Marginalized Off-Policy Evaluation for Reinforcement Learning

Tengyang Xie[*]
UMass Amehrst
txie@cs.umass.edu

Yu-Xiang Wang[*]
UC Santa Barbara
yuxiangw@cs.ucsb.edu

Yifei Ma
Amazon AI
yifeim@amazon.com

Abstract

Off-policy evaluation is concerned with evaluating the performance of a policy using the historical data obtained by different behavior policies. In the real-world application of reinforcement learning, acting a policy can be costly and dangerous, and off-policy evaluation usually plays as a crucial step. Currently, the existing methods for off-policy evaluation are mainly based on the Markov decision process (MDP) model of discrete tree MDPs, and they suffer from high variance due to the cumulative product of importance weights. In this paper, we propose a new off-policy evaluation approach directly based on the discrete directed acyclic graph (DAG) MDPs. Our approach can be applied to most of the estimators of off-policy evaluation without modification and could reduce the variance dramatically. We also provide a theoretical analysis of our approach and evaluate it by empirical results.

1 Introduction

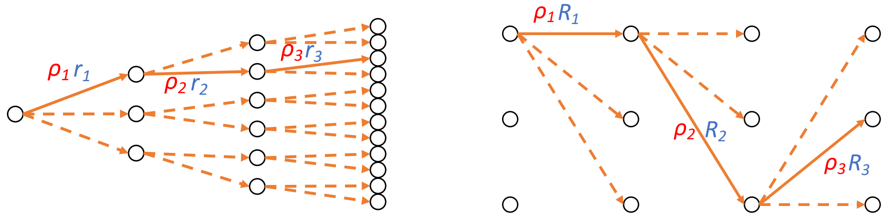
The problem of *off-policy evaluation* (OPE), which predicts the performance of policy with data only sampled by a behavior policy [Sutton and Barto, 1998], is crucial for using *reinforcement learning* (RL) algorithms responsibly in many real-world applications. For the setting that RL algorithm has been already deployed, e.g., applying RL for improving the targeting of advertisements and marketing [Bottou et al., 2013; Tang et al., 2013; Chapelle et al., 2015; Theocharous et al., 2015; Thomas et al., 2017] or for the medical treatment [Murphy et al., 2001; Raghu et al., 2017], online estimation for the policy value is usually expensive, risky, or even illegal. Also, using a bad policy in these applications is dangerous and could lead to severe consequences. Thus, the problem of OPE is essential for the practical application of RL.

To tackle the problem of OPE, the idea of importance sampling (IS) corrects the mismatch in the distributions under the behavior policy and estimated policy. It also provides typically unbiased or strongly consistent estimators [Precup et al., 2000]. IS-based off-policy evaluation methods have also seen a lot of interest recently especially for short-horizon problems, including contextual bandits [Murphy et al., 2001; Hirano et al., 2003; Dudík et al., 2011; Wang et al., 2017]. However, the variance of IS based approaches [Precup et al., 2000; Thomas et al., 2015; Jiang and Li, 2016; Thomas and Brunskill, 2016; Guo et al., 2017; Farajtabar et al., 2018] tends to be too high to be useful [Mandel et al., 2014] for long-horizon problems, because the variance of the cumulative product of importance weights is exploding exponentially as the horizon goes long.

Given this high-variance issue, it is necessary to find a IS-based approach without relying heavily on the cumulative product of importance weights. In this paper, we propose a marginalized method for off-policy evaluation based on the DAG MDPs, while previous methods are all based on the discrete tree MDPs [Jiang and Li, 2016]. Given the key difference of the graph MDPs from the tree MDPs,

[*] Most of this work performed at Amazon AI.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.



(a) Tree MDPs $v(\pi) = \mathbb{E}[\prod_{t'=0}^{t} \rho'_t(\pi) r_t]$ (b) DAG MDPs $v(\pi) = \sum_t \mathbb{E}[w_t(\pi)\rho_t(\pi)R_t]$

Figure 1: Tree MDPs vs DAG MDPs. DAG MDPs avoid multiplicative factors.

where one state can be possibly visited more than once, we use the model of graph MDPs to avoid the cumulative product of importance weights. Figure 1 illustrates the difference between tree MDPs and DAG MDPs.

Our approach can be used in any OPE estimators that leverage IS based estimators to become marginalized OPE estimators, such as doubly robust [Jiang and Li, 2016], MAGIC [Thomas and Brunskill, 2016], MRDR [Farajtabar et al., 2018]. The marginalized OPE estimators work in the space of possible states , instead of the space of trajectories, resulting in a significant potential for variance reduction. We also evaluate a number of marginalized OPE estimators against to their original version empirically. The experimental results demonstrate the effectiveness of the marginalized OPE estimators across a number of problems.

Here is a roadmap for the rest of the paper. Section 2 provides the preliminaries of the problem of off-policy evaluation. In Section 3, we summary the previous importance-sampling based off-policy estimators, and we offer a new framework for the marginalized estimator. We provide the empirical results in Section 3, and more detailed preliminary theoretical analysis in Section 5. At last, Section 6 provides the conclusion and future work.

2 Preliminaries

We consider the problem of off-policy evaluation for an MDP, which is a tuple defined by $M = (\mathcal{S}, \mathcal{A}, T, R, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the *transition function* with $T(s' \mid s, a)$ defined by probability of achieving state s' after taking action a in state s, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the expected reward function with $\mathcal{R}(s, a)$ defined by the mean of immediate received reward after taking action a in state s, and $\gamma \in [0, 1]$ is the discount factor.

Let $|\mathcal{S}|, |\mathcal{A}|$ be the cardinality of $\mathcal{S}$ and $\mathcal{A}$ and h be the time horizon. We denote $T_t \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|\times|\mathcal{A}|}$ the transition matrix at time t, $T_t[s, s', a]$ denotes the probability of transitioning from s into s' when action a is taken at time t. When the an MDP is time-invariant, let $T = T_t, \forall\, 0 \leq t \leq H - 1$.

We use $\mathbb{P}[E]$ to denote the probability of an event E and $p(x)$ denotes p.m.f. (or pdf) of the random variable X taking value x. Let $\mu, \pi : \mathcal{S} \to \mathbb{P}_{\mathcal{A}}$ be policies which output a distribution of actions given an observed state. For notation convenience we denote $\mu(a_t|s_t)$ and $\pi(a_t|s_t)$ the p.m.f of actions given state at time t. Moreover, we denote $d_t^{\mu}(s_t)$ and $d_t^{\pi}(s_t)$ the induced state distribution at time t. They are functions of not just the policies themselves but also the unknown underlying transition dynamics.

We call μ the logging policy (a.k.a. behavioral policy) and π the target policy. μ is used to collect data in forms of $(s_t^i, a_t^i, r_t^i) \in \mathcal{S} \times \mathcal{A} \times \mathbb{R}$ for time index $t = 0, \ldots, H - 1$ and episode index $i = 1, ..., n$. π is the target policy that we are interested to evaluate. Also, let $\mathcal{D}$ to denote the historical data, which contains n trajectories in total.

