Towards Optimal Off-Policy Evaluation for Reinforcement Learning with Marginalized Importance Sampling

Tengyang Xie*
Dept. of Computer Science
UIUC
Urbana, IL 61801
tx10@illinois.edu

Yifei Ma
AWS AI Labs
Amazon.com Services, Inc.
East Palo Alto, CA 94303
yifeim@amazon.com

Yu-Xiang Wang
Dept. of Computer Science,
UC Santa Barbara
Santa Barbara, CA 93106
yuxiangw@cs.ucsb.edu

Abstract

Motivated by the many real-world applications of reinforcement learning (RL) that require safe-policy iterations, we consider the problem of off-policy evaluation (OPE) — the problem of evaluating a new policy using the historical data obtained by different behavior policies — under the model of nonstationary episodic Markov Decision Processes (MDP) with a long horizon and a large action space. Existing importance sampling (IS) methods often suffer from large variance that depends exponentially on the RL horizon H. To solve this problem, we consider a marginalized importance sampling (MIS) estimator that recursively estimates the state marginal distribution for the target policy at every step. MIS achieves a mean-squared error of

$$\frac{1}{n}\sum_{t=1}^{H}\mathbb{E}_{\mu}\left[\frac{d_t^{\pi}(s_t)^2}{d_t^{\mu}(s_t)^2}\mathrm{Var}_{\mu}\left[\frac{\pi_t(a_t|s_t)}{\mu_t(a_t|s_t)}\left(V_{t+1}^{\pi}(s_{t+1})+r_t\right)\middle|s_t\right]\right]+\tilde{O}(n^{-1.5})$$

where μ and π are the logging and target policies, $d_t^{\mu}(s_t)$ and $d_t^{\pi}(s_t)$ are the marginal distribution of the state at tth step, H is the horizon, n is the sample size and V_{t+1}^{π} is the value function of the MDP under π. The result matches the Cramer-Rao lower bound in Jiang and Li [2016] up to a multiplicative factor of H. To the best of our knowledge, this is the first OPE estimation error bound with a polynomial dependence on H. Besides theory, we show empirical superiority of our method in time-varying, partially observable, and long-horizon RL environments.

1 Introduction

The problem of *off-policy evaluation* (OPE), which predicts the performance of a 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 already been 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. 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 target policy. It also provides typically unbiased or

*The research was partially conducted when TX was visiting YW and YM at Amazon AWS AI Labs during his internship in Summer 2018.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

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 [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 provide informative results, for long-horizon problems [Mandel et al., 2014], since the variance of the product of importance weights may grow exponentially as the horizon goes long. There are also model-based approaches for solving OPE problems [Liu et al., 2018b; Gottesman et al., 2019], where the value of the target policy is estimated directly using the approximated MDP.

Given this high-variance issue, it is necessary to find an IS-based approach without relying heavily on the cumulative product of importance weights from the whole trajectories. While the benefit of cumulative products is to allow unbiased estimation even without any state observability assumptions, reweighing the entire trajectories may not be necessary if some intermediate states are directly observable. For the latter, based on Markov independence assumptions, we can aggregate all trajectories that share the same state transition patterns to directly estimate the state distribution shifts after the change of policies from the behavioral to the target. We call this approach marginalized importance sampling (MIS), because it computes the *marginal* state distribution shifts at every single step, instead of the product of policy weights.

Related work [Liu et al., 2018a] tackles the high variance issue due to the cumulative product of importance weights. They apply importance sampling on the average visitation distribution of state-action pairs, based on an estimation of the mixed state distribution. Hallak and Mannor [2017] and Gelada and Bellemare [2019] also leverage the same fact in time-invariant MDPs, where they use the stationary ratio of state-action pairs to replace the trajectory weights. However, these methods may not directly work in finite-horizon MDPs, where the state distributions may not mix.

In contrast to the prior work, the first goal of our paper is to study the sample complexity and optimality of the marginalized approach. Specifically, we provide the first finite sample error bound on the mean-square error for our MIS off-policy evaluation estimator under the episodic tabular MDP setting (with potentially continuous action space). Our MSE bound is the exact calculation up to low order terms. Comparing to the Cramer-Rao lower bound established in [Jiang and Li, 2016, Theorem 3] for DAG-MDP, our bound is larger by at most a factor of H and we have good reasons to believe that this additional factor is required for any OPE estimators in this setting.

In addition to the theoretical results, we empirically evaluate our estimator against a number of strong baselines from prior work in a number of time-invariant/time-varying, fully observable/partially observable, and long-horizon environments. Our approach can also 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).

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 offer the design of our marginalized estimator, and we study its information-theoretical optimality in Section 4. We present the empirical results in a number of RL tasks in Section 5. At last, Section 6 concludes the paper.

2 Problem formulation

Symbols and notations. We consider the problem of off-policy evaluation for a finite horizon, nonstationary, episodic MDP, which is a tuple defined by $M = (\mathcal{S}, \mathcal{A}, T, r, H)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $T_t : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the *transition function* with $T_t(s'|s, a)$ defined by probability of achieving state s' after taking action a in state s at time t, and $r_t : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the expected reward function with $r_t(s, a, s')$ defined by the mean of immediate received reward after taking action a in state s and transitioning into s', and H denotes the finite horizon. 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. $\mathbb{E}[\cdot]$ and $\mathbb{E}[\cdot|E]$ denotes the expectation and conditional expectation given E, respectively.

Let $\mu, \pi : \mathcal{S} \to \mathbb{P}_{\mathcal{A}}$ be policies which output a distribution of actions given an observed state. We call μ the behavioral policy and π the target policy. 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. The expectation operators in this paper will either be indexed with π or μ, which denotes that all random variables coming from roll-outs from

the specified policy. Moreover, we denote $d_t^\mu(s_t)$ and $d_t^\pi(s_t)$ the induced state distribution at time t. When $t = 1$, the initial distributions are identical $d_1^\mu = d_1^\pi = d_1$. For $t > 1$, $d_t^\mu(s_t)$ and $d_t^\pi(s_t)$ are functions of not just the policies themselves but also the unknown underlying transition dynamics, i.e., for π (and similarly μ), recursively define

$$d_t^\pi(s_t) = \sum_{s_{t-1}} P_t^\pi(s_t|s_{t-1}) d_{t-1}^\pi(s_{t-1}),$$
$$\text{where } P_t^\pi(s_t|s_{t-1}) = \sum_{a_{t-1}} T_t(s_t|s_{t-1}, a_{t-1})\pi(a_{t-1}|s_{t-1}). \tag{2.1}$$

We denote $P_{i,j}^\pi \in \mathbb{R}^{S\times S}\ \forall j < i$ as the state-transition probability from step j to step i under a sequence of actions taken by π. Note that $P_{t+1,t}^\pi(s'|s) = \sum_a P_{t+1,t}(s'|s,a)\pi_t(a|s) = T_{t+1}(s'|s, \pi_t(s))$.

Behavior 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 $t = 1, \ldots, H$ and episode index $i = 1, ..., n$. Target policy π is what we are interested to evaluate. Also, let $\mathcal{D}$ to denote the historical data, which contains n episode trajectories in total. We also define $\mathcal{D}_h = \{(s_t^{(i)}, a_t^{(i)}, r_t^{(i)}) : i \in [n], t \le h\}$ to be roll-in realization of n trajectories up to step h.

Throughout the paper, probability distributions are often used in their vector or matrix form. For instance, d_t^π without an input is interpreted as a vector in a S-dimensional probability simplex and $P_{i,j}^\pi$ is then a stochastic transition matrix. This allows us to write (2.1) concisely as $d_{t+1}^\pi = P_{t+1,t}^\pi d_t^\pi$.

Also note that while s_t, a_t, r_t are usually used to denote fixed elements in set $\mathcal{S}, \mathcal{A}$ and $\mathbb{R}$, in some cases we also overload them to denote generic random variables $s_t^{(1)}, a_t^{(1)}, r_t^{(1)}$. For example, $\mathbb{E}_\pi[r_t] = \mathbb{E}_\pi[r_t^{(1)}] = \sum_{s_t,a_t,s_{t+1}} d^\pi(s_t, a_t, s_{t+1}) r_t(s_t, a_t, s_{t+1})$ and $\mathrm{Var}_\pi[r_t(s_t, a_t, s_{t+1})] = \mathrm{Var}_\pi[r_t(s_t^{(1)}, a_t^{(1)}, s_{t+1}^{(1)})]$. The distinctions will be clear in each context.

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

$$v^\pi = \sum_{t=1}^{H} \sum_{s_t} d_t^\pi(s_t) \sum_{a_t} \pi(a_t|s_t) \sum_{s_{t+1}} T_t(s_{t+1}|s_t, a_t) r_t(s_t, a_t, s_{t+1}), \tag{2.2}$$

where we assume knowledge about $\mu(a|s)$ and $\pi(a|s)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, but *do not observe* $r_t(s_t, a_t, s_{t+1})$ for any actions other than a noisy version of it the evaluated actions. Nor do we observe the state distributions $d_t^\pi(s_t) \forall t > 1$ implied by the change of policies. Nonetheless, our goal is to find an estimator to minimize the mean-square error (MSE): $\mathrm{MSE}(\pi, \mu, M) = \mathbb{E}_\mu[(\hat{v}^\pi - v^\pi)^2]$, using the observed data and the known action probabilities. 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 words, this is a setting where we do not have enough data points to estimate the state-action-state transition dynamics, but we do observe the states and can estimate the distribution of the states after the change of policies, which is our main strategy.

Assumptions: We list the technical assumptions we need and provide necessary justification.

- A1. $\exists R_{\max}, \sigma < +\infty$ such that $0 \le \mathbb{E}[r_t|s_t, a_t, s_{t+1}] \le R_{\max}, \mathrm{Var}[r_t|s_t, a_t, s_{t+1}] \le \sigma^2$ for all t, s_t, a_t.
- A2. Behavior policy μ obeys that $d_m := \min_{t,s_t} d_t^\mu(s_t) > 0 \quad \forall t, s_t$ such that $d_t^\pi(s_t) > 0$.
- A3. Bounded weights: $\tau_s := \max_{t,s_t} \frac{d_t^\pi(s_t)}{d_t^\mu(s_t)} < +\infty$ and $\tau_a := \max_{t,s_t,a_t} \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)} < +\infty$.

Assumption A1 is assumed without loss of generality. The σ bound is required even for on-policy evaluation and the assumption on the non-negativity and $R_{\max}$ can always be obtained by shifting and rescaling the problem. Assumption A2 is necessary for any consistent off-policy evaluation estimator. Assumption A3 is also necessary for discrete state and actions, as otherwise the second moments of the importance weight would be unbounded. For continuous actions, $\tau_a < +\infty$ is stronger than we need and should be considered a simplifying assumption for the clarity of our presentation. Finally, we comment that the dependence in the parameter d_m, τ_s, τ_a do not occur in the leading $O(1/n)$ term of our MSE bound, but only in simplified results after relaxation.

3 Marginalized Importance Sampling Estimators for OPE

In this section, we present the design of marginalized IS estimators for OPE. For small action spaces, we may directly build models by the estimated transition function $T_t(s_t|s_{t-1}, a_{t-1})$ and the reward function $r_t(s_t, a_t, s_{t+1})$ from empirical data. However, the models may be inaccurate in large action spaces, where not all actions are frequently visited. Function approximation in the models may cause additional biases from covariate shifts due to the change of policies. Standard importance sampling estimators (including the doubly robust versions)[Dudík et al., 2011; Jiang and Li, 2016] avoid the need to estimate the model's dynamics but rather directly approximating the expected reward:

$$\widehat{v}^{\pi}_{\text{IS}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H}\left[\prod_{t=1}^{h}\frac{\pi(a_t^{(i)}|s_t^{(i)})}{\mu(a_t^{(i)}|s_t^{(i)})}\right] r_h^{(i)}.$$

To adjust for the differences in the policy, importance weights are used and it can be shown that this is an unbiased estimator of v^π (See more detailed discussion of IS and the doubly robust version in Appendix C). The main issue of this approach, when applying to the episodic MDP with large action space is that the variance of the importance weights grows exponentially in H [Liu et al., 2018a], which makes the sample complexity exponentially worse than the model-based approaches, when they are applicable. We address this problem by proposing an alternative way of estimating the importance weights which achieves the same sample complexity as the model-based approaches while allowing us to achieve the same flexibility and interpretability as the IS estimator that does not explicitly require estimating the state-action dynamics T_t. We propose the Marginalized Importance Sampling (MIS) estimator:

$$\widehat{v}^{\pi}_{\text{MIS}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{H}\frac{\widehat{d}^{\pi}_t(s_t^{(i)})}{\widehat{d}^{\mu}_t(s_t^{(i)})}\widehat{r}^{\pi}_t(s_t^{(i)}). \tag{3.1}$$

Clearly, if $\widehat{d}^{\pi} \to d^{\pi}_t$, $\widehat{d}^{\mu} \to d^{\mu}_t$, $\widehat{r}^{\pi}_t \to \mathbb{E}_\pi[R_t(s_t, a_t)|s_t]$, then $\widehat{v}^{\pi}_{\text{MIS}} \to v^\pi$.

It turns out that if we take $\widehat{d}^{\mu}_t(s_t) := \frac{1}{n}\sum_i \mathbf{1}(s_t^{(i)} = s_t)$ — the empirical mean — and define $\widehat{d}^{\pi}_t(s_t)/\widehat{d}^{\mu}_t(s_t) = 0$ whenever $n_{s_t} = 0$, then (3.1) is equivalent to $\sum_{t=1}^{H}\sum_{s_t}\widehat{d}^{\pi}_t(s_t)\widehat{r}^{\pi}(s_t)$ – the direct plug-in estimator of (2.2). It remains to specify $\widehat{d}^{\pi}_t(s_t)$ and $\widehat{r}^{\pi}(s_t)$. $\widehat{d}^{\pi}_t(s_t)$ is estimated recursively using

$$\widehat{d}^{\pi}_t = \widehat{P}^{\pi}_t\widehat{d}^{\pi}_{t-1}, \text{ where } \widehat{P}^{\pi}_t(s_t|s_{t-1}) = \frac{1}{n_{s_{t-1}}}\sum_{i=1}^{n}\frac{\pi(a_{t-1}^{(i)}|s_{t-1})}{\mu(a_{t-1}^{(i)}|s_{t-1})}\mathbf{1}((s_{t-1}^{(i)}, s_t^{(i)}) = (s_{t-1}, s_t));$$

$$\text{and } \widehat{r}^{\pi}_t(s_t) = \frac{1}{n_{s_t}}\sum_{i=1}^{n}\frac{\pi(a_t^{(i)}|s_t)}{\mu(a_t^{(i)}|s_t)}r_t^{(i)}\mathbf{1}(s_t^{(i)} = s_t), \tag{3.2}$$

where n_{s_τ} is the empirical visitation frequency to state s_τ at time τ. Note that our estimator of $r^{\pi}_t(s_t)$ is the standard IS estimators we use in bandits [Li et al., 2015], which are shown to be optimal (up to a universal constant) when A is large [Wang et al., 2017].

The advantage of MIS over the naive IS estimator is that the variance of the importance weight need not depend exponentially in H. A major theoretical contribution of this paper is to formalize this argument by characterizing the dependence on π, μ as well as parameters of the MDP M. Note that MIS estimator does not dominate the IS estimator. In the more general setting when the state is given by the entire history of observations, Jiang and Li [2016] establishes that no estimators can achieve polynomial dependence in H. We give a concrete example later (Example 1) about how IS estimator suffers from the "curse of horizon" [Liu et al., 2018a]. MIS estimator can be thought of as one that exploits the state-observability while retaining properties of the IS estimators to tackle the problem of large action space. As we illustrate in the experiments, MIS estimator can be modified to naturally handle *partially observed* states, e.g., when s is only observed every other step.

Finally, when available, model-based approaches can be combined into importance-weighted methods [Jiang and Li, 2016; Thomas and Brunskill, 2016]. We defer discussions about these extensions in Appendix C to stay focused on the scenarios where model-based approaches are not applicable.

