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 (OPE) considers evaluating the performance of a policy using the historical data obtained by different behavior policies. In the real-world application of reinforcement learning (RL), acting a policy can be costly and dangerous, and OPE usually plays as a crucial step. Vanilla model-free OPE methods use importance sampling to unbiasedly estimate the expectation of the new-policy values, which, however, may suffer from large variance that depends exponentially on the RL horizons. In this paper, we view the model-free approaches as distribution-shift estimators for random walks on the Markov decision trees (Tree MDPs) induced by rolling in different policies. We propose a new OPE approach directly based on the discrete directed acyclic graph Markov decision processes (DAG MDPs), which leverages the observable states to reduce variance to a polynomial function of the horizons. Besides, different from model-based approaches, our DAG-MDPs estimate the state distributions via marginalized importance ratios, which is robust to partially observable states and model biases. Our approach can be applied to most of the estimators of OPE without change to 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. In many settings where RL algorithms have been already deployed, e.g., targeted advertising and marketing [Bottou et al., 2013; Tang et al., 2013; Chapelle et al., 2015; Theocharous et al., 2015; Thomas et al., 2017] or medical treatments [Murphy et al., 2001; Ernst et al., 2006; Raghu et al., 2017], online policy evaluation is usually expensive, risky, or even unethical. Also, using a bad policy in these applications is dangerous and could lead to severe consequences. Thus, solving OPE is often the starting point in many RL applications.

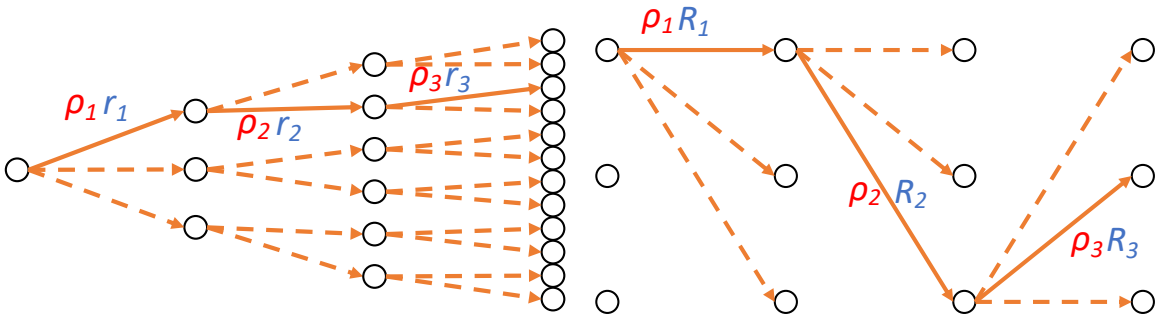
To tackle the problem of OPE, the idea of importance sampling (IS) corrects the mismatch in the distributions under the behavior policy and estimated policy. It also provides typically unbiased or strongly consistent estimators [Precup et al., 2000]. IS-based off-policy evaluation methods have also seen lots 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 tends to be too high to be useful [Precup et al., 2000; Thomas et al.,

*Most of this work performed at Amazon AI.

2015; Jiang and Li, 2016; Thomas and Brunskill, 2016; Guo et al., 2017; Farajtabar et al., 2018], especially for long-horizon problems [Mandel et al., 2014], because the variance of the cumulative product of importance weights may grow exponentially as the horizon goes long.

Given this high-variance issue, it is necessary to find an IS-based approach without relying heavily on the cumulative product of importance weights. Most state-of-the-art off-policy evaluation methods use the cumulative product of importance weights to re-weight the distribution of the whole trajectories. If we view each single trajectory as a depth-first random walk on a Markov decision tree, we call it Tree-MDP, where each level defines a single time step and each node defines a state at the specific time, and we may notice that no information can be shared between different time levels. I.e., re-weighting of the entire trajectory, which is required by the Tree-MDP models, may not be necessary, if some states are directly observable, allowing Markov independence assumptions. One simple example of the alternative is when the states are directly observable, yet the transitions are time-variant. In this case, we only need to use IS to correct for the biases in the transition models.

Generalizing this observation, we define discrete directed-acyclic graph Markov decision process (DAG MDPs) as a relaxation of Tree MDPs. Just as the Tree MDPs, the regions of reachable state space of DAG-MDPs in different time step are also disjoint, but DAG MDPs make the Markov assumption by allowing trajectories that separate in early steps to reunions at some states [Jiang and Li, 2016]. Markov assumption is the key to avoiding the cumulative product of importance weights, and it could reduce the working space from the trajectory space to the (different time step's) reachable state space. Figure 1 illustrates the difference between Tree MDPs and DAG MDPs.



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

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

Related work [Liu et al., 2018a] also tackles the high variance issue due to the cumulative product of importance weights. The key idea of their approach is to apply importance sampling on the average visitation distribution of state-action pairs, instead of the distribution of the whole trajectories. Their method addressed the high-variance issue in long-horizon MDPs. The main difference between this prior work and ours is that our approach covers the time-varying or partially observable MDP setting [Zinkevich et al., 2006; White and Bowling, 2009] if there exist observable states, while the methods based on estimating the average visitation distribution mainly focus on the time-invariant and fully observable MDPs. The partial observability can also be the obstacle of model-based approaches [Liu et al., 2018b], where the approximated MDP model may not converge to the true MDP due to the partial observation [Thomas and Brunskill, 2016].

In this paper, we propose a marginalized method for off-policy evaluation based on the DAG MDPs, while most of the previous methods are all based on the Tree MDPs. Our approach can be used in most of OPE estimators that leverage IS-based estimators, such as doubly robust [Jiang and Li, 2016], MAGIC [Thomas and Brunskill, 2016], MRDR [Farajtabar et al., 2018] under mild

assumptions (Markov assumption). 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. Our theoretical analysis of our marginalized estimators shows that our approach only depends polynomially on the horizon, while the IS-based methods usually have exponential dependencies. Our experimental results demonstrate the effectiveness of the marginalized OPE estimators across a number of problems.

Here is a road map for the rest of the paper. Section 2 provides the preliminaries of the problem of off-policy evaluation. In Section 3, we summarize the previous IS-based off-policy estimators, and we offer a new framework for the marginalized estimators. We present the details of our estimation methods of marginalized density ratio in Section 4 with its theoretical analysis, and provide the empirical results in Section 5. At last, Section 6 provides the conclusion and future work.

2 Preliminaries

We consider the problem of off-policy evaluation for a 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 $T(s'|s, a)$ 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 $\mathcal{A}$, respectively, 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, such that $T_t[s, s', a]$ is the probability to transition from s into s' when action a is taken at time t. When the an MDP is time-invariant, let $T = T_t, \forall\ 0 \le t \le H - 1$.

We use $\mathbb{P}[E]$ to denote the probability of an event E and $p(x)$ the 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 the form 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 episode 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

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

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 these 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 $\rho_t^i \coloneqq \frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}$ be the importance ratio at time step t of i-th trajectory, and $\rho_{0:t}^i \coloneqq \prod_{t'=0}^{t} \frac{\pi(a_{t'}^i|s_{t'}^i)}{\mu(a_{t'}^i|s_{t'}^i)}$ be the cumulative importance ratio for the i-th trajectory. We also use $\rho_t(s_t, a_t)$ to denote $\pi(a_t|s_t)/\mu(a_t|s_t)$ over this paper. The generic framework of IS-based estimators can be expressed as follows

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

where $\phi_t : \mathbb{R}_+^n \to \mathbb{R}_+$ are the "self-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:

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

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

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

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:

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

and the WDR estimator:

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

Note that, the DR estimator reduced the variance from the stochasticity of action by using the technique of control variate $f_t(s_t^i, a_t^i, s_{t+1}^i)$ in value-related function, and the WDR estimators reducing variance by the bias-variance trade-off using self-normalization, especially in the presence of weight clipping [Bottou et al., 2013]. However, both could still suffer large variance, because the cumulative importance ratio $\rho_{0:t}^i$ always appear directly in this framework, which makes the variance to increase exponentially as 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., $s_t = h_t \coloneqq o_1 a_1 \dots o_{t-1} a_{t-1} o_t$, where o_i is the observation at step $i (1 \le i \le t)$.*
- *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 MDPs 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) MDPs, 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 $d_t^\pi(s) \coloneqq \mathbb{P}[s_t = s|\pi]$, $d_t^\mu(s) \coloneqq \mathbb{P}[s_t = s|\mu]$ be the marginal distribution of state s at time step t, then given the conditional independence following from the Markov property, we can obtain the

following results

$$
\begin{aligned}
w_t(s_t) :=& \frac{d_t^\pi(s_t)}{d_t^\mu(s_t)} \\
\Rightarrow v(\pi) =& \sum_{t=0}^{H-1} \mathbb{E}_{\tau\sim\pi}[r_t] = \sum_{t=0}^{H-1} \mathbb{E}_{(s_t,a_t)\sim\pi}[r_t] \\
=& \sum_{t=0}^{H-1} \mathbb{E}_{(s_t,a_t)\sim\mu}[w_t(s_t) r_t]
\end{aligned} \tag{3.2}
$$

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 $\widehat{w}_t(s)$. 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:

$$
\begin{aligned}
\widehat{v}_M(\pi) =& \frac{1}{n}\sum_{i=1}^{n} g(s_0^i) \\
&+ \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{H-1} \widehat{w}_t(s_t^i)\rho_t^i\gamma^t(r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i)).
\end{aligned} \tag{3.3}
$$

Note that the "self-normalization" function ϕ has not appeared in the framework above is because we can implement the self-normalization within the estimate of $w_t(s)$. We will discuss this property in detail in the next section.

