Deeply-Debiased Off-Policy Interval Estimation

Chengchun Shi [*1] **Runzhe Wan** [*2] **Victor Chernozhukov** [3] **Rui Song** [2]

Abstract

Off-policy evaluation learns a target policy's value with a historical dataset generated by a different behavior policy. In addition to a point estimate, many applications would benefit significantly from having a confidence interval (CI) that quantifies the uncertainty of the point estimate. In this paper, we propose a novel deeply-debiasing procedure to construct an efficient, robust, and flexible CI on a target policy's value. Our method is justified by theoretical results and numerical experiments. A Python implementation of the proposed procedure is available at `https://github.com/RunzheStat/D2OPE`.

1. Introduction

Reinforcement learning (RL, Sutton & Barto, 2018) is a general technique in sequential decision making that learns an optimal policy to maximize the average cumulative reward. Prior to adopting any policy in practice, it is crucial to know the impact of implementing such a policy. In many real domains such as healthcare (Murphy et al., 2001; Luedtke & van der Laan, 2017; Shi et al., 2020a), robotics (Andrychowicz et al., 2020) and autonomous driving (Sallab et al., 2017), it is costly, risky, unethical, or even infeasible to evaluate a policy's impact by directly running this policy. This motivates us to study the off-policy evaluation (OPE) problem that learns a target policy's value with pre-collected data generated by a different behavior policy.

In many applications (e.g., mobile health studies), the number of observations is limited. Take the OhioT1DM dataset (Marling & Bunescu, 2018) as an example, only a few thousands observations are available (Shi et al., 2020b). In these cases, in addition to a point estimate on a target policy's value, it is crucial to construct a confidence interval (CI) that quantifies the uncertainty of the value estimates.

This paper is concerned with the following question: *is it possible to develop a robust and efficient off-policy value estimator, and provide rigorous uncertainty quantification under practically feasible conditions?* We will give an affirmative answer to this question.

Overview of the OPE Literature. There is a growing literature for OPE. Existing works can be casted into as direct method (see e.g., Le et al., 2019; Shi et al., 2020c; Feng et al., 2020), importance sampling-based method (IS, Precup, 2000; Thomas et al., 2015b; Hanna et al., 2016; Liu et al., 2018; Nachum et al., 2019; Dai et al., 2020) and doubly robust method (Jiang & Li, 2016; Thomas & Brunskill, 2016; Farajtabar et al., 2018; Tang et al., 2019; Uehara et al., 2019; Kallus & Uehara, 2020; Jiang & Huang, 2020). Direct method derives the value estimates by learning the system transition matrix or the Q-function under the target policy. IS estimates the value by re-weighting the observed rewards with the density ratio of the target and behavior policies. Both direct method and IS have their own merits. In general, IS-type estimators might suffer from a large variance due to the use of the density ratio, whereas direct method might suffer from a large bias due to the potential misspecification of the model. Doubly robust methods combine both for more robust and efficient value evaluation.

Despite the popularity of developing a point estimate of a target policy's value, less attention has been paid to constructing its CI, which is the focus of this paper. Among those available, Thomas et al. (2015b) and Hanna et al. (2016) derived the CI by using bootstrap or concentration inequality applied to the stepwise IS estimator. These methods suffer from the curse of horizon (Liu et al., 2018), leading to very large CIs. Feng et al. (2020) applied the Hoeffding's inequality to derive the CI based on a kernel-based Q-function estimator. Similar to the direct method, their estimator might suffer from a large bias. Dai et al. (2020) reformulated the OPE problem using the generalized estimating equation approach and applied the empirical likelihood approach (see e.g., Owen, 2001) to CI estimation. They derived the CI by assuming the data transactions are i.i.d. However, observations in reinforcement learning are typically time-dependent. Directly applying the empirical likelihood method to weakly

[*]Equal contribution [1]Department of Statistics, London School of Economics and Political Science, London, United Kingdom [2]Department of Statistics, North Carolina State University, Raleigh, USA [3]Department of Economics, Massachusetts Institute of Technology, Cambridge, USA. Correspondence to: Rui Song <rsong@ncsu.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

dependent data would fail without further adjustment (Kitamura et al., 1997; Duchi et al., 2016). The resulting CI might not be valid. We discuss this in detail in Appendix D.

Recently, Kallus & Uehara (2019) made an important step forward for OPE, by developing a double reinforcement learning (DRL) estimator that achieves the semiparametric efficiency bound (see e.g., Tsiatis, 2007). Their method learns a Q-function and a marginalized density ratio and requires either one of the two estimators to be consistent. When both estimators converge at certain rates, DRL is asymptotically normal, based on which a Wald-type CI can be derived. However, these convergence rates might not be achievable in complicated RL tasks with high-dimensional state variables, resulting in an asymptotically biased value estimator and an invalid CI. See Section 2.2 for details.

Finally, we remark that our work is also related to a line of research on statistical inference in bandits (Van Der Laan & Lendle, 2014; Deshpande et al., 2018; Zhang et al., 2020; Hadad et al., 2021). However, these methods are not applicable to our setting.

Advances of the Proposed Method. Our proposal is built upon the DRL estimator to achieve sample efficiency. To derive a valid CI under weaker and practically more feasible conditions than DRL, we propose to learn a conditional density ratio estimator and develop a deeply-debiasing process that iteratively reduces the biases of the Q-function and value estimator. Debiasing brings additional robustness and flexibility. In a contextual bandit setting, our proposal shares similar spirits to the minimax optimal estimating procedure that uses higher order influence functions for learning the average treatment effects (see e.g., Robins et al., 2008; 2017; Mukherjee et al., 2017; Mackey et al., 2018). As such, the proposed method is:

- **robust** as the proposed value estimator is more robust than DRL and can converge to the true value in cases where neither the Q-function nor the marginalized density ratio estimator is consistent. More specifically, it is "triply-robust" and requires the Q-function, marginalized density ratio, or conditional density ratio estimator to be consistent. See Theorem 1 for a formal statement.

- **efficient** as we can show it achieves the semiparametric efficiency bound as DRL. This in turn implies that the proposed CI is tight. See Theorem 2 for details.

- **flexible** as it requires much weaker and practically more feasible conditions to achieve nominal coverage. Specifically, our procedure allows the Q-estimator and marginalized density ratio to converge at an arbitrary rate. See Theorem 3 for details.

2. Preliminaries

We first formulate the OPE problem. We next review the DRL method, as it is closely related to our proposal.

2.1. Off-Policy Evaluation

We assume the data in OPE follows a Markov Decision Process (MDP, Puterman, 2014) model defined by a tuple $(\mathcal{S}, \mathcal{A}, p, r, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $p : \mathcal{S}^2 \times \mathcal{A} \to [0, 1]$ is the Markov transition matrix that characterizes the system transitions, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, and $\gamma \in (0, 1)$ is a discounted factor that balances the immediate and future rewards. To simplify the presentation, we assume the state space is discrete. Meanwhile, the proposed method is equally applicable to continuous state space as well.