4 Theoretical Analysis of the MIS Estimator

Motivated by the challenge of curse of horizon with naive IS estimators, similar to [Liu et al., 2018a], we show that the sample complexity of our MIS estimator reduces to $O(H^3)$. To the best of our knowledge, this is first sample complexity guarantee under this setting, which also matches the Cramer-Rao lower bound for DAG-MDP [Jiang and Li, 2016] as $n \to \infty$ up to a factor of H.

Example 1 (Curse of horizon)**.** *Assume a MDP with i.i.d. state transition models over time and assume that $\frac{\pi_t}{\mu_t}$ is bounded from both sides for all t. Suppose the reward is a constant 1 only shown at the last step, such that naive IS becomes $\widehat{v}^{\pi}_{\text{IS}} = \frac{1}{n}\sum_{i=1}^{n}\left[\prod_{t=1}^{H}\frac{\pi(a_t^{(i)}|s_t^{(i)})}{\mu(a_t^{(i)}|s_t^{(i)})}\right]$. For every trajectory, $\prod_{t=1}^{H}\frac{\pi_t}{\mu_t} = \exp\left[\sum_{t=1}^{H}\log\frac{\pi_t}{\mu_t}\right]$; let $E_{\log} = \mathbb{E}[\log\frac{\pi_t}{\mu_t}]$ and $V_{\log} = \text{Var}[\log\frac{\pi_t}{\mu_t}]$. By Central Limit Theorem, $\sum_{t=1}^{H}\log\frac{\pi_t}{\mu_t}$ asymptotically follows a normal distribution with parameters $\left(-HE_{\log}, HV_{\log}\right)$. In other words, $\prod_{t=1}^{H}\frac{\pi_t}{\mu_t}$ asymptotically follows $\text{LogNormal}\left(-HE_{\log}, HV_{\log}\right)$, whose variance is exponential in horizon: $\left(\exp\left(HV_{\log}\right)-1\right)$. On the other hand, MIS estimates the state distributions recursively, yielding variance that is polynomial in horizon and small OPE errors.*

We now formalize the sample complexity bound in Theorem 4.1.

Theorem 4.1. *Let the value function under π be defined as follows:*

$$V_h^{\pi}(s_h) := \mathbb{E}_{\pi}\left[\sum_{t=h}^{H} r_t(s_t^{(1)}, a_t^{(1)}, s_{t+1}^{(1)})\middle| s_h^{(1)} = s_h\right] \in [0, V_{\max}],\ \forall h \in \{1, 2, ..., H\}.$$

For the simplicity of the statement, define boundary conditions: $r_0(s_0) \equiv 0$, $\sigma_0(s_0, a_0) \equiv 0$, $\frac{d_0^{\pi}(s_0)}{d_0^{\mu}(s_0)} \equiv 1$, $\frac{\pi(a_0|s_0)}{\mu(a_0|s_0)} \equiv 1$ and $V_{H+1}^{\pi} \equiv 0$. Moreover, let $\tau_a := \max_{t,s_t,a_t}\frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}$ and $\tau_s := \max_{t,s_t}\frac{d_t^{\pi}(s_t)}{d_t^{\mu}(s_t)}$. If the number of episodes n obeys that

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

for all $t = 2, ..., H$, then the our estimator $\widehat{v}^{\pi}_{\text{MIS}}$ with an additional clipping step obeys that

$$\begin{aligned}\mathbb{E}[(\mathcal{P}\widehat{v}^{\pi}_{\text{MIS}} - v^{\pi})^2] \leq & \frac{1}{n}\sum_{h=0}^{H}\sum_{s_h}\frac{d_h^{\pi}(s_h)^2}{d_h^{\mu}(s_h)}\text{Var}_{\mu}\left[\frac{\pi(a_h^{(1)}|s_h)}{\mu(a_h^{(1)}|s_h)}(V_{h+1}^{\pi}(s_{h+1}^{(1)}) + r_h^{(1)})\middle| s_h^{(1)} = s_h\right] \\ & \cdot\left(1 + \sqrt{\frac{16\log n}{n\min_{t,s_t} d_t^{\mu}(s_t)}}\right) + \frac{19\tau_a^2\tau_s^2 SH^2(\sigma^2 + R_{\max}^2 + V_{\max}^2)}{n^2}.\end{aligned}$$

Corollary 1. *In the familiar setting when $V_{\max} = HR_{\max}$, then the same conditions in Theorem 4.1 implies that:*

$$\mathbb{E}[(\mathcal{P}\widehat{v}^{\pi}_{\text{MIS}} - v^{\pi})^2] \leq \frac{4}{n}\tau_a\tau_s(H\sigma^2 + H^3R_{\max}^2).$$

We make a few remarks about the results in Theorem 4.1.

Dependence on S, A and the weights. The leading term in the variance bound very precisely calculates the MSE of a clipped version of our estimator $\widehat{v}_{\text{MIS}}$[1] modulo a $(1 + O(n^{-1/2}))$ multiplicative factor and an $O(1/n^2)$ additive factor. Specifically, our bound does not explicitly depend on S and A but instead on how similar π and μ are. This allows the method to handle the case when the action space is continuous. The dependence on τ_a, τ_s only appear in the low-order terms, while the leading term depends only on the second moments of the importance weights.

Dependence on H. In general, our sample complexity upper bound is proportional to H^3, as Corollary 1 indicates. Our bound reveals that in several cases it is possible to achieve a smaller

[1]The clipping step to $[0, HR_{\max}]$ or $[0, V_{\max}]$ should not be alarming. It is required only for technical reasons, and the clipped estimator is a valid estimator to begin with. Since the true policy value must be within the range, the clipping step is only going to improve the MSE.

exponent on H for specific triplets of (M, π, μ). For instance, when $\pi \approx \mu$, such that $\tau_a, \tau_s = 1 + O(1/H)$, the variance bound gives $O((V_{\max}^2 + H\sigma^2)/n)$ or $O((H^2 R_{\max}^2 + H\sigma^2)/n)$, which matches the MSE bound (up to a constant) of the simple-averaging estimator that knows $\pi = \mu$ a-priori. (See Remark 3 in the Appendix for more details). If $V_{\max}$ is a constant that does not depend on H (this is often the case in games when there is a fixed reward at the end), then the sample complexity is only $O(H)$.

Optimality. Comparing to the Cramer-Rao lower bound of the Theorem 3 in [Jiang and Li, 2016], which we paraphrase below

$$\frac{1}{n}\sum_{h=1}^{H}\sum_{s_h}\frac{d_h^\pi(s_h)^2}{d_h^\mu(s_h)}\sum_{a_h}\frac{\pi_h(a_h|s_h)^2}{\mu_h(a_h|s_h)}\mathrm{Var}\left[V_{h+1}^\pi(s_{h+1}^{(1)}) + r_h^{(1)}\Big| s_h^{(1)} = s_h, a_h^{(1)} = a_h\right], \tag{4.1}$$

the MSE of our estimator is asymptotically bigger by an additive factor of

$$\frac{1}{n}\sum_{h=1}^{H}\sum_{s_h}\frac{d_h^\pi(s_h)^2}{d_h^\mu(s_h)}\mathrm{Var}_\mu\left[\frac{\pi_h(a_h^{(1)}|s_h)}{\mu_h(a_h^{(1)}|s_h)}Q_h^\pi(s_h, a_h^{(1)})\right], \tag{4.2}$$

where $Q_h^\pi(s_h, a_h) := \mathbb{E}\big[(V_{h+1}^\pi(s_{h+1}^{(1)}) + r_h^{(1)})\big| s_h^{(1)} = s_h, a_h^{(1)} = a_h\big]$ is the standard Q-function the MDP. The gap is significant as the CR lower bound (4.1) itself only has a worst-case bound of $H^2\tau_s\tau_a/n$ [2], while (4.2) is proportional to $H^3\tau_s\tau_a/n$. This implies that our estimator is optimal up to a factor of H. See Remark 4 for more details in the appendix.

It is an intriguing open question whether this additional factor of H can be removed. Our conjecture is that the answer is negative and what we established in Theorem 4.1 matches the *correct* information-theoretic limit for any methods in the cases when the action space $\mathcal{A}$ is continuous (or significantly larger than n). This conjecture is consistent with an existing lower bound in the simpler contextual bandits setting, where Wang et al. [2017] established that a variance of expectation term analogous to the one above cannot be removed, and no estimators can asymptotically attain the CR lower bound for all problems in the large state/action space setting.

4.1 Proof Sketch

In this section, we briefly describe the main technical components in the proof of Theorem 4.1. More detailed arguments are deferred to the full proof in Appendix B.

Recall that (3.1) is equivalent to $\sum_{t=1}^{H}\sum_{s_t}\widehat{d}_t^\pi(s_t)\widehat{r}^\pi(s_t)$, where $\widehat{r}^\pi(s_t)$ is estimated with importance sampling and $\widehat{d}_t^\pi(s_t)$ is recursively estimated using $\widehat{d}_{t-1}^\pi(s_{t-1})$ and the importance sampling estimator of the transition matrix $P_t^\pi(s_t|s_{t-1})$ under π. While the MIS estimator is easy to state, it is not straightforward to analyze. We highlight three challenges below.

1. *Dependent data and complex estimator:* While the episodes are independent, the data within each episode are not. Each time step of our MIS estimator uses the data from all episodes up to that time step.
2. *An annoying bias:* There is a non-zero probability that some states s_t at time t is not visited at all in all n episodes. This creates a bias in the estimator of $\hat{d}_h^\pi$ for all time $h > t$. While the probability of this happening is extremely small, conditioning on the high probability event breaks the natural conditional independences, which makes it hard to analyze.
3. *Error propagation:* The recursive estimator $\hat{d}_t^\pi$ is affected by all estimation errors in earlier time steps. Naive calculation of the error with a constant slack in each step can lead to a "snowball" effect that causes an exponential blow-up.

All these issues require delicate handling because otherwise the MSE calculation will not be tight. Our solutions are as follows.

Defining the appropriate filtration. The first observation is that we need to have a convenient representation of the data. Instead of considering the n episodes as independent trajectories, it is more useful to think of them all together as a Markov chain of multi-dimensional observations

[2] This is somewhat surprising as each of the H summands in the expression can be as large as H^2.

of n *state, action, reward* triplets. Specifically, we define the "cumulative" data up to time t by $\text{Data}_t := \left\{ s_{1:t}^{(i)}, a_{1:t-1}^{(i)}, r_{1:t-1}^{(i)} \right\}_{i=1}^{n}$. Also, we observe that the state of the Markov chain at time t can be summarized by n_{s_t} — the number of times state s_t is visited.

Fictitious estimator technique. We address the bias issue by defining a fictitious estimator $\widetilde{v}^\pi$. The fictitious estimator is constructed by, instead of $\hat{d}_t^\pi$ and $\hat{r}_t^\pi$, the fictitious version of these estimators $\tilde{d}_t^\pi$ and $\tilde{r}_t^\pi$, where $\tilde{d}_t^\pi$ is constructed recursively using

$$\tilde{d}_t^\pi(s_t) = \sum_{s_{t-1}} \tilde{P}^\pi(s_t|s_{t-1}) \tilde{d}_{t-1}^\pi(s_{t-1}).$$

The key difference is that whenever $n_{s_t} < \mathbb{E}_\mu n_{s_t}(1-\delta)$ for some $0 < \delta < 1$, we assign $\tilde{P}^\pi(s_{t+1}|s_t) = P^\pi(s_{t+1}|s_t)$ and $\tilde{r}^\pi(s_t) = \mathbb{E}_\pi[r_t|s_t]$ — the true values of interest. This ensures that the fictitious estimator is always unbiased (see Lemma B.2). Note that this fictitious estimator cannot be implemented in practice. It is used as a purely theoretical construct that simplifies the analysis of the (biased) MIS estimator. In Lemma B.1, we show that the $\tilde{v}^\pi$ and $\hat{v}^\pi$ are exponentially close to each other.

Disentangling the dependency by backwards peeling. The fictitious estimator technique reduces the problem of calculating the MSE of the MIS estimator to a variance analysis of the fictitious estimator. By recursively applying the law of total variance backwards that peels one item at a time from Data_t, we establish an exact linear decomposition of the variance of the fictitious estimator (Lemma B.3):

$$\text{Var}[\widetilde{v}^\pi] = \sum_{h=0}^{H} \sum_{s_h} \mathbb{E}\left[\frac{\widetilde{d}_h^\pi(s_h)^2}{n_{s_h}} \mathbf{1}\left(n_{s_h} \geq \frac{n d_h^\mu(s_h)}{(1-\delta)^{-1}} \right) \right] \text{Var}_\mu \left[\frac{\pi(a_h^{(1)}|s_h)}{\mu(a_h^{(1)}|s_h)} (V_{h+1}^\pi(s_{h+1}^{(1)}) + r_h^{(1)}) \middle| s_h^{(1)} = s_h \right].$$

Observe that the value function V_t^π shows up naturally. This novel decomposition can be thought of as a generalization of the celebrated Bellman-equation of variance [Sobel, 1982] in the off-policy, episodic MDP setting with a finite sample and can be of independent interest.

Characterizing the error propagation in $\tilde{d}_h^\pi(s_h)$. Lastly, we bound the error term in the state distribution estimation as follows

$$\mathbb{E}\left[\frac{\widetilde{d}_h^\pi(s_h)^2}{n_{s_h}} \mathbf{1}\left(n_{s_h} \geq \frac{n d_h^\mu(s_h)}{(1-\delta)^{-1}} \right) \right] \leq \frac{(1-\delta)^{-1}}{n} \left(\frac{d_h^\pi(s_h)^2}{d_h^\mu(s_h)} + \text{Var}\left[\widetilde{d}_h^\pi(s_h) \right] \right),$$

which reduces the problem to bounding $\text{Var}[\widetilde{d}_h^\pi(s_h)]$. We show (in Theorem B.1) that instead of an exponential blow-up as will a concentration-inequality based argument imply, the variance increases at most linearly in h: $\text{Var}[\widetilde{d}_h^\pi(s_h)] \leq \frac{2(1-\delta)^{-1} h d_h^\pi(s_h)}{n}$. The proof uses a novel decomposition of $\text{Cov}(\widetilde{d}_h^\pi)$ (Lemma B.5), which is derived using a similar backwards peeling argument as before. Finally, Theorem 4.1 is established by appropriately choosing $\delta = O(\sqrt{\log n / n \min_{t,s_t} d_t^\mu(s_t)})$.

Due to space limits, we can only highlight a few key elements of the proof. We invite the readers to check out a more detailed exposition in Appendix B.

5 Experiments

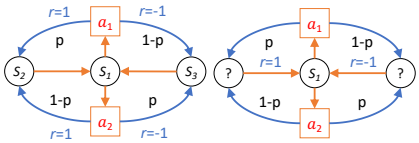
Throughout this section, we present the empirical results to illustrate the comparison among different estimators. We demonstrate the effectiveness of our proposed marginalized estimator by comparing it with different classic estimators on several domains.

The methods we compare in this section are: *direct method* (DM), *importance sampling* (IS), *weighted importance sampling* (WIS), *importance sampling with stationary state distribution* (SSD-IS), and *marginalized importance sampling* (MIS). DM uses the model-based approach to estimate $T_t(s_t|s_{t-1}, a_{t-1}), r_t(s_t, a_t)$ by enumerating all tuples of (s_{t-1}, a_{t-1}, s_t), IS is the step-wise importance sampling method, WIS uses the step-wise weighted (self-normalized) importance sampling method, SSD-IS denotes the method of importance sampling with stationary state distribution proposed by [Liu et al., 2018a][3], and MIS is our proposed marginalized method. Note that our MIS

[3]Our implementation of SSD-IS for the discrete state case is described in Appendix D.3.

also uses the trick of self-normalization to obtain better performance, but the MIS normalization is different: we normalize the estimate $\widehat{d}_t^\pi$ to the probability simplex, whereas WIS normalizes the importance weights. We provide further results by comparing doubly robust estimator, weighted doubly robust estimator, and our estimators in Appendix D. We use logarithmic scales in all figures and include 95% confidence intervals as error bars from 128 runs. Our metric is the relative root mean squared error (**Relative-RMSE**), which is **the ratio of RMSE and the true value** v^π.

