Marginalized Off-Policy Evaluation for Reinforcement Learning

Anonymous Author(s)
Affiliation
Address
email

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 with a long horizon and 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 $O(H^2 R_{\max}^2 \sum_{t=1}^{H} \mathbb{E}_\mu[(w_{\pi,\mu}(s_t, a_t))^2]/n)$ for large n, where $w_{\pi,\mu}(s_t, a_t)$ is the ratio of the marginal distribution of tth step under π and μ, H is the horizon, $R_{\max}$ is the maximal rewards, and n is the sample size. The result nearly matches the Cramer-Rao lower bounds for DAG MDP in Jiang and Li [2016] for most non-trivial regimes. To the best of our knowledge, this is the first OPE estimator with provably optimal dependence in H and the second moments of the importance weight. Besides theoretical optimality, we empirically demonstrate the 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 strongly consistent estimators [Precup et al., 2000]. IS-based off-policy evaluation methods have also seen lots of interest recently especially for short-horizon problems, including contextual bandits [Murphy et al., 2001; Hirano et al., 2003; Dudík et al., 2011; Wang et al., 2017]. However, the variance of IS-based approaches tends to be too high to be useful [Precup et al., 2000; Thomas et al., 2015; Jiang and Li, 2016; Thomas and Brunskill, 2016; Guo et al., 2017; Farajtabar et al., 2018], especially for long-horizon problems [Mandel et al., 2014], since the variance of the product of importance weights may grow exponentially as the horizon goes long. In contrast to the IS-based

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

approaches, solving OPE problems can also use the model-based approaches Liu et al. [2018b]; Gottesman et al. [2019], where the value of target policy is estimated by building whole MDP model.

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, in stead 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, instead of the distribution of the whole trajectories, which provides an approach to breaking the curse of horizon time-invariant MDPs. [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.

In contrast to the prior works, the first goal of our paper is to study the optimality of the marginalized approach. Jiang and Li [2016] studied the hardness of off-policy problems, and presented a Cramer-Rao lower Bound for all the off-policy evaluation methods. In this paper, we provide a finite sample bound on the mean-squared error of our method. We also show that our estimator achieves the optimal rate in sample complexity with respect to the information-theoretical lower-bound proposed by Jiang and Li [2016]. In addition to the theoretical optimality, 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 $r_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the expected reward function with $r_t(s, a)$ defined by the mean of immediate received reward after taking action a in state 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 $d^\mu_t(s_t)$ and $d^\pi_t(s_t)$ the induced state distribution at time t. When $t = 1$, the initial distributions are identical $d^\mu_1 = d^\pi_1 = d_1$. For $t > 1$, $d^\mu_t(s_t)$ and $d^\pi_t(s_t)$ are functions of not just the policies themselves but also the unknown underlying transition dynamics, i.e., for π (and similarly μ), recursively define

$$d^\pi_t(s_t) = \sum_{s_{t-1}} P^\pi_t(s_t|s_{t-1}) d^\pi_{t-1}(s_{t-1}),$$

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

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

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 $P^\pi_{i,j}$ is then a stochastic transition matrix. This allows us to write (2.1) concisely as $d^\pi_{t+1} = P^\pi_{t+1,t} d^\pi_t$.

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, $\mathbb{E}_\pi[r_t] = \mathbb{E}_\pi[r_t^{(1)}] = \sum_{s_t,a_t} d^\pi(s_t)\pi(a_t|s_t) r_t(s_t,a_t)$ and $\mathrm{Var}_\pi[r_t(s_t,a_t)] = \mathrm{Var}_\pi[r_t(s_t^{(1)}, a_t^{(1)})]$. 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

$$v^\pi = \mathbb{E}_\pi\left[\sum_{t=1}^H r_t(s_t,a_t)\right] = \mathbb{E}_\pi\left[\sum_{t=1}^H \sum_{s_t} d_t^\pi(s_t) \sum_{a_t} \pi(a_t|s_t) r_t(s_t,a_t)\right], \tag{2.2}$$

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* $r_t(s_t,a_t)$ for any actions other than the evaluated actions or 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 $0 \le \mathbb{E}[r_t|s_t,a_t] \le R_{\max}(s_t,a_t), \mathrm{Var}[r_t|s_t,a_t] \le \sigma^2(s_t,a_t)$ 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 $r_t(s_t,a_t)$ 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:

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

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 estimator:

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

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

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

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

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

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

The advantage of marginalization 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]. Our 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, even in the partially observable setting, the MIS estimator remains a competitive approximation in cases when H, A are large.

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 a polynomial of H. 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 constant.