Let $\{(S_t, A_t, R_t)\}_{t\geq 0}$ denote a trajectory generated from the MDP model where (S_t, A_t, R_t) denotes the state-action-reward triplet at time t. Throughout this paper, we assume the following Markov assumption (MA) and the conditional mean independence assumption (CMIA) hold:

$$\mathbb{P}(S_{t+1} = s|\{S_j, A_j, R_j\}_{0\leq j\leq t}) = p(s; A_t, S_t), \text{ (MA)},$$
$$\mathbb{E}(R_t|S_t, A_t, \{S_j, A_j, R_j\}_{0\leq j<t}) = r(A_t, S_t), \text{ (CMIA)}.$$

These two assumptions guarantee the existence of an optimal *stationary* policy (see e.g., Puterman, 2014). Following a given stationary policy π, the agent will select action a with probability $\pi(a|s)$ at each decision time. The corresponding state value function and state-action value function (better known as the Q-function) are given as follows:

$$V^\pi(s) = \sum_{t=0}^{+\infty} \gamma^t \mathbb{E}^\pi(R_t|S_0 = s),$$
$$Q^\pi(a, s) = \sum_{t=0}^{+\infty} \gamma^t \mathbb{E}^\pi(R_t|A_0 = a, S_0 = s),$$

where the expectation $\mathbb{E}^\pi$ is defined by assuming the system follows the policy π.

The observed data consists of n i.i.d. trajectories, and can be summarized as $\{(S_{i,t}, A_{i,t}, R_{i,t}, S_{i,t+1})\}_{0\leq t<T_i, 1\leq i\leq n}$ where T_i denotes the termination time of the ith trajectory. Without loss of generality, we assume $T_1 = \cdots = T_n = T$ and the immediate rewards are uniformly bounded. We consider evaluating the value of a given *target policy* π with respect to a given reference distribution $\mathbb{G}$, defined as

$$\eta^\pi = \mathbb{E}_{s\sim\mathbb{G}} V^\pi(s).$$

In applications such as video games where a large number of trajectories are available, one may set $\mathbb{G}$ to the initial state distribution and approximate it by the empirical distribution

of $\{S_{i,0}\}_{1\le i\le n}$. In applications such as mobile health studies, the number of trajectories is limited. For instance, the OhioT1DM dataset contains data for six patients (trajectories) only. In these cases, $\mathbb{G}$ shall be manually specified. In this paper, we primarily focus on the latter case with a prespecified $\mathbb{G}$. Meanwhile, the proposed method is equally applicable to the former case as well.

2.2. Double Reinforcement Learning

We review the DRL estimator in this section. We first define the marginalized density ratio under the target policy π as

$$\omega^{\pi}(a,s)=\frac{(1-\gamma)\sum_{t=0}^{+\infty}\gamma^t p_t^{\pi}(a,s)}{p_{\infty}(a,s)}, \tag{1}$$

where $p_t^{\pi}(a,s)$ denotes the probability of $(A_t,S_t)=(a,s)$ following policy π with $S_0\sim\mathbb{G}$, and p_{∞} denotes the limiting distribution of the stochastic process $\{(A_t,S_t)\}_{t\ge 0}$. Such a marginalized density ratio plays a critical role in breaking the curse of horizon.

Let $\widehat{Q}$ and $\widehat{\omega}$ be some estimates for Q^{π} and ω^{π}, respectively. Kallus & Uehara (2019) proposed to construct the following estimating function for every i and t:

$$\psi_{i,t}\equiv\frac{1}{1-\gamma}\widehat{\omega}(A_{i,t},S_{i,t})\{R_{i,t}-\widehat{Q}(A_{i,t},S_{i,t})+\gamma\mathbb{E}_{a\sim\pi(\cdot|S_{i,t+1})}\widehat{Q}(a,S_{i,t+1})\}+\mathbb{E}_{s\sim\mathbb{G},a\sim\pi(\cdot|s)}\widehat{Q}(a,s). \tag{2}$$

The resulting value estimator is given by

$$\widehat{\eta}_{\text{DRL}}=\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=0}^{T-1}\psi_{i,t}.$$

One can show that $\widehat{\eta}_{\text{DRL}}$ is consistent when either $\widehat{Q}$ or $\widehat{\omega}$ is consistent. This is referred to as the doubly-robustness property. In addition, when both $\widehat{Q}$ and $\widehat{\omega}$ converge at a rate faster than $(nT)^{-1/4}$, $\sqrt{nT}(\widehat{\eta}_{\text{DRL}}-\eta^{\pi})$ converges weakly to a normal distribution with mean zero and variance

$$\frac{1}{(1-\gamma)^2}\mathbb{E}\left[\omega^{\pi}(A,S)\{R+\gamma V^{\pi}(S')-Q^{\pi}(A,S)\}\right]^2, \tag{3}$$

where the tuple (S,A,R,S') follows the limiting distribution of the process $\{(S_t,A_t,R_t,S_{t+1})\}_{t\ge 0}$. See Theorem 11 of Kallus & Uehara (2019) for a formal proof. A consistent estimator for (3) can be derived based on the observed data. A Wald-type CI for η^{π} can thus be constructed.

Moreover, it follows from Theorem 5 of Kallus & Uehara (2019) that (3) is the *semiparametric efficiency bound* for infinite-horizon OPE. Informally speaking, a semiparametric efficiency bound can be viewed as the nonparametric extension of the Cramer–Rao lower bound in parametric models (Bickel et al., 1993). It provides a lower bound of the asymptotic variance among all regular estimators (Van der Vaart, 2000). Many other OPE methods such as Liu et al. (2018), are statistically inefficient in that the variance of their value estimator is strictly larger than this bound. As such, CIs based on these methods are not tight.

2.3. Limitations of DRL

We end this section by discussing the limitations of DRL. As we commented in Section 2.2, the validity of the Wald-type CI based on DRL requires both nuisance function estimators to converge at a rate faster than $(nT)^{-1/4}$. When this assumption is violated, the resulting CI cannot achieve nominal coverage.

To elaborate this, we design a toy example with three states (denote by A, B and C) arranged on a circle. The agent can move either clockwise or counter-clockwise. The reward is 1 if the agent reaches state A and 0 otherwise. We set the behaviour policy to a random policy. The target policy is very close to the optimal one. We inject some random errors to the true Q-function and marginalized density ratio to construct the CI based on DRL. It can be seen from Figure 1 that DRL is valid when the nuisance estimators are $(nT)^{-1/2}$-consistent but fails when they are $(nT)^{-1/4}$- or $(nT)^{-1/6}$-consistent. See Appendix C for details.

We remark that the convergence rate assumption required by DRL is likely to be violated in complicated RL tasks with high-dimensional state variables. Take the Q-function estimator as an example. Suppose the true Q-function is Hölder smooth with exponent β and $\widehat{Q}$ is computed via the deep Q-network (Mnih et al., 2015) algorithm. Then similar to Theorem 4.4 of Fan et al. (2020), we can show that $\widehat{Q}$ converges at a rate of $(nT)^{-\beta/(2\beta+d)}$ where d denotes the dimension of the state. When $d\ge 2\beta$, it is immediate to see that the assumption on $\widehat{Q}$ is violated. Learning the marginalized density ratio is even more challenging than learning the Q-function. We expect that the convergence rate assumption on $\widehat{\omega}$ would be violated as well.

This motivates us to derive a valid CI under weaker and practically more feasible conditions. Our proposal requires to specify a hyper-parameter that determines the order of our value estimator. The larger this parameter, the weaker assumption our method requires. As an illustration, it can be seen from Figure 1 that our CI (denote by TR) achieves nominal coverage when the nuisance estimators are $(nT)^{-1/4}$- or even $(nT)^{-1/6}$-consistent.