The problem of off-policy evaluation is about finding an estimator $\widehat{v} : (\mathcal{S} \times \mathcal{A} \times \mathbb{R})^{H\times n} \to \mathbb{R}$ that makes use of the data collected by running μ to estimate

$$v(\pi) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{H-1} \gamma^t r_t(s_t, a_t)\right].$$

Note that we assume that $\mu(a|s)$ and $\pi(a|s)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ are known to us and can be used in the estimator but we do *not* observe the state-transition distributions therefore do not have $\mu(s_t), \pi(s_t)$. The corresponding minimax risk for square error is

$$R(\pi, \mu, T, r_{\max}, \sigma^2) = \inf_{\widehat{v}} \sup_{\left\{\substack{r(s,a)\in\mathcal{S}\times\mathcal{A}\to\mathbb{P}_{\mathbb{R}} \text{ s.t. } \forall s\in\mathcal{S}, a\in\mathcal{A} \\ \mathbb{E}[r(s,a)]\le r_{\max}, \mathrm{Var}[r(s,a)]\le\sigma^2}\right\}} \mathbb{E}[(\widehat{v}(\pi) - v(\pi))^2]$$

Different from previous studies, we focus on the case where S is sufficiently small but S^2A is too large for a reasonable sample size. In other word, this is a setting where we do not have enough data points to estimate the state-transition dynamics, but we do observe the states and can estimate the distribution of the states.

3 Marginalized Estimators for OPE

In this section, we show some example of popular IS based estimators for the problem of OPE and provide the reason for its difficulties. After that, we discuss the way to get around that difficulties and present our new design for the OPE estimators.

3.1 Generic IS-Based Estimators Setup

The IS based estimators usually provide an biased or consistent estimate of the value of target policy π [Thomas, 2015]. We first provide a generic framework of IS-based estimators, and analysis the similarity and difference between different IS-based estimators. This framework could give us insight into the design of IS-based estimators, and is useful to understand the limitation of them.

Let $\rho_t^i := \frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}$ be the importance ratio at time step t of i-th trajectory, and $\rho_{0:t}^i := \prod_{t'=0}^{t} \frac{\pi(a_t^{\prime i}|s_t^{\prime i})}{\mu(a_t^{\prime i}|s_t^{\prime i})}$ be the cumulative importance ratio for the i-th trajectory. The generic framework of IS-based estimators can be expressed as follows

$$\widehat{v}(\pi) = \frac{1}{n}\sum_{i=1}^{n} g(s_0^i) + \sum_{i=1}^{n}\sum_{t=0}^{H-1} \frac{\rho_{0:t}^i}{\phi_t(\rho_{0:t}^{1:n})}\gamma^t(r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i)), \tag{3.1}$$

where $\phi_t : \mathbb{R}_+^n \to \mathbb{R}_+$ are the "normalization" functions for $\rho_{0:t}^i$, $g : \mathcal{S} \to \mathbb{R}$ and $f_t : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}$ are the "value-related" functions. Note $\mathbb{E}f_t = 0$. For the unbiased IS-based estimators, it usually has $\phi_t(\rho_{0:t}^{1:n}) = n$, and we first observe that the importance sampling (IS) estimator [Precup et al., 2000] falls in this framework using:

$$\text{(IS)}: \qquad g(s_0^i) = 0;\ \phi_t(\rho_{0:t}^{1:n}) = n;\ f_t(s_t^i, a_t^i, s_{t+1}^i) = 0.$$

For the doubly tobust (DR) estimator [Jiang and Li, 2016], the normalization function and value-related functions are:

$$\text{(DR)}: \qquad g(s_0^i) = \widehat{V}^\pi(s_0);\ \phi_t(\rho_{0:t}^{1:n}) = n;\ f_t(s_t^i, a_t^i, s_{t+1}^i) = -\widehat{Q}^\pi(s_t^i, a_t^i) + \gamma\widehat{V}^\pi(s_{t+1}^i).$$

Self-normalized estimators such as weight importance sampling (WIS) and weighted doubly robust (WDR) estimators [Thomas and Brunskill, 2016] are popular consistent estimators to achieve better bias-variance trade-off. The critical difference of consistent self-normalized estimators is to use $\sum_{j=1}^n \rho_{0:t}^j$ as normalization function ϕ_t rather than n. Thus, the WIS estimator is using the following normalization and value-related functions:

$$\text{(WIS)}: \qquad g(s_0^i) = 0;\ \phi_t(\rho_{0:t}^{1:n}) = \sum_{j=1}^{n}\rho_{0:t}^j;\ f_t(s_t^i, a_t^i, s_{t+1}^i) = 0,$$

and the WDR estimator:

$$\text{(WDR)}: \quad g(s_0^i) = \widehat{V}^\pi(s_0);\ \phi_t(\rho_{0:t}^{1:n}) = \sum_{j=1}^{n}\rho_{0:t}^j;\ f_t(s_t^i, a_t^i, s_{t+1}^i) = -\widehat{Q}^\pi(s_t^i, a_t^i) + \gamma\widehat{V}^\pi(s_{t+1}^i).$$

Note that, the DR estimator reduced the variance from the stochasticity of action by using the technique of control variants in value-related function, and the WDR estimators reducing variance by the bias-variance trade-off using self-normalization. However, it could still suffer a large variance, because the cumulative importance ratio $\rho_{0:t}^i$ always appear directly in this framework, and it makes the variance increase exponentially with the horizon goes long. This kind of high-variance issue is inherited from the hardness of OPE problem for discrete tree MDPs which is defined as follows:

Definition 1 (Discrete Tree MDPs)**.** *If an MDP satisfies:*

- *The state is represented by history, i.e.,* $s_t = h_t := o_1 a_1 \ldots o_{t-1} a_{t-1} o_t$*, where* o_i *is the observation at step* $i (1 \leq i \leq t)$*.*
- *The observations and actions are discrete.*
- *The initial state takes the form of* $s_0 = o_0$*. After taking action* a *at state* $s = h$*, the next can be only expressed in the form of* $s' = hao$*, with probability* $\mathbb{P}(o|h, a)$*.*

Then, this MDP is a discrete tree MDP.

It can be proved that the variance of any unbiased OPE estimator for discrete tree MDP can be lower bounded by $\sum_{t=0}^{H-1} \mathbb{E}\left[\rho_{0:t-1}^2 \mathbb{V}_t[V(s_t)]\right]$ [Jiang and Li, 2016]. However, the condition of discrete tree MDP can be related to discrete directed acyclic graph (DAG) MDP, which could reduce the lower bound of variance dramatically.

Definition 2 (Discrete DAG MDPs)**.** *If an MDP satisfies:*

- *The state space and action space are finite.*
- *Each state can only occur at a particular time step.*

Then, this MDP is a discrete DAG MDP.

The lower variance bound of any unbiased OPE estimator under discrete DAG MDPs is $\sum_{t=0}^{H-1} \mathbb{E}\left[\frac{(d_t^\pi(s_{t-1})\pi(a_{t-1}|s_{t-1}))^2}{(d_t^\mu(s_{t-1})\mu(a_{t-1}|s_{t-1}))^2} \mathbb{V}_t[V(s_t)]\right]$ [Jiang and Li, 2016]. In this paper, we will mainly design OPE estimators based on this relaxed case.