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 $\widehat{v}^{\pi}_{IS} = \frac{1}{n}\sum_{i=1}^{n}\left[\prod_{t=1}^{H}\frac{\pi(a_t^{(i)}|s_t^{(i)})}{\mu(a_t^{(i)}|s_t^{(i)})}\right]$. 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 $V_{\log} = \mathrm{Var}[\log\frac{\pi_t}{\mu_t}]$. By Central Limit Theorem, $\sum_{t=1}^{H}\log\frac{\pi_t}{\mu_t}$ asymptotically follows a normal distribution with parameters $\big(-HE_{\log}, HV_{\log}\big)$. In other words, $\prod_{t=1}^{H}\frac{\pi_t}{\mu_t}$ asymptotically follows $LogNormal\big(-HE_{\log}, HV_{\log}\big)$, whose variance is exponential in horizon: $\big(\exp\left(HV_{\log}\right)-1\big)$. 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:*

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

For the simplicity of the statement, define boundary conditions: $r_0(s_0) \equiv \sigma_0(s_0, a_0) \equiv 0$, $\frac{d_0^{\pi}(s_0)}{d_0^{\mu}(s_0)} \equiv \frac{\pi(a_0|s_0)}{\mu(a_0|s_0)} \equiv 1$ 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

$$n \geq \max\left\{\frac{4t\tau_a\tau_s}{\min_{t,s_t}\max\{d_t^{\pi}(s_t), d_t^{\mu}(s_t)\}}, \frac{4\log_{e/2}(nS)}{\min_{t,s_t} d_t^{\mu}(s_t)}\right\}$$

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

$$\begin{aligned}
&\mathbb{E}[(\mathcal{P}\widehat{v}^{\pi}_{MIS} - v^{\pi})^2] \\
\leq& \frac{4}{n}\sum_{h=0}^{H}\sum_{s_h}\frac{d_h^{\pi}(s_h)^2}{d_h^{\mu}(s_h)}\sum_{a_h}\frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)}(\|V_{h+1}^{\pi}\|^2_{T_{h+1}(\cdot|s_h,a_h)} + \sigma_h^2(s_h,a_h) + r_h(s_h,a_h)^2) \\
&+ \frac{19\tau_a^2\tau_s^2 SH^2(\sigma^2 + R_{\max}^2 + V_{\max}^2)}{n^2}.
\end{aligned}$$

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

$$\mathbb{E}_{\mu}[(\mathcal{P}\widehat{v}^{\pi}_{MIS} - v^{\pi})^2] \leq \frac{8}{n}\tau_a\tau_s(H\sigma^2 + H^3R_{\max}^2).$$

The proof in the appendix involves a little more than direct application of the above theorem.

We make a few remarks about the results in Theorem 4.1. First, the leading term does not depend on τ_a, τ_s and it has explicit dependence on the moments of the importance weights. Second, the leading term nearly matches the Cramer-Rao lower bound of the Theorem 3 in [Jiang and Li, 2016] for *every instance* up to a constant factor when the importance weights have a second moment substantially greater than 1 [1]. Finally, although our results do not directly imply an off-line learning methods, a high-probability extension of our results (which can be obtained via Bernstein-McDiarmid inequality) will allow us to achieve an entirely off-policy learning bound in the Tabular MDPs setting with sample complexity (number of episodes) $O(H^3SA/\epsilon^2)$, or a regret lower bound of $\sqrt{H^3SAn}$. This matches the corresponding lower bounds in Dann and Brunskill [2015]; Azar et al. [2017]; Jin et al. [2018]. Formalizing these optimality statements are left to a longer version of the work.

4.1 Proof Sketch

We overview the insight of the proof of Theorem 4.1 in this section. Our key insight is to break the curse of horizon via error propagation calculation, which can be thought of as the off-policy version

[1] We acknowledge that the bound is not tight for the case when $\pi = \mu$ (naive averages give H^2/n). This is an artifact of the proof that is easily fixable, although the current statement is cleaner.

of the celebrated Bellman equation for Variance. We show a linear decomposition of the total variance via a peeling argument, using the filtration of events to recursively separate the expectation of the variance in every step (Lemma 4.1). Additionally, the single-step variance is inversely proportional to the empirical state visitation count n_{s_t}, which converges to $nd_t^\mu(s_t) \asymp O(n), \forall t, s_t$ exponentially fast (Lemma B.1). Compared with naive IS which ignores the state distribution, our MIS estimates the state distribution with variance that is linear in horizon H (Theorem B.1). This results in the final MSE bound (Theorem 4.1), considering the maximal value function is of order $O(HR_{\max})$.

One challenge we encounter is that $\hat{v}_{MIS}$ is not an unbiased estimator, due to non-zero probability of observing $n_{s_t} = 0$ for some s_t. We address this by defining a fictitious estimator $\tilde{v}$ that behaves perfectly when $n_{s_t} < \mathbb{E}_\mu n_{s_t}/2$, which makes it unbiased. We establish that the fictitious estimator is extremely similar to the $\hat{v}_{MIS}$ hence reduces the problem to a slightly simpler problem.

For variance decomposition, we compare with Bellman equation $V_t^\pi(s_t) = r_t^\pi(s_t) + \sum_{s_{t+1}} P_t^\pi(s_{t+1}|s_t) V_{t+1}^\pi(s_{t+1})$, where $V_t^\pi(s_t)$ denotes the value function under π, and use a peeling argument

$$\begin{aligned}\mathrm{Var}[\widetilde{v}^\pi] &= \mathrm{Var}\left[\langle \widetilde{d}_{h+1}^\pi, V_{h+1}^\pi\rangle + \sum_{t=1}^{h} \langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle\right] \\ &= \mathbb{E}\left[\mathrm{Var}\left[\langle \widetilde{d}_{h+1}^\pi, V_{h+1}^\pi\rangle + \langle \widetilde{d}_h^\pi, \widetilde{r}_h^\pi\rangle \middle| \mathrm{Data}_h\right]\right] + \mathrm{Var}\left[\langle \widetilde{d}_h^\pi, V_h^\pi\rangle + \sum_{t=1}^{h-1} \langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle\right],\end{aligned} \tag{4.1}$$

where the second part is the variance of the expectation, which reduces to the true value function due to the unbiasedness of the fictitious estimator. Further calculation yields Lemma 4.1.