3. Deeply-Debiased OPE

3.1. An Overview of Our Proposal

We first present an overview of our algorithm. Our procedure is composed of the following four steps, including data splitting, estimation of nuisance functions, debias iteration

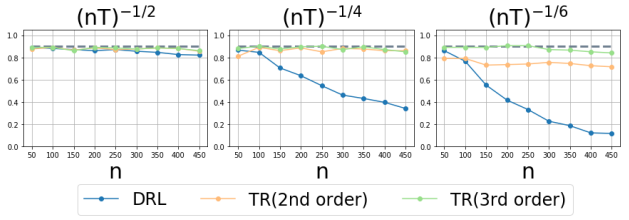

Figure 1. Empirical coverage probabilities for CIs based on DRL and the proposed triply-robust (TR) estimator, aggregated over 200 replications in the toy example. The nominal level is 90% and $\gamma = 0.95$. From left to right, we inject noises to the true Q-function and marginalized density ratio with standard errors proportional to $(nT)^{-1/2}$, $(nT)^{-1/4}$, and $(nT)^{-1/6}$, respectively. We vary the number of trajectories n and fix $T = 50$.

and construction of the CI.

Step 1. Data Splitting. We randomly divide the indices of all trajectories $\{1, \cdots, n\}$ into $\mathbb{K} \geq 2$ disjoint subsets. Denote the kth subset by $\mathbb{I}_k$ and let $\mathbb{I}_k^c = \{1, \cdots, n\} - \mathbb{I}_k$. Data splitting allows us to use one part of data ($\mathbb{I}_k^c$) to train RL models and the remaining part ($\mathbb{I}_k$) to do the estimation of the main parameter, i.e., η^π. We could aggregate the resulting estimates over different k to get full efficiency. This allows us to establish the limiting distribution of the value estimator under minimal conditions. Data splitting has been commonly used in statistics and machine learning (see e.g., Chernozhukov et al., 2017; Kallus & Uehara, 2019; Shi & Li, 2021).

Step 2. Estimation of Nuisance Functions. This step is to estimate three nuisance functions, including the Q-function Q^π, the marginalized density ratio ω^π, and a conditional density ratio τ^π. Several algorithms in the literature can be applied to learn Q^π and ω^π, e.g., Le et al. (2019); Liu et al. (2018); Kallus & Uehara (2019); Uehara et al. (2019). The conditional density ratio can be learned from the observed data in a similar fashion as ω^π. See Section 3.3 for more details. We use $\widehat{Q}_k$, $\widehat{\omega}_k$ and $\widehat{\tau}_k$ to denote the corresponding estimators, computed based on each data subset in $\mathbb{I}_k^c$.

Step 3. Debias Iteration. This step is the key to our proposal. It recursively reduces the biases of the initial Q-estimator, allowing us to derive a valid CI for the target value under weaker and more practically feasible conditions on the estimated nuisance functions. Specifically, our CI allows the nuisance function estimator to converge at arbitrary rates. See Section 3.2 for details.

Step 4. Construction of the CI. Based on the debiased Q-estimator obtained in Step 3, we construct our value estimate and obtain a consistent estimator for its variance. A Wald-type CI can thus be derived. See Section 3.4 for details.

In the following, we detail some major steps. We first introduce the debias iteration, as it contains the main idea of our proposal. We next detail Step 2 and Step 4.

3.2. Debias Iteration

3.2.1. The intuition for debias

To motivate the proposed debias iteration, let us take a deeper look at DRL. Note that the second term on the right-hand-side of (2) is a plug-in estimator of the value based on the initial Q-estimator. The first term corresponds to an augmentation term. The purpose of adding this term is to offer additional protection against potential model misspecification of the Q-function. The resulting estimator's consistency relies on either Q^π or ω^π to be correctly specified. As such, (2) can be understood as a de-biased version of the plug-in value estimator $\mathbb{E}_{s\sim\mathbb{G},a\sim\pi(\cdot|s)}\widehat{Q}(a,s)$.

Similarly, we can debias the initial Q-estimator $\widehat{Q}(a_0, s_0)$ for any (a_0, s_0). Toward that end, we introduce the conditional density ratio. Specifically, by setting $\mathbb{G}(\bullet)$ to a Dirac measure $\mathbb{I}(\bullet = s_0)$ and further conditioning on an initial action a_0, the marginalized density ratio in (1) becomes a conditional density ratio $\tau^\pi(a, s, a_0, s_0)$, defined as

$$\frac{(1-\gamma)\{\mathcal{I}(a = a_0, s = s_0) + \sum_{t=1}^{+\infty}\gamma^t p_t^\pi(a, s|a_0, s_0)\}}{p_\infty(a, s)},$$

where $p_t^\pi(a, s|a_0, s_0)$ denotes the probability of $(A_t, S_t) = (a, s)$ following policy π conditional on the event that $\{A_0 = a_0, S_0 = s_0\}$, and $\mathcal{I}(\cdot)$ denotes the indicator function. By definition, the numerator corresponds to the discounted conditional visitation probability following π given that the initial state-action pair equals (s_0, a_0). In addition, one can show that $\omega^\pi(a, s) = \mathbb{E}_{s_0\sim\mathbb{G},a_0\sim\pi(\cdot|s_0)}\tau^\pi(a, s, a_0, s_0)$.

By replacing $\widehat{\omega}_k$ in (2) with some estimated conditional density ratio $\widehat{\tau}_k$, we obtain the following estimation function

$$\mathcal{D}_k^{(i,t)}Q(a,s) = Q(a,s) + \frac{1}{1-\gamma}\widehat{\tau}_k(A_{i,t}, S_{i,t}, a, s) \times\{R_{i,t} + \gamma\mathbb{E}_{a'\sim\pi(\cdot|S_{i,t+1})}Q(a', S_{i,t+1}) - Q(A_{i,t}, S_{i,t})\}, \quad (4)$$

for any Q. Here, we refer $\mathcal{D}_k^{(i,t)}$ as the *individual debiasing operator*, since it debiases any Q based on an individual data tuple $(S_{i,t}, A_{i,t}, R_{i,t}, S_{i,t+1})$.

Similar to (2), the augmentation term in (4) is to offer protection against potential model misspecification of the Q-function. As such, $\mathcal{D}_k^{(i,t)}Q(a,s)$ is unbiased to $Q^\pi(a,s)$ whenever $Q = Q^\pi$ or $\widehat{\tau}_k = \tau^\pi$.

3.2.2. The two-step debias iteration

Based on the above discussions, a debiased version of the Q-estimator is given by averaging $\mathcal{D}_k^{(i,t)}\widehat{Q}_k$ over the data

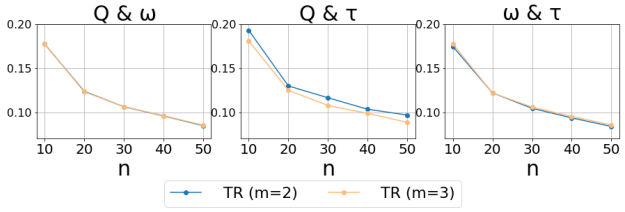

Figure 2. Root mean squared error (RMSE) of the proposed estimators in the toy example, computed over 200 replications. From left to right, we inject non-degenerate noises into Q^π and ω^π, Q^π and τ^π, ω^π and τ^π, respectively. It can be seen that the RMSE decays as the sample size increases, when one of the three models is correctly specified.

tuples in $\mathbb{I}_k$, i.e.,

$$\widehat{Q}_k^{(2)} = \frac{1}{|\mathbb{I}_k|T} \sum_{i\in\mathbb{I}_k} \sum_{0\le t<T} \mathcal{D}_k^{(i,t)} \widehat{Q}_k.$$

The bias of this estimator will decay at a faster rate than the initial Q-estimator $\widehat{Q}_k$, as shown in the following lemma.

