitations in our design of TMIS estimator and based on that we are able to achieve an asymptotic lower Mean Square Error (MSE) bound. The main components are the followings.

Fictitious Tabular-MIS estimator. Fictitious Tabular-MIS estimator $\widetilde{v}^{\pi}_{\mathrm{TMIS}}$ is a modified version of $\widehat{v}^{\pi}_{\mathrm{TMIS}}$ with $\widehat{P}^{\pi}_{t+1}(\cdot|s_t, a_t)$, $\widehat{r}^{\pi}_t(s_t, a_t)$ replaced by the underlying true $P^{\pi}_{t+1}(\cdot|s_t, a_t)$, $r^{\pi}_t(s_t, a_t)$ when the visitation frequency of state-action pair (s_t, a_t) is insufficient (*e.g.* $n_{s_t,a_t} < O(nd^{\mu}_t(s_t, a_t))$). Specifically, fictitious Tabular-MIS estimator $\widetilde{v}^{\pi}_{\mathrm{TMIS}}$ remains every part of $\widehat{v}^{\pi}_{\mathrm{TMIS}}$ unchanged except the following:

$$\widetilde{r}_t(s_t, a_t) = \begin{cases} \widehat{r}_t(s_t, a_t) & \text{if } n_{s_t,a_t} \geq nd^{\mu}_t(s_t, a_t)(1-\theta) \\ r_t(s_t, a_t) & \text{otherwise;} \end{cases} \tag{10}$$

and

$$\widetilde{P}_{t+1,t}(\cdot|s_t, a_t) = \begin{cases} \widehat{P}_{t+1,t} & \text{if } n_{s_t,a_t} \geq nd^{\mu}_t(s_t, a_t)(1-\theta) \\ P_{t+1,t} & \text{otherwise,} \end{cases} \tag{11}$$

where θ is the parameter constrained by $0 < \theta < 1$, which we will choose later in the proof.

This slight modification makes $\widetilde{v}^{\pi}_{\mathrm{TMIS}}$ no longer implementable using the logging data $\mathcal{D}$, but it does provide an unbiased estimator of v^{π} (Lemma C.5 in appendix) and, most importantly, it is easier to do theoretical analysis on $\widetilde{v}^{\pi}_{\mathrm{TMIS}}$ than on $\widehat{v}^{\pi}_{\mathrm{TMIS}}$. Moreover, Multiplicative Chernoff bound (Lemma A.2 in appendix) helps to find the connection between $\widetilde{v}^{\pi}_{\mathrm{TMIS}}$ and $\widehat{v}^{\pi}_{\mathrm{TMIS}}$.

Peeling arguments using the total law of variance (expectation). The core idea in analyzing the variance of $\widetilde{v}^{\pi}$ is to peel the variance from behind (start from time H to 1) and the peeling tool we used here is through marriaging the standard Bellman equations with the total law of variance. Lemma C.2 (in appendix) shows this spirit and it is used repeatedly

throughout the whole analysis. Beyond that, the peeling argument can be used to prove the dependence in H is only H^2 for our Tabular-MIS estimator. This result explicates that H^2 is enough for TMIS to evaluate a particular policy and this is different from SMIS, which in general requires the dependence of H^3 for off-policy evaluation.

3.2 A high-probability bound with data-splitting TMIS.

Tabular-MIS estimator provides the asymptotic optimal variance bound of order $O(H^2SA/n)$ and based on that it is natural to ask the related learning question: whether TMIS can further achieve a high probability bound with the same sample complexity? We figure out that the standard concentration inequalities (*e.g.* Hoeffding's inequality, Bernstein inequality) cannot be directly applied because of the highly correlated structures of the Tabular-MIS estimator. To address this problem we design the following data split version of TMIS and as we will see, the original TMIS is essentially a special case of data-splitting TMIS.

Data splitting Tabular-MIS estimator. Assume the total number of episodes n can be factorized as $n = M \cdot N$, where $M, N > 1$ are two integers,[6] and we can partition the data $\mathcal{D}$ into N folds with each fold $\mathcal{D}^{(i)}$ $(i = 1, ..., N)$ has M different episodes, or in other words, we split the n episodes evenly. Then by the i.i.d. nature of n episodes, we have $\mathcal{D}^{(1)}, \mathcal{D}^{(2)}, ..., \mathcal{D}^{(N)}$ are independent collections.