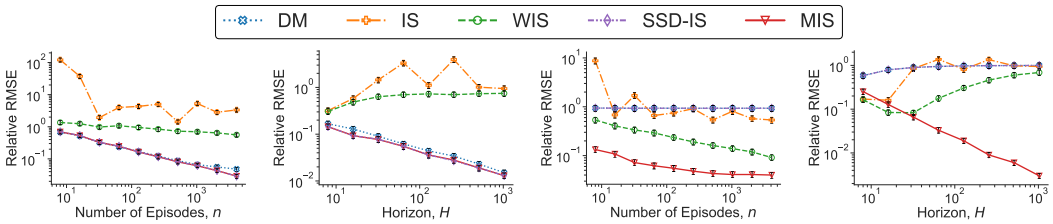
Time-invariant MDPs We first test our methods on the standard ModelWin and ModelFail models with time-invariant MDPs, first introduced by Thomas and Brunskill [2016]. The **ModelWin** domain simulates a fully observable MDP, depicted in Figure 1(a). On the other hand, the **ModelFail** domain (Figure 1(b)) simulates a partially observable MDP, where the agent can only tell the difference between s_1 and the "other" unobservable states. A detailed description of these two domains can be found in Appendix D. For both problems, the target policy π is to always select a_1 and a_2 with probabilities 0.2 and 0.8, respectively, and the behavior policy μ is a uniform policy.



(a) ModelWin (b) ModelFail

Figure 1: MDPs of OPE domains.

We provide two types of experiments to show the properties of our marginalized approach. The first kind is with different numbers of episodes, where we use a fixed horizon $H = 50$. The second kind is with different horizons, where we use a fixed number of episodes $n = 1024$. We use MIS only with observable states and the partial trajectories between them. Details about applying MIS with partial observability can be found in Appendix C. While this approach is general in more complex applications, for ModelFail, the agent always visits s_1 at every other step and we can simply replace $\frac{\pi(a_t^{(i)}|s_t^{(i)})}{\mu(a_t^{(i)}|s_t^{(i)})}$ with $\frac{\pi(a_{2\tau}^{(i)}|s_{2\tau}^{(i)}=?)}{\mu(a_{2\tau}^{(i)}|s_{2\tau}^{(i)}=?)}\frac{\pi(a_{2\tau-1}^{(i)}|s_{2\tau-1}^{(i)})}{\mu(a_{2\tau-1}^{(i)}|s_{2\tau-1}^{(i)})}$ for $t = 2\tau - 1$ in (3.2).



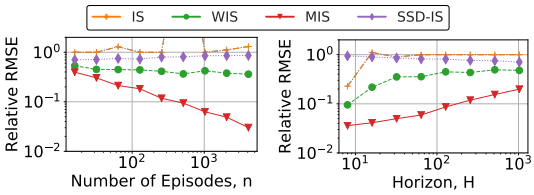
(a) ModelWin with different number of episodes n. (b) ModelWin with different horizon H. (c) ModelFail with different number of episodes n. (d) ModelFail with different horizon H.

Figure 2: Results on Time-invariant MDPs. MIS matches DM on ModelWin and outperforms IS/WIS on ModelFail, both of which are the best existing methods on their respective domains.

Figure 2 shows the results in the time-invariant ModelWin MDP and ModelFail MDP. The results clearly demonstrate that MIS maintains a polynomial dependence on H and matches the best alternatives such as DM in Figure 2(b) and IS at the beginning of Figure 2(d). Notably, the IS in Figure 2(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.

Time-varying, non-mixing MDPs with continuous actions. We also test our approach in simulated MDP environments where the states are binary, the actions are continuous between [0,1] and the state transition models are time-varying with a finite horizon H. The agent starts at State 1. At every step, the environment samples a random parameter $p \in [0.5/H, 0.5 - 0.5/H]$. Any agent in State 1 will transition to State 0 if and only if it samples an action between $[p - 0.5/H, p + 0.5/H]$. On the other hand, State 0 is a sinking state. The agent collects rewards at State 0 in the latter half of the steps ($t \geq H/2$). Thus, the agent wants to transition to State 0, but the transition probability is inversely proportional to the horizon H for uniform action policies. We pick the behavior policy to be uniform on $[0, 1]$ and the target policy to be uniform on $[0, 0.5]$ with 95% total probability and 5% chance uniformly distributed on $[0.5, 1]$.

Figure 3(a) shows the asymptotic convergence rates of RMSE with respect to the number of episodes, given fixed horizon $H = 64$. MIS converges at a $O(1/\sqrt{n})$ rate from the very beginning. In comparison, neither IS or MIS has entered their asymptotic $n^{-1/2}$ regime yet with $n \leq 4,096$. SSD-IS does not improve as n gets larger, because the stationary state distribution (a point mass on State 0) is not a good approximation of the average probability of visiting State 0 for $t \in [H/2, H]$. We exclude DM because it requires additional model assumptions to apply to continuous action spaces.



(a) Dependency on n (b) Dependency on H

Figure 3: Time-varying MDPs

Figure 3(b) shows the Relative RMSE dependency in H, fixing the number of episodes $n = 1024$. We see that as H gets larger, the Relative RMSE scales as $O(\sqrt{H})$ for MIS and stays roughly constant for SSD-IS. Since the true reward $v^\pi \propto H$, the result matches the worst-case bound of a $O(H^3)$ MSE in Corollary 1. SSD-IS saves a factor of H in variance, as it marginalizes over the H steps, but introduces a large bias as we have seen in Figure 3(a). IS and WIS worked better for small H, but quickly deteriorates as H increases. Together with Figure 3(a), we may conclude that In conclusion, MIS is the only method, among the alternatives in this example, that produces a consistent estimator with low variance.

Mountain Car. Finally, we benchmark our estimator on the Mountain Car domain [Singh and Sutton, 1996], where an under-powered car drives up a steep valley by "swinging" on both sides to gradually build up potential energy. 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 1.25. Note that this is a finite-horizon MDP with continuous state. We apply MIS by discretizing the state space as in [Jiang and Li, 2016].

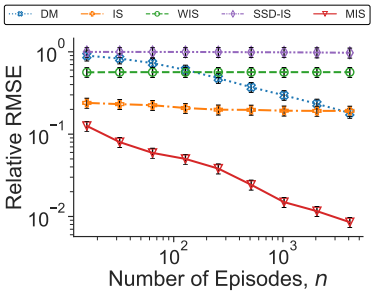


Figure 4: Mountain Car with different number of episodes.

The results, shown in Figure 4, demonstrate the effectiveness of our approach in a common benchmark control task, where the ability to evaluate under long horizons is required for success. Note that Mountain Car is an episodic environment with a absorbing state, so it is not a setting that SSD-IS is designed for. We include the the detailed description on the experimental setup and discussion on the results in Appendix D.

6 Conclusions

In this paper, we propose a marginalized importance sampling (MIS) method for the problem of off-policy evaluation in reinforcement learning. Our approach gets rid of the burden of horizon by using an estimated marginal state distribution of the target policy at every step instead of the cumulative product of importance weights.

Comparing to the pioneering work of Liu et al. [2018a] that uses a similar philosophy, this paper focuses on the finite state episodic setting with an potentially infinite action space. We proved the first finite sample error bound for such estimators with polynomial dependence in all parameters. The error bound is tight in that it matches the asymptotic variance of a fictitious estimator that has access to oracle information up to a low-order additive factor. Moreover, it is within a factor of $O(H)$ of the Cramer-Rao lower bound of this problem in [Jiang and Li, 2016]. We conjecture that this additional factor of H is required for any estimators in the *infinite action* setting.

Our experiments demonstrate that the MIS estimator is effective in practice as it achieves substantially better performance than existing approaches in a number of benchmarks.

Acknowledgement

The authors thank Yu Bai, Murali Narayanaswamy, Lin F. Yang, Nan Jiang, Phil Thomas, Ying Yang for helpful discussion and Amazon internal review committee for the feedback on an early version of the paper. We also acknowledge the NeurIPS area chair, anonymous reviewers for helpful comments and Ming Yin for carefully proofreading the paper.

YW was supported by a start-up grant from UCSB CS department, NSF-OAC 1934641 and a gift from AWS ML Research Award.

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.

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

Dudík, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning. In *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 *International Conference on Machine Learning (ICML-18)*, volume 80, pages 1447–1456, Stockholmsmässan, Stockholm Sweden. PMLR.

Gelada, C. and Bellemare, M. G. (2019). Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *AAAI Conference on Artificial Intelligence (AAAI-19)*, volume 33, pages 3647–3655.

Gottesman, O., Liu, Y., Sussex, S., Brunskill, E., and Doshi-Velez, F. (2019). Combining parametric and nonparametric models for off-policy evaluation. In *International Conference on Machine Learning (ICML-19)*.

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 (NIPS-17)*, pages 2492–2501.

Hallak, A. and Mannor, S. (2017). Consistent on-line off-policy evaluation. In *International Conference on Machine Learning (ICML-17)*, pages 1372–1383. JMLR. org.

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

Jiang, N. and Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning (ICML-16)*, pages 652–661. JMLR. org.

Li, L., Munos, R., and Szepesvari, C. (2015). Toward minimax off-policy value estimation. In *Artificial Intelligence and Statistics (AISTATS-15)*, pages 608–616.

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 (NeurIPS-18)*, 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 (NeurIPS-18)*, 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 *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 *International Conference on Machine Learning (ICML-00)*, 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.

Sobel, M. J. (1982). The variance of discounted markov decision processes. *Journal of Applied Probability*, 19(4):794–802.

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 *ACM International Conference on Information & Knowledge Management (CIKM-13)*, 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 *International Joint Conferences on Artificial Intelligence (IJCAI-15)*, pages 1806–1812.

Thomas, P. and Brunskill, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning (ICML-16)*, 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 Conference on Artificial Intelligence (AAAI-15)*, 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 Conference on Artificial Intelligence (AAAI-17)*, 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 (ICML-17)*, pages 3589–3597.

Appendix

A Concentration inequalities

Lemma A.1 (Multiplicative Chernoff bound [Chernoff et al., 1952]). *Let X be a Binomial random variable with parameter p, n. For any $\delta > 0$, we have that*

$$\mathbb{P}[X > (1+\delta)pn] < \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{np}$$

and

$$\mathbb{P}[X < (1-\delta)pn] < \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{np}.$$

A slightly weaker bound that suffices for our propose is the following:

$$\mathbb{P}[X < (1-\delta)pn] < e^{-\frac{\delta^2 pn}{2}}$$

If we take $\delta = \sqrt{\frac{20\log(n)}{pn}}$,

$$\mathbb{P}[X < (1-\delta)pn] < n^{-10}.$$

B Theoretical analysis of the marginalized IS estimator

Recall that the marginalized IS estimators are of the following form:

$$\widehat{v}^{\pi} = \sum_{t=1}^{H}\sum_{s_t} \widehat{d}_t^{\pi}(s_t)\widehat{r}_t^{\pi}(s_t),$$

where we recursively estimate the state-marginal under the target policy π using

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

We focus on the setting where the number of actions is large and possibly unbounded, in which case, we use importance sampling based estimators of $\widehat{P}_{t-1,t}^{\pi}$ and $\widehat{r}_t^{\pi}(s_t)$ instead to get bounds that are independent to A. Specifically, we use:

$$\widehat{P}_{t-1}^{\pi}(s_t|s_{t-1}) = \frac{1}{n_{s_{t-1}}}\sum_{i=1}^{n}\frac{\pi(a_{t-1}^{(i)}|s_{t-1})}{\mu(a_{t-1}^{(i)}|s_{t-1})}\mathbf{1}(s_{t-1}^{(i)} = s_{t-1}, a_{t-1}^{(i)}, s_t^{(i)} = s_t).$$

and

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

The main challenge in analyzing these involves finding a way to decompose the error in the face of the complex recursive structure, as well as to deal with the bias of the estimator.

Constructing a fictitious estimator. Our proof makes novel use of a fictitious estimator $\widetilde{v}^{\pi}$ which uses $\widetilde{d}_t^{\pi} = \widetilde{P}_{t+1,t}^{\pi}\widetilde{d}_{t-1}^{\pi}$ and $\widetilde{r}_t^{\pi}$ instead of $\widehat{d}_t^{\pi} = \widehat{P}_{t+1,t}^{\pi}(\cdot|s_t)\widehat{d}_{t-1}^{\pi}$ and $\widehat{r}_t^{\pi}$ in the original estimator $\widehat{v}^{\pi}$. To write it down more formally,

$$\widetilde{v}^{\pi} := \sum_{t=1}^{H}\sum_{s_t} \widetilde{d}_t^{\pi}(s_t)\widetilde{r}_t^{\pi}(s_t)$$

where $\widetilde{d}_t^{\pi}(s_t)$ is constructed recursively using

$$\widetilde{d}_t^{\pi} = \widetilde{\mathbb{P}}_{t,t-1}^{\pi}\widetilde{d}_{t-1}^{\pi}$$

as in our regular estimator for $t = 2, 3, 4, ..., H$, and $\widetilde{d}_1^\pi = \widehat{d}_1$. In particular,

$$\widetilde{r}_t^\pi(s_t) = \begin{cases} \widehat{r}_t^\pi(s_t) & \text{if } n_{s_t} \geq n d_t^\mu(s_t)(1-\delta) \\ r_t^\pi(s_t) & \text{otherwise;} \end{cases}$$

and

$$\widetilde{\mathbb{P}}_{t,t-1}^\pi(\cdot|s_{t-1}) = \begin{cases} \widehat{\mathbb{P}}_{t,t-1}^\pi & \text{if } n_{s_{t-1}} \geq n d_t^\mu(s_{t-1})(1-\delta) \\ \mathbb{P}_{t,t-1}^\pi & \text{otherwise.} \end{cases}$$

In the above, $0 < \delta < 1$ is a parameter that we will choose later.

This estimator $\widetilde{v}^\pi$ is fictitious because it is *not implementable* using the data[4], but it is somewhat easier to work with and behaves essentially the same as our actual estimator $\widehat{v}^\pi$. As a result, we can analyze our estimator through analyzing $\widetilde{v}^\pi$. The following lemma formalizes the idea.