3.2 Design of Marginalized Estimators

Based on the analysis in the last section, the critical part of the variance of estimators fall in the framework (3.1) is the cumulative importance ratio $\rho_{0:t}^i$. Note that, this ratio is used for re-weighting the probability of the trajectory $\tau_{0:t}$, i.e.,

$$v(\pi) = \sum_{t=0}^{H-1} \mathbb{E}_{\tau \sim \pi}[r_t] = \sum_{t=0}^{H-1} \mathbb{E}_{\tau \sim \mu}[\rho_{0:t} r_t].$$

Let $d_t^\pi(s) := \mathbb{P}[s_t = s|\pi]$, $d_t^\mu(s) := \mathbb{P}[s_t = s|\mu]$ be the marginal distribution of state s at time step t, then given the conditional independence following from the Markov property; define

$$w_t(s_t) := \frac{d_t^\pi(s_t)}{d_t^\mu(s_t)} \quad \Rightarrow \quad v(\pi) = \sum_{t=0}^{H-1} \mathbb{E}_{\tau \sim \pi}[r_t] = \sum_{t=0}^{H-1} \mathbb{E}_{s_t \sim \pi}[r_t] = \sum_{t=0}^{H-1} \mathbb{E}_{s_t \sim \mu}[w_t(s_t) r_t]$$

For the moment let us assume that we have an unbiased or consistent estimator of $w_t(s) = d_t^\pi(s)/d_t^\mu(s)$ called $\widehat{w}_t^n(s)$. Designing such an estimator is the main technical contribution of the paper but we will start by assuming we have such an estimator as a black box and use it to construct off-policy evaluation methods. Thus, we obtain a generic framework of marginalized IS-based estimators as:

$$\widehat{v}_M(\pi) = \frac{1}{n} \sum_{i=1}^{n} g(s_0^i) + \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{H-1} \widehat{w}_t^n(s_t^i) \rho_t^i \gamma^t (r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i)). \tag{3.2}$$

Note that the "normalization" function ϕ has not appeared in the framework above is because it can be a part in $\widehat{w}_t(s)$.

We first show the equivalence between framework (3.1) with $\phi_t(\rho_{0:t}^{1:n}) = n$ and (3.2) in expectation if $\widehat{w}_t^n(s)$ is exactly $w_t(s)$.

Lemma 1. *If $\phi_t(\rho_{0:t}^{1:n}) = n$ in framework (3.1) and $\widehat{w}_t^n(s) = w_t(s)$ in framework (3.2), then these two frameworks are equal in expectation, i.e.,*

$$\mathbb{E}\left[w_t(s_t^i)\rho_t^i(r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i))\right] = \mathbb{E}\left[\rho_{0:t}^i(r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i))\right]$$

holds for all i and t.

Full proof of Lemma 1 can be found in the appendix.

We now show that if we have that unbiased or consistent $\widehat{w}_t^n$, all the OPE estimator that leverage IS-based estimators can use $\widehat{w}_t^n$ directly by replacing $\prod_{t'=0}^{t-1} \frac{\pi(a_{t'}|s_{t'})}{\mu(a_{t'}|s_{t'})}$ with $\widehat{w}_t^n(s_t)$, and keep unbiasedness or consistency.

Theorem 1. *Let $\phi_t(\rho_{0:t}^{1:n}) = n$ in framework (3.1), then framework (3.2) could keep the unbiasedness and consistency same as in framework (3.1) if $\widehat{w}_t^n(s)$ is a unbiased or consistent estimator for marginalized ratio $w_t(s)$ for all t:*

1. *If an unbiased estimator falls in framework (3.1), then its marginalized estimator in framework (3.2) is also a unbiased estimator of $v(\pi)$ given unbiased estimator $\widehat{w}_t^n(s)$ for all t.*
2. *If a consistent estimator falls in framework (3.1), then its marginalized estimator in framework (3.2) is also a consistent estimator of $v(\pi)$ given consistent estimator $\widehat{w}_t^n(s)$ for all t.*

Full proof of Theorem 1 can be found in the appendix.

Next, we propose a new marginalized IS estimator to further improve the data efficiency and reduce variance. Since DR only reduces the variance from the stochasticity of action [Jiang and Li, 2016] and our marginalized estimator (3.2) reduce the variance from the cumulative importance weights, it is also possible to reduce the variance the stochasticity of reward function.

Based on the definition of MDP, we know that r_t is the random variable that only determined by s_t, a_t. Thus, if $\widehat{R}(s, a)$ is a unbiased and consistent estimator for $R(s, a)$, r_t^i can be in framework (3.2) can be replaced by that $\widehat{R}(s_t^i, a_t^i)$, and keep unbiasedness or consistency same as using r_t^i.

Note that we can use a unbiased and consistent Monte-Carlo based estimator

$$\widehat{R}(s_t, a_t) = \frac{\sum_{i=1}^n r_t^{(i)} \mathbf{1}(s_t^{(i)} = s_t, a_t^{(i)} = a_t)}{\sum_{i=1}^n \mathbf{1}(s_t^{(i)} = s_t, a_t^{(i)} = a_t)},$$

and then we obtain a better marginalized framework

$$\widehat{v}_{BM}(\pi) = \frac{1}{n}\sum_{i=1}^n g(s_0^i) + \frac{1}{n}\sum_{i=1}^n \sum_{t=0}^{H-1} \widehat{w}_t^n(s_t^i)\rho_t^i\gamma^t(\widehat{R}(s_t^i, a_t^i) + f_t(s_t^i, a_t^i, s_{t+1}^i)). \tag{3.3}$$

Remark 1. *Note that, the only difference between (3.2) and (3.3) is r_t^i and $\widehat{R}(s_t^i, a_t^i)$. Thus, the unbiasedness or consistency of (3.3) can be obtained similarly by following Theorem 1 and its proof.*

One interesting observation is that when each (s_t, a_t)-pair is observed only once in n iterations, then framework (3.3) reduces to (3.2). Note that when this happens, we could still potentially estimate $\widehat{w}_t^n(s_t)$ well if A is large but S is relative small, in which case we can still afford to observe each potential values of s_t many times.

3.3 Estimating $w_t(s)$

We now present our estimators $w_t(s)$. Note that $w_t(s)$ is defined by the marginalized ratio $d_t^\pi(s)/d_t^\mu(s)$ at time step t. Considering the cumulative importance ratio is actually calculating the ratio for the probability of the whole episode, i.e.,

$$\prod_{t'=0}^{t} \frac{\pi(a_{t'}^i|s_{t'}^i)}{\mu(a_{t'}^i|s_{t'}^i)} = \frac{\mathbb{P}(s_0^i, a_0^i, \dots, s_t^i, a_t^i|\pi)}{\mathbb{P}(s_0^i, a_0^i, \dots, s_t^i, a_t^i|\mu)},$$

for all i and t, thus, we can obtain the unbiased and consistent Monte-Carlo based estimator for $w_t(s)$ as

$$\widehat{w}_t^n(s) = \frac{\sum_{i=1}^n \prod_{t'=0}^{t-1} \frac{\pi(a_{t'}^i|s_{t'}^i)}{\mu(a_{t'}^i|s_{t'}^i)} \mathbf{1}(s_t^i = s)}{\sum_{i=1}^n \mathbf{1}(s_t^i = s)}. \tag{3.4}$$

We then show the unbiasedness of $\widehat{w}_t^n$ in the next theorem.