For each $\mathcal{D}^{(i)}$, we can create a Tabular-MIS estimator $\widehat{v}^{\pi(i)}_{\text{TMIS}}$ (for notation simplicity we use $\widehat{v}^{\pi}_{(i)}$ to denote $\widehat{v}^{\pi(i)}_{\text{TMIS}}$ in the future discussions) using its own M episodes. Then $\widehat{v}^{\pi}_{(1)}, \widehat{v}^{\pi}_{(2)}, ..., \widehat{v}^{\pi}_{(N)}$ are independent of each other and we can use the empirical mean to define the data splitting Tabular-MIS estimatorand the corresponding fictitious version:

$$\widehat{v}^{\pi}_{\text{split}} = \frac{1}{N}\sum_{i=1}^{N} \widehat{v}^{\pi}_{(i)}, \qquad \widetilde{v}^{\pi}_{\text{split}} = \frac{1}{N}\sum_{i=1}^{N} \widetilde{v}^{\pi}_{(i)}, \tag{12}$$

where each $\widetilde{v}^{\pi}_{(i)}$ is the fictitious estimator of $\widehat{v}^{\pi}_{(i)}$.

The data splitting TMIS estimator explicitly characterizes the independence of n different episodes by grouping them into N chunks. Chunks are independent of each other and taking the average over all $\widehat{v}^{\pi}_{(i)}$ $i = 1, ..., N$ will guarantee the validity of using concentration inequalities.

More importantly, the data splitting TMIS estimator holds the same information-theoretical variance lower bound as the non-data splitting TMIS estimator, which is not surprising since the non-data splitting TMIS estimator is just the special case of the data splitting Tabular-MIS estimator with $N = 1$. This idea is summarized into the following theorem:

Theorem 3.6. *Using n i.i.d. episodic data from a near-uniform[7] logging policy μ and suppose M, the number of episodes for each $\mathcal{D}^{(i)}$, satisfies:*

$$M > \max\left[O(SA \cdot \text{Polylog}(S, H, A, n)), O(H\tau_a\tau_s)\right],$$

then the data splitting Tabular-MIS estimator obeys:

$$\mathbb{E}[(\widehat{v}^{\pi}_{\text{split}} - v^{\pi})^2] \leq O(\frac{H^2SA}{n}). \tag{13}$$

Remark 3.7. *The condition in Theorem 3.6 is achieveable. For example, choose $M \approx \sqrt{n}$, then the condition holds when n is sufficiently large.*

High probability bound. By coupling the data splitting techniques with the boundedness of Tabular-MIS estimator (*i.e.* $\widehat{v}^{\pi} \leq HR_{\max}, \widetilde{v}^{\pi} \leq HR_{\max}$, see Lemma C.3 in appendix), we can apply concentration inequalities to show the difference between $\widehat{v}^{\pi}_{\text{split}}$ and v^{π} is bounded by order $\widetilde{O}(\sqrt{H^2SA/n})$, which is summarized into the following theorem.

Theorem 3.8. *Suppose n i.i.d. episodic historical data comes from a near-uniform logging policy μ and suppose M, the number of episodes in each $\mathcal{D}^{(i)}$, satisfies: $\widetilde{O}(\sqrt{n \cdot SA}) \geq M$ and $M > \max\left[O(SA \cdot \text{Polylog}(S, H, A, n, 1/\delta)), O(H\tau_a\tau_s)\right]$. Then we have with probability $1 - \delta$, the data splitting Tabular-MIS estimator obeys:*

$$|\widehat{v}^{\pi}_{\text{split}} - v^{\pi}| \leq \widetilde{O}(\sqrt{\frac{H^2SA}{n}}).$$

The proof Theorem 3.8 relies on bounding the difference between $\widehat{v}^{\pi}_{\text{split}}$ and $\widetilde{v}^{\pi}_{\text{split}}$ using Multiplicative Chernoff bound and bounding the difference between $\widetilde{v}^{\pi}_{\text{split}}$ and v^{π} using Bernstein inequality. During the process of bounding $|\widehat{v}^{\pi}_{\text{split}} - \widetilde{v}^{\pi}_{\text{split}}|$ we observe that a stronger uniform bound can be derived. In fact, this bound is 0. We formalize it into the following lemma.