We first show the equivalence between framework (3.1) and framework (3.3) in expectation if $\phi_t(\rho_{0:t}^{1:n}) = n$ and $\widehat{w}_t(s) = w_t(s)$.

Lemma 1. *If $\phi_t(\rho_{0:t}^{1:n}) = n$ in framework (3.1) and $\widehat{w}_t(s) = w_t(s)$ in framework (3.3), then these two frameworks are equal in expectation, i.e.,*

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

holds for all i and t.

Full proof of Lemma 1 can be found in Appendix A.

Next, we show that if we have an unbiased or consistent estimate $\widehat{w}_t$ of w_t, the IS-based OPE estimators that simply replace $\prod_{t'=0}^{t-1}\frac{\pi(a_{t'}|s_{t'})}{\mu(a_{t'}|s_{t'})}$ with $\widehat{w}_t(s_t)$ will remain unbiased or consistent.

Theorem 1. *Let $\phi_t(\rho_{0:t}^{1:n}) = n$ in framework (3.1), then framework (3.3) could keep the unbiasedness and consistency same as in framework (3.1) if $\widehat{w}_t(s)$ 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 (3.3) is also a unbiased estimator of $v(\pi)$ given unbiased estimator $\widehat{w}_t(s)$ for all t.*

2. *If a consistent estimator falls in framework (3.1), then its marginalized estimator in framework (3.3) is also a consistent estimator of $v(\pi)$ given consistent estimator $\widehat{w}_t(s)$ for all t.*

Full proof of Theorem 1 can be found in Appendix A.

In partially observable MDPs (POMDPs), we may not be able to obverse all states. However, if there exist any observable states, our marginalized approach could leverage these observable states to reduce variance. That is, we use the partial trajectory from the closest observable states to the current time step to represent the current state. Assume the current time step is t and the closest observable states is s_{t-L} at time step $t-L$, then we can use $\frac{d_t^\pi(s_{t-L})}{d_t^\mu(s_{t-L})}\prod_{i=t-L}^{t-1}\frac{\pi(a_i|s_i)}{\mu(a_i|s_i)}$ as $w_t(s_t)$, while other IS-based methods are equivalent to using $\prod_0^{t-1}\frac{\pi(a_i|s_i)}{\mu(a_i|s_i)}$ as $w_t(s_t)$. The observable states in POMDPs can be considered as the states that can be reunioned at in the DAG MDPs. If there is no observable state in POMDPs, then it is equivalent that DAG MDPs is reduced to tree MDPs.

Finally, 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 (3.3) 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 MDPs, 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 (3.3) 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

$$\begin{aligned}\widehat{v}_{BM}(\pi) &= \frac{1}{n}\sum_{i=1}^n g(s_0^i) \\ &+ \frac{1}{n}\sum_{i=1}^n\sum_{t=0}^{H-1}\widehat{w}_t(s_t^i)\rho_t^i\gamma^t(\widehat{R}(s_t^i,a_t^i) + f_t(s_t^i,a_t^i,s_{t+1}^i)).\end{aligned} \tag{3.4}$$

Remark 1. *Note that, the only difference between (3.3) and (3.4) is r_t^i and $\widehat{R}(s_t^i,a_t^i)$. Thus, the unbiasedness or consistency of (3.4) 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 (3.4) reduces to (3.3). Note that when this happens, we could still potentially estimate $\widehat{w}_t^n(s_t)$ well if $|\mathcal{A}|$ is large but $|\mathcal{S}|$ is relative small, in which case we can still afford to observe each potential values of s_t many times.

4 Marginalized Density Ratio Estimation

In this section, we present our estimators for $w_t(\cdot)$ which is defined by the marginalized ratio $d_t^\pi(\cdot)/d_t^\mu(\cdot)$ at time step t. Since the initial state distribution $d_0(\cdot)$ and the marginalized state distribution of behavior policy $d_t^\mu(\cdot)$ are same as the data distribution, which allows us to obtain the on-policy estimation for $d_0(\cdot)$ and $d_t^\mu(\cdot)$ directly, we will focus on the estimate of $d_t^\pi(\cdot)$ in this section. In the following parts, we use $(s_t,a_t)\sim d_t^\pi$ to denote drawing from distribution $d_t^\pi(s_t,a_t) \coloneqq d_t^\pi(s_t)\pi(a_t|s_t)$, and we also use $(s_t,a_t)\sim d_t^\mu$ similarly. We define $T_t^\pi(s'|s) \coloneqq \sum_a \pi(a|s)T_t^\pi(s'|s,a)$ as the state transition probability from s to s' at time step t under policy π.

Our idea to estimate $d_t^\pi(\cdot)$ is to use the relationship between $d_t^\pi(\cdot)$ and $d_{t+1}^\pi(\cdot)$ to optimize the data efficiency. Given the definition of $T_t^\pi(s'|s)$, we have $d_{t+1}^\pi(\cdot)$ is satisfying

$$\begin{aligned} d_t^\pi(s_{t+1}) =& \mathbb{E}_{s_t \sim d_t^\pi}\left[T_t^\pi(s_{t+1}|s_t)\right] \\ =& \mathbb{E}_{(s_t,a_t)\sim d_t^\pi}\left[T_t(s_{t+1}|s_t,a_t)\right] \\ =& \mathbb{E}_{(s_t,a_t)\sim d_t^\mu}\left[\frac{d_t^\pi(s_t)}{d_t^\mu(s_t)}\frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}T_t(s_{t+1}|s_t,a_t)\right]. \end{aligned}$$

The equality above can be leveraged to estimate $d_t^\pi(\cdot)$ recursively. Note that we have $d_0^\mu(\cdot) = d_0^\mu(\cdot) = d_0(\cdot)$, and then our first estimator for $d_t^\pi(\cdot)$ is as follows

$$\widehat{d}_{t+1}^\pi(s) = \frac{1}{n}\sum_{i=1}^n \frac{\widehat{d}_t^\pi(s_t^i)}{\widehat{d}_t^\mu(s_t^i)}\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}\mathbf{1}(s_{t+1}^i = s). \tag{4.1}$$

Since $d_t^\pi(\cdot)$ always satisfies the normalization property (i.e. $\sum_{s\in\mathcal{S}} d_t^\pi(s) = 1$), we can obtain another estimator of $d_t^\pi(\cdot)$ using self-normalization as follows

$$\widetilde{d}_{t+1}^\pi(s) = \frac{\sum_{i=1}^n \frac{\widehat{d}_t^\pi(s_t^i)}{\widehat{d}_t^\mu(s_t^i)}\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}\mathbf{1}(s_{t+1}^i = s)}{\sum_{i=1}^n \frac{\widehat{d}_t^\pi(s_t^i)}{\widehat{d}_t^\mu(s_t^i)}\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}}. \tag{4.2}$$

Estimators (4.1) and (4.2) above provide the off-policy estimate of $\widehat{d}_t^\pi(\cdot)$, which can then be plugged in to (3.2) to estimate $\widehat{w}_t(s) = \widehat{d}_t^\pi(s)/\widehat{d}_t^\mu(s)$ recursively and $\widehat{v}(\pi)$ eventually. Notice, the state distributions induced by the behavior policies can be obtained via direct counting $\widehat{d}_t^\mu(s) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(s_t^i = s)$. Further, when the MDPs induced by the behavior policies have small mixing diameters, the $\widehat{d}_t^\mu$ can be estimated rather accurately. Algorithm 1 in Appendix C shows the complete DAG-MDP estimators in pseudo codes.

4.1 Theoretical Analysis

In this section, we provide a theoretical analysis of our estimators of $w_t(\cdot)$. We study the error propagation of our recursive estimators. Our theoretical results show that the dependency of our recursive estimators is polynomial on the horizon. This is in contrast to the Tree-IS estimators, which may have exponential dependencies. Example 1 shows one case of the exponential dependencies of the Tree-IS estimators.

Example 1. *Vanilla Tree-IS weights are products of single-step IS weights, i.e.,* $\log \rho_{0:H} = \sum_{t=0}^{H-1} \log \rho_t$. *Let* $E_{\log} = \mathbb{E}\log\rho_t$ *and* $V_{\log} = \mathbb{V}\log\rho_t$ *be the mean and variance of the single-step log-IS weights and, for simplicity, suppose the policies are iid at different times. By Central Limit Theorem,* $\log\rho_{0:H}$ *asymptotically follows a normal distribution with parameters* $\left(-HE_{\log}, HV_{\log}\right)$. *As a result, the Tree-IS weights asymptotically follow a log-normal distribution with variance* $\left(\exp\left(HV_{\log}\right) - 1\right)$, *which is exponential in the horizon* H.