Theorem 2. *The estimator of $w_t^n(s)$ in (3.4) is unbiased estimator.*

Full proof of Theorem 2 can be found in the appendix.

Self-normalization is a popular strategy in the problem of off-policy evaluation, since the cumulative product of important weight $\prod_{t'=0}^{t-1} \frac{\pi(a_{t'}^i|s_{t'}^i)}{\mu(a_{t'}^i|s_{t'}^i)}$ is usually distributed with heavy upper tails [Thomas et al., 2015]. By following this insight of self-normalization, we can obtain the following consistent estimator of $w_t(s)$:

$$\widetilde{w}_t^n(s) = \frac{\sum_{i=1}^n \prod_{t'=0}^{t-1} \frac{\pi(a_{t'}^i|s_{t'}^i)}{\mu(a_{t'}^i|s_{t'}^i)} \mathbf{1}(s_t^i = s)}{\frac{1}{n}\sum_{i=1}^n \prod_{t'=0}^{t-1} \frac{\pi(a_{t'}^i|s_{t'}^i)}{\mu(a_{t'}^i|s_{t'}^i)} \sum_{j=1}^n \mathbf{1}(s_t^j = s)} = \frac{n}{\sum_{i=1}^n \prod_{t'=0}^{t-1} \frac{\pi(a_{t'}^i|s_{t'}^i)}{\mu(a_{t'}^i|s_{t'}^i)}} \widehat{w}_t^n(s). \tag{3.5}$$

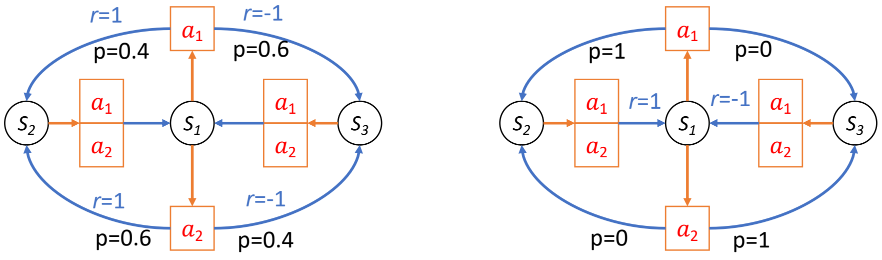
It is clear intuitively that this estimator is consistent since $\mathbb{E}[\prod_{t'=0}^{t-1} \frac{\pi(a_{t'}^i|s_{t'}^i)}{\mu(a_{t'}^i|s_{t'}^i)}] = 1$, so using Slutsky's theorem [Slutsky, 1925] we know (3.5) convergences to the same result as (3.4).

Given the unbiasedness or consistency of the estimator (3.4) and (3.5), we can obtain unbiased or consistent marginalized estimators for OPE based on the framework (3.2) and Theorem 1.

4 Experiments

Throughout this section, we present the empirical results to illustrate the comparison among different estimators. We demonstrate the effectiveness of the proposed marginalized approaches by comparing the classical estimator with their marginalized version on several different domains and their variants. The estimators we compared are as follows:

1. IS: importance sampling
2. WIS: weighted importance sampling
3. DR: doubly robust
4. WDR: weighted doubly robust
5. MIS: marginalized importance sampling (importance sampling with $\widehat{w}_t^n$ in (3.4))
6. W-MIS: marginalized importance sampling (importance sampling with $\widetilde{w}_t^n$ in (3.5))
7. MDR: marginalized doubly robust (doubly robust with $\widehat{w}_t^n$ in (3.4))
8. W-MDR: marginalized doubly robust (doubly robust with $\widetilde{w}_t^n$ in (3.5))



(a) ModelWin MDP

(b) ModelFail MDP

Figure 2: Domains from Thomas and Brunskill [2016]

We first provide results on the variants of standard domains, ModelWin and ModelFail, which are first introduced by Thomas and Brunskill [2016]. In the ModelWin domain (as showed in Figure 2(a)), the agent always begins in s_1, where it must select between two actions. The first action, a_1, causes the agent to transition to s_2 with probability p and s_3 with probability $1 - p$. The second action, a_2, does the opposite: the agent transitions to s_2 with probability $1 - p$ and s_3 with probability p. We set $p = 0.4$. If the agent transitions to s_2, then it receives a reward of 1, and if it transitions to s_3 it receives a reward of -1. In states s_2 and s_3, the agent also has two possible actions a_1 and a_2, but both always produce a reward of zero and a deterministic transition back to s_1. In our ModelFail Domain (as showed in Figure 2(b)), we set the state transition is same as the ModelWin domain, with $p = 1$, but we delay the reward receiving – the agent receives a reward of 1 when it arrive s_1 from s_2, and receives a reward of -1 when it arrive s_1 from s_3. The discount factor $\gamma = 1$. The evaluated policy π is selecting action a_1 and a_2 with probabilities 0.2 and 0.8 respectively everywhere for both of these domains, and behavior policy μ is uniform policy. To show the effectiveness of our approach in the long-horizon domain, we set the horizon H of these two domains to vary among $\{8, 16, 32, 64, 128, 256, 512, 1024\}$. Note that the sacle of $v(\pi)$ is in direct proportion to H, so we provide results for both RMSE and relative RMSE where relative RMSE is defined by the ratio between RMSE and $v(\pi)$. The results on the Modelwin domain is in Table 1 (RMSE) and Table 2 (Relative RMSE), and the results on the Modelwin domain is in Table 3 (RMSE) and Table 4 (Relative RMSE). All results are from 1024 runs.

Horizon	IS	WIS	DR	WDR	MIS	W-MIS	MDR	W-MDR
8	0.152	0.149	0.425	0.425	0.079	**0.075**	0.395	0.394
16	0.510	0.415	1.093	0.985	0.142	**0.104**	0.804	0.802
32	3.394	1.187	5.099	2.160	1.551	**0.205**	2.269	1.646
64	14.937	2.714	9.280	4.442	5.221	**0.205**	5.832	3.312
128	19.651	5.418	34.416	8.229	6.576	**0.291**	5.978	6.569
256	48.638	11.029	22.593	15.546	30.472	**0.412**	18.642	13.228
512	46.670	22.443	38.884	29.287	33.821	**0.580**	20.575	26.497
1024	66.632	46.402	49.578	58.071	58.340	**0.806**	15.552	52.931

Table 1: RMSE on the ModelWin Domain

Horizon	IS	WIS	DR	WDR	MIS	W-MIS	MDR	W-MDR
8	0.317	0.310	0.885	0.885	0.164	**0.156**	0.822	0.821
16	0.532	0.433	1.138	1.026	0.148	**0.108**	0.838	0.836
32	1.768	0.618	2.655	1.125	0.808	**0.073**	1.182	0.857
64	3.89	0.707	2.417	1.157	1.360	**0.053**	1.159	0.862
128	2.559	0.706	4.481	1.071	0.856	**0.038**	0.778	0.855
256	3.167	0.718	1.471	1.012	1.310	**0.027**	2.268	0.861
512	1.519	0.731	1.266	0.953	1.101	**0.019**	0.670	0.863
1024	3.793	0.755	6.656	0.945	2.072	**0.013**	1.853	0.862

Table 2: Relative RMSE on the ModelWin Domain.