Lemma B.1. *Let $\widehat{v}^\pi$ be our MIS estimator and $\mathcal{P}$ be the projection operator to $[0, HR_{\max}]$ and $\widetilde{v}^\pi$ be the unbiased fictitious estimator that we described above with parameter δ. The MSE of the clipped version of our MIS estimator obeys*

$$\mathbb{E}[(\mathcal{P}\widehat{v}^\pi - v^\pi)^2] \leq \mathbb{E}[(\widetilde{v}^\pi - v^\pi)^2] + 3H^3 S R_{\max}^2 e^{-\frac{\delta^2 n \min_{t,s_t} d_t^\mu(s_t)}{2}}$$

Proof of Lemma B.1. Let E denotes the event of $\{\exists t, s_t, \text{ s.t. } n_{s_t} < n d_t^\mu(s_t)(1-\delta)\}$. Let $\mathcal{P}_E$ be the *conditional* projection operator that clips the value to $[0, HR_{\max}]$ whenever E is true. Note that for any $x \in \mathbb{R}$, we have $\mathcal{P}(\mathcal{P}_E x) = \mathcal{P}x$. By the non-expansiveness of $\mathcal{P}$,

$$\begin{aligned}
&\mathbb{E}[(\mathcal{P}\widehat{v}^\pi - v^\pi)^2] \leq \mathbb{E}[(\mathcal{P}_E\widehat{v}^\pi - v^\pi)^2] = \mathbb{E}[(\mathcal{P}_E\widehat{v}^\pi - \mathcal{P}_E\widetilde{v}^\pi + \mathcal{P}_E\widetilde{v}^\pi - v^\pi)^2] \\
=&\mathbb{E}[(\mathcal{P}_E\widehat{v}^\pi - \mathcal{P}_E\widetilde{v}^\pi)^2] + 2\mathbb{E}[(\mathcal{P}_E\widehat{v}^\pi - \mathcal{P}_E\widetilde{v}^\pi)(\mathcal{P}_E\widetilde{v}^\pi - v^\pi)] + \mathbb{E}[(\mathcal{P}_E\widetilde{v}^\pi - v^\pi)^2] \\
=&\mathbb{P}[E]\mathbb{E}\big[(\mathcal{P}_E\widehat{v}^\pi - \mathcal{P}_E\widetilde{v}^\pi)^2 + 2(\mathcal{P}_E\widehat{v}^\pi - \mathcal{P}_E\widetilde{v}^\pi)(\mathcal{P}_E\widetilde{v}^\pi - v^\pi)\big|E\big] + \mathbb{P}[E^c]\cdot 0 + \mathbb{E}[(\mathcal{P}_E\widetilde{v}^\pi - \mathcal{P}_E v^\pi)^2] \\
\leq&3\mathbb{P}[E]H^2R_{\max}^2 + \mathbb{E}[(\widetilde{v}^\pi - v^\pi)^2].
\end{aligned}$$

The third line is by the law of total expectation and the fact that whenever E is not true, $\widehat{v}^\pi = \widetilde{v}^\pi$. The last line uses the fact that $\mathcal{P}_E\widehat{v}^\pi, \mathcal{P}_E\widetilde{v}^\pi, v^\pi$ are all within $[0, HR_{\max}]$ when conditioning on E as well as the non-expansiveness of the projection operator which implies that

$$\mathbb{E}[(\mathcal{P}_E(\widetilde{v}^\pi - v^\pi))^2] \leq \mathbb{E}[(\widetilde{v}^\pi - v^\pi)^2].$$

It remains to bound $\mathbb{P}[E]$. By the multiplicative Chernoff bound (Lemma A.1 in the Appendix) we get that

$$\mathbb{P}\left[n_{s_t} < n d_t^\mu(s_t)(1-\delta)\right] \leq e^{-\frac{\delta^2 n d_t^\mu(s_t)}{2}}$$

By a union bound over each t and s_t, we have

$$\mathbb{P}[E] \leq \sum_t \sum_{s_t} \mathbb{P}[n_{s_t,t} < n d_t^\mu(s_t)(1-\delta)] \leq HSe^{-\frac{\delta^2 n \min_{t,s_t} d_t^\mu(s_t)}{2}}$$

as stated. □

Lemma B.1 establishes that when $n \geq \frac{\text{polylog}(S,H,n)}{\min_{t,s_t} d_t^\mu(s_t)}$, we can bound the MSE of a projected version of our estimator using the MSE of the fictitious estimator. The projection to $[0, HR_{\max}]$ is a post-processing that we needed in our proof for technical reasons, and we know that $\mathbb{E}[(\mathcal{P}\widehat{v}^\pi - v^\pi)^2] \leq \mathbb{E}[(\widehat{v}^\pi - v^\pi)^2]$ so it only improves the performance.

Properties of the Fictitious Estimator. Now let us prove that $\widetilde{v}^\pi$ is unbiased and also analyze its variance. Recall that the estimator is the following:

$$\widetilde{v}^\pi = \sum_{t=1}^H \sum_{s_t} \widetilde{d}_t^\pi(s_t)\widetilde{r}_t^\pi(s_t) = \sum_{t=1}^H \langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi \rangle$$

where we denote quantities $\widetilde{d}_t^\pi, \widetilde{r}_t^\pi$ in vector forms in $\mathbb{R}^S$.

In the remainder of this section, we will use E_t as a short hand to denote the event such that $\{n_{s_t} \geq n d_t^\mu(s_t)(1-\delta)\}$, and $\mathbf{1}(E_t)$ be the corresponding indicator function.

[4] It depends on unknown information such as d_t^μ, $\mathbb{P}_{t,t-1}^\pi$, exact conditional expectation of the reward r_t^π and so on.

Lemma B.2 (Unbiasedness of $\widetilde{v}^\pi$)**.** $\mathbb{E}[\widetilde{v}^\pi] = v^\pi$ *for all* $\delta < 1$.

Proof of Lemma B.2. The idea of the proof is to recursively apply the Law of Total Expectation backwards from the last round by taking conditional expectations. For simplicity of the proof we will denote

$$\text{Data}_t := \left\{ s_{1:t}^{(i)}, a_{1:t-1}^{(i)}, r_{1:t-1}^{(i)} \right\}_{i=1}^n .$$

Also, in the base case, let's denote $\text{Data}_1 := \left\{ s_1^{(i)} \right\}_{i=1}^n$ and that $r_t^\pi(s_t) := \mathbb{E}_\pi[r_t^{(1)} | s_t^{(1)} = s_t]$

We first making a few observations that will be useful in the arguments that follow. Firstly, $\widetilde{d}_t^\pi$ and $\widetilde{r}_{t-1}^\pi$ are deterministic given Data_t. Secondly,

$$\mathbb{E}[\widetilde{P}_{t,t-1}^\pi | \text{Data}_{t-1}] = P_{t,t-1}^\pi, \quad \text{and} \quad \mathbb{E}[\widetilde{r}_t^\pi | \text{Data}_t] = r_t^\pi .$$

These observations are true for all $t = 1, ..., H$. To see the unbiasedness of the conditional expectation, note that when $n_{s_t} \geq n d_t^\mu(s_t)(1-\delta)$, the estimators are just empirical mean, which are unbiased and when $n_{s_t} < n d_t^\mu(s_t)(1-\delta)$, we also have an unbiased estimator by the construction of the fictitious estimator. For all $\delta < 1$, the case $n_{s_t} = 0$ is ruled out.Thirdly, we write down the standard Bellman equation for policy π

$$V_h(s_h) = r_h^\pi(s_h) + \sum_{s_{h+1}} P_{h+1,h}^\pi(s_{h+1}|s_h) V_{h+1}(s_{h+1}).$$

where $V_h(s_h) := \mathbb{E}_\pi \left[\sum_{t=h}^H r_t^{(1)} \middle| s_t^{(1)} = s_h \right]$ or in a matrix form

$$V_h = r_h^\pi + [P_{h+1,h}^\pi]^T V_{h+1}.$$

These observations together allow us to write the following recursion:

$$\begin{aligned}
&\mathbb{E}\left[\langle \widetilde{d}_h^\pi, V_h^\pi \rangle + \sum_{t=1}^{h-1} \langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi \rangle \middle| \text{Data}_{h-1} \right] \\
=&\langle \mathbb{E}[\widetilde{P}_{h,h-1}^\pi | \text{Data}_{h-1}] \widetilde{d}_{h-1}^\pi, V_h^\pi \rangle + \langle \widetilde{d}_{h-1}^\pi, \mathbb{E}\left[\widetilde{r}_{h-1}^\pi \middle| \text{Data}_{h-1} \right] \rangle + \sum_{t=1}^{h-2} \langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi \rangle \\
=&\langle \widetilde{d}_{h-1}^\pi, [P_{h,h-1}^\pi]^T V_h^\pi + r_{h-1}^\pi \rangle + \sum_{t=1}^{h-2} \langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi \rangle \\
\underset{\substack{\uparrow \\ \text{Bellman equation}}}{=}&\langle \widetilde{d}_{h-1}^\pi, V_{h-1}^\pi \rangle + \sum_{t=1}^{h-2} \langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi \rangle .
\end{aligned}$$

Finally, by taking (full) expectation and chaining the above recursions together, we get

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^H \langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi \rangle \right] &= \mathbb{E}\left[\langle \widetilde{d}_H^\pi, V_H^\pi \rangle + \sum_{t=1}^{H-1} \langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi \rangle \right] \\
&= \mathbb{E}\left[\langle \widetilde{d}_{H-1}^\pi, V_{H-1}^\pi \rangle + \sum_{t=1}^{H-2} \langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi \rangle \right] \\
&= \dots \\
&= \mathbb{E}\left[\langle \widetilde{d}_1^\pi, V_1^\pi \rangle \right] = v^\pi,
\end{aligned}$$

which concludes the proof. □

Now let's tackle the variance of the fictitious estimator.

Lemma B.3 (Variance decomposition)**.**

$$\begin{aligned}\mathrm{Var}[\widetilde{v}^\pi] =& \frac{\mathrm{Var}[V_1^\pi(s_1^{(1)})]}{n} \\ &+ \sum_{h=1}^{H}\sum_{s_h}\mathbb{E}\left[\frac{\widetilde{d}_h^\pi(s_h)^2}{n_{s_h}}\mathbf{1}(E_h)\right]\mathrm{Var}_\mu\left[\frac{\pi(a_h^{(1)}|s_h)}{\mu(a_h^{(1)}|s_h)}(V_{h+1}^\pi(s_{h+1}^{(1)})+r_h^{(1)})\middle| s_h^{(1)}=s_h\right].\end{aligned}$$

where $V_t^\pi(s_t)$ denotes the value function under π which satisfies the Bellman equation

$$V_t^\pi(s_t) = r_t^\pi(s_t) + \sum_{s_{t+1}} P_t^\pi(s_{t+1}|s_t)V_{t+1}^\pi(s_{t+1}).$$

Remark 1. *The decomposition of variance is very interpretable. The first part of the variance is coming from estimating the initial state. The second part is coming from the conditional variance of estimating $P_{t+1,t}^\pi(s_t)$ and $r_t^\pi(s_t)$ using importance sampling over a_t.*

Proof of Lemma B.3. The proof uses a peeling argument that recursively applies the law of total variance from the last time point backwards.

The key of the argument relies upon the following identity that holds for all $h = 1, ..., H-1$.

$$\begin{aligned}\mathrm{Var}\left[\langle\widetilde{d}_{h+1}^\pi, V_{h+1}^\pi\rangle + \sum_{t=1}^{h}\langle\widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle\right] =& \mathbb{E}\left[\mathrm{Var}\left[\langle\widetilde{d}_{h+1}^\pi, V_{h+1}^\pi\rangle + \langle\widetilde{d}_h^\pi, \widetilde{r}_h^\pi\rangle\middle|\mathrm{Data}_h\right]\right] \\ &+ \mathrm{Var}\left[\langle\widetilde{d}_h^\pi, V_h^\pi\rangle + \sum_{t=1}^{h-1}\langle\widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle\right].\end{aligned} \tag{B.1}$$

Note that in (B.1), when we condition on Data_h, $\widetilde{d}_h^\pi$ is fixed. Also, $\widetilde{P}_{h+1,h}(\cdot, s_h)$ and $\widetilde{r}_h^\pi(s_h)$ for each s_h are conditionally independent given Data_h, since Data_h partitions the n episodes into S disjoint sets according to the states $s_h^{(i)}$ at time h. These observations imply that

$$\begin{aligned}&\mathbb{E}\left[\mathrm{Var}\left[\langle\widetilde{d}_{h+1}^\pi, V_{h+1}^\pi\rangle + \langle\widetilde{d}_h^\pi, \widetilde{r}_h^\pi\rangle\middle|\mathrm{Data}_h\right]\right] \\ =&\mathbb{E}\left[\sum_{s_h}\mathrm{Var}\left[\widetilde{d}_h^\pi(s_h)\langle\widetilde{P}_{h+1,h}(\cdot,s_h), V_{h+1}^\pi\rangle + \widetilde{d}_h^\pi(s_h)\cdot\widetilde{r}_h^\pi(s_h)\middle|\mathrm{Data}_h\right]\right] \\ =&\mathbb{E}\left[\sum_{s_h}\mathbf{1}(E_h)\mathrm{Var}\left[\widetilde{d}_h^\pi(s_h)\langle\widetilde{P}_{h+1,h}(\cdot,s_h), V_{h+1}^\pi\rangle + \widetilde{d}_h^\pi(s_h)\cdot\widetilde{r}_h^\pi(s_h)\middle|\mathrm{Data}_h\right]\right] \\ =&\mathbb{E}\left[\sum_{s_h}\mathbf{1}(E_h)\mathrm{Var}\left[\left\langle\frac{\widetilde{d}_h^\pi(s_h)}{n_{s_h}}\sum_{i|s_h^{(i)}=s_h}\frac{\pi(a_h^{(i)}|s_h)}{\mu(a_h^{(i)}|s_h)}\mathbf{e}_{s_{h+1}^{(i)}}, V_{h+1}^\pi\right\rangle + \frac{\widetilde{d}_h^\pi(s_h)}{n_{s_h}}\sum_{i|s_h^{(i)}=s_h}\frac{\pi(a_h^{(i)}|s_h)}{\mu(a_h^{(i)}|s_h)}r_h^{(i)}\middle|\mathrm{Data}_h\right]\right] \\ =&\mathbb{E}\left[\sum_{s_h}\widetilde{d}_h^\pi(s_h)^2\mathbf{1}(E_h)\mathrm{Var}\left[\frac{1}{n_{s_h}}\sum_{i|s_h^{(i)}=s_h}\frac{\pi(a_h^{(i)}|s_h)}{\mu(a_h^{(i)}|s_h)}(V_{h+1}^\pi(s_{h+1}^{(i)})+r_h^{(i)})\middle|\mathrm{Data}_h\right]\right] \\ =&\sum_{s_h}\mathbb{E}\left[\frac{\widetilde{d}_h^\pi(s_h)^2}{n_{s_h}}\mathbf{1}(E_h)\right]\mathrm{Var}\left[\frac{\pi(a_h^{(1)}|s_h)}{\mu(a_h^{(1)}|s_h)}(V_{h+1}^\pi(s_{h+1}^{(1)})+r_h^{(1)})\middle| s_h^{(1)}=s_h\right].\end{aligned} \tag{B.2}$$

The second line uses the conditional independence we mentioned above. The third line uses that when $n_{s_h} < nd_h^\mu(s_h)$, the conditional variance is 0. The fourth and fifth line apply the definition of the importance sampling estimators and finally the last line uses that the episodes are iid.