Lemma 1 *For any k, suppose $\widehat{Q}_k$ and $\widehat{\tau}_k$ converge in L_2-norm to Q^π and τ^π at a rate of $(nT)^{-\alpha_1}$ and $(nT)^{-\alpha_2}$, respectively. With weakly dependent data (see Condition (A1) in Section 4 in detail), we have*

$$\mathbb{E}_{(a,s)\sim p_\infty} |\mathbb{E}\widehat{Q}_k^{(2)}(a,s) - Q(a,s)| = O\{(nT)^{-(\alpha_1+\alpha_2)}\}.$$

To save space, we defer the detailed definition of L_2-norm convergence rate in Appendix A.2. Suppose the square bias and variance of $\widehat{Q}_k$ are of the same order. Then we can show that the aggregated bias $\mathbb{E}_{(a,s)\sim p_\infty}|\mathbb{E}\widehat{Q}_k(a,s) - Q(a,s)|$ decays at a rate of $(nT)^{-\alpha_1}$. Consequently, Lemma 1 implies that the bias of $\widehat{Q}_k^{(2)}$ decays faster than $\widehat{Q}_k$.

To illustrate the usefulness of the debiased Q-estimator $\widehat{Q}_k^{(2)}$, we propose to construct an estimating function $\psi_{i,t}^{(2)}$ for any $(i,t) \in \mathbb{I}_k$ by replacing $\widehat{Q}$ in (2) with $\widehat{Q}_k^{(2)}$. This yields our second-order estimator

$$\widehat{\eta}_{\text{TR}}^{(2)} = (nT)^{-1} \sum_{i,t} \psi_{i,t}^{(2)}.$$

As we will show in Theorem 1, the proposed estimator $\widehat{\eta}_{\text{TR}}^{(2)}$ converges to the true value when one model for Q^π, ω^π or τ^π is correctly specified. As such, it is *triply-robust*. See Figure 2 as an illustration. In addition, similar to Lemma 1, the bias of $\widehat{\eta}_{\text{TR}}^{(2)}$ decays at a faster rate than the DRL estimator. Specifically, we have the following results.

Lemma 2 *Suppose the conditions in Lemma 1 hold and $\widehat{\omega}_k$ converges in L_2-norm to ω^k at a rate of $(NT)^{-\alpha_3}$ for any k. Let $\alpha = \min(1, \alpha_1+\alpha_2+\alpha_3)$. Then*

$$|\mathbb{E}\widehat{\eta}_{TR}^{(2)} - \eta^\pi| = O\{(nT)^{-\alpha}\}.$$

In contrast, the bias of the DRL estimator decays at a rate of $(nT)^{-\alpha_1-\alpha_3}$. To ensure the resulting CI achieves valid coverage, we require the bias to decay at a rate faster than its variance which is typically of the order $O\{(nT)^{-1/2}\}$. As such, DRL requires $\min(\alpha_1,\alpha_3) > 1/4$ whereas our second-order triply robust estimator relaxes this condition by requiring $\min(\alpha_1,\alpha_2,\alpha_3) > 1/6$, as shown in Figure 1.

3.2.3. The m-step debias iteration

To further relax the convergence rate requirement, we can iteratively debias the Q-estimator to construct higher-order value estimates. Specifically, for any order $m \ge 2$, we iteratively apply the debiasing operator to the initial Q-estimator $m-1$ times and average over all individual tuples, leading to the following estimator,

$$\widehat{Q}_k^{(m)} = \binom{|\mathbb{I}_k|T}{(m-1)}^{-1} \sum \mathcal{D}_k^{(i_1,t_1)} \cdots \mathcal{D}_k^{(i_{m-1},t_{m-1})} \widehat{Q}_k.$$

Here, the sum is taken over all possible combinations of disjoint tuples $(i_1,t_1),(i_2,t_2),\cdots,(i_{m-1},t_{m-1})$ in the set $\{(i,t) : i \in \mathbb{I}_k, 0 \le t < T\}$. Note that the definition involves repeated compositions of debiasing operator. For $m=3$, we present the detailed form in the appendix. In general, $\widehat{Q}_k^{(m)}(a,s)$ corresponds to an order $(m-1)$ U-statistic (see e.g., Lee, 2019) for any (a,s). The resulting value estimator $\widehat{\eta}_{\text{TR}}^{(m)}$ is given by $(nT)^{-1}\sum_{i,t}\psi_{i,t}^{(m)}$ where for any $(i,t)\in\mathbb{I}_k$, the estimating function $\psi_{i,t}^{(m)}$ is obtained by replacing $\widehat{Q}$ in (2) with $\widehat{Q}_k^{(m)}$.

We make a few remarks. First, when $m=1$, $\widehat{Q}_k^{(m)}$ corresponds to the initial Q-estimator. As such, the proposed estimator reduces to the DRL estimator. When $m=2$, the definition here is consistent to the second-order triply-robust estimator.

Second, for large m, calculating $\widehat{Q}_k^{(m)}$ is computationally intensive. In practice, we may approximate it using the incomplete U-statistics (Lee, 2019; Chen et al., 2019) to facilitate the computation. For instance, to calculate $\widehat{Q}_{-k}^{(3)}(a,s)$, we could approximate it by averaging $\widehat{\mathcal{D}}_k^{(i_1,t_1)}\widehat{\mathcal{D}}_k^{(i_2,t_2)}\widehat{Q}_k(a,s)$ over M pairs sampled from the set $\{(i_1,t_1,i_2,t_2) : i_1,i_2 \in \mathbb{I}_k, (i_1,t_1) \ne (i_2,t_2)\}$. We require M to diverge with nT such that the approximation error is asymptotically negligible. The computational complexity of our whole algorithm is analyzed in Appendix C.4 in the supplement.

Third, the bias of the Q-estimator and that of the resulting value decrease as the order m increases. Specifically, we have the following results.

Lemma 3 *Suppose the conditions in Lemma 2 hold. Let* $\alpha_* = \alpha_1+(m-1)\alpha_2$ *and* $\alpha = \min(1, \alpha_1+(m-1)\alpha_2+\alpha_3)$. *Then* $\mathbb{E}_{(a,s)\sim p_\infty}|\mathbb{E}\widehat{Q}_k^{(m)}(a,s) - Q(a,s)| = O\{(nT)^{-(\alpha_*)}\}$ *and* $|\mathbb{E}\widehat{\eta}_{TR}^{(m)} - \eta^\pi| = O\{(NT)^{-\alpha}\}$.