Lemma 4.1 (Variance decomposition)**.**

$$\mathrm{Var}[\widetilde{v}^\pi] \le \frac{\|V_1^\pi\|_{d_1(\cdot)}^2}{n} + 2\sum_{h=1}^{H}\sum_{s_h} \mathbb{E}\left[\frac{\widetilde{d}_h^\pi(s_h)^2}{n_{s_h}} \mathbf{1}\left(n_{s_h} \ge \frac{nd_h^\mu(s_h)}{2}\right)\right] \sum_{a_h}\left[\frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)} \left(\|V_{h+1}^\pi\|_{T_{h+1}(\cdot|s_h,a_h)}^2 + \sigma^2(s_h,a_h) + r_h(s_h,a_h)^2\right)\right],$$

where we used $\|x\|_w^2 := \sum_i w[i]x[i]^2$ to denote squared weighted Euclidean norm.

Finally, we bound the error term in the state distribution estimation $\mathbb{E}\left[\frac{\widetilde{d}_h^\pi(s_h)^2}{n_{s_h}} \mathbf{1}\left(n_{s_h} \ge \frac{nd_h^\mu(s_h)}{2}\right)\right] \le \frac{1}{2nd_h^\mu(s_h)}\left(d_h^\pi(s_h)^2 + \mathrm{Var}\left[\widetilde{d}_h^\pi(s_h)\right]\right)$. For the variance term, we again apply a peeling argument $\mathrm{Cov}[\widetilde{d}_h^\pi] = \mathbb{E}\left[\mathrm{Cov}\left[\widetilde{P}_h^\pi \widetilde{d}_{h-1}^\pi \middle| \mathrm{Data}_{h-1}\right]\right] + P_h^\pi \mathrm{Cov}\left[\widetilde{d}_{h-1}^\pi\right][P_h^\pi]^\top$, where we overload P_h^π as the state-transition matrix under policy π at step h. Recursion on the second term adds a multiplicative $P_{h,t}^\pi = \prod_{\tau=t}^{h} P_\tau^\pi$, which is still a proper transition matrix and therefore contracts the (co-)variance. Additional calculation on the expectation of covariance completes the proof of Theorem 4.1.

Appendix B shows the complete details of the proofs. While the main story is that marginalized state distribution estimation breaks curse of horizon, detailed variance decomposition recovers correct rates with respect to information-theoretic lower-bounds. Besides avoiding dependency on the action space (ergodicity only requires sufficient visitation to all states), our IS-based approach also has additional benefits to handle, e.g., partially observable states, shown in our experiments.

5 Experiments

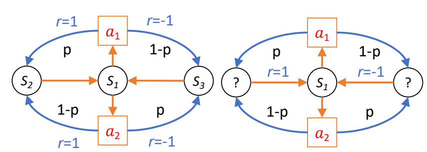
We use this section to empirically showcase the benefits of MIS on key properties including sample complexity with respect to MDP horizons, adaptivity to partially observable states — an additional empirical property inherited from IS-approaches, time-varying state transition models, and the combination of them. We first borrow the synthetic ModelWin and ModelFail MDPs from [Thomas and Brunskill, 2016] to verify the horizon-dependency and adaptivity to partially observable states. We then modify the MDPs to time-varying domains, where our episodic approach is more appropriate than other related infinite-horizon solutions. We lastly show Mountain Car experiments, which have primarily long-horizon problems but also all of the issues combined.

The methods we compare in this section are DM, IS, WIS, SSD-IS, and MIS. DM denotes 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 denotes the importance sampling method based on the whole trajectories, WIS denotes the weighted (self-normalized) importance sampling method, SSD-IS denotes the method of importance sampling with stationary state distribution proposed by [Liu et al., 2018a], and MIS is our proposed marginalized approach. Note that our MIS also uses the trick of self-normalization to obtain better performance, but the MIS normalization is different: we project 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 the results include confidence intervals from 128 runs. Our metric is the relative root of mean squared error (Relative-RMSE) with error bars, which is the ratio of RMSE and true cumulative reward, typically on the order of $O(H)$.