To demonstrate the scalability of the proposed marginalized approaches, we also test all estimators in the Mountain Car domain [Singh and Sutton, 1996], with the horizon of $H = 100$, the initial state distributed uniformly randomly, and same state aggregations as Jiang and Li [2016]. To construct the stochastic behavior policy μ and stochastic evaluated policy π, we first compute the optimal Q-function using Q-learning, and use its softmax policy of the optimal Q-function as evaluated policy π (with temperature of 1). For the behavior policy μ, we also use the softmax policy of the optimal Q-function, but we set the temperature as 0.75. The results on the Mountain Car domain is in Table 5, and all results are from 1024 runs.

Horizon	IS	WIS	DR	WDR	MIS	W-MIS	MDR	W-MDR
8	0.190	0.083	2.071	2.065	0.133	**0.062**	2.066	2.064
16	0.760	0.265	4.239	4.150	0.497	**0.087**	4.176	4.147
32	5.630	1.014	10.349	8.290	6.146	**0.127**	9.942	8.282
64	54.604	3.672	35.948	16.538	30.800	**0.172**	37.63	16.527
128	85.255	12.764	50.726	33.002	55.125	**0.25**	61.085	32.980
256	99.324	36.698	92.37	65.967	76.782	**0.339**	59.914	65.909
512	377.751	91.585	82.85	131.724	289.502	**0.497**	279.660	131.593
1024	289.616	215.780	70.795	262.955	288.790	**0.709**	45.124	262.890

Table 3: RMSE on the ModelFail Domain

Horizon	IS	WIS	DR	WDR	MIS	W-MIS	MDR	W-MDR
8	0.079	0.035	0.863	0.861	0.056	**0.026**	0.861	0.860
16	0.158	0.055	0.883	0.864	0.104	**0.018**	0.870	0.864
32	0.586	0.106	1.078	0.864	0.64	**0.013**	1.036	0.863
64	2.844	0.191	1.872	0.861	1.604	**0.009**	1.960	0.861
128	2.220	0.332	1.321	0.859	1.436	**0.006**	1.591	0.859
256	1.293	0.478	1.203	0.859	1.000	**0.004**	0.780	0.858
512	2.459	0.596	0.539	0.858	1.885	**0.003**	1.821	0.857
1024	0.943	0.702	0.230	0.856	0.940	**0.002**	0.147	0.856

Table 4: Relative RMSE on the ModelFail Domain

5 Theoretical Analysis

We provide the preliminary theoretical analysis in this section. We use the high-confidence bound to measure properties of $\widehat{w}_t^n(s)$ and $\widetilde{w}_t^n(s)$. We first show the high-confidence bound of unbiased estimator $\widehat{w}_t^n(s)$ in (3.4).

Theorem 3. *If there exists an $\eta > 0$ such that $\mu(a|s) \geq \eta\pi(a|s)$ for every s and a, then with probability at least $1-\delta$, we have*

$$|\widehat{w}_t^n(s) - w_t(s)| \leq \frac{1}{\eta^t}\sqrt{\frac{\log(4/\delta)\sqrt{2n}}{2\left(nd_t^\mu(s)\sqrt{2n} - \sqrt{\log(2/\delta)}\right)}},$$

for $\widehat{w}_t^n$ in (3.4).

Full proof of Theorem 3 can be found in the appendix.

The following theorem provides the high-confidence bound of estimator $\widetilde{w}_t^n(s)$ in (3.5), which illustrate the consistency of estimator $\widetilde{w}_t^n(s)$.

Data Size	IS	WIS	DR	WDR	MIS	W-MIS	MDR	W-MDR
16	534.51	24.55	724.48	22.81	533.67	**5.39**	695.25	27.85
32	398.98	24.63	504.99	23.81	398.84	**4.12**	497.50	24.48
64	335.24	24.63	404.75	23.42	335.21	**3.06**	400.46	20.2
128	307.73	24.63	357.06	21.93	307.63	**2.40**	344.29	15.75
256	275.34	24.63	260.40	18.93	275.19	**2.04**	252.05	11.00
512	256.75	24.63	210.47	13.91	256.63	**1.80**	205.58	9.34
1024	249.97	24.63	166.67	6.71	249.94	**1.70**	163.30	11.62
2048	246.08	24.63	108.34	6.13	246.07	**1.64**	106.86	17.46

Table 5: RMSE on the Mountain Car Domain

Theorem 4. *If there exists an $\eta > 0$ such that $\mu(a|s) \geq \eta\pi(a|s)$ for every s and a, then with probability at least $1-\delta$, we have*

$$\begin{aligned}&|\widetilde{w}_t^n(s) - w_t(s)| \\ \leq &\left(\frac{n\sqrt{2n}}{\eta^t n\sqrt{2n} - \sqrt{\log(6/\delta)}} \sqrt{\frac{\log(6/\delta)\sqrt{2n}}{2\left(nd_t^\mu(s)\sqrt{2n} - \sqrt{\log(3/\delta)}\right)}} + \frac{w_t(s)\sqrt{\log(6/\delta)}}{\eta^t n\sqrt{2n} - \sqrt{\log(6/\delta)}} \right) \wedge \\ &\left(w_t(s) \vee \left(\frac{\sqrt{2n}}{d_t^\mu(s)\sqrt{2n} - \sqrt{\log(2/\delta) - w_t(s \in \mathcal{S}_t)}} - w_t(s) \right) \right),\end{aligned}$$

for $\widetilde{w}_t^n$ in (3.5).

Full proof of Theorem 4 can be found in the appendix.

Note that the results in this section is preliminary. Although we are able to obtain the high-confidence bounds of $\widehat{w}_t^n(s)$ and $\widetilde{w}_t^n(s)$ from Theorem 3 and Theorem 4, those high-confidence bound do not perfectly matching the experimental results. The bound in Theorem 3 can be tighter than the bound in Theorem 4. However, the experimental results in Section 4 shows that estimators with $\widetilde{w}_t^n(s)$ always outperform those with $\widehat{w}_t^n(s)$. Intuitively, the reason is the correlation in the self-normalization, i.e., $\prod_{t'=0}^{t-1} \frac{\pi(a_{t'}^i|s_{t'}^i)}{\mu(a_{t'}^i|s_{t'}^i)}$ in both numerator and denominator of $\widetilde{w}_t^n(s)$, but the high-confidence bound is not able to fully express that correlation.