To ensure $\alpha < 1/2$, it suffices to require $\alpha_1 + (m-1)\alpha_2 + \alpha_3 > 1/2$. As long as $\alpha_1, \alpha_2, \alpha_3 > 0$, there exists some m that satisfies this condition. As such, the resulting bias decays faster than $(nT)^{-1/2}$. This yields the flexibility of our estimator as it allows the nuisance function estimator to converge at an arbitrary rate. When $m = 2$, Lemmas 1 and 2 are directly implied by Lemma 3.

3.3. Learning the Nuisance Functions

This step is to estimate the nuisance functions used in our algorithm, including Q^π, ω^π, and τ^π, based on each data subset $\mathbb{I}_k^c$, for $k = 1, \cdots, \mathbb{K}$.

The Q-function. There are multiple learning methods available to produce an initial estimator for Q^π. We employ the fitted Q-evaluation method (Le et al., 2019) in our implementation. Based on the Bellman equation for Q^π (see Equation (4.6), Sutton & Barto, 2018), it iteratively solves the following optimization problem,

$$\widehat{Q}^\ell = \arg\min_Q \sum_{i\in\mathbb{I}_k^c}\sum_{t<T}\{\gamma\mathbb{E}_{a'\sim\pi(\cdot|S_{i,t+1})}\widehat{Q}^{\ell-1}(a', S_{i,t+1}) + R_{i,t} - Q(A_{i,t}, S_{i,t})\}^2,$$

for $\ell = 1, 2, \cdots$, until convergence.

We remark that the above optimization problem can be conveniently solved via supervised learning algorithms. In our experiments, we use random forest (Breiman, 2001) to estimate Q^π.

The Marginalized Density Ratio. We next discuss the method for learning ω^π. In our implementation, we employ the method of Uehara et al. (2019). The following observation forms the basis of the method: when the process $\{(S_t, A_t)\}_{t\geq 0}$ is stationary, ω^π satisfies the equation $\mathbb{E}L(\omega^\pi, f) = 0$ for any function f, where $L(\omega^\pi, f)$ equals

$$\Big[\mathbb{E}_{a\sim\pi(\cdot|S_{t+1})}\{\omega^\pi(A_t, S_t)(\gamma f(a, S_{t+1}) - f(A_t, S_t))\} + (1-\gamma)\mathbb{E}_{s\sim\mathbb{G}, a\sim\pi(\cdot|s)}f(a, s). \tag{5}$$

As such, ω^π can be learned by solving the following mini-max problem,

$$\arg\min_{\omega\in\Omega}\sup_{f\in\mathcal{F}}\{\mathbb{E}L(\omega, f)\}^2, \tag{6}$$

for some function classes Ω and $\mathcal{F}$. The expectation in (6) is approximated by the sample mean. To simplify the calculation, we choose $\mathcal{F}$ to be a reproducing kernel Hilbert space (RKHS). This yields a closed form expression for $\sup_{f\in\mathcal{F}}\{\mathbb{E}L(\omega, f)\}^2$. Consequently, ω^π can be learned by solving the outer minimization via stochastic gradient descent. To save space, we defer the details to Appendix B.2 in the supplementary article.

The Conditional Density Ratio. Finally, we develop a method to learn τ^π based on the observed data. Note that τ^π can be viewed as a version of ω^π by conditioning on the initial state-action pair. Similar to (5), we have

$$\mathbb{E}\Big[\mathbb{E}_{a\sim\pi(\cdot|S_{t+1})}\tau^\pi(A_t, S_t, a_0, s_0)\{\gamma g(a, S_{t+1}) - g(A_t, S_t)\}\Big] + (1-\gamma)g(a_0, s_0) = 0,$$

for any g and state-action pair (a_0, s_0), or equivalently,

$$\mathbb{E}\Big[\mathbb{E}_{a\sim\pi(\cdot|S_{t+1})}\tau^\pi(A_t, S_t, a_0, s_0)\{\gamma f(a, S_{t+1}, a_0, s_0) - f(A_t, S_t, a_0, s_0)\}\Big] + (1-\gamma)f(a_0, s_0, a_0, s_0) = 0, \tag{7}$$

for any function f and (a_0, s_0). Integrating (a_0, s_0) on the left-hand-side of (7) with respect to the stationary state-action distribution p_∞, we obtain the following lemma.

Lemma 4 *Suppose the process* $\{(A_t, S_t)\}_{t\geq 0}$ *is strictly stationary. For any function* f, τ^π *satisfies the equation* $h(\tau^\pi, f) = 0$ *where* $h(\tau^\pi, f)$ *is given by*

$$\mathbb{E}\Big[(1-\gamma)f(A_{i_1,t_1}, S_{i_1,t_1}, A_{i_1,t_1}, S_{i_1,t_1}) - \tau^\pi(A_{i_2,t_2}, S_{i_2,t_2}, A_{i_1,t_1}, S_{i_1,t_1})\times \{f(A_{i_2,t_2}, S_{i_2,t_2}, A_{i_1,t_1}, S_{i_1,t_1}) - \gamma\mathbb{E}_{a\sim\pi(\cdot|S_{i_2,t_2+1})}f(S_{i_2,t_2+1}, a; S_{i_1,t_1}, A_{i_1,t_1})\}\Big],$$

for any $i_1 \neq i_2$ *such that* $(S_{i_1,t_1}, A_{i_1,t_1}, S_{i_1,t_1+1})$ *and* $(S_{i_2,t_2}, A_{i_2,t_2}, S_{i_2,t_2+1})$ *are independent.*

Similar to Lemma 15 of Kallus & Uehara (2019), we can also show that τ^π is the only function that satisfies Lemma 4. Motivated by this lemma, τ^π can be learned by solving the following mini-max optimization problem

$$\arg\min_{\tau\in\mathcal{T}}\sup_{f\in\mathcal{F}} h^2(\tau, f), \tag{8}$$

for some function classes $\mathcal{T}$ and $\mathcal{F}$. For any τ and f, we estimate $h(\tau, f)$ based on the observed data. Setting $\mathcal{F}$ to an RKHS and $\mathcal{T}$ to a class of deep neural networks, the above optimization can be solved in a similar fashion as (6). We defer the details to Appendix B.3 to save space.

3.4. Construction of the CI

In this step, we construct a CI based on $\widehat{\eta}_{\text{TR}}^{(m)}$. Specifically, under mild assumptions, the asymptotic variance

of $\sqrt{nT}\widehat{\eta}_{\text{TR}}^{(m)}$ can be consistently estimated by the sampling variance estimator of $\{\psi_{i,t}^{(m)}\}_{i,t}$ (denote by $\{\widehat{\sigma}^{(m)}\}^2$). For a given significance level α, the corresponding two-sided CI is given by $[\widehat{\eta}_{\text{TR}}^{(m)} - z_{\alpha/2}(nT)^{-1/2}\widehat{\sigma}^{(m)}, \widehat{\eta}_{\text{TR}}^{(m)} + z_{\alpha/2}(nT)^{-1/2}\widehat{\sigma}^{(m)}]$ where z_α corresponds to the upper αth quantile of a standard normal random variable.

4. Robustness, Efficiency and Flexibility

We first summarize our results. Theorem 1 establishes the triply-robust property of our value estimator $\widehat{\eta}^{(m)}$. Theorem 2 shows the asymptotic variance of $\widehat{\eta}^{(m)}$ achieves the semiparametric efficiency bound (3). As such, our estimator is sample efficient. Theorem 3 implies that our CI achieves nominal coverage under weaker and much practically feasible conditions than DRL. All of our theoretical guarantees are derived under the asymptotic framework that requires either the number of trajectories n or the number of decision points T per trajectory to diverge to infinity. Results of this type provide useful theoretical guarantees for different types of applications, and are referred as bidirectional theories.