Lemma 3.9. *Suppose n i.i.d. episodic historical data comes from a near-uniform logging policy μ and suppose M, the number of episodes in each $\mathcal{D}^{(i)}$, satisfies: $M > \max\left[O(SA \cdot \text{Polylog}(S, H, A, N, 1/\delta)), O(H\tau_a\tau_s)\right]$. Then we have with probability $1 - \delta$,*

$$\sup_{\pi \in \prod} |\widehat{v}^{\pi}_{\text{split}} - \widetilde{v}^{\pi}_{\text{split}}| = 0$$

Since $n = N \cdot M$, therefore let $N = 1$, $M = n$, then if

$$M > \max\left[O(SA \cdot \text{Polylog}(S, H, A, 1/\delta)), O(H\tau_a\tau_s)\right],$$

[6] In general this might not be true, *e.g.* if n is prime number. However, we can resolve it by choosing $M = \lfloor n/N \rfloor$.

[7] Near-uniform here means: $\min_{t,s_t,a_t} d^{\mu}_t(s_t, a_t) > \Omega(1/(SA))$.

$$\sup_{\pi\in\prod} |\widehat{v}^{\pi}_{\mathrm{TMIS}} - \widetilde{v}^{\pi}_{\mathrm{TMIS}}| = 0$$

holds with probability $1-\delta$*, where* $\prod$ *consists of all the* H*-step nonstationary policies.*

Remark 3.10. *The uniform difference bound between* $\widehat{v}^{\pi}_{\mathrm{TMIS}}$ *and* $\widetilde{v}^{\pi}_{\mathrm{TMIS}}$ *is obtained by observing the construction of fictitious estimator* (10) *and* (11) *are independent of the specific target policy* π*. This result tells the* $\sup_{\pi\in\prod} |\widehat{v}^{\pi}_{\mathrm{TMIS}} - \widetilde{v}^{\pi}_{\mathrm{TMIS}}|$ *can be arbitrarily small with high probability and therefore does not depend on* H *factor. This fact will help us to derive the correct dependence in* H *for uniform convergence problem, see Section 5.*

Algorithm 1 Tabular MIS Off-Policy Evaluation

Input: Logging data $\mathcal{D} = \{\{s_t^{(i)}, a_t^{(i)}, r_t^{(i)}\}_{t=1}^{H}\}_{i=1}^{n}$ from the behavior policy μ. A target policy π which we want to evaluate its cumulative reward.

1: Calculate the on-policy estimation of initial distribution $d_1(\cdot)$ by $\widehat{d}_1(s) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(s_1^{(i)} = s)$, and set $\widehat{d}_1^{\mu}(\cdot) := \widehat{d}_1(\cdot)$, $\widehat{d}_1^{\pi}(s) := \widehat{d}_1(\cdot)$.

2: **for** $t = 2, 3, \ldots, H$ **do**

3: Choose all transition data at time step t, $\{s_t^{(i)}, a_t^{(i)}, r_t^{(i)}\}_{i=1}^{n}$.

4: Calculate the on-policy estimation of $d_t^{\mu}(\cdot)$ by $\widehat{d}_t^{\mu}(s) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(s_t^{(i)} = s)$.

5: Set the off-policy estimation of $\widehat{P}_t(s_t|s_{t-1}, a_{t-1})$:

$$\widehat{P}_t(s_t|s_{t-1}, a_{t-1}) := \frac{\sum_{i=1}^{n} \mathbf{1}[(s_t^{(i)}, a_{t-1}^{(i)}, s_{t-1}^{(i)}) = (s_t, s_{t-1}, a_{t-1})]}{n_{s_{t-1}, a_{t-1}}}$$

when $n_{s_{t-1},a_{t-1}} > 0$. Otherwise set it to be zero.

6: Estimate the reward function

$$\widehat{r}_t(s_t, a_t) := \frac{\sum_{i=1}^{n} r_t^{(i)} \mathbf{1}(s_t^{(i)} = s_t, a_t^{(i)} = a_t)}{\sum_{i=1}^{n} \mathbf{1}(s_t^{(i)} = s_t, a_t^{(i)} = a_t)}.$$

when $n_{s_t,a_t} > 0$. Otherwise set it to be zero.