6 Conclusions and Future Work

In this paper, we proposed the marginalized approach to the problem of off-policy evaluation in reinforcement learning. This is the first approach based on the model of DAG MDPs. It achieves substantially better performance than existing approaches. The theoretical analysis of the marginalized approach is still an open problem, and we will extend our theoretical results to fully explain all these marginalized approaches.

References

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. (2013). Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260.

Chapelle, O., Manavoglu, E., and Rosales, R. (2015). Simple and scalable response prediction for display advertising. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(4):61.

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. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1447–1456, Stockholmsmässan, Stockholm Sweden. PMLR.

Guo, Z., Thomas, P. S., and Brunskill, E. (2017). Using options and covariance testing for long horizon off-policy policy evaluation. In *Advances in Neural Information Processing Systems*, pages 2492–2501.

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.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30.

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.

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.

Precup, D., Sutton, R. S., and Singh, S. P. (2000). Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 759–766. Morgan Kaufmann Publishers Inc.

Raghu, A., Komorowski, M., Celi, L. A., Szolovits, P., and Ghassemi, M. (2017). Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Machine Learning for Healthcare Conference*, pages 147–163.

Singh, S. P. and Sutton, R. S. (1996). Reinforcement learning with replacing eligibility traces. *Machine learning*, 22(1-3):123–158.

Slutsky, E. (1925). Uber stochastische asymptoten und grenzwerte. *Metron*, 5(3):3–89.

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

Tang, L., Rosales, R., Singh, A., and Agarwal, D. (2013). Automatic ad format selection via contextual bandits. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1587–1594. ACM.

Theocharous, G., Thomas, P. S., and Ghavamzadeh, M. (2015). Personalized ad recommendation systems for life-time value optimization with guarantees. In *IJCAI*, pages 1806–1812.

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. (2015). *Safe reinforcement learning*. PhD thesis, University of Massachusetts Amherst.

Thomas, P. S., Theocharous, G., and Ghavamzadeh, M. (2015). High-confidence off-policy evaluation. In *AAAI*, pages 3000–3006.

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 *AAAI*, pages 4740–4745.

Wang, Y.-X., Agarwal, A., and Dudık, M. (2017). Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597.

Appendix

A Proof Details

In this section, we provide the full proofs for our lemmas and theorems. We first provide the proof of lemma 1.

Proof of Lemma 1. Given the conditional independence in the Markov property, we have

$$\begin{aligned}\mathbb{E}\left[\rho_{0:t}^i(r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i))\right] =& \mathbb{E}\left[\mathbb{E}\left[\rho_{0:t}^i(r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i))|s_t^i\right]\right] \\ =& \mathbb{E}\left[\mathbb{E}\left[\rho_{0:t-1}^i|s_t^i\right]\mathbb{E}\left[\rho_t^i(r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i))|s_t^i\right]\right] \\ =& \mathbb{E}\left[w_t(s_t^i)\mathbb{E}\left[\rho_t^i(r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i))|s_t^i\right]\right] \\ =& \mathbb{E}\left[w_t(s_t^i)\rho_t^i(r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i))\right],\end{aligned}$$

where the first equation follows from the law of total expectation, the second equation follows from the conditional independence from the Markov property. This completes the proof. □

We then provide the proof of theorem 1.

Proof of Theorem 1. We first provide the proof of the first part of unbiasedness. Given $\mathbb{E}[\widehat{w}_t^n(s)|s] = w_t(s)$ for all t, then

$$\begin{aligned}\mathbb{E}\left[\widehat{w}_t^n(s_t^i)\rho_t^i\gamma^t(r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i))\right] =& \mathbb{E}\left[\mathbb{E}\left[\widehat{w}_t^n(s_t^i)\rho_t^i\gamma^t(r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i))|s_t^i\right]\right] \\ =& \mathbb{E}\left[\mathbb{E}\left[\widehat{w}_t^n(s_t^i)|s_t^i\right]\mathbb{E}\left[\rho_t^i\gamma^t(r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i))|s_t^i\right]\right] \\ =& \mathbb{E}\left[w_t(s_t^i)\mathbb{E}\left[\rho_t^i\gamma^t(r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i))|s_t^i\right]\right] \\ =& \mathbb{E}\left[w_t(s_t^i)\rho_t^i(r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i))\right] \\ =& \mathbb{E}\left[\rho_{0:t}^i(r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i))\right],\end{aligned} \tag{A.1}$$

where the the first equation follows from the law of total expectation, the second equation follows from the conditional independence of the Markov property, the last equation follows from Lemma 1. Since the original estimator falls in framework (3.1) is unbiased, summing (A.1) over i and t completes the proof of the first part.

We now prove the second part of consistency. Since we have

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{H-1}\widehat{w}_t^n(s_t^i)\rho_t^i\gamma^t(r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i)) = \sum_{t=0}^{H-1}\gamma^t\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\widehat{w}_t^n(s_t^i)\rho_t^i(r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i)),$$

then, to prove the consistency, it is sufficient to show

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\widehat{w}_t^n(s_t^i)\rho_t^i(r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i)) = \operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\rho_{0:t}^i(r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i)), \tag{A.2}$$

given $\operatorname{plim}_{n\to\infty}\widehat{w}_t^n(s) = w_t(s)$ for all $s \in \mathcal{S}$. Note that $d_t^\mu(s)$ is the state distribution under behavior policy μ at time step t, then for the left hand side of (A.2), we have

$$\begin{aligned}
&\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\widehat{w}_t^n(s_t^i)\rho_t^i(r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i)) \\
=&\sum_{s\in\mathcal{S}} d_t^\mu(s)\operatorname*{plim}_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n}\widehat{w}_t^n(s)\frac{\pi(a_t^i|s)}{\mu(a_t^i|s)}\mathbf{1}(s_t^i = s)(r_t^i + f_t(s, a_t^i, s_{t+1}^i))\right] \\
=&\sum_{s\in\mathcal{S}} d_t^\mu(s)\operatorname*{plim}_{n\to\infty}\left[\widehat{w}_t^n(s)\frac{1}{n}\sum_{i=1}^{n}\frac{\pi(a_t^i|s)}{\mu(a_t^i|s)}\mathbf{1}(s_t^i = s)(r_t^i + f_t(s, a_t^i, s_{t+1}^i))\right] \\
=&\sum_{s\in\mathcal{S}} d_t^\mu(s)\left[\operatorname*{plim}_{n\to\infty}(\widehat{w}_t^n(s))\operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\pi(a_t^i|s)}{\mu(a_t^i|s)}\mathbf{1}(s_t^i = s)(r_t^i + f_t(s, a_t^i, s_{t+1}^i))\right)\right] \\
=&\sum_{s\in\mathcal{S}} d_t^\mu(s)w_t(s)\operatorname*{plim}_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\pi(a_t^i|s)}{\mu(a_t^i|s)}\mathbf{1}(s_t^i = s)(r_t^i + f_t(s, a_t^i, s_{t+1}^i))\right] \\
=&\sum_{s\in\mathcal{S}} d_t^\mu(s)w_t(s)\mathbb{E}\left[\frac{\pi(a_t|s)}{\mu(a_t|s)}(r_t + f_t(s, a_t, s_{t+1}))\middle| s_t = s\right], \qquad \text{(A.3)}
\end{aligned}$$