5.1 Time-invariant MDPs

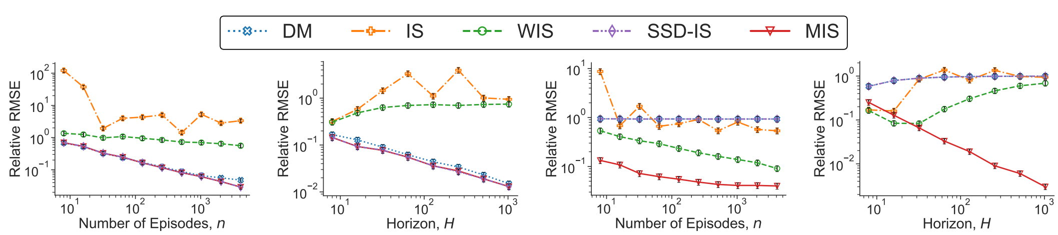
We test our methods on the standard ModelWin and ModelFail models with time-invariant MDPs, first introduced by Thomas and Brunskill [2016]. The **ModelWin** domain simulates a fully observable MDP, depicted in Figure 1(a). The agent always begins in s_1, where it must select between two actions. The first action a_1 causes the agent to transition to s_2 with probability p and s_3 with probability $1-p$. The second action a_2 does the opposite. 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. 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. The dynamics of ModelFail MDP 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. 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 $n = 1024$. Note that the rewards in ModelFail do not depend on the current states and actions, but those of the previous steps; we use MIS only with observable states and the partial trajectories between them. 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).



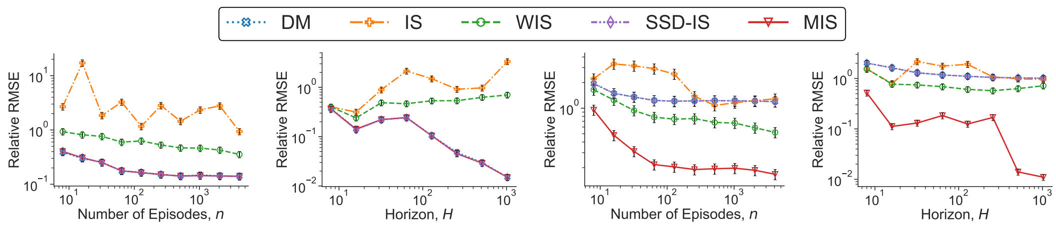
(a) ModelWin with different number of episodes n. (b) ModelWin with different horizon H. (c) ModelFail with different number of episodes n. (d) ModelFail with different horizon H.

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.

5.2 Time-varying MDPs

We also test our approach in the time-varying MDPs. The time-varying MDPs we use in this section are also modified on the standard domains introduced by Thomas and Brunskill [2016]. We use the similar dynamic of ModelWin MDP and ModelFail MDP, but we set the transition probability p_t to be varying over time t for both MDPs, where p_t is sampled from a uniform distribution $\mathcal{U}(0.2, 0.5)$ for each t.



(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 3: Results on time-varying MDPs. Besides amplifying the time-invariant results, MIS outperforms SSD-ID, which is the best existing method with infinite-horizon MDPs.

Figure 3 shows the relative RMSE in the time-varying ModelWin MDP and ModelFail MDP. We observe the results of Figure 3 are similar to the time-invariant case, which demonstrate the effectiveness of our approach in the time-varying domains. Particularly, we show that MIS outperforms SSD-ID, which is the best existing method with infinite-horizon MDPs. SSD-ID is inferior because the stationary state distribution it finds does not agree with the true time-varying state distributions and SSD-ID cannot aggregate only on the partially observed states as MIS.

5.3 Mountain Car

To demonstrate the scalability of the proposed marginalized approaches, we also test all estimators in the Mountain Car domain [Singh and Sutton, 1996], where an under-powered car drives up a steep valley by "swinging" on both sides to gradually build up potential energy. We use a horizon of $H = 100$, a uniform initial state distribution, and the same state aggregations as Jiang and Li [2016]. To construct the stochastic behavior policy μ and stochastic evaluated policy π, we first compute the optimal Q-function using Q-learning and use its softmax policy of the optimal Q-function as evaluated policy π (with the temperature of 1). For the behavior policy μ, we also use the softmax policy of the optimal Q-function but set the temperature to 1.33.

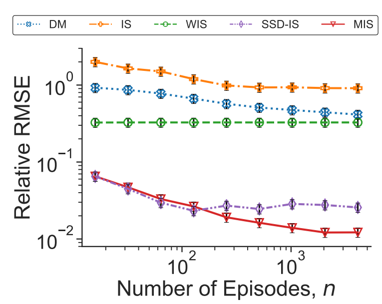


Figure 4: Mountain Car with different number of episodes.

The results on the Mountain Car domain is in Figure 4, which demonstrate the effectiveness of our approach in the common benchmark control task, where the ability to evaluate under long horizons is required for success.

6 Conclusions

In this paper, we propose a marginalized approach to solve the problem of off-policy evaluation in reinforcement learning. Our approach gets rid of the burden of horizon by using the the target state distribution at every step instead of the cumulative product of importance weights. Further more, we provide the theoretical analysis of our estimator and it shows that our approach matches the information-theoretical optimal rate of the OPE problem. Our experiments demonstrate the effectiveness of our approach. It achieves substantially better performance than existing approaches.

References

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

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.

Chao, M.-T. and Strawderman, W. (1972). Negative moments of positive random variables. *Journal of the American Statistical Association*, 67(338):429–431.

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.

Dann, C. and Brunskill, E. (2015). Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826.

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

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

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. (2018). More robust doubly robust off-policy evaluation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1447–1456, Stockholmsmässan, Stockholm Sweden. PMLR.

Gelada, C. and Bellemare, M. G. (2019). Off-policy deep reinforcement learning by bootstrapping the covariate shift. *arXiv preprint arXiv:1901.09455*.