We first consider $\widehat{d}_0(\cdot)$ and $\widehat{d}_t^\mu(\cdot)$, since we can use on-policy data to estimate these distributions. Hoeffding's inequality implies the following fact.

Fact 1. *For* $\widehat{d}_0(s)$ *and* $\widehat{d}_t^\mu(s_t^i)$, $\forall s \in \mathcal{S}$, *w.p.* $1-\delta$, *we have*

$$\widehat{d}_0(s) = d_0(s) + \varepsilon_0(s) = d_0(s) + \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\right)$$
$$\widehat{d}_t^\mu(s) = d_t^\mu(s) + \varepsilon_t^\mu(s) = d_t^\mu(s) + \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\right).$$

Error propagation We now discuss the error propagation of our first recursive estimator for $d_t^\pi(\cdot)$ (defined in (4.1)). Over this section, we assume there exists a $\eta > 0$, such that $\pi(a|s)/\mu(a|s) \le 1/\eta$ for any (s, a). We also assume there exists $\{\eta_t > 0\}_{t=0}^{H-1}$, such that $d_t^\mu(s) \ge \eta_t$ for all s and t. Thus, we have theorems for error propagation as follows.

Theorem 2. *Let $\widehat{d}_t^\pi(\cdot)$ be the estimator of $d_t^\pi(\cdot)$, defined in (4.1). Let $\widehat{\varepsilon}_t^\pi(s) \coloneqq \widehat{d}_t^\pi(s) - d_t^\pi(s)$, then we have*

$$\sum_s |\widehat{\varepsilon}_{t+1}^\pi(s)| \le \left(1 + \widetilde{\mathcal{O}}\left(|\mathcal{S}|\sqrt{\frac{\eta^2}{n}}\right)\right) \sum_{s_t} |\varepsilon_t^\pi(s_t)| + \widetilde{\mathcal{O}}\left(|\mathcal{S}|\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}}\right).$$

We provide the detailed proof of Theorem (2) in Appendix B. Theorem 2 provides the properties of error propagation for the recursive estimator (4.1). It implies that $\sum_s |\widehat{\varepsilon}_{t+1}^\pi(s)|$ only has polynomial dependency on the horizon if n is large enough. We discuss this dependency later in this section.

The error propagation for the recursive estimator (4.2) is as follows

Theorem 3. *Let $\widetilde{d}_t^\pi(\cdot)$ be the estimator of $d_t^\pi(\cdot)$, defined in (4.2). Let $\widetilde{\varepsilon}_t^\pi(s) \coloneqq \widetilde{d}_t^\pi(s) - d_t^\pi(s)$, and $\varepsilon_{t+1}^\pi(s) \coloneqq \frac{1}{n}\sum_{i=1}^n \frac{\widetilde{d}_t^\pi(s_t^i)}{\widetilde{d}_t^\mu(s_t^i)} \frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)} \mathbf{1}(s_{t+1}^i = s) - d_{t+1}^\pi(s)$, then we have*

$$\begin{aligned}
&\sum_s |\widetilde{\varepsilon}_{t+1}^\pi(s)| \\
\le &\frac{1 + \widetilde{\mathcal{O}}\left(|\mathcal{S}|\sqrt{\frac{\mathbb{E}_\mu[\rho_t^2|s_t]}{n\eta_t}} + \frac{|\mathcal{S}|}{n\eta\eta_t}\right)}{1 + \sum_s \varepsilon_{t+1}^\pi(s)} \sum_{s_t} |\widetilde{\varepsilon}_t^\pi(s_t)| \\
&+ \frac{1}{1 + \sum_s \varepsilon_{t+1}^\pi(s)} \widetilde{\mathcal{O}}\left(|\mathcal{S}|\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}}\right).
\end{aligned}$$

where

$$\begin{aligned}
\sum_s \varepsilon_{t+1}^\pi(s) &= \widetilde{\mathcal{O}}\left(|\mathcal{S}|\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}}\right) \\
&+ \widetilde{\mathcal{O}}\left(\sum_s \sum_{s_t} \left(\sqrt{\frac{\mathbb{E}_\mu[\rho_t^2|s_t]}{n\eta_t}} + \frac{1}{n\eta\eta_t}\right) \widetilde{\varepsilon}_t^\pi(s_t)\right).
\end{aligned}$$

We provide the detailed proof of Theorem (3) in Appendix B. The error propagation of the estimator with self-normalization is quantitatively the same as the estimator without self-normalization. But it strictly reduces the error if $\sum_s \varepsilon_{t+1}^\pi(s) > 0$.

The next Corollary discuss the estimated error $\sum_s |\widehat{\varepsilon}_t^\pi(s)|$ based on the Theorems about error propagation. It shows the linear dependency on horizon H if n is large enough.

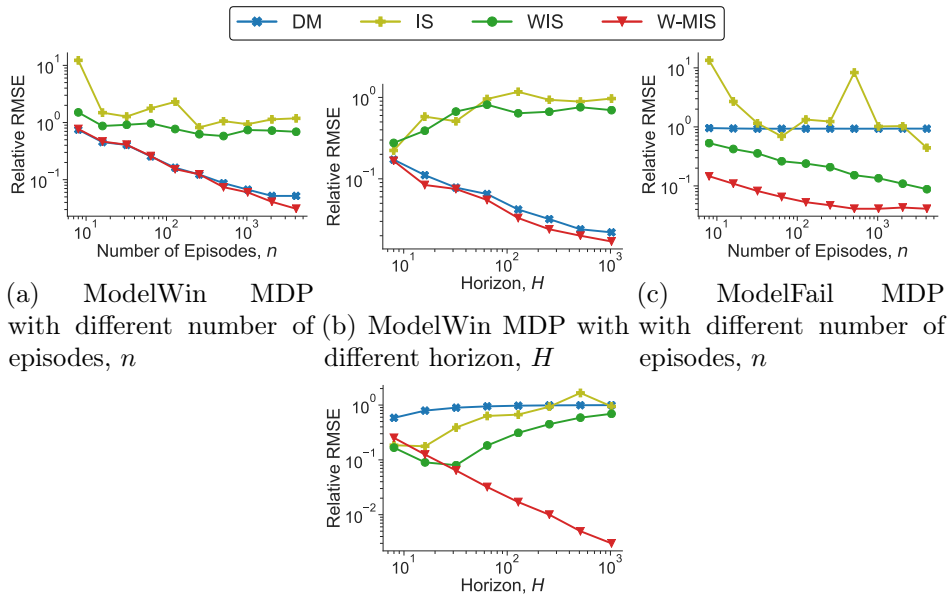
Corollary 1. *Let $\rho_t(s_t, a_t) \le 1/\eta$. Let $\widehat{\varepsilon}_t^\pi(s) \coloneqq \widehat{d}_t^\pi(s) - d_t^\pi(s)$.*

If $n \gg \frac{|S|^2 H^2}{\eta^2}$, then with high probability,

$$\sum_s |\widehat{\varepsilon}_t^\pi(s)| = \widetilde{\mathcal{O}}\left(\frac{H|S|}{\eta\sqrt{n}}\right).$$

The detailed proof of Corollary 1 is provided in Appendix B.

Remark 2. *Corollary 1 provide an upper bound of estimated error* $\sum_s |\widehat{\varepsilon}_t^{\pi}(s)|$ *based on its error propagation. Based on that Corollary above, we can obtain an absolute error bound of our estimator for the off-policy policy evaluation problem. If the reward boundedness and we can estimate the reward function* $\mathbb{E}[r_t|s_t, a_t]$ *perfectly, our bounds on marginalized distribution* $d_t^{\mu}(\cdot)$ *and* $d_t^{\pi}(\cdot)$ *will give the off-policy policy evaluation an absolute error bound of* $\widetilde{\mathcal{O}}\left(\frac{H^2|S|}{\eta\sqrt{n}}\right)$.