7: Set $\widehat{d}_t^{\pi}(\cdot)$ according to $\widehat{d}_t^{\pi} = \widehat{P}_t^{\pi} \widehat{d}_{t-1}^{\pi}$, with $\widehat{P}_t^{\pi}$ defined according to (7). Also, set $\widehat{r}_t^{\pi}(\cdot)$ according to (7).

8: **end for**

9: Substitute the all estimated values above into (1) to obtain $\widehat{v}^{\pi}$, the estimated value of π.

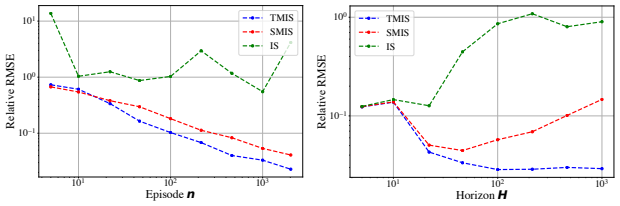
3.3 Some interpretations.

Logging policy free algorithm. We point out the implementation of Tabular-MIS estimator does not require the knowledge of logging policy μ, as shown in Algorithm 1,2.[8] This is critical in the sense that in the real-world sequential decision making problems, it is very likely the complete information about logging policy is not provided. This may happen due to misrecords or the lack of maintainance. By only using the historical data, tabular MIS off-policy evaluation is able to achieve the asymptotic efficiency. In contrast, the state MIS estimator always requires the full information about the logging policy.

Connection to approximate MDP estimation. Our TMIS is essentially an approximate MDP estimator (with the non-stationary dynamic transitions P_t estimated by *maximum likelihood estimator* (MLE)) except that we marginalize out the action in both $\widehat{r}_t^{\pi}(s)$ and $\widehat{d}_t^{\pi}(s)$ and provide an importance sampling interpretation. To the best of our knowledge, existing analysis of the fully model-based approach does not provide tight bounds. We give two examples. The seminal simulation lemma in Kearns and Singh (2002) together with a naive concentration-type analysis gives only an $\widetilde{O}(\sqrt{H^4S^3A/n})$ bound in our setting. In a very recent compilation of improvements over this bound (Jiang, 2018), this bound can be improved to either $\widetilde{O}(\sqrt{H^4S^2A/n})$ or $\widetilde{O}(\sqrt{H^6SA/n})$. Our result is the first that achieves the optimal $\widetilde{O}(\sqrt{H^2SA/n})$ rate regardless of whether it is the model-based or model-free approach.

4 Experiments

In this section, we present some empirical studies to demonstrate that our main theoretical results about Tabular-MIS estimator in Theorem 3.1 are empirically verified.



(a) Different Episode n (b) Different Horizon H

Figure 1: Relative RMSE ($\sqrt{\mathrm{MSE}}/v^{\pi}$) on Non-stationary Non-mixing MDP

Time-varying, non-mixing Tabular MDP. We test our approach in simulated MDP environment where both the states and the actions are binary. Concretely, there are two states s_0 and s_1 and two actions a_1 and a_2. State s_0 always has probability

[8] Algorithm 2 is deferred to appendix due to space constraint.

1 going back to itself, regardless of the actions, *i.e.* $P_t(s_0|s_0,a_1) = 1$ and $P_t(s_0|s_0,a_2) = 1$. For state s_1, at each time step there is one action (we call it a) that has probability $2/H$ going to s_0 and the other action (we call it a') has probability 1 going back to s_1, *i.e.* $P_t(s_0|s_1,a) = 2/H = 1 - P_t(s_1|s_1,a)$ and $P_t(s_1|s_1,a') = 1$. Moreover, which action will make state s_1 go to state s_0 with probability $2/H$ is decided by a random parameter $p_t \in [0,1]$. If $p_t < 0.5$, $a = a_1$ and if $p_t \geq 0.5$, $a = a_2$. One can receive reward 1 at each time step if $t > H/2$ and is in state s_0, and will receive reward 0 otherwise. Lastly, for state s_0, we set $\mu(\cdot|s_0) = \pi(\cdot|s_0)$; for state s_1, we set $\mu(a_1|s_1) = \mu(a_2|s_1) = 1/2$ and $\pi(a_1|s_1) = 1/4 = 1 - \pi(a_2|s_1)$.