Apply (B.1) recursively

$$
\begin{aligned}
\mathrm{Var}[\widetilde{v}^\pi] =& \mathbb{E}\mathrm{Var}[\widetilde{v}^\pi|\mathrm{Data}_H] + \mathrm{Var}[\mathbb{E}[\widetilde{v}^\pi|\mathrm{Data}_H]] \\
=& \mathbb{E}\left[\mathrm{Var}[\langle \widetilde{d}_H^\pi, \widetilde{r}_H^\pi\rangle|\mathrm{Data}_H]\right] + \mathrm{Var}[\mathbb{E}[\langle \widetilde{d}_H^\pi, \widetilde{r}_H^\pi\rangle|\mathrm{Data}_H] + \sum_{t=1}^{H-1}\langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle] \\
=& \mathbb{E}\left[\mathrm{Var}[\langle \widetilde{d}_H^\pi, \widetilde{r}_H^\pi\rangle|\mathrm{Data}_H]\right] + \mathrm{Var}[\langle \widetilde{d}_H^\pi, r_H^\pi\rangle + \sum_{t=1}^{H-1}\langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle] \\
=& \mathbb{E}\left[\mathrm{Var}[\langle \widetilde{d}_H^\pi, \widetilde{r}_H^\pi\rangle|\mathrm{Data}_H]\right] + \mathrm{Var}[\langle \widetilde{d}_H^\pi, V_H^\pi\rangle + \sum_{t=1}^{H-1}\langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle] \\
=& \mathbb{E}\left[\mathrm{Var}[\langle \widetilde{d}_H^\pi, \widetilde{r}_H^\pi\rangle|\mathrm{Data}_H]\right] + \mathbb{E}\left[\mathrm{Var}\left[\langle \widetilde{d}_H^\pi, V_H^\pi\rangle + \langle \widetilde{d}_{H-1}^\pi, \widetilde{r}_{H-1}^\pi\rangle\middle|\mathrm{Data}_{H-1}\right]\right] \\
&+ \mathrm{Var}\left[\langle \widetilde{d}_{H-1}^\pi, V_{H-1}^\pi\rangle + \sum_{t=1}^{H-2}\langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle\right] \\
=& \mathbb{E}\left[\mathrm{Var}[\langle \widetilde{d}_H^\pi, \widetilde{r}_H^\pi\rangle|\mathrm{Data}_H]\right] + \sum_{h=H-1}^{H}\mathbb{E}\left[\mathrm{Var}\left[\langle \widetilde{d}_h^\pi, V_h^\pi\rangle + \langle \widetilde{d}_{h-1}^\pi, \widetilde{r}_{h-1}^\pi\rangle\middle|\mathrm{Data}_{h-1}\right]\right] \\
&+ \mathrm{Var}\left[\langle \widetilde{d}_{H-2}^\pi, V_{H-2}^\pi\rangle + \sum_{t=1}^{H-3}\langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle\right] \\
=& \mathbb{E}\left[\mathrm{Var}[\langle \widetilde{d}_H^\pi, \widetilde{r}_H^\pi\rangle|\mathrm{Data}_H]\right] + \sum_{h=2}^{H}\mathbb{E}\left[\mathrm{Var}\left[\langle \widetilde{d}_h^\pi, V_h^\pi\rangle + \langle \widetilde{d}_{h-1}^\pi, \widetilde{r}_{h-1}^\pi\rangle\middle|\mathrm{Data}_{h-1}\right]\right] + \mathrm{Var}\left[\langle \widetilde{d}_1^\pi, V_1^\pi\rangle\right]
\end{aligned}
$$

Use the boundary condition $V_{H+1} \equiv 0$ as stated in the theorem and apply (B.2), we get that

$$
\begin{aligned}
\mathrm{Var}[\widetilde{v}^\pi] =& \frac{\mathrm{Var}[V_1^\pi(s_1^{(1)})]}{n} \\
&+ \sum_{h=1}^{H}\sum_{s_h}\mathbb{E}\left[\frac{\widetilde{d}_h^\pi(s_h)^2}{n_{s_h}}\mathbf{1}(E_h)\right]\mathrm{Var}\left[\frac{\pi(a_h^{(1)}|s_h)}{\mu(a_h^{(1)}|s_h)}(V_{h+1}^\pi(s_{h+1}^{(1)}) + r_h^{(1)})\middle| s_h^{(1)} = s_h\right].
\end{aligned}
$$

This completes the proof. □

Bounding the importance weights It remains to show that for all h, s_h,

$$
\mathbb{E}\left[\frac{\widetilde{d}_h^\pi(s_h)^2}{n_{s_h}}\mathbf{1}(E_h)\right] \approx \frac{d_h^\pi(s_h)^2}{n d_h^\mu(s_h)}.
$$

By the non-negativity of $\widetilde{d}_h^\pi(s_h)^2$

$$
\mathbb{E}\left[\frac{\widetilde{d}_h^\pi(s_h)^2}{n_{s_h}}\mathbf{1}(E_h)\right] \le \frac{(1-\delta)^{-1}}{n d_h^\mu(s_h)}\mathbb{E}\left[\widetilde{d}_h^\pi(s_h)^2\right] = \frac{(1-\delta)^{-1}}{n d_h^\mu(s_h)}(d_h^\pi(s_h)^2 + \mathrm{Var}[\widetilde{d}_h^\pi(s_h)]). \tag{B.3}
$$

where the last identity is true because $\widetilde{d}_h^\pi$ is an unbiased estimator of $d_h^\pi(s_h)$ as the following lemma establishes.

Lemma B.4 (Unbiasedness of $\widetilde{d}_h^\pi$)**.** *For all $h = 1, ..., H$, the fictitious state marginal estimators are unbiased, that is,*

$$
\mathbb{E}[\widetilde{d}_h^\pi] = d_h^\pi.
$$

Proof of Lemma B.4. Recall the recursive relationship by construction

$$
\widetilde{d}_h^\pi = \widetilde{\mathbb{P}}_{h,h-1}^\pi \widetilde{d}_{h-1}^\pi.
$$

We will prove by induction on h. First, take the base case $h = 1$: $\mathbb{E}[\widetilde{d}_1^\pi] = \mathbb{E}[\widehat{d}_1^\pi] = d_1^\pi$. Now if $\mathbb{E}[\widetilde{d}_{h-1}^\pi] = d_{h-1}^\pi$, then by the law of total expectation:

$$\begin{aligned}\mathbb{E}[\widetilde{d}_h^\pi] &= \mathbb{E}\left[\mathbb{E}[\widetilde{\mathbb{P}}_{h,h-1}^\pi \widetilde{d}_{h-1}^\pi | \text{Data}_{h-1}]\right] \\ &= \mathbb{P}_{h,h-1}^\pi \mathbb{E}\left[\widetilde{d}_{h-1}^\pi\right] = \mathbb{P}_{h,h-1}^\pi d_{h-1}^\pi = d_h^\pi.\end{aligned}$$

This completes the proof for all h. □

So the problem reduces to bounding $\mathrm{Var}[\widetilde{d}_h^\pi(s_h)]$. We will prove something more useful by bounding the covariance matrix of $\widetilde{d}_h^\pi(s_h)$ in semidefinite ordering.

Lemma B.5 (Covariance of $\widetilde{d}_h^\pi$)**.**

$$\begin{aligned}&\mathrm{Cov}(\widetilde{d}_h^\pi) \\ &\preceq \frac{(1-\delta)^{-1}}{n} \sum_{t=2}^{h} \mathbb{P}_{h,t}^\pi \mathrm{diag}\left[\sum_{s_{t-1}} \frac{d_{t-1}^\pi(s_{t-1})^2 + \mathrm{Var}(\widetilde{d}_{t-1}^\pi(s_{t-1}))}{d_{t-1}^\mu(s_{t-1})} \sum_{a_{t-1}} \frac{\pi(a_{t-1}|s_{t-1})^2}{\mu(a_{h-1}|s_{t-1})} \mathbb{P}_{t,t-1}(\cdot|s_{t-1}, a_{t-1})\right] \left[\mathbb{P}_{h,t}^\pi\right]^T \\ &\quad + \frac{1}{n} \mathbb{P}_{h,1}^\pi \mathrm{diag}\left[d_1^\pi\right] \left[\mathbb{P}_{h,1}^\pi\right]^T.\end{aligned}$$

where $\mathbb{P}_{h,t}^\pi = \mathbb{P}_{h,h-1}^\pi \cdot \mathbb{P}_{h-1,h-2}^\pi \cdot \ldots \cdot \mathbb{P}_{t+1,t}^\pi$ — the transition matrices under policy π from time t to h (define $\mathbb{P}_{h,h}^\pi := I$).

Before proving the result, let us connect it to what we need in (B.3).

Corollary 2. *For $h = 1$, we have:*

$$\mathrm{Var}[\widetilde{d}_1^\pi(s_1)] = \frac{1}{n}(d_h^\pi(s_1) - d_h^\pi(s_1)^2).$$

For $h = 2, 3, \ldots, H$, we have:

$$\mathrm{Var}[\widetilde{d}_h^\pi(s_h)] \leq \frac{(1-\delta)^{-1}}{n} \sum_{t=2}^{h} \sum_{s_t} \mathbb{P}_{h,t}^\pi(s_h|s_t)^2 \varrho(s_t) + \frac{1}{n} \sum_{s_1} \mathbb{P}_{h,1}^\pi(s_h|s_1)^2 d_1(s_1)$$

where $\varrho(s_t) := \sum_{s_{t-1}} \left(\frac{d_{t-1}^\pi(s_{t-1})^2 + \mathrm{Var}(\widetilde{d}_{t-1}^\pi(s_{t-1}))}{d_{t-1}^\mu(s_{t-1})} \sum_{a_{t-1}} \frac{\pi(a_{t-1}|s_{t-1})^2}{\mu(a_{t-1}|s_{t-1})} \mathbb{P}_{t,t-1}(s_t|s_{t-1}, a_{t-1}) \right)$.

Note that we have $\mathrm{Var}[\widetilde{d}_{t-1}^\pi(s_{t-1})]$ on the RHS of the equation, which suggests that we in fact need to recursively apply our bounds from $h = 1$ to obtain the overall bound.

Theorem B.1 (Error propagation)**.** *Let $\tau_a := \max_{t,s_t,a_t} \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}$ and $\tau_s := \max_{t,s_t} \frac{d_t^\pi(s_t)}{d_t^\mu(s_t)}$*[5]*. If $n \geq \frac{2(1-\delta)^{-1} t \tau_a \tau_s}{\max\{d_t^\pi(s_t), d_t^\mu(s_t)\}}$ for all $t = 2, \ldots, H$, then for all $h = 1, 2, \ldots, H$ and s_h, we have that:*

$$\mathrm{Var}[\widetilde{d}_h^\pi(s_h)] \leq \frac{2(1-\delta)^{-1} h \tau_a \tau_s}{n} d_h^\pi(s_h).$$

Proof of Theorem B.1. We prove by induction. The base case for $h = 1$ is trivially true because

$$\mathrm{Var}[\widetilde{d}_1^\pi(s_1)] = \frac{1}{n}(d_1^\pi(s_1) - d_1^\pi(s_1)^2) \leq \frac{2(1-\delta)^{-1} \tau_a \tau_s}{n} d_1^\pi(s_1).$$

since $\tau_a \geq 1$ and $\tau_s \geq 1$ by construction.

Assume $\mathrm{Var}[\widetilde{d}_t^\pi(s_t)] \leq \frac{2(1-\delta)^{-1} t \tau_a \tau_s}{n} d_t^\pi(s_t)$ is true for all $t = 1, \ldots, h-1$, then by our assumption on n and that $h \leq H$, we obtain that

$$\mathrm{Var}[\widetilde{d}_t^\pi(s_t)] \leq d_t^\pi(s_t) \max\{d_t^\pi(s_t), d_t^\mu(s_t)\}$$

[5] These are really not in more precise calculations but are assumed to simplify the statement of our results.

for all $t = 1, ..., h-1$. Plug this into Corollary 2, we get that

$$\begin{aligned}
\varrho(s_t) \leq & \sum_{s_{t-1}} \left(d^{\pi}_{t-1}(s_{t-1}) \frac{2\max\{d^{\pi}_{t-1}(s_{t-1}), d^{\mu}_{t-1}(s_{t-1})\}}{d^{\mu}_{t-1}(s_{t-1})} \sum_{a_{h-1}} \frac{\pi(a_{t-1}|s_{t-1})^2}{\mu(a_{t-1}|s_{t-1})} \mathbb{P}_{t,t-1}(s_t|s_{t-1}, a_{t-1}) \right) \\
\leq & 2\tau_s\tau_a \sum_{s_{t-1}} d^{\pi}_{t-1}(s_{t-1}) \sum_{a_{h-1}} \pi(a_{t-1}|s_{t-1}) \mathbb{P}_{t,t-1}(s_t|s_{t-1}, a_{t-1}) \\
= & 2\tau_s\tau_a d^{\pi}_t(s_t),
\end{aligned}$$

and that

$$\begin{aligned}
\mathrm{Var}[\widetilde{d}^{\pi}_h(s_h)] \leq & \frac{2(1-\delta)^{-1}\tau_s\tau_a}{n} \sum_{t=2}^{h} \sum_{s_t} \mathbb{P}^{\pi}_{h,t}(s_h|s_t)^2 d^{\pi}_t(s_t) + \frac{1}{n} \sum_{s_1} \mathbb{P}^{\pi}_{h,1}(s_h|s_1)^2 d_1(s_1) \\
\leq & \frac{2(1-\delta)^{-1}\tau_s\tau_a}{n} \sum_{t=1}^{h} \sum_{s_t} \mathbb{P}^{\pi}_{h,t}(s_h|s_t)^2 d^{\pi}_t(s_t) \\
\leq & \frac{2(1-\delta)^{-1}\tau_s\tau_a}{n} \sum_{t=1}^{h} \sum_{s_t} \mathbb{P}^{\pi}_{h,t}(s_h|s_t) d^{\pi}_t(s_t) \\
= & \frac{2(1-\delta)^{-1}h\tau_s\tau_a}{n} d^{\pi}_h(s_h)
\end{aligned}$$

The second inequality uses that $\tau_s, \tau_a \geq 1$, the third inequality uses that $0 \leq \mathbb{P}^{\pi}_{h,t}(s_h|s_t) \leq 1$. □

Note that the bound is tight and it implies that the error propagation is moderate. Instead of increasing exponentially, the error increases only linearly in time horizon, as long as n is at least linear in h.

Proof of Lemma B.5. We start by applying the law of total variance to obtain the following recursive equation

$$\begin{aligned}
\mathrm{Cov}[\widetilde{d}^{\pi}_h] &= \mathbb{E}\left[\mathrm{Cov}\left[\widetilde{\mathbb{P}}^{\pi}_{h,h-1}\widetilde{d}^{\pi}_{h-1}\middle|\mathrm{Data}_{h-1}\right]\right] + \mathrm{Cov}\left[\mathbb{E}\left[\widetilde{\mathbb{P}}^{\pi}_{h,h-1}\widetilde{d}^{\pi}_{h-1}\middle|\mathrm{Data}_{h-1}\right]\right] \\
&= \mathbb{E}\left[\mathrm{Cov}\left[\sum_{s_{h-1}} \widetilde{\mathbb{P}}^{\pi}_{h,h-1}(\cdot|s_{h-1})\widetilde{d}^{\pi}_{h-1}(s_{h-1})\middle|\mathrm{Data}_{h-1}\right]\right] + \mathrm{Cov}\left[\mathbb{E}\left[\widetilde{\mathbb{P}}^{\pi}_{h,h-1}\widetilde{d}^{\pi}_{h-1}\middle|\mathrm{Data}_{h-1}\right]\right] \\
&= \underbrace{\mathbb{E}\left[\sum_{s_{h-1}} \mathrm{Cov}\left[\widetilde{\mathbb{P}}^{\pi}_{h,h-1}(\cdot|s_{h-1})\middle|\mathrm{Data}_{h-1}\right] \widetilde{d}^{\pi}_{h-1}(s_{h-1})^2\right]}_{(***)} + \mathbb{P}^{\pi}_{h,h-1}\mathrm{Cov}[\widetilde{d}^{\pi}_{h-1}][\mathbb{P}^{\pi}_{h,h-1}]^T.
\end{aligned} \tag{B.4}$$

The decomposition of the covariance in the third line uses that $\mathrm{Cov}(X+Y) = \mathrm{Cov}(X) + \mathrm{Cov}(Y)$ when X and Y are statistically independent. Note that $n_{s_{h-1}}$, $\widetilde{d}^{\pi}_{h-1}(s_{h-1})$ are fixed and the columns

of $\widetilde{\mathbb{P}}_{h,h-1}$ are independent when conditioning on Data_{h-1}.

$$
\begin{aligned}
(***) =& \mathbb{E}\left[\sum_{s_{h-1}} \text{Cov}\left[\frac{1}{n_{s_{h-1}}}\sum_{i=1}^{n}\frac{\pi(a_{h-1}^{(i)}|s_{h-1}^{(i)})}{\mu(a_{h-1}^{(i)}|s_{h-1}^{(i)})}\mathbf{1}(s_{h-1}^{(i)}=s_{h-1})\mathbf{e}_{s_h^{(i)}}\middle|\text{Data}_{h-1}\right]\mathbf{1}(E_{h-1})\widetilde{d}_{h-1}^{\pi}(s_{h-1})^2\right] \\
=& \mathbb{E}\left[\sum_{s_{h-1}}\frac{1}{n_{s_{h-1}}}\text{Cov}\left[\frac{\pi(a_{h-1}^{(1)}|s_{h-1})}{\mu(a_{h-1}^{(1)}|s_{h-1})}\mathbf{e}_{s_h^{(1)}}\middle|s_{h-1}^{(1)}=s_{h-1}\right]\mathbf{1}(E_{h-1})\widetilde{d}_{h-1}^{\pi}(s_{h-1})^2\right] \\
=& \sum_{s_{h-1}}\left\{\mathbb{E}\left[\frac{1}{n_{s_{h-1}}}\mathbf{1}(E_{h-1})\widetilde{d}_{h-1}^{\pi}(s_{h-1})^2\right]\left(\sum_{a_{h-1}}\frac{\pi(a_{h-1}|s_{h-1})^2}{\mu(a_{h-1}|s_{h-1})}\text{diag}[\mathbb{P}_{h,h-1}(\cdot|s_{h-1},a_{h-1})]\right.\right. \\
&\left.\left.-\mathbb{P}_{h,h-1}^{\pi}(\cdot|s_{h-1})[\mathbb{P}_{h,h-1}^{\pi}(\cdot|s_{h-1})]^T\right)\right\} \\
\prec& \sum_{s_{h-1}}\left\{\frac{d_{h-1}^{\pi}(s_{h-1})^2+\text{Var}[\widetilde{d}_{h-1}^{\pi}(s_{h-1})]}{nd_{h-1}^{\mu}(s_{h-1})(1-\delta)}\sum_{a_{h-1}}\frac{\pi(a_{h-1}|s_{h-1})^2}{\mu(a_{h-1}|s_{h-1})}\text{diag}[\mathbb{P}_{h,h-1}(\cdot|s_{h-1},a_{h-1})]\right\}
\end{aligned}
\tag{B.5}
$$