(a) ModelWin MDP with different number of episodes, n

(b) ModelWin MDP with different horizon, H

(c) ModelFail MDP with different number of episodes, n

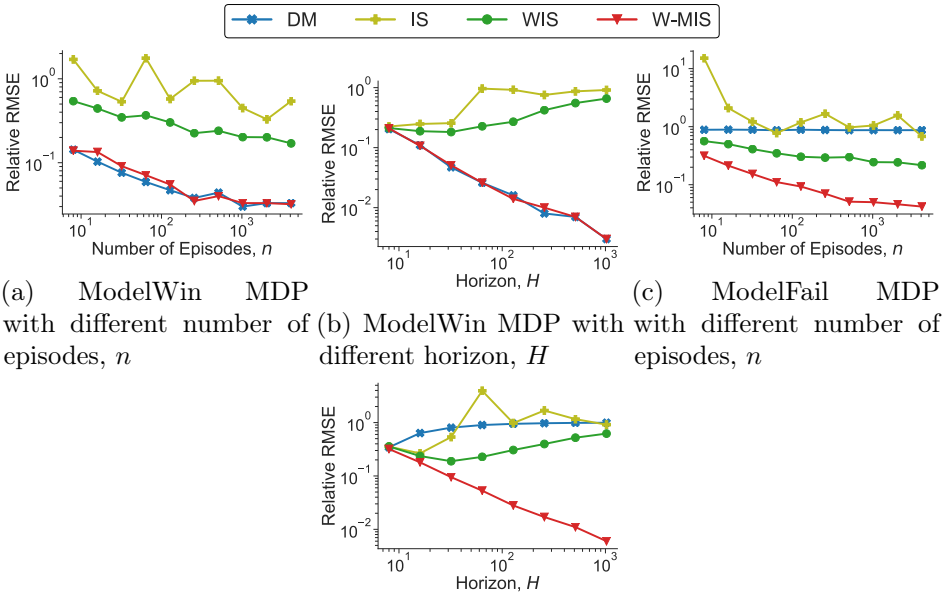
(d) ModelFail MDP with different horizon, H

Figure 3: Results on Time-invariant MDPs. Over this paper, both axes always use a logarithmic scale, and the results are from 128 runs. This figure shows the relative root of mean squared error (relative RMSE, the ratio of RMSE and true cumulative reward) of different estimators as n, the number of episodes, increases; and the relative RMSE of different estimators as H, the length of the horizon, increases.

5 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 estimators with their marginalized versions on several common domains and their variants.

Since our marginalized approaches are based on the DAG-MDPs model, which leverages the observable states to reduce variance for even POPMPs and time-varying MDPs, our experiments include two main parts: time-invariant MDPs and time-varying MDPs. The methods we compared in this section are DM, IS, WIS and W-MIS, where DM denotes the direct method of approximating MDP, IS denotes the importance sampling method, WIS denote the weighted (self-normalized) importance sampling method, and W-MIS is the marginalized importance sampling method with $\widehat{w}_t$ using (4.2) which empirically achieve better performance than $\widehat{w}_t$ using (4.1).



(a) ModelWin MDP with different number of episodes, n

(b) ModelWin MDP with different horizon, H

(c) ModelFail MDP with different number of episodes, n

(d) ModelFail MDP with different horizon, H

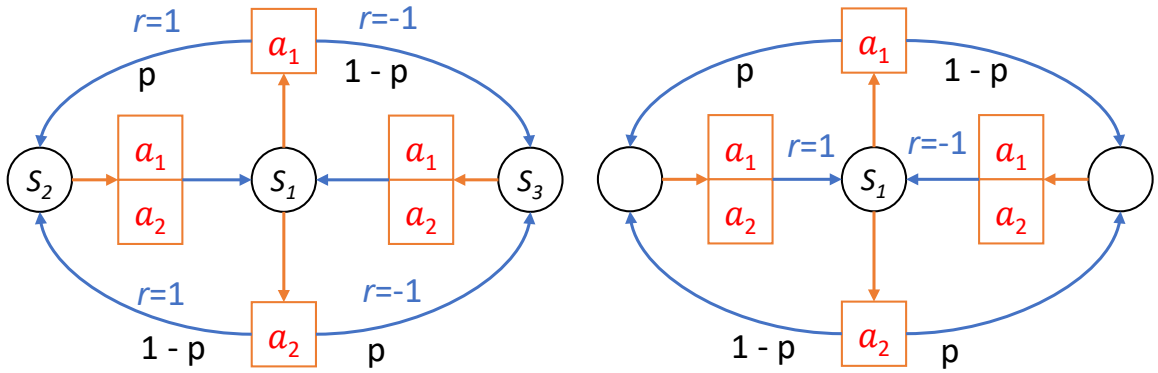
Figure 4: Results on time-varying MDPs. This figure shows the relative RMSE of different estimators as n, the number of episodes, increases; and the relative RMSE of different estimators as H, the length of the horizon, increase.

5.1 Time-invariant MDPs

In this section, we present the experimental results in the time-invariant MDPs. We first provide results on MDPs that modified on some standard domains, ModelWin and ModelFail, which are first introduced by Thomas and Brunskill [2016].

The **ModelWin** domain is depicted in Figure 2(a), simulates the case of fully observed MDPs. The agent could observe the true underlying states of the ModelWin MDP. It 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. The evaluated policy π for the ModelWin MDP 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 a uniform policy. The discount factor γ in the ModelWin MDP is 1.

The **ModelFail** domain is depicted in Figure Figure 2(b), where the agent can only tell the difference between $s1$ and other states. The dynamic of the ModelFail MDP is similar to the ModelWin MDP, with $p = 1$, but we delay the reward receiving – the agent receives a reward of 1 when it arrives s_1 from the left-most state, and receives a reward of -1 when it arrives s_1 from the right-most state. In this case, model-based estimators would fail because the reward is not related to the current state, but to the previous state, instead. The discount factor $\gamma = 1$. The evaluated policy π for the ModelFail MDP 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 a uniform policy. The discount factor



(a) ModelWin MDP (b) ModelFail MDP

Figure 2: Domains based on [Thomas and Brunskill, 2016]

γ in the ModelFail MDP is 1.

We provide two types of experiments to show the properties of our marginalized approach. The first kind of experiments is with different number of episodes, where we use the horizon of $H = 50$. The second one is with different horizon, where we use the number of episodes $n = 1024$. Since we consider partial observation in the ModelFail MDP where the agent can only tell the difference between s_1 and all unobservable "others" as a single state, we define a "trajectory-state", that is the partial trajectory starting from s_1 until reaching s_1 again as a single state, as we discussed in Section 3.

Figure 3 shows the results in the time-invariant ModelWin MDP and ModelFail MDP. Note that we use the ratio of RSME and true cumulative reward as relative RMSE. In both time-invariant ModelWin MDP and ModelFain MDP, the true return is directly proportional to H. The results clearly show that W-MIS maintains a polynomial dependency on H and matches the best alternatives such as DM in Figure 3(b) and IS at the beginning of Figure 3(d). Notably, the IS in Figure 3(d) reflects a bias-variance trade-off, that its RMSE is smaller at short horizons due to unbiasedness yet larger at long horizons due to high variance.

A careful study of Figure 3(b) find the slope of W-MIS to approach -0.5, i.e., the mean square error of our approach may depend linearly on H. The rate is better than our theoretical prediction because the policies are iid in time and the reward distributions are independent of the state distributions (which are uniform). Our theoretical analysis is targeted at more general RL problems where these dependencies may happen adversarially, yet the exponential-to-polynomial reduction still applies.

5.2 Time-varying MDPs

We also test our approach in the time-varying MDPs. The time-varying MDPs we used in this section are also modified on the standard domains introduced by Thomas and Brunskill [2016]. We use the similar dynamic of ModelWin MDP and ModelFail MDP, but we set the transition probability p to be varying over time. In the time-varying ModelWin MDP, we use $p = 0.5t/H$ as its transition probability, and we use $p = 1 - 0.5t/H$ for the time-varying ModelFail MDP. We use the "trajectory-state" for the ModelFail MDP as before, and the behavior policy μ and evaluated policy π we used are same as the time-invariant case.

Figure 4 shows the relative RMSE in the time-varying ModelWin MDP and ModelFail MDP. We observe the results of Figure 3 are similar to the time-invariant case, which demonstrate the

effectiveness of our approach in the time-varying domains.

5.3 Mountain Car

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 the temperature of 1). For the behavior policy μ, we also use the softmax policy of the optimal Q-function but set the temperature to 1.33.

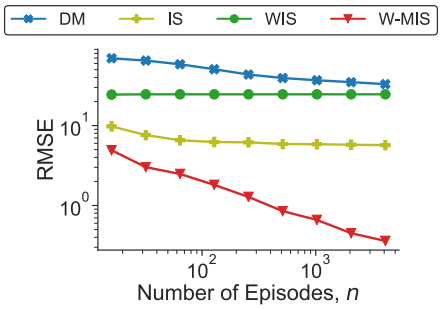


Figure 5: Mountain Car with different number of episodes.

The results on the Mountain Car domain is in Figure 5, which demonstrate the effectiveness of our approach in the common benchmark control task.

6 Conclusions

In this paper, we propose a marginalized approach to solve the problem of off-policy evaluation in reinforcement learning. This is the first approach designed based on the model of DAG MDPs. Our method address the high variance issue of most IS-based approaches in the long horizon problems. The theoretical analysis of the marginalized approach shows that the dependency of our estimators are polynomial on the horizon, while the traditional IS-based methods may have exponential dependencies. Our experiments demonstrate the effectiveness of our approach. It achieves substantially better performance than existing 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.

Ernst, D., Stan, G.-B., Goncalves, J., and Wehenkel, L. (2006). Clinical data based optimal sti strategies for hiv: a reinforcement learning approach. In *Decision and Control, 2006 45th IEEE Conference on*, pages 667–672. IEEE.

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.

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.

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

Liu, Y., Gottesman, O., Raghu, A., Komorowski, M., Faisal, A. A., Doshi-Velez, F., and Brunskill, E. (2018b). Representation balancing mdps for off-policy policy evaluation. In *Advances in Neural Information Processing Systems 31*, pages 2649–2658.

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.

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.