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

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

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

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

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

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

Li, L., Munos, R., and Szepesvari, C. (2015). Toward minimax off-policy value estimation. In *Artificial Intelligence and Statistics*, 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*, pages 5361–5371.

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

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

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

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

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

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

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

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

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

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

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

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

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

Thomas, P. S., Theocharous, G., Ghavamzadeh, M., Durugkar, I., and Brunskill, E. (2017). Predictive off-policy policy evaluation for nonstationary decision problems, with applications to digital marketing. In *AAAI*, pages 4740–4745.

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

Appendix

A Concentration inequalities and other technical lemmas

Lemma A.1 ([Chao and Strawderman, 1972])**.** *Let X be a Binomial random variable with parameter p, n, we have that* $\mathbb{E}[1/(X+1)] = \frac{2}{p(n+1)}(1-(1-p)^{n+1})$*,*

Lemma A.2 (Negative moment of Binomial R.V.)**.** *Let X be a Binomial r.v. with parameter p, n.*

$$\mathbb{E}[\frac{1}{X}\mathbf{1}_{\{X>0\}}] \leq \frac{2}{pn}.$$

Proof. By Lemma A.1 due to we have that $\mathbb{E}[1/(X+1)] = \frac{2}{p(n+1)}(1-(1-p)^{n+1})$, which implies that

$$\begin{aligned}\mathbb{E}[\frac{1}{X}\mathbf{1}_{\{X>0\}}] &\leq \mathbb{E}[\frac{2}{1+X}\mathbf{1}_{\{X>0\}}] = 2\mathbb{E}[\frac{1}{1+X}] - 2(1-p)^n \\ &= \frac{2}{p(n+1)}(1-(1-p)^{n+1}) - 2(1-p)^n \leq \frac{2}{pn}.\end{aligned}$$

□

Lemma A.3 (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

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

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,

$$\widetilde{r}_t^\pi(s_t) = \begin{cases} \widehat{r}_t^\pi(s_t) & \text{if } n_{s_t} \geq nd_t^\mu(s_t)/2 \\ r_t^\pi(s_t) & \text{otherwise;} \end{cases}$$

and

$$\widetilde{\mathbb{P}}_{t,t-1}^\pi(\cdot|s_{t-1}) = \begin{cases} \widehat{\mathbb{P}}_{t,t-1}^\pi & \text{if } n_{s_{t-1}} \geq nd_t^\mu(s_{t-1})/2 \\ \mathbb{P}_{t,t-1}^\pi & \text{otherwise.} \end{cases}$$

This estimator $\widetilde{v}^\pi$ is fictitious because it is *not implementable* using the data[2], 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 IS 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. The MSE of*

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

Proof of Lemma B.1. Let E denotes the event of $\{\exists t, s_t, \text{ s.t. } n_{s_t} < nd_t^\mu(s_t)/2\}$. 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 prove that $\mathbb{P}[E] \leq SHe^{-n\min_{s,t} d_t^\mu(s_t)}$. By the multiplicative Chernoff bound (Lemma A.3 in the Appendix) with $\delta = 0.5$, we get that

$$\mathbb{P}\left[n_{s_t} < \frac{nd_t^\mu(s_t)}{2}\right] \leq \left(\frac{2}{e}\right)^{\frac{nd_t^\mu(s_t)}{2}} \leq \left(\frac{2}{e}\right)^{\frac{n\min_{t,s_t} d_t^\mu(s_t)}{2}}.$$

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

$$\mathbb{P}[E] \leq \sum_t\sum_{s_t}\mathbb{P}[n_{s_t,t} < \frac{nd_t^\mu(s_t)}{2}] \leq HS\left(\frac{2}{e}\right)^{\frac{n\min_{t,s_t} d_t^\mu(s_t)}{2}}.$$

as stated. ☐

[2] 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.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 $\mathbb{E}[(\mathcal{P}\widehat{v}^\pi - v^\pi)^2] \leq \mathbb{E}[(\widehat{p}^\pi - v^\pi)^2]$ so it ionly 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 $\mathbb{R}^S$.

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

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 $\text{Data}_1 := \left\{ s_{1:t}^{(i)} \right\}_{i=1}^{n}$ 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 $n_{s_t} > 0$, the estimators are just empirical mean, which are unbiased and when $n_{s_t} = 0$, we also have an unbiased estimator by the construction of the fictitious estimator. 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 4.1 (Variance decomposition)**.**

$$\begin{aligned}\mathrm{Var}[\widetilde{v}^\pi]\le&\frac{\|V_1^\pi\|_{d_1(\cdot)}^2}{n}+2\sum_{h=1}^{H}\sum_{s_h}\mathbb{E}\left[\frac{\widetilde{d}_h^\pi(s_h)^2}{n_{s_h}}\mathbf{1}_{\{n_{s_h}\ge\frac{nd_h^\mu(s_h)}{2}\}}\right]\sum_{a_h}\frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)}\big(\|V_{h+1}^\pi\|_{P_{h+1,h}(\cdot|s_h,a_h)}^2\\ &+\sigma^2(s_h,a_h)+r_h(s_h,a_h)^2\big).\end{aligned}$$

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

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

and we used $\|x\|_w^2:=\sum_i w[i]x[i]^2$ to denote squared weighted Euclidean norm.