The second line uses the fact that $(s_h^{(i)}, a_h^{(i)})$ are i.i.d over i given $s_{h-1}^{(i)} = s_{h-1}$. The third line uses law of total variance over $a_{h-1}^{(1)}$ as follows

$$
\begin{aligned}
&\text{Cov}\left[\frac{\pi(a_{h-1}^{(1)}|s_{h-1})}{\mu(a_{h-1}^{(1)}|s_{h-1})}\mathbf{e}_{s_h^{(1)}}\middle|s_{h-1}^{(1)}=s_{h-1}\right] \\
=& \mathbb{E}\left[\left(\frac{\pi(a_{h-1}^{(1)}|s_{h-1})}{\mu(a_{h-1}^{(1)}|s_{h-1})}\right)^2\text{Cov}\left[\mathbf{e}_{s_h^{(1)}}\middle|a_{h-1}^{(1)},s_{h-1}^{(1)}=s_{h-1}\right]\middle|s_{h-1}^{(1)}=s_{h-1}\right] \\
&+\text{Cov}\left[\frac{\pi(a_{h-1}^{(1)}|s_{h-1})}{\mu(a_{h-1}^{(1)}|s_{h-1})}\mathbb{E}\left[\mathbf{e}_{s_h^{(1)}}\middle|a_{h-1}^{(1)},s_{h-1}^{(1)}=s_{h-1}\right]\middle|s_{h-1}^{(1)}=s_{h-1}\right] \\
=& \sum_{a_{h-1}}\frac{\pi(a_{h-1}|s_{h-1})^2}{\mu(a_{h-1}|s_{h-1})}\left[\text{diag}(\mathbb{P}_{h,h-1}(\cdot|s_{h-1},a_{h-1}))-\mathbb{P}_{h,h-1}(\cdot|s_{h-1},a_{h-1})\mathbb{P}(\cdot|s_{h-1},a_{h-1})^T\right] \\
&+\sum_{a_{h-1}}\frac{\pi(a_{h-1}|s_{h-1})^2}{\mu(a_{h-1}|s_{h-1})}\mathbb{P}_{h,h-1}(\cdot|s_{h-1},a_{h-1})\mathbb{P}_{h,h-1}(\cdot|s_{h-1},a_{h-1})^T-\mathbb{P}_{h,h-1}^{\pi}(\cdot|s_{h-1})[\mathbb{P}_{h,h-1}^{\pi}(\cdot|s_{h-1})]^T \\
=& \sum_{a_{h-1}}\frac{\pi(a_{h-1}|s_{h-1})^2}{\mu(a_{h-1}|s_{h-1})}\text{diag}(\mathbb{P}_{h,h-1}(\cdot|s_{h-1},a_{h-1}))-\mathbb{P}_{h,h-1}^{\pi}(\cdot|s_{h-1})[\mathbb{P}_{h,h-1}^{\pi}(\cdot|s_{h-1})]^T
\end{aligned}
$$

The last line (B.5) follows from the fact that $\mathbb{P}_{h,h-1}^{\pi}(\cdot|s_{h-1})[\mathbb{P}_{h,h-1}^{\pi}(\cdot|s_{h-1})]^T$ is positive semidefinite and that $\mathbb{E}[X^2] = \text{Var}[X] + (\mathbb{E}[X])^2$. Combining (B.4) and (B.5) and by recursively apply them, we get the stated results. □

Combine Lemma B.1, (B.3) and Theorem B.1 with an appropriately chosen δ, we get our final result:

Theorem 4.1 (Main Theorem, restated)**.** *Let the immediate expected reward, its variance and the value function be defined as follows (for all $h = 1, 2, 3, ..., H$):*

$$
\begin{aligned}
r_h(s_h, a_h, s_{h+1}) &:= \mathbb{E}_\pi\left[r_h^{(1)}\middle|s_h^{(1)}=s_h, a_h^{(1)}=a_h, s_{h+1}^{(1)}=s_{h+1}\right] \in [0, R_{\max}] \\
\sigma_h(s_h, a_h, s_{h+1}) &:= \text{Var}_\pi\left[r_h^{(1)}\middle|s_h^{(1)}=s_h, a_h^{(1)}=a_h, s_{h+1}^{(1)}=s_{h+1}\right]^{1/2} \le \sigma \\
V_h^\pi(s_h) &:= \mathbb{E}_\pi\left[\sum_{t=h}^{H} r_t(s_t^{(1)}, a_t^{(1)})\middle|s_h^{(1)}=s_h\right] \in [0, V_{\max}].
\end{aligned}
$$

For the simplicity of the statement, define boundary conditions: $r_0(s_0) \equiv 0$, $\sigma_0(s_0, a_0) \equiv 0$, $\frac{d_0^\pi(s_0)}{d_0^\mu(s_0)} \equiv 1$, $\frac{\pi(a_0|s_0)}{\mu(a_0|s_0)} \equiv 1$ *and* $V_{H+1}^\pi \equiv 0$. *Moreover, let* $\tau_a := \max_{t,s_t,a_t} \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}$ *and* $\tau_s := \max_{t,s_t} \frac{d_t^\pi(s_t)}{d_t^\mu(s_t)}$. *If the number of episodes* n *obeys that*

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

for all $t = 2, ..., H$, *then the our estimator* $\widehat{v}_{\text{MIS}}^\pi$ *with an additional clipping step obeys that*

$$\mathbb{E}[(\mathcal{P}\widehat{v}_{\text{MIS}}^\pi - v^\pi)^2] \leq \frac{1}{n}\sum_{h=0}^{H}\sum_{s_h}\frac{d_h^\pi(s_h)^2}{d_h^\mu(s_h)}\text{Var}\left[\frac{\pi(a_h^{(1)}|s_h)}{\mu(a_h^{(1)}|s_h)}(V_{h+1}^\pi(s_{h+1}^{(1)}) + r_h^{(1)})\middle| s_h^{(1)} = s_h\right] \cdot \left(1 + \sqrt{\frac{16\log n}{n\min_{t,s_t} d_t^\mu(s_t)}}\right) + \frac{19\tau_a^2\tau_s^2 SH^2(\sigma^2 + R_{\max}^2 + V_{\max}^2)}{n^2}.$$

Proof of Theorem 4.1. Choose $\delta = \sqrt{4\log(n)/(n\min_{t,s_t} d_t^\mu(s_t))}$. Lemma B.2, Lemma B.3 and Theorem B.1 provide an MSE bound of the fictitious estimator and then by substituting the resulting bound to Lemma B.1, we obtain:

$$\begin{aligned}
&\mathbb{E}[(\mathcal{P}\widehat{v}_{\text{MIS}}^\pi - v^\pi)^2] \\
\leq &\frac{\text{Var}[V_1^\pi(s_1^{(1)})]}{n} + \frac{(1-\delta)^{-1}}{n}\sum_{h=1}^{H}\sum_{s_h}\frac{d_h^\pi(s_h)^2}{d_h^\mu(s_h)}\text{Var}\left[\frac{\pi(a_h^{(1)}|s_h)}{\mu(a_h^{(1)}|s_h)}(V_{h+1}^\pi(s_{h+1}^{(1)}) + r_h^{(1)})\middle| s_h^{(1)} = s_h\right] \\
&+ \frac{(1-\delta)^{-1}}{n}\sum_{h=1}^{H}\sum_{s_h}\frac{2(1-\delta)^{-1}h\tau_a\tau_s}{n}\frac{d_h^\pi(s_h)}{d_h^\mu(s_h)}\text{Var}\left[\frac{\pi(a_h^{(1)}|s_h)}{\mu(a_h^{(1)}|s_h)}(V_{h+1}^\pi(s_{h+1}^{(1)}) + r_h^{(1)})\middle| s_h^{(1)} = s_h\right] \qquad \text{(B.6)} \\
&+ \frac{3}{n^2}HSV_{\max}^2.
\end{aligned}$$

The first assumption on n ensures that $\delta < 1/2$, which allows us to write $(1-\delta)^{-1} \leq (1+2\delta)$ in the leading term and $(1-\delta)^{-1} \leq 2$ in the subsequent terms. The second assumption on n ensures that we can apply Theorem B.1 with parameter $\delta < 1/2$.

Then to obtain the simplified expression as stated in the theorem, we simply bound $d_h^\pi(s_h)/d_h^\mu(s_h) \leq \tau_s$ in (B.6), and then use the following bound

$$\begin{aligned}
&\text{Var}\left[\frac{\pi(a_h^{(1)}|s_h)}{\mu(a_h^{(1)}|s_h)}(V_{h+1}^\pi(s_{h+1}^{(1)}) + r_h^{(1)})\middle| s_h^{(1)} = s_h\right] \\
=&\mathbb{E}\text{Var}\left[\frac{\pi(a_h^{(1)}|s_h)}{\mu(a_h^{(1)}|s_h)}(V_{h+1}^\pi(s_{h+1}^{(1)}) + r_h^{(1)})\middle| s_h^{(1)} = s_h, a_h^{(1)}, s_{h+1}^{(1)}\right] \\
&+ \text{Var}\left[\frac{\pi(a_h^{(1)}|s_h)}{\mu(a_h^{(1)}|s_h)}(V_{h+1}^\pi(s_{h+1}^{(1)}) + r_h(s_h, a_{h+1}^{(1)}, s_{h+1}^{(1)}))\middle| s_h^{(1)} = s_h\right] \\
\leq&\mathbb{E}_\pi\left[\frac{\pi(a_h^{(1)}|s_h)^2}{\mu(a_h^{(1)}|s_h)^2}\middle| s_h^{(1)} = s_h\right]\sigma^2 + \text{Var}_\mu\left[\frac{\pi(a_h^{(1)}|s_h)}{\mu(a_h^{(1)}|s_h)}(V_{h+1}^\pi(s_{h+1}^{(1)}) + r_h(s_h, a_{h+1}^{(1)}, s_{h+1}^{(1)}))\middle| s_h^{(1)} = s_h\right] \\
\leq&\mathbb{E}_\pi\left[\frac{\pi(a_h^{(1)}|s_h)}{\mu(a_h^{(1)}|s_h)}\middle| s_h^{(1)} = s_h\right]\sigma^2 + \mathbb{E}_\pi\left[\frac{\pi(a_h^{(1)}|s_h)}{\mu(a_h^{(1)}|s_h)}(V_{h+1}^\pi(s_{h+1}^{(1)}) + r_h(s_h, a_{h+1}^{(1)}, s_{h+1}^{(1)}))^2\middle| s_h^{(1)} = s_h\right] \\
\leq&\tau_a(\sigma^2 + 2V_{\max}^2 + 2R_{\max}^2).
\end{aligned}$$

The second line uses the law of total expectation, the third line replaces the variance with an upper bound σ^2, the fourth line uses $\text{Var}[X] \leq \mathbb{E}[X^2]$ and a change of measure from μ to π. The last line takes the upper bound τ_a, $R_{\max}$ and $V_{\max}$.

The proof is complete by combining the bounds of the second and the third term. □

Proof of Corollary 1. The results in Corollary 1 requires a slightly different bound of (B.6) then the one we derived above. We use the assumption on n to ensure that

$$\frac{4h\tau_a\tau_s}{n}\frac{d_h^\pi(s_h)}{d_h^\mu(s_h)} \leq \frac{d_h^\pi(s_h)\max\{d_h^\pi(s_h), d_h^\mu(s_h)\}}{d_h^\mu(s_h)} \leq \frac{d_h^\pi(s_h)^2}{d_h^\mu(s_h)} + d_h^\pi(s_h),$$

which gives us an upper bound of proportional to $n^{-1}H(\tau_a\tau_s + \tau_a)(\sigma^2 + H^2\mathbb{R}_{\max}^2)$. □

Remark 2 (Sample complexity in the finite action case)**.** *The result implies a sample complexity upper bound (in terms of the number of episodes) of H^3SA/ϵ^2 for evaluating a fixed target policy by running an exploration policy that visits every state and action pair with probability $\Omega(1/(SA))$.*

The Cramer-Rao lower bound for the discrete DAG-MDP model Jiang and Li [2016, Theorem 3] implies a lower bound of H^2SA/ϵ^2, which suggests that our bound is optimal up to a factor of H even for the cases where A is small. In the settings where A is unbounded. Based on our insight with the contextual bandits setting[Wang et al., 2017], we conjecture that the additional dependence on H in our $H^3\tau_a\tau_s/\epsilon^2$ bound is required.

The comparison with the CR lower bound is a lot more delicate and interesting. We defer more detailed discussion on that to Remark 4.

Remark 3 (When $\pi \approx \mu$)**.** *It is not entirely straightforward to see how Theorem 4.1 gives a H^2/n bound in the case of $\pi \approx \mu$ rather than the H^3/n bound that we describe in Corollary 1. We make it explicit here in this remark. First the variance term in the bound can be expanded using* $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

$$\begin{aligned}
&\sum_{s_h}\frac{d^\pi(s_h)^2}{d^\mu(s_h)}\mathrm{Var}\left[\frac{\pi(a_h^{(1)}|s_h)}{\mu(a_h^{(1)}|s_h)}(V_{h+1}^\pi(s_{h+1}^{(1)}) + r_h^{(1)})\middle| s_h^{(1)} = s_h\right] \\
=&\sum_{s_h}\frac{d^\pi(s_h)^2}{d^\mu(s_h)}\sum_{a_h}\frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)}\Bigg(\mathbb{E}[V_{h+1}^\pi(s_{h+1})^2 + r_h(s_h,a_h,s_h')^2 + \sigma^2(s_h,a_h,s_h')|s_h,a_h] \\
&\qquad + 2\mathbb{E}[V_{h+1}^\pi(s_{h+1})r_h(s_h,a_h,s_h')|s_h,a_h]\Bigg) - \sum_{s_h}\frac{d^\pi(s_h)^2}{d^\mu(s_h)}V_h^\pi(s_h)^2 \\
=&\sum_{s_h,a_h,s_{h+1}}\frac{d^\pi(s_h,a_h,s_{h+1})^2}{d^\mu(s_h,a_h,s_{h+1})}\left(V_{h+1}^\pi(s_{h+1})^2 + [r_h^2 + \sigma_h^2 + 2r_hV_{h+1}^\pi](s_h,a_h,s_{h+1})\right) - \sum_{s_h}\frac{d^\pi(s_h)^2}{d^\mu(s_h)}V_h^\pi(s_h)^2.
\end{aligned}$$

If we substitute the above bound into Theorem 4.1, we can see that the negative part of the bound getting combined with $\sum_{s_{h-1},a_{h-1},s_h}\frac{d^\pi(s_{h-1},a_{h-1},s_h)^2}{d^\mu(s_h,a_h,s_{h+1})}V_h^\pi(s_h)^2$ from the previous time point, which gives the following more interpretable upper bound of the leading term below

$$\begin{aligned}
&\frac{1}{n}\sum_{h=0}^H\Bigg[\sum_{s_{h+1}}\left(\sum_{s_h,a_h}\frac{d^\pi(s_h,a_h,s_{h+1})^2}{d^\mu(s_h,a_h,s_{h+1})} - \frac{d^\pi(s_{h+1})^2}{d^\mu(s_{h+1})}\right)V_{h+1}^\pi(s_{h+1})^2 \\
&+ \sum_{s_h,a_h,s_{h+1}}\frac{d^\pi(s_h,a_h,s_{h+1})^2}{d^\mu(s_h,a_h,s_{h+1})}\left([r_h^2 + \sigma_h^2 + 2r_hV_{h+1}^\pi](s_h,a_h,s_{h+1})\right)\Bigg].
\end{aligned}$$

When $\pi = \mu$, the first term goes away and the above can be bounded by

$$\frac{1}{n}\sum_{h=0}^H\sum_{s_h,a_h,s_{h+1}}d^\pi(s_h,a_h,s_{h+1})(R_{\max}r_h + \sigma^2 + 2V_1^\pi r_h) \leq \frac{1}{n}(R_{\max}V_1^\pi + H\sigma^2 + 2[V_1^\pi]^2) \leq \frac{3V_{\max}^2 + H\sigma^2}{n}.$$

Check that when π and μ are sufficiently close such that $\sum_{s_{h+1}}\left(\sum_{s_h,a_h}\frac{d^\pi(s_h,a_h,s_{h+1})^2}{d^\mu(s_h,a_h,s_{h+1})} - \frac{d^\pi(s_{h+1})^2}{d^\mu(s_{h+1})}\right) = O(1/H)$, then we get the same H^2/n rate as above.