White, M. and Bowling, M. H. (2009). Learning a value analysis tool for agent evaluation. In *IJCAI*, pages 1976–1981.

Zinkevich, M., Bowling, M., Bard, N., Kan, M., and Billings, D. (2006). Optimal unbiased estimators for evaluating agent performance. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 573. Citeseer.

Appendix

A Proof Details of Section 3

In this section, we provide the full proofs for our lemmas and theorems in Section 3. Since we are showing the asymptotic results in this section, we use superscript n for $\widehat{w}_t$ to denote the estimations via n episodes training data over this section.

We first provide the proof of lemma 1.

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

$$\begin{aligned}\mathbb{E}\left[\rho^i_{0:t}(r^i_t + f_t(s^i_t, a^i_t, s^i_{t+1}))\right] =& \mathbb{E}\left[\mathbb{E}\left[\rho^i_{0:t}(r^i_t + f_t(s^i_t, a^i_t, s^i_{t+1}))|s^i_t\right]\right] \\ =& \mathbb{E}\left[\mathbb{E}\left[\rho^i_{0:t-1}|s^i_t\right]\mathbb{E}\left[\rho^i_t(r^i_t + f_t(s^i_t, a^i_t, s^i_{t+1}))|s^i_t\right]\right] \\ =& \mathbb{E}\left[w_t(s^i_t)\mathbb{E}\left[\rho^i_t(r^i_t + f_t(s^i_t, a^i_t, s^i_{t+1}))|s^i_t\right]\right] \\ =& \mathbb{E}\left[w_t(s^i_t)\rho^i_t(r^i_t + f_t(s^i_t, a^i_t, s^i_{t+1}))\right],\end{aligned}$$

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 $\mathbb{E}[\widehat{w}^n_t(s)|s] = w_t(s)$ for all t, then

$$\begin{aligned}\mathbb{E}\left[\widehat{w}^n_t(s^i_t)\rho^i_t\gamma^t(r^i_t + f_t(s^i_t, a^i_t, s^i_{t+1}))\right] =& \mathbb{E}\left[\mathbb{E}\left[\widehat{w}^n_t(s^i_t)\rho^i_t\gamma^t(r^i_t + f_t(s^i_t, a^i_t, s^i_{t+1}))|s^i_t\right]\right] \\ =& \mathbb{E}\left[\mathbb{E}\left[\widehat{w}^n_t(s^i_t)|s^i_t\right]\mathbb{E}\left[\rho^i_t\gamma^t(r^i_t + f_t(s^i_t, a^i_t, s^i_{t+1}))|s^i_t\right]\right] \\ =& \mathbb{E}\left[w_t(s^i_t)\mathbb{E}\left[\rho^i_t\gamma^t(r^i_t + f_t(s^i_t, a^i_t, s^i_{t+1}))|s^i_t\right]\right] \\ =& \mathbb{E}\left[w_t(s^i_t)\rho^i_t(r^i_t + f_t(s^i_t, a^i_t, s^i_{t+1}))\right] \\ =& \mathbb{E}\left[\rho^i_{0:t}(r^i_t + f_t(s^i_t, a^i_t, s^i_{t+1}))\right], \qquad (A.1)\end{aligned}$$

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

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{H-1}\widehat{w}^n_t(s^i_t)\rho^i_t\gamma^t(r^i_t + f_t(s^i_t, a^i_t, s^i_{t+1})) = \sum_{t=0}^{H-1}\gamma^t\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\widehat{w}^n_t(s^i_t)\rho^i_t(r^i_t + f_t(s^i_t, a^i_t, s^i_{t+1})),$$

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

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\widehat{w}^n_t(s^i_t)\rho^i_t(r^i_t + f_t(s^i_t, a^i_t, s^i_{t+1})) = \operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\rho^i_{0:t}(r^i_t + f_t(s^i_t, a^i_t, s^i_{t+1})), \qquad (A.2)$$

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

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

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

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

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. □

B Proof Details of Section 4

In this section, we provide the details proof in the part of marginalized density ratio estimation.

Before we start our proof, we review the assumptions we made about $\pi, \mu, d_t^\mu, d_t^\pi$. Assume there exists $\eta > 0$, such that $\pi(a|s)/\mu(a|s) \le 1/\eta$ for any (s, a). We also assume there exists $\{\eta_t > 0\}_{t=0}^{H-1}$, such that $d_t^\mu(s) \ge \eta_t$ for all s and t. we first prove Theorem 2.

Proof of Theorem 2. Given the definition of the non-normalized estimation $\widehat{d}_t^{\pi}(s_t^i)$ in (4.1), we have

$$
\begin{aligned}
&\widehat{d}_{t+1}^{\pi}(s) \\
=&\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{d}_t^{\pi}(s_t^i)}{\widehat{d}_t^{\mu}(s_t^i)}\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}\mathbf{1}(s_{t+1}^i=s) \\
=&\frac{1}{n}\sum_{i=1}^{n}\frac{d_t^{\pi}(s_t^i)+\varepsilon_t^{\pi}(s_t^i)}{d_t^{\mu}(s_t^i)+\varepsilon_t^{\mu}(s_t^i)}\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}\mathbf{1}(s_{t+1}^i=s) \\
=&\frac{1}{n}\sum_{i=1}^{n}\left(\frac{(d_t^{\pi}(s_t^i)+\varepsilon_t^{\pi}(s_t^i))(d_t^{\mu}(s_t^i)-\varepsilon_t^{\mu}(s_t^i))}{(d_t^{\mu}(s_t^i))^2-(\varepsilon_t^{\mu}(s_t^i))^2}\right)\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}\mathbf{1}(s_{t+1}^i=s) \\
=&\frac{1}{n}\sum_{i=1}^{n}\left(\frac{(d_t^{\pi}(s_t^i)+\varepsilon_t^{\pi}(s_t^i))(d_t^{\mu}(s_t^i)-\varepsilon_t^{\mu}(s_t^i))}{(d_t^{\mu}(s_t^i))^2}+\frac{d_t^{\pi}(s_t^i)}{(d_t^{\mu}(s_t^i))^3}\widetilde{\mathcal{O}}\left(\frac{1}{n}\right)\right)\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}\mathbf{1}(s_{t+1}^i=s) \\
=&\underbrace{\frac{1}{n}\sum_{i=1}^{n}\frac{d_t^{\pi}(s_t^i)}{d_t^{\mu}(s_t^i)}\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}\mathbf{1}(s_{t+1}^i=s)\left(1+\widetilde{\mathcal{O}}\left(\frac{1}{n}\right)\right)}_{\text{(a)}}+\underbrace{\frac{1}{n}\sum_{i=1}^{n}\frac{\varepsilon_t^{\pi}(s_t^i)}{d_t^{\mu}(s_t^i)}\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}\mathbf{1}(s_{t+1}^i=s)}_{\text{(b)}} \\
&+\underbrace{\frac{1}{n}\sum_{i=1}^{n}\frac{d_t^{\pi}(s_t^i)\varepsilon_t^{\mu}(s_t^i)}{(d_t^{\mu}(s_t^i))^2}\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}\mathbf{1}(s_{t+1}^i=s)}_{\text{(c)}}+\underbrace{\frac{1}{n}\sum_{i=1}^{n}\frac{\varepsilon_t^{\pi}(s_t^i)\varepsilon_t^{\mu}(s_t^i)}{(d_t^{\mu}(s_t^i))^2}\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}\mathbf{1}(s_{t+1}^i=s)}_{\text{(d)}},
\end{aligned}
$$

where the all inequalities above are calculated by the direct simplification.

We now analyze these four terms above separately. For the term (a):

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\frac{d_t^{\pi}(s_t^i)}{d_t^{\mu}(s_t^i)}\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}\mathbf{1}(s_{t+1}^i=s)\left(1+\widetilde{\mathcal{O}}\left(\frac{1}{n}\right)\right) \\
=&\left(d_{t+1}^{\pi}(s)+\widetilde{\mathcal{O}}\left(\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}}\right)\right)\left(1+\widetilde{\mathcal{O}}\left(\frac{1}{n}\right)\right) \\
=&d_{t+1}^{\pi}(s)+\widetilde{\mathcal{O}}\left(\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}}\right),
\end{aligned}
\tag{B.1}
$$

where $w_t(s)=d_t^{\pi}(s)/d_t^{\mu}(s)$.