where the first equation follows from the weak law of large number. Similarly, for the right hand side of (A.2), we have

$$\begin{aligned}
&\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\rho_{0:t}^i(r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i)) \\
=&\sum_{s\in\mathcal{S}} d_t^\mu(s)\operatorname*{plim}_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n}\prod_{t'=0}^{t-1}\frac{\pi(a_{t'}^i|s_{t'}^i)}{\mu(a_{t'}^i|s_{t'}^i)}\mathbf{1}(s_t^i = s)\frac{\pi(a_t^i|s)}{\mu(a_t^i|s)}(r_t^i + f_t(s, a_t^i, s_{t+1}^i))\right] \\
=&\sum_{s\in\mathcal{S}} d_t^\mu(s)\mathbb{E}\left[\prod_{t'=0}^{t-1}\frac{\pi(a_{t'}|s_{t'})}{\mu(a_{t'}|s_{t'})}\frac{\pi(a_t|s)}{\mu(a_t|s)}(r_t + f_t(s, a_t, s_{t+1}))\middle| s_t = s\right] \\
=&\sum_{s\in\mathcal{S}} d_t^\mu(s)\mathbb{E}\left[\prod_{t'=0}^{t-1}\frac{\pi(a_{t'}|s_{t'})}{\mu(a_{t'}|s_{t'})}\middle| s_t = s\right]\mathbb{E}\left[\frac{\pi(a_t|s)}{\mu(a_t|s)}(r_t + f_t(s, a_t, s_{t+1}))\middle| s_t = s\right] \\
=&\sum_{s\in\mathcal{S}} d_t^\mu(s)w_t(s)\mathbb{E}\left[\frac{\pi(a_t|s)}{\mu(a_t|s)}(r_t + f_t(s, a_t, s_{t+1}))\middle| s_t = s\right], \qquad \text{(A.4)}
\end{aligned}$$

where the first equation follows from the weak law of large number and the third equation follows from the conditional independence of the Markov property. Thus, we have (A.3) equal to (A.4). This completes the proof of the second half. □

We now provide the proof of theorem 2.

Proof of Theorem 2. The unbiasedness of $\widehat{w}_t^n$ can be obtained directed by taking expectation as follows.

$$\begin{aligned}
\mathbb{E}\left[\widehat{w}_t^n(s)\right] =& \mathbb{E}\left[\frac{\sum_{i=1}^n \prod_{t'=0}^{t-1} \frac{\pi(a_{t'}^i|s_{t'}^i)}{\mu(a_{t'}^i|s_{t'}^i)} \mathbf{1}(s_t^i = s)}{\sum_{i=1}^n \mathbf{1}(s_t^i = s)}\right] \\
=& \mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{i=1}^n \prod_{t'=0}^{t-1} \frac{\pi(a_{t'}^i|s_{t'}^i)}{\mu(a_{t'}^i|s_{t'}^i)} \mathbf{1}(s_t^i = s)}{\sum_{i=1}^n \mathbf{1}(s_t^i = s)}\middle| s_t^{1:n}\right]\right] \\
=& \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^n \prod_{t'=0}^{t-1} \frac{\pi(a_{t'}^i|s_{t'}^i)}{\mu(a_{t'}^i|s_{t'}^i)} \mathbf{1}(s_t^i = s)\middle| s_t^{1:n}\right] \frac{1}{\sum_{i=1}^n \mathbf{1}(s_t^i = s)}\right] \\
=& \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^n \frac{d_t^\mu(s_t^i)}{d_t^\pi(s_t^i)} \mathbf{1}(s_t^i = s)\middle| s_t^{1:n}\right] \frac{1}{\sum_{i=1}^n \mathbf{1}(s_t^i = s)}\right] \\
=& \mathbb{E}\left[\frac{1}{\sum_{i=1}^n \mathbf{1}(s_t^i = s)} \sum_{i=1}^n \frac{d_t^\mu(s_t^i)}{d_t^\pi(s_t^i)} \mathbf{1}(s_t^i = s)\right] \\
=& \mathbb{E}\left[\frac{1}{\sum_{i=1}^n \mathbf{1}(s_t^i = s)} \sum_{i=1}^n \frac{d_t^\mu(s)}{d_t^\pi(s)} \mathbf{1}(s_t^i = s)\right] \\
=& \mathbb{E}\left[\frac{d_t^\mu(s)}{d_t^\pi(s)} \frac{1}{\sum_{i=1}^n \mathbf{1}(s_t^i = s)} \sum_{i=1}^n \mathbf{1}(s_t^i = s)\right] \\
=& \frac{d_t^\mu(s)}{d_t^\pi(s)},
\end{aligned}$$

where the first equation follows from the law of total expectation, the other equations are all follows from the conditional independence. Thus, we complete the proof of unbiasedness. □

Proof of Theorem 3. For the case of $\widehat{w}_t^n$ is (3.4), we have the following inequality according to the Hoeffding's inequality [Hoeffding, 1963]

$$\mathbb{P}\left(\left|\widehat{w}_t^n(s \in \mathcal{S}_t) - w_t(s \in \mathcal{S}_t)\right| \geq \varepsilon_1\right) \leq 2\exp\left(-2n_t(s)\varepsilon_1^2\eta^{2t}\right). \tag{A.5}$$

Since $\mathbb{E}[n_t(s)] = nd_t^\mu(s)$, we also have the following inequality according to the Hoeffding's inequality

$$\mathbb{P}\left(nd_t^\mu(s) - n_t(s) \leq \varepsilon_2\right) \leq \exp\left(-2n\varepsilon_2^2\right). \tag{A.6}$$

By setting $2\exp\left(-2n_t(s)\varepsilon_1^2\eta^{2t}\right) = \exp\left(-2n\varepsilon_2^2\right) = \delta/2$, we can obtain

$$\begin{aligned}
2\exp\left(-2n_t(s)\varepsilon_1^2\eta^{2t}\right) =& \frac{\delta}{2} \\
2n_t(s)\varepsilon_1^2\eta^{2t} =& \log(4/\delta) \\
\varepsilon_1 =& \frac{1}{\eta^t}\sqrt{\frac{\log(4/\delta)}{2n_t(s)}},
\end{aligned} \tag{A.7}$$

and

$$\begin{aligned}
\exp\left(-2n\varepsilon_2^2\right) =& \frac{\delta}{2} \\
2n\varepsilon_2^2 =& \log(2/\delta) \\
\varepsilon_2 =& \sqrt{\frac{\log(2/\delta)}{2n}}.
\end{aligned} \tag{A.8}$$

By combining (A.6) and (A.8), we know that with probability at least $1-\delta/2$, we have

$$n_t(s) \geq nd_t^{\mu}(s) - \sqrt{\frac{\log(2/\delta)}{2n}}. \tag{A.9}$$

Thus, substitute (A.9) into (A.7), we can obtain the following inequality holds with probability at least $1-\delta/2$

$$\begin{aligned}\varepsilon_1 \leq & \frac{1}{\eta^t}\sqrt{\frac{\log(4/\delta)}{2\left(nd_t^{\mu}(s) - \sqrt{\frac{\log(2/\delta)}{2n}}\right)}} \\ = & \frac{1}{\eta^t}\sqrt{\frac{\log(4/\delta)\sqrt{2n}}{2\left(nd_t^{\mu}(s)\sqrt{2n} - \sqrt{\log(2/\delta)}\right)}}.\end{aligned}$$

Then, based on the inequality (A.5), we completes the proof. □

At last, we provide the proof of theorem 4.