Remark 1. *The decomposition is very interpretable. The first part of the variance is coming from estimating the initial state. The second part ($\|V_{h+1}\|_{P_{h+1,h}(\cdot|s_h,a_h)}^2$) is coming from the conditional variance of estimating $P_{t,t-1}^\pi$ using importance sampling over a_t given all observations up to $t-1$. The third part ($\sigma^2(s_h,a_h)+r_h(s_h,a_h)^2$) is coming from the conditional variance of estimating r_t^π using importance sampling over a_t given all observations up to time t.*

Proof of Lemma 4.1. 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 $h=1,...,H-1$.

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

Note that in (4.1), when we condition on Data_h, $\widetilde{d}_h^\pi$ is fixed, but $\widetilde{d}_{h+1}^\pi$ and $\widetilde{r}_h^\pi$ are not independent given Data_h. By the inequality that $\mathrm{Var}[X+Y]\le 2\mathrm{Var}[X]+2\mathrm{Var}[Y]$, we have that

$$\begin{aligned}&\mathbb{E}\left[\mathrm{Var}\left[\langle\widetilde{d}_{h+1}^\pi,V_{h+1}^\pi\rangle+\langle\widetilde{d}_h^\pi,\widetilde{r}_h^\pi\rangle\middle|\mathrm{Data}_h\right]\right]\\ \le&2\mathbb{E}\left[\mathrm{Var}\left[\langle\widetilde{d}_{h+1}^\pi,V_{h+1}^\pi\rangle\middle|\mathrm{Data}_h\right]+\mathrm{Var}\left[\langle\widetilde{d}_h^\pi,\widetilde{r}_h^\pi\rangle\middle|\mathrm{Data}_h\right]\right]\\ =&2\underbrace{[V_{h+1}^\pi]^T\mathbb{E}\left[\mathrm{Cov}\left[\widetilde{d}_{h+1}^\pi\middle|\mathrm{Data}_h\right]\right]V_{h+1}^\pi}_{(*)}+2\underbrace{\mathbb{E}\left[[\widetilde{d}_h^\pi]^T\mathrm{Cov}\left[\widetilde{r}_h^\pi|\mathrm{Data}_h\right]\widetilde{d}_h^\pi\right]}_{(**)}\end{aligned}\tag{B.1}$$

We now work out upper bounds of $(*)$ and $(**)$.

Bounding $(*)$. $(*)$ can be written as a different form:

$$\begin{aligned}
(*) =&\mathbb{E}\left[[V_{h+1}^{\pi}]^T\mathbb{E}\left[\widetilde{d}_{h+1}^{\pi}[\widetilde{d}_{h+1}^{\pi}]^T\middle|\text{Data}_h\right]V_{h+1}^{\pi}-\mathbb{E}\left[\langle V_{h+1}^{\pi},\widetilde{d}_{h+1}^{\pi}\rangle\middle|\text{Data}_h\right]^2\right]\\
=&\mathbb{E}\left[\text{Var}[\langle V_{h+1}^{\pi},\widetilde{d}_{h+1}^{\pi}\rangle|\text{Data}_h]\right]=\mathbb{E}\left[\text{Var}[\langle V_{h+1}^{\pi},\widetilde{P}_{h+1,h}^{\pi}\widetilde{d}_{h}^{\pi}\rangle|\text{Data}_h]\right]\\
=&\mathbb{E}\left[\sum_{s_h}\frac{\widetilde{d}_h^{\pi}(s_h)^2}{n_{s_h}}\text{Var}\left[\frac{\pi(a_h^{(1)}|s_h)}{\mu(a_h^{(1)}|s_h)}V_{h+1}^{\pi}(s_{h+1}^{(1)})\middle|s_h^{(1)}=s_h\right]\mathbf{1}_{\{n_{s_h}\geq\frac{nd_h^{\mu}(s_h)}{2}\}}\right]\\
\leq&\mathbb{E}\left[\sum_{s_h}\frac{\widetilde{d}_h^{\pi}(s_h)^2}{n_{s_h}}\sum_{a_h}\frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)}\sum_{s_{h+1}}P_{h+1,h}(s_{h+1}|s_h,a_h)V_{h+1}^{\pi}(s_{h+1})^2\mathbf{1}_{\{n_{s_h}\geq\frac{nd_h^{\mu}(s_h)}{2}\}}\right] \quad \text{(B.2)}\\
=&\sum_{s_h}\mathbb{E}\left[\frac{\widetilde{d}_h^{\pi}(s_h)^2}{n_{s_h}}\mathbf{1}_{\{n_{s_h}\geq\frac{nd_h^{\mu}(s_h)}{2}\}}\right]\sum_{a_h}\frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)}\sum_{s_{h+1}}P_{h+1,h}(s_{h+1}|s_h,a_h)V_{h+1}^{\pi}(s_{h+1})^2\\
=&\sum_{s_h}\mathbb{E}\left[\frac{\widetilde{d}_h^{\pi}(s_h)^2}{n_{s_h}}\mathbf{1}_{\{n_{s_h}\geq\frac{nd_h^{\mu}(s_h)}{2}\}}\right]\sum_{a_h}\frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)}\|V_{h+1}^{\pi}\|^2_{P_{h+1,h}(\cdot|s_h,a_h)} \quad \text{(B.3)}
\end{aligned}$$