In the above derivation, we applied Hoeffding's inequality on

$$
\frac{1}{n}\sum_{i=1}^{n}\frac{d_t^{\pi}(s_t^i)}{d_t^{\mu}(s_t^i)}\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}\mathbf{1}(s_{t+1}^i=s)-\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{d_t^{\pi}(s_t^i)}{d_t^{\mu}(s_t^i)}\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}\mathbf{1}(s_{t+1}^i=s)\right].
$$

We first calculate the expectation, where a_t^i, s_t^i, s_{t+1}^i are sampled by behavior policy μ.

$$\begin{aligned}
\mathbb{E}_\mu\left[\frac{d_t^\pi(s_t^i)}{d_t^\mu(s_t^i)}\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}\mathbf{1}(s_{t+1}^i=s)\right] &= \int p_\mu(s_t^i, s_{t+1}^i, a_t^i)\frac{d_t^\pi(s_t^i)}{d_t^\mu(s_t^i)}\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}\mathbf{1}(s_{t+1}^i=s)d(s_t^i, s_{t+1}^i, a_t^i)\\
&= \int p(s_{t+1}^i|s_t^i, a_t^i)p_\mu(s_t^i)\mu(a_t^i|s_t^i)\frac{d_t^\pi(s_t^i)}{d_t^\mu(s_t^i)}\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}\mathbf{1}(s_{t+1}^i=s)d(s_t^i, s_{t+1}^i, a_t^i)\\
&= \int p(s_{t+1}^i|s_t^i, a_t^i)d_t^\pi(s_t^i)\pi(a_t^i|s_t^i)\mathbf{1}(s_{t+1}^i=s)d(s_t^i, s_{t+1}^i, a_t^i)\\
&= \mathbb{E}_\pi[\mathbf{1}(s_{t+1}^i=s)]\\
&= d_{t+1}^\pi(s)
\end{aligned}$$

Then, by Hoeffding's inequality we have that with probability $1-\delta$

$$\left|\frac{1}{n}\sum_{i=1}^n \frac{d_t^\pi(s_t^i)}{d_t^\mu(s_t^i)}\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}\mathbf{1}(s_{t+1}^i=s) - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \frac{d_t^\pi(s_t^i)}{d_t^\mu(s_t^i)}\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}\mathbf{1}(s_{t+1}^i=s)\right]\right| \le O\left(\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}}\log(1/\delta)\right)$$

For the term (b):

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \frac{\varepsilon_t^\pi(s_t^i)}{d_t^\mu(s_t^i)}\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}\mathbf{1}(s_{t+1}^i=s)\\
=&\sum_{s_t}\frac{1}{n}\sum_{i=1}^n \frac{\pi(a_t^i|s_t)}{\mu(a_t^i|s_t)}\mathbf{1}(s_t^i=s_t, s_{t+1}^i=s)\frac{\varepsilon_t^\pi(s_t)}{d_t^\mu(s_t)}\\
=&\sum_{s_t} d_t^\mu(s_t)\left[\mathbb{P}_\pi(s|s_t) + \widetilde{\mathcal{O}}\left(\sqrt{\frac{\mathbb{E}_\mu[\rho_t^2|s_t]}{n\eta_t}} + \frac{1}{n\eta\eta_t}\right)\right]\frac{\varepsilon_t^\pi(s_t)}{d_t^\mu(s_t)} \qquad \text{(B.2)}
\end{aligned}$$

In the above derivation, we applied Bernstein's inequality on

$$\frac{1}{n}\sum_{i=1}^n \frac{\pi(a_t^i|s_t)}{\mu(a_t^i|s_t)}\mathbf{1}(s_t^i=s_t, s_{t+1}^i=s) - \mathbb{E}\left[\frac{\pi(a_t^i|s_t)}{\mu(a_t^i|s_t)}\mathbf{1}(s_t^i=s_t, s_{t+1}^i=s)\right].$$

Let's first calculate the expectation (noted that a_t^i, s_t^i, s_{t+1}^i are drawn under policy μ).

$$\begin{aligned}
\mathbb{E}_\mu\left[\frac{\pi(a_t^i|s_t)}{\mu(a_t^i|s_t)}\mathbf{1}(s_t^i=s_t, s_{t+1}^i=s)\right] &= \int p_\mu(s_t^i, s_{t+1}^i, a_t^i)\frac{\pi(a_t^i|s_t)}{\mu(a_t^i|s_t)}\mathbf{1}(s_t^i=s_t, s_{t+1}^i=s)d(s_t^i, s_{t+1}^i, a_t^i)\\
&= \int p(s_{t+1}^i|s_t^i, a_t^i)p_\mu(s_t^i)\mu(a_t^i|s_t^i)\frac{\pi(a_t^i|s_t)}{\mu(a_t^i|s_t)}1(s_t^i=s_t, s_{t+1}^i=s)d(s_t^i, s_{t+1}^i, a_t^i)\\
&= \int [\sum_{a_t^i} p(s_{t+1}^i|s_t^i, a_t^i)\pi(a_t^i|s_t)]p_\mu(s_t^i)\mathbf{1}(s_t^i=s_t, s_{t+1}^i=s)d(s_t^i, s_{t+1}^i)\\
&= \sum_{a_t^i} p(s_{t+1}^i=s|s_t, a_t^i)\pi(a_t^i|s_t)]p_\mu(s_t)\\
&= p_\pi(s|s_t)p_\mu(s_t).
\end{aligned}$$

Let's now bound the variance of this random variable.

$$\begin{aligned}\mathrm{Var}_\mu\left[\frac{\pi(a_t^i|s_t)}{\mu(a_t^i|s_t)}\mathbf{1}(s_t^i=s_t,s_{t+1}^i=s)\right]&\le\mathbb{E}_\mu\left[(\frac{\pi(a_t^i|s_t)}{\mu(a_t^i|s_t)})^2\mathbf{1}(s_t^i=s_t,s_{t+1}^i=s)\right]\\&=\mathbb{E}_\mu\left[(\frac{\pi(a_t^i|s_t)}{\mu(a_t^i|s_t)})^2\middle|s_t^i=s_t,s_{t+1}^i=s\right]p_\mu\left[s_t\right]p_\mu\left[s|s_t\right]\\&\le\mathbb{E}_\mu\left[(\frac{\pi(a_t^i|s_t)}{\mu(a_t^i|s_t)})^2\mathbf{1}(s_t^i=s_t)\right]\le\mathbb{E}\left[(\frac{\pi(a_t^i|s_t)}{\mu(a_t^i|s_t)})^2\middle|s_t\right]p_\mu(s_t).\end{aligned}$$

By Bernstein's inequality we have that with probability $1-\delta$

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^n\frac{\pi(a_t^i|s_t)}{\mu(a_t^i|s_t)}\mathbf{1}(s_t^i=s_t,s_{t+1}^i=s)-p_\pi(s|s_t)p_\mu(s_t)\right|\\\le&O\left(\left(\sqrt{\frac{\mathbb{E}\left[\rho_t^2\middle|s_t\right]\mathbb{P}_\mu(s_t)}{n}}+\frac{1}{\eta n}\right)\log(1/\delta)\right)\end{aligned}$$

For the term (c):

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n\frac{d_t^\pi(s_t^i)\varepsilon_t^\mu(s_t^i)}{(d_t^\mu(s_t^i))^2}\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}\mathbf{1}(s_{t+1}^i=s)\\=&\frac{1}{n}\sum_{i=1}^n\frac{d_t^\pi(s_t^i)}{d_t^\mu(s_t^i)}\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}\mathbf{1}(s_{t+1}^i=s)\frac{\varepsilon_t^\mu(s_t^i)}{d_t^\mu(s_t^i)}\\=&\left(d_{t+1}^\pi(s)+\widetilde{\mathcal{O}}\left(\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}}\right)\right)\frac{\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\right)}{d_t^\mu(s_t^i)},\end{aligned}\tag{B.3}$$

where the last equality follow from (B.1) and Fact 1.

For the term (d):

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n\frac{\varepsilon_t^\pi(s_t^i)\varepsilon_t^\mu(s_t^i)}{(d_t^\mu(s_t^i))^2}\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}\mathbf{1}(s_{t+1}^i=s)\\=&\frac{1}{n}\sum_{i=1}^n\frac{\varepsilon_t^\pi(s_t^i)}{d_t^\mu(s_t^i)}\frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}\mathbf{1}(s_{t+1}^i=s)\frac{\varepsilon_t^\mu(s_t^i)}{d_t^\mu(s_t^i)}\\=&\sum_{s_t}\left[\mathbb{P}_\pi(s|s_t)+\widetilde{\mathcal{O}}\left(\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}}\right)\right]\frac{\varepsilon_t^\pi(s_t)\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\right)}{d_t^\mu(s_t^i)}\end{aligned}\tag{B.4}$$

where the last equality follow from (B.2) and Fact 1.