Figure 1(a) shows the asymptotic convergence rates of relative RMSE with respect to the number of episodes, given fixed horizon $H = 100$. Both SMIS and TMIS has a $O(1/\sqrt{n})$ convergence rate. The saving of $\sqrt{H}$ of TMIS over SMIS in this log-log plot is reflected in the intercept. Figure 1(b) has fixed $n = 1024$ with varying horizon H. Note since $v^\pi \approx O(H)$, therefore for TMIS our theoretical result implies $\sqrt{\text{MSE}}/v^\pi = O(\sqrt{H^2}/H) = O(1)$, which is consistent with the horizontal line when H is large. Moreover, for SMIS $\sqrt{\text{MSE}}/v^\pi = O(\sqrt{H^3}/H) = O(\sqrt{H})$, so after taking the $\log(\cdot)$ we should have asymptotic linear trend with coefficient 1/2. The red line in Figure 1(b) empirically verifies this result. More empirical study discussions are deferred to Appendix E.

5 Discussion

From off-policy evaluation to offline learning. A real offline reinforcement learning system is equipped with both offline learning algorithms and off-policy evaluation algorithms. The decision maker should first run the offline learning algorithm to find a near optimal policy and then use off-policy evaluation methods to check if the obtained policy is good enough. Under our tabular MDP setting, we point out it is possible to find a ϵ-optimal policy in near optimal time and sample complexity $O(H^3SA/\epsilon^2)$using the Q-value iteration (QVI) based algorithm designed by Sidford et al. (2018). Their QVI algorithm assumes a generative model which can provide independent sample of the next state s' given any current state-action (s,a). At a first glance, this assumption seems too strong for offline learning since we cannot force the agent to stay in any arbitrary location. In fact, the Assumption 2.2 on μ actually reveals that the underlying logging policy can be considered as the surrogate of the generative model. As n goes large, the visitation frequency of any (s_t,a_t) will be large enough with high probability, as guaranteed by Multiplicative Chernoff bound.

From off-policy evaluation to uniform off-policy evaluation. The high probability result achieves $\widetilde{O}(\sqrt{H^2SA/n})$ complexity. Following this discovery line, then it is natural to ask whether uniform convergence over a class of policies (*e.g.* all deterministic policies) can be achieved with optimal sample complexity. This problem is interesting since it will guarantee the strong performance of off-policy evaluation methods over all policies in certain policy class $\prod$. By a direct application of union bound, we can obtain the following result:

Theorem 5.1. *Let $\prod$ contains all the deterministic H-step policies. Then under the same condition as Theorem 3.8, the data splitting Tabular-MIS estimator satisfies:*

$$\sup_{\pi\in\Pi} |\widehat{v}^\pi_{\text{split}} - v^\pi| \leq \widetilde{O}(\sqrt{\frac{H^3S^2A}{n}}),$$

with probability $1-\delta$.

The uniform convergence bound implies that the empirical best policy $\hat{\pi} = \text{argmax}_\pi \widehat{v}^\pi_{\text{split}}$ is within $\epsilon = O(\sqrt{\frac{H^3S^2A}{n}})$ of the optimal policy. This matches the sample complexity lower bound for learning the optimal policy (Azar et al., 2013) in all parameters except a factor of S.

Open problem: 1. H^3 vs H^2 in the infinite $\mathcal{A}$ setting. Finally, we note that the conjecture posed by Xie et al. (2019) remains unsolved. In the infinite $\mathcal{A}$ case, we can never observe any (s,a) pairs more than once, hence not able to estimate the transition dynamics or the expected reward. The minimax lower bound in (Wang et al., 2017) (for the contextual bandit setting) already establishes that the Cramer-Rao lower bound is not achievable in this setting even if $H = 1$ and $S = 1$. It remains open whether H^3 is required.
2. Better dependence for stationary transition case. We conjecture the dependence of H^2 can be further reduced in the stationary transition case. Our current analysis cannot further reduce the dependence for stationary transition setting.