We next introduce some conditions.

(A1) The process $\{(S_t, A_t, R_t)\}_{t\geq 0}$ is strictly stationary and exponentially β-mixing (see e.g., Bradley, 2005, for a detailed explanation of this definition).

(A2) For any k, $\widehat{Q}_k$, $\widehat{\tau}_k$ and $\widehat{\omega}_k$ converge in L_2-norm to Q^π, τ^π and ω^π at a rate of $(nT)^{-\alpha_1}$, $(nT)^{-\alpha_2}$ and $(nT)^{-\alpha_3}$ for any α_1, α_2 and $\alpha_3 > 0$, respectively.

(A3) τ^π and ω^π are uniformly bounded away from infinity.

Condition (A1) allows the data observations to be weakly dependent. When the behavior policy is not history-dependent, the process $\{(S_t, A_t, R_t)\}_{t\geq 0}$ forms a Markov chain. The exponential β-mixing condition is automatically satisfied when the Markov chain is geometrically ergodic (see Theorem 3.7 of Bradley, 2005). Geometric ergodicity is less restrictive than those imposed in the existing reinforcement learning literature that requires observations to be independent (see e.g., Dai et al., 2020) or to follow a uniform-ergodic Markov chain (see e.g., Bhandari et al., 2018; Zou et al., 2019). We also remark that the stationarity assumption in (A1) is assumed for convenience, since the Markov chain will eventually reach stationarity.

Condition (A2) characterizes the theoretical requirements on the nuisance function estimators. This assumption is mild as we require these estimators to converge at any rate. When using kernels or neural networks for function approximation, the corresponding convergence rates of $\widehat{Q}_k$ and $\widehat{\omega}_k$ are provided in Fan et al. (2020); Liao et al. (2020). The convergence rate for $\widehat{\tau}_k$ can be similarly derived as $\widehat{\omega}_k$.

Condition (A3) essentially requires that any state-action pair supported by the density function $(1-\gamma)\sum_{t\geq 0}\gamma^t p_t^\pi$ is supported by the stationary behavior density function as well. This assumption is similar to the sequential overlap condition imposed by Kallus & Uehara (2020).

Theorem 1 (Robustness) *Suppose (A1) and (A3) hold, and $\widehat{Q}_k$, $\widehat{\tau}_k$, $\widehat{\omega}_k$ are uniformly bounded away from infinity almost surely. Then for any m, as either n or T diverges to infinity, our value estimator $\widehat{\eta}_{TR}^{(m)}$ is consistent when $\widehat{Q}_k$, $\widehat{\tau}_k$ or $\widehat{\omega}_k$ converges in L_2-norm to Q^π, τ^π or ω^π for any k.*

Theorem 1 does not rely on Condition (A2). It only requires one of the three nuisance estimators to converge. As such, it is more robust than existing doubly-robust estimators.

Theorem 2 (Efficiency) *Suppose (A1) and (A2) hold, and $\widehat{Q}_k$, $\widehat{\tau}_k$, $\widehat{\omega}_k$, τ^π, ω^π are uniformly bounded away from infinity almost surely. Then for any m, as either n or T approaches infinity, $\sqrt{nT}(\widehat{\eta}_{TR}^{(m)} - \mathbb{E}\widehat{\eta}_{TR}^{(m)}) \xrightarrow{d} N(0, \sigma^2)$ where σ^2 corresponds to the efficiency bound in (3).*

We make some remarks. In the proof of Theorem 2, we show that $\widehat{\eta}_{\text{TR}}^{(m)}$ is asymptotically equivalent to an mth order U-statistic. According to the Hoeffding decomposition (Hoeffding, 1948), we can decompose the U-statistic into the sum $\eta^\pi + \sum_{j=1}^m \widehat{\eta}_j$, where η^π is the main effect term that corresponds to the asymptotic mean of the value estimator, $\widehat{\eta}_1$ is the first-order term

$$\frac{1}{nT(1-\gamma)}\sum_{i=1}^{n}\sum_{t=0}^{T-1}\omega^\pi(A_{i,t}, S_{i,t})\{R_{i,t} + \gamma\mathbb{E}_{a\sim\pi(\cdot|S_{i,t+1})}Q^\pi(a, S_{i,t+1}) - Q^\pi(A_{i,t}, S_{i,t})\},$$

and $\widehat{\eta}_j$ corresponds to a jth order degenerate U-statistic for any $j \geq 2$. See Part 3 of the proof of Theorem 2 for details. Note that the DRL estimator is asymptotically equivalent to $\eta^\pi + \widehat{\eta}_1$. Under (A1), these $\widehat{\eta}_j$s are asymptotically uncorrelated. As such, the variance of our estimator is asymptotically equivalent to

$$\sum_{j=1}^{m}\text{Var}(\widehat{\eta}_j) = \sum_{j=1}^{m}\binom{nT}{j}^{-1}\sigma_j^2,$$

where σ_j^2s are bounded. When $j = 1$, we have $\sigma_j^2 = \sigma^2$. For $j \geq 2$, $\text{Var}(\widehat{\eta}_j)$ decays at a faster rate than $\text{Var}(\widehat{\eta}_1) = \sigma^2(nT)^{-1}$. As such, the variance of our estimator is asymptotically equivalent to that of DRL.

However, in finite sample, the variance of the proposed estimator is strictly larger than DRL, due to the presence of high-order variance terms. This is consistent with our experiment results (see Section 5) where we find the proposed CI is usually slightly wider than that based on DRL. This reflects a bias-variance trade-off. Specifically, our procedure alleviates the bias of the DRL estimator to obtain valid

uncertainty quantification. The resulting estimator would have a strictly larger variance than DRL in finite samples, although the difference is asymptotically negligible. We also remark that in interval estimation, the first priority is to ensure the CI has nominal coverage. This requires an estimator's bias to decay faster than its variance. The second priority is to shorten the length of CI (the variance of the estimator) if possible. In that sense, variance is less significant than bias.

Theorem 3 (Flexibility) *Suppose the conditions in Theorem 2 hold. Then as long as m satisfies $\alpha_1 + (m-1)\alpha_2 + \alpha_3 > 1/2$, the proposed CI achieves nominal coverage.*

Theorem 3 implies that our CI allows the nuisance functions to diverge at an arbitrary rate for sufficiently large m.

5. Experiments

In this section, we evaluate the empirical performance of our method using two synthetic datasets: CartPole from the OpenAI Gym environment (Brockman et al., 2016) and a simulation environment (referred to as Diabetes) to simulate the OhioT1DM data (Shi et al., 2020b). In the second environment, the goal is to learn an optimal policy as a function of patients' time-varying covariates to improve their health status. In both settings, following Uehara et al. (2019), we first learn a near-optimal policy as the target policy, and then apply softmax on its Q-function divided by a temperature parameter τ to set the action probabilities to define a behaviour policy. A larger τ implies a larger difference between the behaviour policy and the target policy.