Combining (B.1), (B.2), (B.3) and (B.4), we obtain

$$\widehat{d}_{t+1}^\pi(s)=d_{t+1}^\pi(s)+\sum_{s_t}\mathbb{P}_\pi(s|s_t)\varepsilon_t^\pi(s_t)+\widetilde{\mathcal{O}}\left(\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}}\right)+\sum_{s_t}\widetilde{\mathcal{O}}\left(\sqrt{\frac{\mathbb{E}_\mu[\rho_t^2|s_t]}{n\eta_t}}+\frac{1}{n\eta\eta_t}\right)\varepsilon_t^\pi(s_t)\tag{B.5}$$

According the definition of $\varepsilon_{t+1}^{\pi}(\cdot)$, we have

$$\varepsilon_{t+1}^{\pi}(s) = \sum_{s_t} \mathbb{P}_{\pi}(s|s_t)\varepsilon_t^{\pi}(s_t) + \widetilde{\mathcal{O}}\left(\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}} + \sum_{s_t}\left(\sqrt{\frac{\mathbb{E}_{\mu}[\rho_t^2|s_t]}{n\eta_t}} + \frac{1}{n\eta\eta_t}\right)\varepsilon_t^{\pi}(s_t)\right)$$

$$|\varepsilon_{t+1}^{\pi}(s)| \leq \sum_{s_t} \mathbb{P}_{\pi}(s|s_t)|\varepsilon_t^{\pi}(s_t)| + \widetilde{\mathcal{O}}\left(\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}} + \sum_{s_t}\left(\sqrt{\frac{\mathbb{E}_{\mu}[\rho_t^2|s_t]}{n\eta_t}} + \frac{1}{n\eta\eta_t}\right)|\varepsilon_t^{\pi}(s_t)|\right) \tag{B.6}$$

Taking summation over s on each side of (B.6), we obtain

$$\begin{aligned}
\sum_s |\varepsilon_{t+1}^{\pi}(s)| &\leq \sum_s \sum_{s_t} \mathbb{P}_{\pi}(s|s_t)|\varepsilon_t^{\pi}(s_t)| + \sum_s \widetilde{\mathcal{O}}\left(\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}} + \sum_{s_t}\left(\sqrt{\frac{\mathbb{E}_{\mu}[\rho_t^2|s_t]}{n\eta_t}} + \frac{1}{n\eta\eta_t}\right)|\varepsilon_t^{\pi}(s_t)|\right) \\
&\leq \sum_{s_t} |\varepsilon_t^{\pi}(s_t)| + \widetilde{\mathcal{O}}\left(|\mathcal{S}|\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}} + |\mathcal{S}|\left(\sqrt{\frac{\mathbb{E}_{\mu}[\rho_t^2|s_t]}{n\eta_t}} + \frac{1}{n\eta\eta_t}\right)\sum_{s_t}|\varepsilon_t^{\pi}(s_t)|\right) \\
&= \left(1 + \widetilde{\mathcal{O}}\left(|\mathcal{S}|\left(\sqrt{\frac{\mathbb{E}_{\mu}[\rho_t^2|s_t]}{n\eta_t}} + \frac{1}{n\eta\eta_t}\right)\right)\right)\sum_{s_t}|\varepsilon_t^{\pi}(s_t)| + \widetilde{\mathcal{O}}\left(|\mathcal{S}|\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}}\right)
\end{aligned}$$

Thus, we complete the proof.

□

The proof of Theorem 3 is proved using the similar technique to that of Theorem 2, but it also considers the self-normalization. The detailed proof is as follows.

Proof of Theorem 3. Based on the idea of self-normalization, we have

$$\widetilde{d}_{t+1}^{\pi}(s_t^i) = \frac{d_{t+1}^{\pi}(s_{t+1}^i) + \varepsilon_{t+1}^{\pi}(s_{t+1}^i)}{1 + \sum_s \varepsilon_{t+1}^{\pi}(s_{t+1}^i)},$$

where

$$\varepsilon_{t+1}^{\pi}(s) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \frac{\widetilde{d}_t^{\pi}(s_t^i)}{\widetilde{d}_t^{\mu}(s_t^i)} \frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)} \mathbf{1}(s_{t+1}^i = s) - d_{t+1}^{\pi}(s).$$

Thus, we can analyze the self-normalized case using the similar method to the Proof of Theorem 2

$$d_{t+1}^{\pi}(s) + \varepsilon_{t+1}^{\pi}(s) = d_{t+1}^{\pi}(s) + \sum_{s_t} \mathbb{P}_{\pi}(s|s_t)\widetilde{\varepsilon}_t^{\pi}(s_t) + \widetilde{\mathcal{O}}\left(\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}}\right) + \sum_{s_t}\widetilde{\mathcal{O}}\left(\sqrt{\frac{\mathbb{E}_{\mu}[\rho_t^2|s_t]}{n\eta_t}} + \frac{1}{n\eta\eta_t}\right)\widetilde{\varepsilon}_t^{\pi}(s_t)$$

Using the definition of $\varepsilon_{t+1}^{\pi}(\cdot)$, we have

$$\varepsilon_{t+1}^{\pi}(s) = \sum_{s_t} \mathbb{P}_{\pi}(s|s_t)\widetilde{\varepsilon}_t^{\pi}(s_t) + \widetilde{\mathcal{O}}\left(\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}} + \sum_{s_t}\left(\sqrt{\frac{\mathbb{E}_{\mu}[\rho_t^2|s_t]}{n\eta_t}} + \frac{1}{n\eta\eta_t}\right)\widetilde{\varepsilon}_t^{\pi}(s_t)\right).$$

We first analyze $\sum_s \varepsilon_{t+1}^{\pi}(s)$

$$\begin{aligned}\sum_s \varepsilon_{t+1}^{\pi}(s) =& \sum_s \sum_{s_t} \mathbb{P}_{\pi}(s|s_t)\widetilde{\varepsilon}_t^{\pi}(s_t) + \sum_s \widetilde{\mathcal{O}}\left(\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}} + \sum_{s_t}\left(\sqrt{\frac{\mathbb{E}_{\mu}[\rho_t^2|s_t]}{n\eta_t}} + \frac{1}{n\eta\eta_t}\right)\widetilde{\varepsilon}_t^{\pi}(s_t)\right) \\ =& \widetilde{\mathcal{O}}\left(|\mathcal{S}|\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}} + \sum_s\sum_{s_t}\left(\sqrt{\frac{\mathbb{E}_{\mu}[\rho_t^2|s_t]}{n\eta_t}} + \frac{1}{n\eta\eta_t}\right)\widetilde{\varepsilon}_t^{\pi}(s_t)\right) \\ \leq& \widetilde{\mathcal{O}}\left(|\mathcal{S}|\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}} + |\mathcal{S}|\left(\sqrt{\frac{\mathbb{E}_{\mu}[\rho_t^2|s_t]}{n\eta_t}} + \frac{1}{n\eta\eta_t}\right)\sum_{s_t}|\widetilde{\varepsilon}_t^{\pi}(s_t)|\right) \end{aligned} \tag{B.7}$$

We can obtain $\widetilde{d}_{t+1}^{\pi}(s)$ satisfies the follows equality based on (B.5)

$$\begin{aligned}\widetilde{d}_{t+1}^{\pi}(s) =& \frac{1}{1+\sum_s \varepsilon_{t+1}^{\pi}(s)}\Bigg(d_{t+1}^{\pi}(s) + \sum_{s_t}\mathbb{P}_{\pi}(s|s_t)\widetilde{\varepsilon}_t^{\pi}(s_t) \\ &+ \widetilde{\mathcal{O}}\left(\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}}\right) + \sum_{s_t}\widetilde{\mathcal{O}}\left(\sqrt{\frac{\mathbb{E}_{\mu}[\rho_t^2|s_t]}{n\eta_t}} + \frac{1}{n\eta\eta_t}\right)\widetilde{\varepsilon}_t^{\pi}(s_t)\Bigg)\end{aligned}$$

Based on the definition of $\varepsilon_{t+1}^{\pi}(\cdot)$, we have

$$\begin{aligned}\Rightarrow \widetilde{\varepsilon}_{t+1}^{\pi}(s) =& \frac{1}{1+\sum_s \varepsilon_{t+1}^{\pi}(s)}\Bigg(d_{t+1}^{\pi}(s) + \sum_{s_t}\mathbb{P}_{\pi}(s|s_t)\widetilde{\varepsilon}_t^{\pi}(s_t) \\ &+ \widetilde{\mathcal{O}}\left(\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}}\right) + \sum_{s_t}\widetilde{\mathcal{O}}\left(\sqrt{\frac{\mathbb{E}_{\mu}[\rho_t^2|s_t]}{n\eta_t}} + \frac{1}{n\eta\eta_t}\right)\widetilde{\varepsilon}_t^{\pi}(s_t)\Bigg) - d_{t+1}^{\pi}(s) \\ =& \left(-1 + \frac{1}{1+\sum_s \varepsilon_{t+1}^{\pi}(s)}\right)d_{t+1}^{\pi}(s) + \frac{\sum_{s_t}\mathbb{P}_{\pi}(s|s_t)\widetilde{\varepsilon}_t^{\pi}(s_t)}{1+\sum_s \varepsilon_{t+1}^{\pi}(s)} \\ &+ \frac{1}{1+\sum_s \varepsilon_{t+1}^{\pi}(s)}\left(\widetilde{\mathcal{O}}\left(\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}}\right) + \sum_{s_t}\widetilde{\mathcal{O}}\left(\sqrt{\frac{\mathbb{E}_{\mu}[\rho_t^2|s_t]}{n\eta_t}} + \frac{1}{n\eta\eta_t}\right)\widetilde{\varepsilon}_t^{\pi}(s_t)\right) \\ \leq& \left(-1 + \frac{1}{1+\sum_s \varepsilon_{t+1}^{\pi}(s)}\right)d_{t+1}^{\pi}(s) + \frac{\sum_{s_t}\mathbb{P}_{\pi}(s|s_t)\widetilde{\varepsilon}_t^{\pi}(s_t)}{1+\sum_s \varepsilon_{t+1}^{\pi}(s)} \\ &+ \frac{1}{1+\sum_s \varepsilon_{t+1}^{\pi}(s)}\left(\widetilde{\mathcal{O}}\left(\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}} + \left(\sqrt{\frac{\mathbb{E}_{\mu}[\rho_t^2|s_t]}{n\eta_t}} + \frac{1}{n\eta\eta_t}\right)\sum_{s_t}|\widetilde{\varepsilon}_t^{\pi}(s_t)|\right)\right)\end{aligned}$$