Proof of Theorem 4. We now provide the proof of the high-confidence bound of $\widetilde{w}_t^n$ in (3.5). We also have the following inequality according to the Hoeffding's inequality [Hoeffding, 1963] for $\widehat{w}_t^n(s)$ and $n_t(s)$.

$$\begin{aligned}\mathbb{P}\left(\left|\widehat{w}_t^n(s \in \mathcal{S}_t) - w_t(s \in \mathcal{S}_t)\right| \geq \varepsilon_1\right) \leq & 2\exp\left(-2n_t(s)\varepsilon_1^2\eta^{2t}\right), \\ \mathbb{P}\left(nd_t^{\mu}(s) - n_t(s) \leq \varepsilon_2\right) \leq & \exp\left(-2n\varepsilon_2^2\right).\end{aligned} \tag{A.10}$$

In addition, for the part $n/\sum_{i=1}^n \prod_{t'=0}^{t-1} \frac{\pi(a_{t'}^i|s_{t'}^i)}{\mu(a_{t'}^i|s_{t'}^i)}$ in (3.5), we have

$$\mathbb{P}\left(\left|\sum_{i=1}^n \prod_{t'=0}^{t-1} \frac{\pi(a_{t'}^i|s_{t'}^i)}{\mu(a_{t'}^i|s_{t'}^i)} - n\right| \geq \varepsilon_3\right) \leq 2\exp\left(-2n\varepsilon_3^2\eta^{2t}\right). \tag{A.11}$$

By setting $2\exp\left(-2n_t(s)\varepsilon_1^2\eta^{2t}\right) = \exp\left(-2n\varepsilon_2^2\right) = 2\exp\left(-2n\varepsilon_3^2\eta^{2t}\right) = \delta/3$, we can obtain

$$\begin{aligned}2\exp\left(-2n_t(s)\varepsilon_1^2\eta^{2t}\right) = & \frac{\delta}{3} \\ 2n_t(s)\varepsilon_1^2\eta^{2t} = & \log(6/\delta) \\ \varepsilon_1 = & \frac{1}{\eta^t}\sqrt{\frac{\log(6/\delta)}{2n_t(s)}},\end{aligned} \tag{A.12}$$

$$\begin{aligned}\exp\left(-2n\varepsilon_2^2\right) = & \frac{\delta}{3} \\ 2n\varepsilon_2^2 = & \log(3/\delta) \\ \varepsilon_2 = & \sqrt{\frac{\log(3/\delta)}{2n}},\end{aligned} \tag{A.13}$$

and

$$\begin{aligned}2\exp\left(-2n\varepsilon_3^2\eta^{2t}\right) = & \frac{\delta}{3} \\ 2n\varepsilon_3^2\eta^{2t} = & \log(6/\delta) \\ \varepsilon_3 = & \frac{1}{\eta^t}\sqrt{\frac{\log(6/\delta)}{2n}}.\end{aligned} \tag{A.14}$$

By combining (A.10) and (A.13), we know that with probability at least $1 - \delta/3$, we have

$$n_t(s) \geq nd_t^{\mu}(s) - \sqrt{\frac{\log(3/\delta)}{2n}}. \tag{A.15}$$

Then, substitute (A.15) into (A.12), we can obtain the following inequality holds with probability at least $1 - \delta/3$

$$\begin{aligned} \varepsilon_1 \leq & \frac{1}{\eta^t} \sqrt{\frac{\log(6/\delta)}{2\left(nd_t^{\mu}(s) - \sqrt{\frac{\log(3/\delta)}{2n}}\right)}} \\ = & \frac{1}{\eta^t} \sqrt{\frac{\log(6/\delta)\sqrt{2n}}{2\left(nd_t^{\mu}(s)\sqrt{2n} - \sqrt{\log(3/\delta)}\right)}}. \end{aligned} \tag{A.16}$$

Based on the definition of $\widetilde{w}_t^n(s)$ in (3.5) and (A.11), we have the following inequality holds with probability at least $1 - 2\delta/3$

$$|\widetilde{w}_t^n(s \in \mathcal{S}_t) - w_t(s \in \mathcal{S}_t)| \leq \frac{w_t(s) + \varepsilon_1}{1 - \varepsilon_3/n} - w_t(s) = \frac{n\varepsilon_1 + w_t(s)\varepsilon_3}{n - \varepsilon_3}. \tag{A.17}$$

Combining (A.17) with the definition of ε_3 in (A.14) and the high probability bound of ε_1 in (A.16), we can obtain our final high probability bound for $\widetilde{w}_t^n(s)$:

$$\begin{aligned} & |\widetilde{w}_t^n(s) - w_t(s)| \\ \leq & \frac{n\sqrt{2n}}{\eta^t n\sqrt{2n} - \sqrt{\log(6/\delta)}} \sqrt{\frac{\log(6/\delta)\sqrt{2n}}{2\left(nd_t^{\mu}(s)\sqrt{2n} - \sqrt{\log(3/\delta)}\right)}} + \frac{w_t(s)\sqrt{\log(6/\delta)}}{\eta^t n\sqrt{2n} - \sqrt{\log(6/\delta)}} \end{aligned}$$

holds with probability at least $1 - \delta$.

Also, since the consistent estimator $\widetilde{w}_t^n(s)$ in (3.5) contains self-normalization, we can also obtain its upper bound as

$$\widetilde{w}_t^n(s) = \frac{\sum_{i=1}^n \prod_{t'=0}^{t-1} \frac{\pi(a_{t'}^i|s_{t'}^i)}{\mu(a_{t'}^i|s_{t'}^i)} \mathbf{1}(s_t^i = s)}{\frac{1}{n}\sum_{i=1}^n \prod_{t'=0}^{t-1} \frac{\pi(a_{t'}^i|s_{t'}^i)}{\mu(a_{t'}^i|s_{t'}^i)} \sum_{j=1}^n \mathbf{1}(s_t^j = s)} \leq \frac{n}{\sum_{j=1}^n \mathbf{1}(s_t^j = s)}.$$

Since we also have

$$\mathbb{P}\left(\left|\sum_{j=1}^n \mathbf{1}(s_t^j = s) - nd_t^{\mu}(s)\right| > \sqrt{\frac{\log(2/\delta)}{2n}}\right) \leq \delta$$

from the Hoeffding's inequality, then $\widetilde{w}_t^n(s)$ can be upper bounded by $\sqrt{2n}/(d_t^{\mu}(s)\sqrt{2n} - \sqrt{\log(2/\delta)})$ with probability at least $1 - \delta$. This completes the proof. □