6 Conclusion

In this paper, we propose a new marginalized importance sampling estimator for the off-policy evaluation (OPE) problem under the episodic tabular setting. We show that the estimator is has a finite sample error bound that matches the exact Cramer-Rao lower bound up to low-order factors. We also provide an extension with high probability error bound. To the best of our knowledge, these results are the first of their kinds. Future work includes resolving the open problems mentioned before and generalizing the results to more practical settings.

References

Agarwal, A., Kakade, S., and Yang, L. F. (2019). On the optimality of sparse model-based planning for markov decision processes. *arXiv preprint arXiv:1906.03804*.

Azar, M. G., Munos, R., and Kappen, H. J. (2013). Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349.

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org.

Chernoff, H. et al. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507.

Dann, C., Lattimore, T., and Brunskill, E. (2017). Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5713–5723.

Dudík, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1097–1104. Omnipress.

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. (2018). More robust doubly robust off-policy evaluation. *arXiv preprint arXiv:1802.03493*.

Gelada, C. and Bellemare, M. G. (2019). Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3647–3655.

Gottesman, O., Liu, Y., Sussex, S., Brunskill, E., and Doshi-Velez, F. (2019). Combining parametric and nonparametric models for off-policy evaluation. *arXiv preprint arXiv:1905.05787*.

Hallak, A. and Mannor, S. (2017). Consistent on-line off-policy evaluation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1372–1383. JMLR. org.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Jiang, N. (2018). Notes on tabular methods.

Jiang, N. and Agarwal, A. (2018). Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pages 3395–3398.

Jiang, N. and Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 652–661. JMLR. org.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873.

Kallus, N. and Uehara, M. (2019a). Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *arXiv preprint arXiv:1908.08526*.

Kallus, N. and Uehara, M. (2019b). Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes. *arXiv preprint arXiv:1909.05850*.

Kearns, M. and Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2-3):209–232.

Kearns, M. J. and Singh, S. P. (1999). Finite-sample convergence rates for q-learning and indirect algorithms. In *Advances in neural information processing systems*, pages 996–1002.

Li, L., Chu, W., Langford, J., and Wang, X. (2011). Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306. ACM.

Li, L., Munos, R., and Szepesvári, C. (2015). Toward minimax off-policy value estimation.

Liu, Q., Li, L., Tang, Z., and Zhou, D. (2018a). Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pages 5361–5371.

Liu, Y., Gottesman, O., Raghu, A., Komorowski, M., Faisal, A. A., Doshi-Velez, F., and Brunskill, E. (2018b). Representation balancing mdps for off-policy policy evaluation. In *Advances in Neural Information Processing Systems*, pages 2644–2653.

Mahmood, A. R., van Hasselt, H. P., and Sutton, R. S. (2014). Weighted importance sampling for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems*, pages 3014–3022.

Mandel, T., Liu, Y.-E., Levine, S., Brunskill, E., and Popovic, Z. (2014). Offline policy evaluation across representations with applications to educational games. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1077–1084. International Foundation for Autonomous Agents and Multiagent Systems.

Murphy, S. A., van der Laan, M. J., Robins, J. M., and Group, C. P. P. R. (2001). Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423.

Puterman, M. L. (1994). Markov decision processes: Discrete stochastic dynamic programming.

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. (2018). Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, pages 5186–5196.

Sridharan, K. (2002). A gentle introduction to concentration inequalities. *Dept. Comput. Sci., Cornell Univ., Tech. Rep.*

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Swaminathan, A., Krishnamurthy, A., Agarwal, A., Dudik, M., Langford, J., Jose, D., and Zitouni, I. (2017). Off-policy evaluation for slate recommendation. In *Advances in Neural Information Processing Systems*, pages 3632–3642.

Theocharous, G., Thomas, P. S., and Ghavamzadeh, M. (2015). Personalized ad recommendation systems for life-time value optimization with guarantees. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Thomas, P. and Brunskill, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148.

Thomas, P. S., Theocharous, G., Ghavamzadeh, M., Durugkar, I., and Brunskill, E. (2017). Predictive off-policy policy evaluation for nonstationary decision problems, with applications to digital marketing. In *Twenty-Ninth IAAI Conference*.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Wang, Y.-X., Agarwal, A., and Dudik, M. (2017). Optimal and adaptive off-policy evaluation in contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3589–3597. JMLR. org.

Xie, T., Ma, Y., and Wang, Y.-X. (2019). Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems*, pages 9665–9675.