Remark 4 (Comparison to the Cramer-Rao lower bound)**.** *Theorem 3 in [Jiang and Li, 2016, Appendix C.] provides a Cramer-Rao lower bound on the variance of any unbiased estimator for a*

simplified setting of an nonstationary episodic MDP where a reward only appear at the end of the episode and the reward is deterministic (i.e.,$\sigma^2 = 0$). Their bound, in our notation, translates into

$$\lim_{n\to\infty} \mathrm{Var}\big[\sqrt{n}(\widehat{v}^\pi - v^\pi)\big] \geq \sum_{t=0}^{H} \mathbb{E}_\mu \left[\frac{d^\pi(s_t^{(1)})^2}{d^\mu(s_t^{(1)})^2} \frac{\pi(a_t^{(1)}|s_t^{(1)})^2}{\mu(a_t^{(1)}|s_t^{(1)})^2} \mathrm{Var}\Big[V_{t+1}^\pi(s_{t+1}^{(1)})\Big| s_t^{(1)}, a_t^{(1)}\Big]\right].$$

Our Theorem 4.1 implies

$$\lim_{n\to\infty} n\mathbb{E}[(\mathcal{P}\widehat{v}_{\mathrm{MIS}}^\pi - v^\pi)^2] \leq \sum_{t=0}^{H} \mathbb{E}_\mu \left[\frac{d^\pi(s_t^{(1)})^2}{d^\mu(s_t^{(1)})^2} \mathrm{Var}_\mu\Big[\frac{\pi(a_t^{(1)}|s_t^{(1)})}{\mu(a_t^{(1)}|s_t^{(1)})} V_{t+1}^\pi(s_{t+1}^{(1)})\Big| s_t^{(1)}\Big]\right].$$

The upper and lower bounds are clearly very similar, with the only difference in where the importance weights of the actions are. We can verify that the upper bound is bigger because

$$\begin{aligned}
&\mathrm{Var}_\mu\Big[\frac{\pi(a_t^{(1)}|s_t^{(1)})}{\mu(a_t^{(1)}|s_t^{(1)})} V_{t+1}^\pi(s_{t+1}^{(1)})\Big| s_t^{(1)}\Big] \\
=&\mathbb{E}_\mu \left[\mathrm{Var}\Big[\frac{\pi(a_t^{(1)}|s_t^{(1)})}{\mu(a_t^{(1)}|s_t^{(1)})} V_{t+1}^\pi(s_{t+1}^{(1)})\Big| s_t^{(1)}, a_t^{(1)}\Big]\Big| s_t^{(1)}\right] + \mathrm{Var}_\mu \left[\mathbb{E}\Big[\frac{\pi(a_t^{(1)}|s_t^{(1)})}{\mu(a_t^{(1)}|s_t^{(1)})} V_{t+1}^\pi(s_{t+1}^{(1)})\Big| s_t^{(1)}, a_t^{(1)}\Big]\Big| s_t^{(1)}\right] \\
=&\mathbb{E}_\mu \left[\frac{\pi(a_t^{(1)}|s_t^{(1)})^2}{\mu(a_t^{(1)}|s_t^{(1)})^2} \mathrm{Var}\Big[V_{t+1}^\pi(s_{t+1}^{(1)})\Big| s_t^{(1)}, a_t^{(1)}\Big]\Big| s_t^{(1)}\right] + \mathrm{Var}_\mu \left[\frac{\pi(a_t^{(1)}|s_t^{(1)})}{\mu(a_t^{(1)}|s_t^{(1)})} \mathbb{E}\Big[V_{t+1}^\pi(s_{t+1}^{(1)})\Big| s_t^{(1)}, a_t^{(1)}\Big]\Big| s_t^{(1)}\right] \\
=&\mathbb{E}_\mu \left[\frac{\pi(a_t^{(1)}|s_t^{(1)})^2}{\mu(a_t^{(1)}|s_t^{(1)})^2} \mathrm{Var}\Big[V_{t+1}^\pi(s_{t+1}^{(1)})\Big| s_t^{(1)}, a_t^{(1)}\Big]\Big| s_t^{(1)}\right] + \mathrm{Var}_\mu \left[\frac{\pi(a_t^{(1)}|s_t^{(1)})}{\mu(a_t^{(1)}|s_t^{(1)})} Q_t^\pi(s_t^{(1)}, a_t^{(1)})\Big| s_t^{(1)}\right].
\end{aligned}$$

Provided that the second term is comparable to the first, then our upper bound is rate-optimal. Both terms can be bounded by $H^2 R_{\max}^2 \mathbb{E}_\mu[\frac{\pi(a_t^{(1)}|s_t^{(1)})^2}{\mu(a_t^{(1)}|s_t^{(1)})^2}]$ and the bound cannot be improved. However, if we consider the overall bounds that sum over the H items, the summation of the first term (the lower bound) is at most $H^2\tau_a\tau_s R_{\max}^2$ (note that, somewhat surprisingly, no additional factors of H is incurred), while the second term can be as large as $H^3 R_{\max}^2 \mathbb{E}_\mu[\frac{\pi(a_t^{(1)}|s_t^{(1)})^2}{\mu(a_t^{(1)}|s_t^{(1)})^2}]$ in some cases. One trivial example of that would be an MDP that gives a constant immediate reward of $R_{\max}/2$ for all $t = H/2+1, H/2+2, ..., H$. Note that in this case, $Q_t^\pi(s_t^{(1)}, a_t^{(1)}) \equiv (H-t)R_{\max}/2$, which ensures that the second term is lower bounded by

$$\frac{1}{16} H^2 R_{\max}^2 \mathrm{Var}_\mu \left[\frac{\pi(a_t^{(1)}|s_t^{(1)})}{\mu(a_t^{(1)}|s_t^{(1)})}\Big| s_t^{(1)}\right] = \frac{1}{16} H^2 R_{\max}^2 \left(\mathbb{E}_\mu \left[\frac{\pi(a_t^{(1)}|s_t^{(1)})^2}{\mu(a_t^{(1)}|s_t^{(1)})^2}\Big| s_t^{(1)}\right] - 1\right)$$

for all $t, s_t^{(1)}$. As we sum over t, this leads to an H^3 term in our upper bound that does not exist in the Cramer-Rao lower bound.

A curious theoretical question is whether such an additional factor of H in the error bound is required for off-policy evaluation in the small $\mathcal{S}$, large $\mathcal{A}$ setting that we considered.

C Application to Other IS-Based Estimators

In this section, we discuss the applications of our marginalized approach to other IS-based estimators. We first unify some popular IS-based estimators, such as importance sampling and weighted doubly robust estimators, using a generic framework of IS-based estimators. Then we show the corresponding marginalized IS-based estimators, and provide the asymptotic unbiasedness and consistency results. At last, we provide details about how to deal with partial observability when applying our marginalized approach.

C.1 Generic IS-Based Estimators Setup

The IS-based estimators usually provide an unbiased or consistent estimate of the value of target policy π [Thomas, 2015]. We first provide a generic framework of IS-based estimators, and analyze 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 $\rho_{0:t}^{(i)} := \prod_{t'=0}^{t} \frac{\pi(a_{t'}^{(i)}|s_{t'}^{(i)})}{\mu(a_{t'}^{(i)}|s_{t'}^{(i)})}$ 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

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

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:

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

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

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

Self-normalized estimators such as weighted 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:

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

and the WDR estimator:

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

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.

C.2 Marginalized IS-Based Estimators

Recall the marginalized IS estimators (2.2), we obtain a generic framework of marginalized IS-based estimators as:

$$\widehat{v}_M(\pi) = \frac{1}{n}\sum_{i=1}^{n} g(s_0^{(i)}) + \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{H-1} \widehat{w}_t(s_t^{(i)})\rho_t^{(i)}(r_t^{(i)} + f_t(s_t^{(i)}, a_t^{(i)}, s_{t+1}^{(i)})). \tag{C.2}$$

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 $w_t(s)$. Thus, the marginalized IS-based estimators can be obtained by applying different g and f_t in Section C.1 into framework (C.2).

We first show the equivalence between framework (C.1) and framework (C.2) in expectation if $\phi_t(\rho_{0:t}^{(1:n)}) = n$ and $\widehat{w}_t(s) = w_t(s)$.

Lemma C.1. *If $\phi_t(\rho_{0:t}^{(1:n)}) = n$ in framework (C.1) and $\widehat{w}_t(s) = w_t(s)$ in framework (C.2), 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.

Proof of Lemma C.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. □

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 C.1. *Let $\phi_t(\rho_{0:t}^{(1:n)}) = n$ in framework (C.1), then framework (C.2) could keep the unbiasedness and consistency same as in framework (C.1) if $\widehat{w}_t(s)$ is an unbiased or consistent estimator for marginalized ratio $w_t(s)$ for all t:*

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

Proof of Theorem C.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

$$\begin{aligned}\mathbb{E}\left[\widehat{w}_t^n(s_t^{(i)})\rho_t^{(i)}(r_t^{(i)} + f_t(s_t^{(i)}, a_t^{(i)}, s_{t+1}^{(i)}))\right] =&\mathbb{E}\left[\mathbb{E}\left[\widehat{w}_t^n(s_t^{(i)})\rho_t^{(i)}(r_t^{(i)} + f_t(s_t^{(i)}, a_t^{(i)}, s_{t+1}^{(i)}))|s_t^{(i)}\right]\right]\\ =&\mathbb{E}\left[\mathbb{E}\left[\widehat{w}_t^n(s_t^{(i)})|s_t^{(i)}\right]\mathbb{E}\left[\rho_t^{(i)}(r_t^{(i)} + f_t(s_t^{(i)}, a_t^{(i)}, s_{t+1}^{(i)}))|s_t^{(i)}\right]\right]\\ =&\mathbb{E}\left[w_t(s_t^{(i)})\mathbb{E}\left[\rho_t^{(i)}(r_t^{(i)} + f_t(s_t^{(i)}, a_t^{(i)}, s_{t+1}^{(i)}))|s_t^{(i)}\right]\right]\\ =&\mathbb{E}\left[w_t(s_t^{(i)})\rho_t^{(i)}(r_t^{(i)} + f_t(s_t^{(i)}, a_t^{(i)}, s_{t+1}^{(i)}))\right]\\ =&\mathbb{E}\left[\rho_{0:t}^{(i)}(r_t^{(i)} + f_t(s_t^{(i)}, a_t^{(i)}, s_{t+1}^{(i)}))\right],\end{aligned} \tag{C.3}$$

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 C.1. Since the original estimator falls in framework (C.1) is unbiased, summing (C.3) over i and t completes the proof of the first part.

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

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

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

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\widehat{w}_t^n(s_t^{(i)})\rho_t^{(i)}(r_t^{(i)}+f_t(s_t^{(i)},a_t^{(i)},s_{t+1}^{(i)}))=\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\rho_{0:t}^{(i)}(r_t^{(i)}+f_t(s_t^{(i)},a_t^{(i)},s_{t+1}^{(i)})), \tag{C.4}$$

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 (C.4), we have

$$\begin{aligned}
&\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\widehat{w}_t^n(s_t^{(i)})\rho_t^{(i)}(r_t^{(i)}+f_t(s_t^{(i)},a_t^{(i)},s_{t+1}^{(i)}))\\
&=\sum_{s\in\mathcal{S}}d_t^\mu(s)\operatorname*{plim}_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n}\widehat{w}_t^n(s)\frac{\pi(a_t^{(i)}|s)}{\mu(a_t^{(i)}|s)}\mathbf{1}(s_t^{(i)}=s)(r_t^{(i)}+f_t(s,a_t^{(i)},s_{t+1}^{(i)}))\right]\\
&=\sum_{s\in\mathcal{S}}d_t^\mu(s)\operatorname*{plim}_{n\to\infty}\left[\widehat{w}_t^n(s)\frac{1}{n}\sum_{i=1}^{n}\frac{\pi(a_t^{(i)}|s)}{\mu(a_t^{(i)}|s)}\mathbf{1}(s_t^{(i)}=s)(r_t^{(i)}+f_t(s,a_t^{(i)},s_{t+1}^{(i)}))\right]\\
&=\sum_{s\in\mathcal{S}}d_t^\mu(s)\left[\operatorname*{plim}_{n\to\infty}(\widehat{w}_t^n(s))\operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\pi(a_t^{(i)}|s)}{\mu(a_t^{(i)}|s)}\mathbf{1}(s_t^{(i)}=s)(r_t^{(i)}+f_t(s,a_t^{(i)},s_{t+1}^{(i)}))\right)\right]\\
&=\sum_{s\in\mathcal{S}}d_t^\mu(s)w_t(s)\operatorname*{plim}_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\pi(a_t^{(i)}|s)}{\mu(a_t^{(i)}|s)}\mathbf{1}(s_t^{(i)}=s)(r_t^{(i)}+f_t(s,a_t^{(i)},s_{t+1}^{(i)}))\right]\\
&=\sum_{s\in\mathcal{S}}d_t^\mu(s)w_t(s)\mathbb{E}\left[\frac{\pi(a_t|s)}{\mu(a_t|s)}(r_t+f_t(s,a_t,s_{t+1}))\Big|s_t=s\right],
\end{aligned}\tag{C.5}$$

where the first equation follows from the weak law of large number. Similarly, for the right hand side of (C.4), we have

$$\begin{aligned}
&\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\rho_{0:t}^{(i)}(r_t^{(i)}+f_t(s_t^{(i)},a_t^{(i)},s_{t+1}^{(i)}))\\
&=\sum_{s\in\mathcal{S}}d_t^\mu(s)\operatorname*{plim}_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n}\prod_{t'=0}^{t-1}\frac{\pi(a_{t'}^{(i)}|s_{t'}^{(i)})}{\mu(a_{t'}^{(i)}|s_{t'}^{(i)})}\mathbf{1}(s_t^{(i)}=s)\frac{\pi(a_t^{(i)}|s)}{\mu(a_t^{(i)}|s)}(r_t^{(i)}+f_t(s,a_t^{(i)},s_{t+1}^{(i)}))\right]\\
&=\sum_{s\in\mathcal{S}}d_t^\mu(s)\mathbb{E}\left[\prod_{t'=0}^{t-1}\frac{\pi(a_{t'}|s_{t'})}{\mu(a_{t'}|s_{t'})}\frac{\pi(a_t|s)}{\mu(a_t|s)}(r_t+f_t(s,a_t,s_{t+1}))\Big|s_t=s\right]\\
&=\sum_{s\in\mathcal{S}}d_t^\mu(s)\mathbb{E}\left[\prod_{t'=0}^{t-1}\frac{\pi(a_{t'}|s_{t'})}{\mu(a_{t'}|s_{t'})}\Big|s_t=s\right]\mathbb{E}\left[\frac{\pi(a_t|s)}{\mu(a_t|s)}(r_t+f_t(s,a_t,s_{t+1}))\Big|s_t=s\right]\\
&=\sum_{s\in\mathcal{S}}d_t^\mu(s)w_t(s)\mathbb{E}\left[\frac{\pi(a_t|s)}{\mu(a_t|s)}(r_t+f_t(s,a_t,s_{t+1}))\Big|s_t=s\right],
\end{aligned}\tag{C.6}$$

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 (C.5) equal to (C.6). This completes the proof of the second half. □

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. Definition of DAG and Tree MDPs can be found in the extended version of [Jiang and Li, 2016].