We denote the proposed method as TR and present results with $m = 2$ and 3. The choice of m represents a trade-off. In theory, m shall be as large as possible to guarantee the validity of our CI. Yet, the computation complexity increases exponentially in m. In our experiments, we find that setting $m = 3$ yields satisfactory performance in general.

For point estimation, we compare the bias and RMSE of our method with DRL and the estimator computed via fitted-Q evaluation (FQE). For interval estimation, we compare the proposed CI with several competing baselines, including CoinDICE (Dai et al., 2020), stepwise IS-based estimator with bootstrapping (Thomas et al., 2015a), stepwise IS-based estimator with Bernstein inequality (Thomas et al., 2015b), and the CI based on DRL. For each method, we report the empirical coverage probability and the average length of the constructed CI.

We set $T = 300$ and $\gamma = 0.98$ for CartPole, and $T = 200$ and $\gamma = 0.95$ for Diabetes. For both environments, we vary the number of trajectories n and the temperature τ to design different settings. Results are aggregated over 200 replications. Note that FQE and DR share the same subroutines with TR, and hence the same hyper-parameters are used. More details about the environments and the implementations can be found in Section C of the supplement.

The results for CartPole and Diabetes are depicted in Figures 3 and 4, respectively. We summarize our findings as follows. In terms of interval estimation, first, the proposed CI achieves nominal coverage in all cases, whereas the CI based on DRL fails to cover the true value. This demonstrates that the proposed method is more robust than DRL. In addition, the average length of our CI is slightly larger than that of DRL in all cases. This reflects the bias-variance tradeoff we detailed in Section 4. Second, CoinDice yields the narrowest CI. However, its empirical coverage probability is well below the nominal level in all cases. As we have commented in the introduction, this is due to that their method requires i.i.d. observations and would fail with weakly dependent data. Please refer to Appendix D for details. Third, the stepwise IS-based estimators suffer from the curse of horizon. The lengths of the resulting CIs are much larger than ours. Moreover, the CI based on bootstrapping the stepwise IS-estimator fails to achieve nominal coverage. This is because the standard bootstrap method is not valid with weakly dependent data.

In terms of point estimation, TR yields smaller bias than DRL in all cases. FQE suffers from the largest bias among the three methods. The RMSEs of DRL and TR are comparable and generally smaller than that of FQE. This demonstrates the efficiency of the proposed estimator.

6. Discussion

6.1. Order Selection

In this paper, we develop a deeply-debiased procedure for off-policy interval estimation. Our proposal relies on the specification of m, the number of the debias iteration. The choice of m represents a trade-off. In theory, m shall be as large as possible to reduce the bias of the value estimator and guarantee the validity of the resulting CI. Yet, the variance of the value estimator and the computation of our procedure increase with m. In the statistics literature, Lepski's method is a data-adaptive procedure for identifying optimal tuning parameter where cross-validation is difficult to implement, as in our setup (see e.g., Su et al., 2020). It can be naturally coupled with the proposed method for order selection, to balance the bias-variance trade-off. Practical version of Lepski's method was developed using bootstrap in Chernozhukov et al. (2014). This idea is worthwhile to explore and we leave it for future research.

6.2. Nonasymptotic Confidence Bound

Non-asymptotic confidence bound is typically obtained by applying concentration inequalities (e.g., Hoeffding's in-

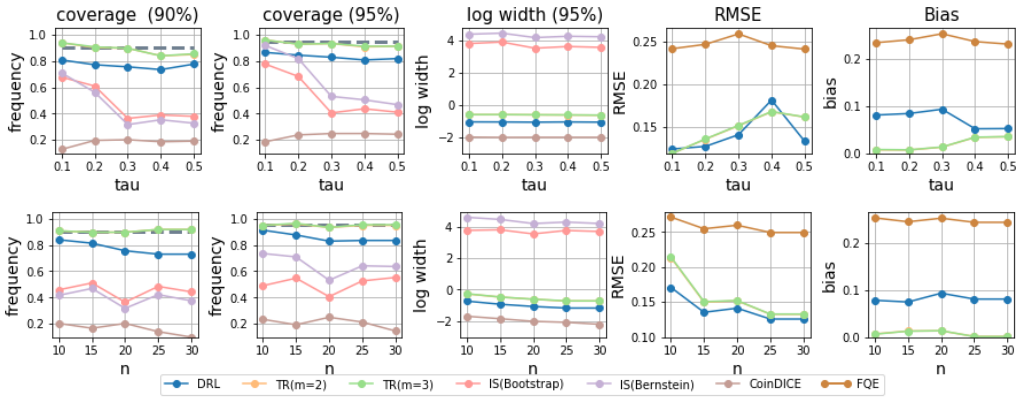

Figure 3. Results for Cartpole. We fix $n = 20$ and vary τ in the upper subplots, and fix $\tau = 0.3$ and vary n in the lower subplots. The subplots from left to right are about the coverage frequency with $\alpha = 0.9$, the coverage frequency with $\alpha = 0.95$, the mean log width of CIs with $\alpha = 0.95$, the RMSE of value estimates, and the bias of value estimates, respectively. The yellow line (TR, $m = 2$) and green line (TR, $m = 3$) are largely overlapped.

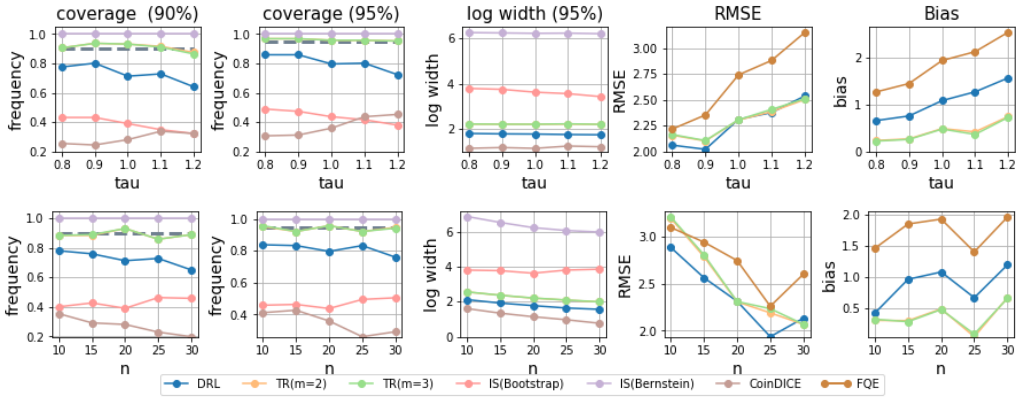

Figure 4. Results for Diabetes. We fix $n = 20$ and vary τ in the upper subplots, and fix $\tau = 1.0$ and vary n in the lower subplots. Same legend as Figure 3. The yellow line (TR, $m = 2$) and green line (TR, $m = 3$) are largely overlapped.

equality or Bernstein inequality Van Der Vaart & Wellner, 1996) to a sum of uncorrelated variables. In our setup, the proposed estimator is a U-statistic. We could apply concentration inequalities to U-statistics (see e.g., Feng et al., 2020) to derive the confidence bound. Alternatively, we may apply self-normalized moderate deviation inequalities (Peña et al., 2008) to derive the non-asymptotic bound. The resulting confidence bound will be wider than the proposed CI. However, it is valid even with small sample size.