We now discuss $\sum_s |\widetilde{\varepsilon}^{\pi}_{t+1}(s)|$ based on the results above

$$\begin{aligned}\sum_s |\widetilde{\varepsilon}^{\pi}_{t+1}(s)| \leq & \sum_s \frac{|\sum_s \varepsilon^{\pi}_{t+1}(s)|}{1+\sum_s \varepsilon^{\pi}_{t+1}(s)} d^{\pi}_{t+1}(s) + \sum_s \frac{\sum_{s_t} \mathbb{P}_{\pi}(s|s_t)|\widetilde{\varepsilon}^{\pi}_t(s_t)|}{1+\sum_s \varepsilon^{\pi}_{t+1}(s)} \\ & + \sum_s \frac{1}{1+\sum_s \varepsilon^{\pi}_{t+1}(s)} \left(\widetilde{\mathcal{O}} \left(\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}} + \left(\sqrt{\frac{\mathbb{E}_{\mu}[\rho_t^2|s_t]}{n\eta_t}} + \frac{1}{n\eta\eta_t} \right) \sum_{s_t} |\widetilde{\varepsilon}^{\pi}_t(s_t)| \right) \right) \\ = & \frac{\sum_s |\widetilde{\varepsilon}^{\pi}_t(s)|}{1+\sum_s \varepsilon^{\pi}_{t+1}(s)} + \frac{|\sum_s \varepsilon^{\pi}_{t+1}(s)|}{1+\sum_s \varepsilon^{\pi}_{t+1}(s)} \\ & + \frac{1}{1+\sum_s \varepsilon^{\pi}_{t+1}(s)} \widetilde{\mathcal{O}} \left(|\mathcal{S}| \sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}} + |\mathcal{S}| \left(\sqrt{\frac{\mathbb{E}_{\mu}[\rho_t^2|s_t]}{n\eta_t}} + \frac{1}{n\eta\eta_t} \right) \sum_{s_t} |\widetilde{\varepsilon}^{\pi}_t(s_t)| \right). \end{aligned}$$

Then, substitute the results about $\sum_s \varepsilon^{\pi}_{t+1}(s)$, (B.7), into the last inequality, we have

$$\begin{aligned}\sum_s |\widetilde{\varepsilon}^{\pi}_{t+1}(s)| \leq & \frac{\sum_s |\widetilde{\varepsilon}^{\pi}_t(s)|}{1+\sum_s \varepsilon^{\pi}_{t+1}(s)} + \frac{1}{1+\sum_s \varepsilon^{\pi}_{t+1}(s)} \widetilde{\mathcal{O}} \left(|\mathcal{S}| \sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}} + |\mathcal{S}| \left(\sqrt{\frac{\mathbb{E}_{\mu}[\rho_t^2|s_t]}{n\eta_t}} + \frac{1}{n\eta\eta_t} \right) \sum_{s_t} |\widetilde{\varepsilon}^{\pi}_t(s_t)| \right) \\ & + \frac{1}{1+\sum_s \varepsilon^{\pi}_{t+1}(s)} \widetilde{\mathcal{O}} \left(|\mathcal{S}| \sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}} + |\mathcal{S}| \left(\sqrt{\frac{\mathbb{E}_{\mu}[\rho_t^2|s_t]}{n\eta_t}} + \frac{1}{n\eta\eta_t} \right) \sum_{s_t} |\widetilde{\varepsilon}^{\pi}_t(s_t)| \right) \\ = & \frac{\sum_s |\widetilde{\varepsilon}^{\pi}_t(s)|}{1+\sum_s \varepsilon^{\pi}_{t+1}(s)} + \frac{1}{1+\sum_s \varepsilon^{\pi}_{t+1}(s)} \widetilde{\mathcal{O}} \left(|\mathcal{S}| \sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}} + |\mathcal{S}| \left(\sqrt{\frac{\mathbb{E}_{\mu}[\rho_t^2|s_t]}{n\eta_t}} + \frac{1}{n\eta\eta_t} \right) \sum_{s_t} |\widetilde{\varepsilon}^{\pi}_t(s_t)| \right) \\ = & \frac{1}{1+\sum_s \varepsilon^{\pi}_{t+1}(s)} \left(1 + \widetilde{\mathcal{O}} \left(|\mathcal{S}| \sqrt{\frac{\mathbb{E}_{\mu}[\rho_t^2|s_t]}{n\eta_t}} + \frac{|\mathcal{S}|}{n\eta\eta_t} \right) \right) \sum_{s_t} |\widetilde{\varepsilon}^{\pi}_t(s_t)| \\ & + \frac{1}{1+\sum_s \varepsilon^{\pi}_{t+1}(s)} \widetilde{\mathcal{O}} \left(|\mathcal{S}| \sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}} \right), \end{aligned}$$

where the $\sum_s \varepsilon^{\pi}_{t+1}(s)$ term in the denominator is defined in (B.7). This completes the proof. □

Next, we provide the proof of corollary 1

Proof of Corollary 1. Denote $x_t := \sum_s |\widehat{\varepsilon}^{\pi}_{t+1}(s)|$ for just this proof. Under our assumption on a η, there is a universal constant C such that

$$x_{t+1} \leq (1 + \frac{C|\mathcal{S}|}{\eta\sqrt{n}}) x_t + \frac{C|\mathcal{S}|}{\eta\sqrt{n}}.$$

Note that $x_0 \leq \frac{C|\mathcal{S}|}{\eta\sqrt{n}}$. By the geometric series, the recursion implies that

$$x_t \leq \sum_{i=0}^{t} \frac{C|\mathcal{S}|}{\eta\sqrt{n}} \left(1 + \frac{C|\mathcal{S}|}{\eta\sqrt{n}}\right)^i = \left(1 + \frac{C|\mathcal{S}|}{\eta\sqrt{n}}\right)^t - 1.$$

Under the assumption that n is large, and $t \leq H$, we can further upper bound the above by

$$1 + \frac{2C|\mathcal{S}|}{\eta\sqrt{n}} - 1 = \frac{2C|\mathcal{S}|}{\eta\sqrt{n}}.$$

□

Algorithm 1 Marginalized Off-Policy Evaluation

Input: Transition data $\mathcal{D} = \{\{s_t^i, a_t^i, r_t^i, s_{t+1}^i\}_{t=0}^{H-1}\}_{i=1}^n$ from the behavior policy μ. A target policy π which we want to evaluate its cumulative reward.

1: Calculate the on-policy estimation of $d_0(\cdot)$ by

$$\widehat{d}_0(s) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(s_0^i == s),$$

and set $\widehat{d}_0^{\mu}(\cdot)$ and $\widehat{d}_0^{\pi}(\cdot)$ as $\widehat{d}_0(s)$.

2: **for** $t = 0, 1, \ldots, H-1$ **do**

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

4: Calculate the on-policy estimation of $d_{t+1}^{\mu}(\cdot)$ by

$$\widehat{d}_{t+1}^{\mu}(s) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(s_{t+1}^i == s).$$

Calculate the off-policy estimation of $d_{t+1}^{\pi}(\cdot)$ by

$$\begin{aligned}
\widehat{d}_{t+1}^{\pi}(s) =& \frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{d}_t^{\pi}(s_t^i)}{\widehat{d}_t^{\mu}(s_t^i)} \frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)} \mathbf{1}(s_{t+1}^i = s) \quad \text{or,} \\
\widetilde{d}_{t+1}^{\pi}(s) =& \frac{\sum_{i=1}^{n} \frac{\widehat{d}_t^{\pi}(s_t^i)}{\widehat{d}_t^{\mu}(s_t^i)} \frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)} \mathbf{1}(s_{t+1}^i = s)}{\sum_{i=1}^{n} \frac{\widehat{d}_t^{\pi}(s_t^i)}{\widehat{d}_t^{\mu}(s_t^i)} \frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}}.
\end{aligned} \tag{B.8}$$

5: Estimate the reward function

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

6: Specify $\widehat{w}_{t+1}(s)$ as $\dfrac{\widehat{d}_{t+1}^{\pi}(s)}{\widehat{d}_{t+1}^{\pi}(s)}$ or $\dfrac{\widetilde{d}_{t+1}^{\pi}(s)}{\widetilde{d}_{t+1}^{\pi}(s)}$.

7: **end for**

8: Substitute the all estimated values above into (3.4) to obtain $\widehat{v}(\pi)$, the estimated cumulative reward of π.

C Algorithm Details

Algorithm 1 summarizes our method of marginalized off-policy evaluation. Note that the W-MIS estimator in Section 5 is using the estimate of $d_t^\pi(\cdot)$ as (B.8) in Algorithm 1.