The third line is true because the columns of $\widetilde{P}_{h+1,h}$ are independent given Data_h, and that when $n_{s_h}<nd_h^{\mu}(s_h)$, the condition variance is 0. The inequality in the fourth row uses that $\text{Var}[X]\leq\mathbb{E}[X^2]$, and we used the shorthand:

$$\|V_{h+1}^{\pi}\|^2_{P_{h+1,h}(\cdot|s_h,a_h)}:=\sum_{s_{h+1}}P_{h+1,h}(s_{h+1}|s_h,a_h)V_{h+1}^{\pi}(s_{h+1})^2.$$

Bounding $(**)$. Using the same argument of the independence of different roll-outs, we observe that $\text{Cov}\left[\widetilde{r}_h^{\pi}|\text{Data}_h\right]$ is diagonal (since they are constructed from disjoint sequences of data). It follows that:

$$\begin{aligned}
(**)&=\mathbb{E}\left[\sum_{s_h}\widetilde{d}_h^{\pi}(s_h)^2\text{Var}\left[\widetilde{r}_h^{\pi}|\text{Data}_h\right]\right]\\
&=\mathbb{E}\left[\sum_{s_h}\frac{\widetilde{d}_h^{\pi}(s_h)^2}{n_{s_h}}\sum_{a_h}\frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)}(\text{Var}[r_h|a_h,s_h]+\mathbb{E}[r_h|a_h,s_h]^2)\mathbf{1}_{\{n_{s_h}\geq\frac{nd_h^{\mu}(s_h)}{2}\}}\right]\\
&\leq\sum_{s_h}\mathbb{E}\left[\frac{\widetilde{d}_h^{\pi}(s_h)^2}{n_{s_h}}\mathbf{1}_{\{n_{s_h}\geq\frac{nd_h^{\mu}(s_h)}{2}\}}\right]\sum_{a_h}\frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)}(\sigma^2(s_h,a_h)+r_h(s_h,a_h)^2). \quad \text{(B.4)}
\end{aligned}$$

The last line is true because when $n_{s_h}<\frac{nd_h^{\mu}(s_h)}{2}$, $\widetilde{r}_h^{\pi}(n_{s_h})=r_h^{\pi}(n_{s_h})$ is a constant, therefore, $\text{Var}[\widetilde{r}_h^{\pi}(n_{s_h})|n_{s_h}<\frac{nd_h^{\mu}(s_h)}{2}]=0$.

Apply (4.1) 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}^{\pi}_H, \widetilde{r}^{\pi}_H\rangle|\mathrm{Data}_H]\right] + \mathrm{Var}[\mathbb{E}[\langle \widetilde{d}^{\pi}_H, \widetilde{r}^{\pi}_H\rangle|\mathrm{Data}_H] + \sum_{t=1}^{H-1}\langle \widetilde{d}^{\pi}_t, \widetilde{r}^{\pi}_t\rangle] \\
=& \mathbb{E}\left[\mathrm{Var}[\langle \widetilde{d}^{\pi}_H, \widetilde{r}^{\pi}_H\rangle|\mathrm{Data}_H]\right] + \mathrm{Var}[\langle \widetilde{d}^{\pi}_H, r^{\pi}_H\rangle + \sum_{t=1}^{H-1}\langle \widetilde{d}^{\pi}_t, \widetilde{r}^{\pi}_t\rangle] \\
=& \mathbb{E}\left[\mathrm{Var}[\langle \widetilde{d}^{\pi}_H, \widetilde{r}^{\pi}_H\rangle|\mathrm{Data}_H]\right] + \mathrm{Var}[\langle \widetilde{d}^{\pi}_H, V^{\pi}_H\rangle + \sum_{t=1}^{H-1}\langle \widetilde{d}^{\pi}_t, \widetilde{r}^{\pi}_t\rangle] \\
=& \mathbb{E}\left[\mathrm{Var}[\langle \widetilde{d}^{\pi}_H, \widetilde{r}^{\pi}_H\rangle|\mathrm{Data}_H]\right] + \mathbb{E}\left[\mathrm{Var}\left[\langle \widetilde{d}^{\pi}_H, V^{\pi}_H\rangle + \langle \widetilde{d}^{\pi}_{H-1}, \widetilde{r}^{\pi}_{H-1}\rangle\middle|\mathrm{Data}_{H-1}\right]\right] \\
&+ \mathrm{Var}\left[\langle \widetilde{d}^{\pi}_{H-1}, V^{\pi}_{H-1}\rangle + \sum_{t=1}^{H-2}\langle \widetilde{d}^{\pi}_t, \widetilde{r}^{\pi}_t\rangle\right] \\
=& \mathbb{E}\left[\mathrm{Var}[\langle \widetilde{d}^{\pi}_H, \widetilde{r}^{\pi}_H\rangle|\mathrm{Data}_H]\right] + \sum_{h=H-1}^{H}\mathbb{E}\left[\mathrm{Var}\left[\langle \widetilde{d}^{\pi}_h, V^{\pi}_h\rangle + \langle \widetilde{d}^{\pi}_{h-1}, \widetilde{r}^{\pi}_{h-1}\rangle\middle|\mathrm{Data}_{h-1}\right]\right] \\
&+ \mathrm{Var}\left[\langle \widetilde{d}^{\pi}_{H-2}, V^{\pi}_{H-2}\rangle + \sum_{t=1}^{H-3}\langle \widetilde{d}^{\pi}_t, \widetilde{r}^{\pi}_t\rangle\right] \\
=& \mathbb{E}\left[\mathrm{Var}[\langle \widetilde{d}^{\pi}_H, \widetilde{r}^{\pi}_H\rangle|\mathrm{Data}_H]\right] + \sum_{h=2}^{H}\mathbb{E}\left[\mathrm{Var}\left[\langle \widetilde{d}^{\pi}_h, V^{\pi}_h\rangle + \langle \widetilde{d}^{\pi}_{h-1}, \widetilde{r}^{\pi}_{h-1}\rangle\middle|\mathrm{Data}_{h-1}\right]\right] + \mathrm{Var}\left[\langle \widetilde{d}^{\pi}_1, V^{\pi}_1\rangle\right]
\end{aligned}$$