6.3. Hardness of Learning of τ^π

Learning τ^π could be much challenging than ω^π. In our current numerical experiments, all the state variables are continuous and it is challenging to obtain the ground truth of the conditional density ratio which involves estimation of a high-dimensional conditional density. As such, we did not investigate the goodness-of-fit of the proposed estimator for τ^π. It would be practically interesting to explore the optimal neural network structure to approximate τ^π and investigate the finite-sample rate of convergence of our estimator. However, this is beyond the scope of the current paper. We leave it for future research.

6.4. Extension to Exploration

Finally, we remark that based on the proposed debiased Q-estimator, a two-sided CI can be similarly to quantify its uncertainty. It allows us to follow the "optimism in the face of uncertainty" principle for online exploration. This is another topic that warrants future investigation.

Acknowledgements

Shi's research was partially supported by the LSE New Research Support Fund. The authors thank the meta reviewer, and the reviewers for their constructive comments, which have led to a significant improvement of the earlier version of this article.

References

Andrychowicz, O. M., Baker, B., Chociej, M., Jozefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A., et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

Bhandari, J., Russo, D., and Singal, R. A finite time analysis of temporal difference learning with linear function approximation. *arXiv preprint arXiv:1806.02450*, 2018.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

Bradley, R. C. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

Breiman, L. Random forests. *Machine learning*, 45(1): 5–32, 2001.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Chen, X., Kato, K., et al. Randomized incomplete u-statistics in high dimensions. *Annals of Statistics*, 47 (6):3127–3156, 2019.

Chernozhukov, V., Chetverikov, D., Kato, K., et al. Anti-concentration and honest, adaptive confidence bands. *Annals of Statistics*, 42(5):1787–1818, 2014.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, 2017.

Dai, B., Nachum, O., Chow, Y., Li, L., Szepesvari, C., and Schuurmans, D. Coindice: Off-policy confidence interval estimation. *Advances in neural information processing systems*, 33, 2020.

Deshpande, Y., Mackey, L., Syrgkanis, V., and Taddy, M. Accurate inference for adaptive linear models. In *International Conference on Machine Learning*, pp. 1194–1203. PMLR, 2018.

Duchi, J., Glynn, P., and Namkoong, H. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.

Fan, J., Wang, Z., Xie, Y., and Yang, Z. A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pp. 486–489. PMLR, 2020.

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

Feng, Y., Ren, T., Tang, Z., and Liu, Q. Accountable off-policy evaluation with kernel bellman statistics. *arXiv preprint arXiv:2008.06668*, 2020.

Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., and Athey, S. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, 118(15), 2021.

Hanna, J. P., Stone, P., and Niekum, S. Bootstrapping with models: Confidence intervals for off-policy evaluation. *arXiv preprint arXiv:1606.06126*, 2016.

Hoeffding, W. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, pp. 293–325, 1948.

Jiang, N. and Huang, J. Minimax value interval for off-policy evaluation and policy optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661. PMLR, 2016.

Kallus, N. and Uehara, M. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *arXiv preprint arXiv:1909.05850*, 2019.

Kallus, N. and Uehara, M. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167): 1–63, 2020.

Kitamura, Y. et al. Empirical likelihood methods with weakly dependent processes. *The Annals of Statistics*, 25(5):2084–2102, 1997.

Le, H. M., Voloshin, C., and Yue, Y. Batch policy learning under constraints. *arXiv preprint arXiv:1903.08738*, 2019.

Lee, A. J. *U-statistics: Theory and Practice*. Routledge, 2019.

Liao, P., Qi, Z., and Murphy, S. Batch policy learning in average reward markov decision processes. *arXiv preprint arXiv:2007.11771*, 2020.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pp. 5356–5366, 2018.

Luedtke, A. R. and van der Laan, M. J. Evaluating the impact of treating the optimal subgroup. *Statistical methods in medical research*, 26(4):1630–1640, 2017.

Mackey, L., Syrgkanis, V., and Zadik, I. Orthogonal machine learning: Power and limitations. In *International Conference on Machine Learning*, pp. 3375–3383. PMLR, 2018.

Marling, C. and Bunescu, R. C. The ohiot1dm dataset for blood glucose level prediction. In *KHD@ IJCAI*, pp. 60–63, 2018.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

Mukherjee, R., Newey, W. K., and Robins, J. M. Semiparametric efficient empirical higher order influence function estimators. *arXiv preprint arXiv:1705.07577*, 2017.

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

Nachum, O., Chow, Y., Dai, B., and Li, L. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, pp. 2318–2328, 2019.

Owen, A. B. *Empirical likelihood*. CRC press, 2001.

Peña, V. H., Lai, T. L., and Shao, Q.-M. *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2008.

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

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Robins, J., Li, L., Tchetgen, E., van der Vaart, A., et al. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, pp. 335–421. Institute of Mathematical Statistics, 2008.

Robins, J. M., Li, L., Mukherjee, R., Tchetgen, E. T., van der Vaart, A., et al. Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics*, 45(5):1951–1987, 2017.

Sallab, A. E., Abdou, M., Perot, E., and Yogamani, S. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 2017(19):70–76, 2017.

Shi, C. and Li, L. Testing mediation effects using logic of boolean matrices. *Journal of the American Statistical Association*, pp. accepted, 2021.

Shi, C., Lu, W., and Song, R. Breaking the curse of nonregularity with subagging—inference of the mean outcome under optimal treatment regimes. *Journal of Machine Learning Research*, 21(176):1–67, 2020a.

Shi, C., Wan, R., Song, R., Lu, W., and Leng, L. Does the markov decision process fit the data: testing for the markov property in sequential decision making. In *International Conference on Machine Learning*, pp. 8807–8817. PMLR, 2020b.

Shi, C., Zhang, S., Lu, W., and Song, R. Statistical inference of the value function for reinforcement learning in infinite horizon settings. *arXiv preprint arXiv:2001.04515*, 2020c.

Su, Y., Srinath, P., and Krishnamurthy, A. Adaptive estimator selection for off-policy evaluation. In *International Conference on Machine Learning*, pp. 9196–9205. PMLR, 2020.

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

Tang, Z., Feng, Y., Li, L., Zhou, D., and Liu, Q. Doubly robust bias reduction in infinite horizon off-policy estimation. In *International Conference on Learning Representations*, 2019.

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

Thomas, P., Theocharous, G., and Ghavamzadeh, M. High confidence policy improvement. In *International Conference on Machine Learning*, pp. 2380–2388, 2015a.

Thomas, P. S., Theocharous, G., and Ghavamzadeh, M. High-confidence off-policy evaluation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015b.

Tsiatis, A. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

Uehara, M., Huang, J., and Jiang, N. Minimax weight and q-function learning for off-policy evaluation. *arXiv preprint arXiv:1910.12809*, 2019.

Van Der Laan, M. J. and Lendle, S. D. Online targeted learning. 2014.

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

Van Der Vaart, A. W. and Wellner, J. A. Weak convergence. In *Weak convergence and empirical processes*, pp. 16–28. Springer, 1996.

Zhang, K. W., Janson, L., and Murphy, S. A. Inference for batched bandits. *arXiv preprint arXiv:2002.03217*, 2020.

Zou, S., Xu, T., and Liang, Y. Finite-sample analysis for sarsa with linear function approximation. In *Advances in Neural Information Processing Systems*, pp. 8665–8675, 2019.