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 (C.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 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 an unbiased and consistent estimator for $R(s,a)$, $r_t^{(i)}$ in framework (C.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 an 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

$$\widehat{v}_{BM}(\pi) = \frac{1}{n}\sum_{i=1}^n g(s_0^{(i)}) + \frac{1}{n}\sum_{i=1}^n\sum_{t=0}^{H-1}\widehat{w}_t(s_t^{(i)})\rho_t^{(i)}(\widehat{r}(s_t^{(i)},a_t^{(i)}) + f_t(s_t^{(i)},a_t^{(i)},s_{t+1}^{(i)})). \quad \text{(C.7)}$$

Remark 5. *Note that, the only difference between (C.2) and (C.7) is $r_t^{(i)}$ and $\widehat{r}(s_t^{(i)}, a_t^{(i)})$. Thus, the unbiasedness or consistency of (C.7) can be obtained similarly by following Theorem C.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 (C.7) reduces to (C.2). 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. Thus, we can also obtain better marginalized IS-based estimators, e.g., the MIS and MDR estimators we use in our experiments, by applying different g and f_t in Section C.1 into framework (C.7).

D Details of Experiments

In this section, we first clarify the experiment settings. We also provide a detailed discussion about the preference of MIS and SSD-IS. Finally, we provide the extended experiential results about applying MIS to doubly robust related approaches.

D.1 Environment Settings

ModelWin MDP As depicted in Figure 1(a), the agent in the ModelWin domain 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. We set $p = 0.4$. The agent receives a reward of 1 every time the state transitions to s_2, -1 to s_3, and 0 otherwise.

ModelFail MDP The dynamics of ModelFail MDP (Figure 1(b)) is similar to ModelWin, but the reward is delayed after the unobservable states — the agent receives a reward of 1 only when it arrives s_1 from the left state and -1 only when it arrives s_1 from the right state. We set $p = 1$ to make the problem easier.

Policy π takes action a_1 and a_2 with probabilities 0.2 and 0.8 when at state s_1 or observing "?". μ take actions uniformly at random.

We remark that the partial-state observability in **ModelFail** is specialized and should be distinguished from the more general partial observability considered in the classical POMDP literature. The two distinctive (and clearly artificial) differences are that

1. There are checkpoints of full state observability every other step.
2. We assume that the action probability is logged when the observation is "?".

The model-based approach clearly fails when "?" is treated as if it is a state when building the model. A standard POMDP that uses just a memory of size 2 will resolve this issue without any trouble. One may also consider an alternative MDP that only takes the checkpoint states s_1 as states, but the two actions that the policies will take are no longer a function of just s_1 (in this example it actually is because the observation is always "?" in the step after s_1).

Finally, both **ModelWin** and **ModelFail** are highly specialized examples with deterministic transitions into the states s_1 that could potentially generate rewards for some actions. Moreover, there are no non-trivial actions involved as we transition from s_2 and s_3 back to s_1. This means that we can perfectly estimate the marginal state-distribution of s_1 with just one data point in all methods. As a result, we do not expect the results to reveal the worst-case dependence on the model parameters such as H. The following example fixes that.

Non-stationary Non-mixing MDP In the time-varying MDP example, we consider the following carefully designed MDP where there are two states and a continuous action in $[0, 1]$. In State 0 the agent always transitions to State 0, regardless of the actions. In State 1, it transitions to State 0 deterministically if the action is taken to be within an unknown subset of measure $1/H$ within $[0, 0.5]$. This subset might be different for different t. When the agent is at State 0, then a reward of 1 is received regardless of the actions taken when the step number is larger than $H/2$; otherwise no reward is received.

The behavior and target policy (probability density on $[0, 1]$) are defined to be

$$\mu(a|s = 1) = 1 \text{ for all } a,$$

and

$$\pi(a|s = 1) = \begin{cases} 1.9 & \text{if } a \in [0, 0.5] \\ 0.1 & \text{otherwise.} \end{cases}$$

and $\pi(a|s = 0) = \mu(a|s = 0) = 1$.

This example is deliberately designed such that we have a non-stationary dynamics[6] that does not really mix beyond a constant factor so an additional factor of H in the sample complexity can potentially appear. Meanwhile, the cumulative reward is proportional to H for the target policy, so we expect to see a $H^1.5$ dependence in the (relative) RMSE curves as we vary H. Finally, due to the non-mixing property of this example, and the importance weight of stationary distributions based on SSD-IS is expected to be biased. All the above observations are consistent with what we see in the experiments presented in Figure 3.

Mountain Car Mountain Car domain is a classic control problem with 2-dimensional state space (position and velocity), 3 discrete one-dimensional actions (push left, no push, push right), and deterministic dynamics. We follow the same dynamic as in [Sutton and Barto, 1998]. The horizon, H, is set to be 100. We use initial state distribution to be uniform in position and 0 in velocity to ensure exploratory. Since our proposed method mainly focuses on the tabular setting, we use the state aggregation for both MIS and SSD-IS to ensure fair comparison: position is multiplied by 2^6 and velocity is multiplied by 2^8, and then we use the rounded integers to be the abstract state (adopted from [Jiang and Li, 2016]). Thus, the (marginalized or stationary) state distribution can be estimated on the tabular abstract states.

[6] The transition matrices on both π and μ are actually stationary.

D.2 Detailed Discussions

The ModelWin domain is only a very special case of episodic fully-observable MDPs. Even if we use the stationary state distribution (estimated by ignoring the within-episode step count in the dataset) instead of marginalized state distribution in (3.1), that value will still happen to be correct in both time-invariant and time-varying case. However, that is not correct in general. As the results we showed in the Mountain Car domain, SSD-IS fails to provide correct evaluation. That is because SSD-IS is designed for the infinite-horizon problems and usually cannot be directly applied to the episodic problems, where SSD-IS uses the stationary distribution ($t \to \infty$) to approximate that for all $t = 1, ..., H$ which is biased and not consistent even as the number of episodes $n \to \infty$ in general. For example, in the Mountain Car domain, the stationary state distribution, $\prod_{t=1}^{\infty} P^{\pi}_{s_{t-1},s_t} d_0$, will converge to the probability mass on the absorbing state with any exploratory policy π.

The result of mountain car experiment in the current version is slightly different from the early version. There are two main modifications in that experiments: 1. In the early version, the on-policy estimated v^{π} for calculating RMSE did not use enough trajectories, so that the curves in the early version are biased. 2. We changed the implementation of SSD-IS in our current version. Previously, we solved Equation (8) and (9) in [Liu et al., 2018a] using an iterative approach. The current implementation solves Equation (8) and (9) in [Liu et al., 2018a] directly by re-formalizing it to be a quadratic programming problem. The current implementation follows the released code provided by the author of [Liu et al., 2018a], and the detailed description of that can be found in Appendix D.3.

We also explain the reason of MIS outperforming MDR in Figure 5 and Figure 6. In the MDR methods, we split our dataset into two halves. We use one half to estimate the marginalized state distribution, and the other half to estimate the Q-function. Intuitively, since Q-function is only used to be a control variant in the estimator, we suppose the statistical error from the marginalized state distribution may dominate the overall statistical error. As MIS uses the whole data to estimate the marginalized state, whereas MDR only uses a half data, the statistical error of MIS could be lass that of MDR. The theoretical explanation of that goes beyond the topic of this paper, and we will leave it as the future work.

D.3 SSD-IS with finite state space.

The pioneering work by Liu et al. [2018a] describes a method — SSD-IS — for estimating the ratio of stationary state distribution under π and μ for an infinite horizon (possibly discounted) MDP.

The estimator is described primarily for the case when the state is a continuous variable, which requires defining a reproducing kernel Hilbert space (RKHS) and solving a mini-max problem.

To be a bit more self-contained in this paper, we provide the concise formula using our notation for estimating the stationary distribution $d^{\pi}_{\infty}(s)$ as well as directly estimating the importance ratio

$$\rho(s) := \frac{d^{\pi}_{\infty}(s)}{d^{\mu}_{1:N-1}(s)}$$

between the $d^{\pi}_{\infty}(s)$ and the marginalized state distribution under μ that measures the average state-visitation within the first N iterations (note that we can observe triplets (s_t, a_t, s_{t+1}) for all $t = 1, ..., N-1$).

Note that a roll-out in an infinite-horizon environment of a fixed length N can be denoted in our notation with a single episode $n = 1$ and horizon $H = N$.

The master equation that we need is the following:

$$d^{\pi}_{\infty}(s') = P^{\pi}(s'|s) d^{\pi}_{\infty}(s) = P^{\pi}(s'|s) d^{\mu}_{1:H-1}(s) \frac{d^{\pi}_{\infty}(s)}{d^{\mu}_{1:H-1}(s)},$$

which, in matrix form, is:

$$d^{\pi}_{\infty} = A^{\pi,\mu} \text{Diag}(d^{\mu}_{1:H-1})^{-1} d^{\pi}_{\infty} \tag{D.1}$$

where $A^{\pi,\mu} \in \mathbb{R}^{S \times S}$ and $A^{\pi,\mu}(s', s)$ measures the joint distribution of $s \sim d^{\mu}_t$ with a randomly chosen t from $1, ..., H-1$ and s' that is obtained by taking an action according to π at s.

By left-multiplying $\text{Diag}(d^{\mu}_{1:H-1})^{-1}$ on both sides of the equation, we also get

$$\rho = \text{Diag}(d^{\mu}_{1:H-1})^{-1} A^{\pi,\mu} \rho. \tag{D.2}$$

Observe that (D.1) and (D.2) are eigenvalue problems of d_∞^π and ρ. They differ only by whether we normalize the $A^{\pi,\mu}$ matrix on the left or on the right by multiplying the diagonal matrix $\text{Diag}(d_{1:H-1}^\mu)^{-1}$.

They suggest that if we can consistently estimate $A^{\pi,\mu}$ and $\text{Diag}(d_{1:H-1}^\mu)^{-1}$, then **the right eigenvector of the corresponding estimated matrices with eigenvalue closest to** 1 will be consistent estimators of d_∞^π and ρ.

Note that we can estimate the joint-distribution $A^{\pi,\mu}[s', s]$ by importance sampling using

$$\hat{A}^{\pi,\mu}[s', s] = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{H-1}\sum_{t=1}^{H-1}\frac{\pi(a_t^{(i)}|s_t^{(i)})}{\mu(a_t^{(i)}|s_t^{(i)})}\mathbf{1}(s_t^{(i)} = s, s_{t+1}^{(i)} = s')$$

with potentially infinite action. And $d_{1:H-1}^\mu$ can be estimated by

$$\hat{d}_{1:H-1}^\mu(s) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{H-1}\sum_{t=1}^{H-1}\mathbf{1}(s_t^{(i)} = s).$$

For the infinite horizon case, we can just take $n = 1$.

We emphasize that while the above results and the spectral estimators were not explicitly presented by Liu et al. [2018a], they are simply a rewriting of Equation (8) and (9) in [Liu et al., 2018a] more explicitly in a more specialized case.

The SSD-IS implementation that we used in the experiments with discrete state space corresponds to this particular version that we described in this section, which is the same as the version of the code released by the authors modulo some boundary conditions[7]. These boundary conditions seem to be important for getting SSD-IS to work correctly for the finite horizon MDPs.

That said, we acknowledge that when the underlying MDP is stationary and H is large enough relative to the mixing rate of the MDP, then using the estimated importance weight ρ to construct importance sampling estimators as in SSD-IS may provide a favorable bias-variance trade-off in finite sample, because its variance is smaller by a factor of H than the standard MIS while its bias on the estimated importance ratio $\frac{d_t^\pi(s)}{d_t^\mu(s)}$ decays exponentially as t gets larger.

D.4 Extended Experimental Studies

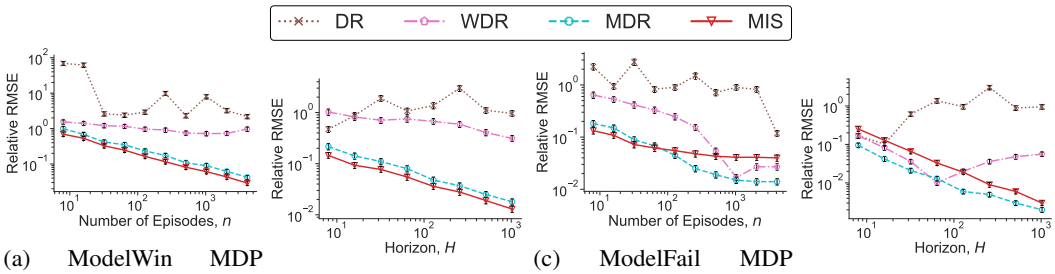
We now present further empirical results. To test the use of our approach in other IS-based estimators, we compared DR, WDR, MDR, and MIS in the same environments, where DR denotes the doubly robust estimator [Jiang and Li, 2016], WDR denotes the weighted doubly robust estimator [Thomas and Brunskill, 2016], MIS denotes the estimator using proposed marginalized approach used with doubly robust, and MIS is our marginalized importance sampling estimator. The estimates of d_t^π and d_t^μ are projected to the probability simplex in our MDR and MIS estimators. The results are obtained in the same environments as Section 5.

The results are in Figure 5 and Figure 6. These demonstrate that other IS based methods can also leverage our marginalized approach to benefit performance dramatically.

E Algorithm Details

Algorithm 1 summarizes our method of marginalized off-policy evaluation. Note that the MIS estimator in Section 5 is using the estimate of $d_t^\pi(\cdot)$ by normalizing (E.1) into the probability simplex for better performance.

[7] In the released code provided by the authors of [Liu et al., 2018a], there is a version of SSD-IS implemented for the discrete state space that first estimates $d_\infty^\pi(s)$ than output the importance weights to be the ratio of this estimate and $\hat{d}_{1:H-1}^\mu$ (see `https://github.com/zt95/infinite-horizon-off-policy-estimation/blob/master/taxi/Density_Ratio_discrete.py`). However, $\hat{d}_\infty^\pi(s)$ is slightly different from the spectral algorithm that we described and it provides a mysterious result that is inconsistent with the stationary distribution that we derived analytically by hands in the example we considered in Figure 3 ($d_\infty^\pi(s = 1) = 1$ with large probability, while the estimated value by running that piece of code is far off).



(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 5: Results on Time-invariant MDPs.

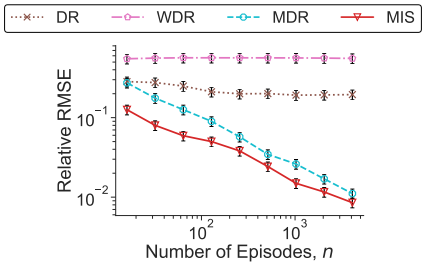


Figure 6: Mountain Car with different number of episodes.

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

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

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: Normalize $d_{t+1}^{\pi}(\cdot)$ into the probability simplex, and specify $\widehat{w}_{t+1}(s)$ as $\dfrac{\widehat{d}_{t+1}^{\pi}(s)}{\widehat{d}_{t+1}^{\mu}(s)}$ for each s.

7: **end for**

8: Substitute the all estimated values above into (C.7) to obtain $\widehat{v}(\pi)$, the estimated cumulative reward of π.