Finally, apply (B.1) with the two bounds (B.3) and (B.4), we get

$$\begin{aligned}
\mathrm{Var}[\widetilde{v}^{\pi}] \le \frac{\|V^{\pi}_1\|^2_{d_1(\cdot)}}{n} + 2\sum_{h=1}^{H}\sum_{s_h}\mathbb{E}\left[\frac{\widetilde{d}^{\pi}_h(s_h)^2}{n_{s_h}}\mathbf{1}_{\{n_{s_h}\ge\frac{nd^{\mu}_h(s_h)}{2}\}}\right]\sum_{a_h}\frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)}(\|V^{\pi}_{h+1}\|^2_{P_{h+1,h}(\cdot|s_h,a_h)} \\
+\sigma^2(s_h,a_h) + r_h(s_h,a_h)^2).
\end{aligned}$$

where $\|V^{\pi}_1\|^2_{d_1(\cdot)} = \sum_{s_1} d_1(s_1)V^{\pi}_1(s_1)^2$. This completes the proof. □

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

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

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

$$\mathbb{E}\left[\frac{\widetilde{d}^{\pi}_h(s_h)^2}{n_{s_h}}\mathbf{1}_{\{n_{s_h}\ge\frac{nd^{\mu}_h(s_h)}{2}\}}\right] \le \frac{1}{2nd^{\mu}_h(s_h)}\mathbb{E}\left[\widetilde{d}^{\pi}_h(s_h)^2\right] = \frac{1}{2nd^{\mu}_h(s_h)}(d^{\pi}_h(s_h)^2 + \mathrm{Var}[\widetilde{d}^{\pi}_h(s_h)]). \tag{B.5}$$

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

Lemma B.3 (Unbiasedness of $\widetilde{d}^{\pi}_h$)**.** *For all $h = 1, ..., H$, the fictitious state marginal estimators are unbiased, that is,*

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

Proof of Lemma B.3. Recall the recursive relationship by construction

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

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.4 (Covariance of $\widetilde{d}_h^\pi$)**.**

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

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.5).

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:

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

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

Note that we have $\mathrm{Var}[\widetilde{d}_h^\pi(s_{h-1})]$ 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 $\tau_s := \max_{t,s_t}\frac{d_t^\pi(s_t)}{d_t^\mu(s_t)}$*[3]*. If $n \ge \frac{4t\tau_a\tau_s}{\max\{d_t^\pi(s_t), d_t^\mu(s_t)\}}$ for all $t = 2, \ldots, H$, then for all $h = 1, 2, \ldots, H$ and s_h, we have that:*

$$\mathrm{Var}[\widetilde{d}_h^\pi(s_h)] \le \frac{4h\tau_a\tau_s}{n}d_h^\pi(s_h).$$

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

$$\mathrm{Var}[\widetilde{d}_1^\pi(s_1)] = \frac{1}{n}(d_h^\pi(s_1) - d_h^\pi(s_1)^2) \le \frac{4\tau_a\tau_s}{n}.$$

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

Assume $\mathrm{Var}[\widetilde{d}_t^\pi(s_t)] \le \frac{4t\tau_a\tau_s}{n}d_t^\pi(s_t)$ is true for all $t = 1, \ldots, h-1$, then by our assumption on n and that $h \le H$, we obtain that

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

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

for all $t = 1, ..., h$. 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

$$\begin{aligned}\mathrm{Var}[\widetilde{d}^{\pi}_h(s_h)] \leq &\frac{4\tau_s\tau_a}{n} \sum_{t=2}^{h} \sum_{s_t} \mathbb{P}^{\pi}_{h,t}(s_h|s_t)^2 d^{\pi}_t(s_t) + \frac{1}{n} \sum_{s_1} \mathbb{P}^{\pi}_{h,1}(s_h|s_1)^2 d_1(s_1) \\ \leq &\frac{4\tau_s\tau_a}{n} \sum_{t=1}^{h} \sum_{s_t} \mathbb{P}^{\pi}_{h,t}(s_h|s_t)^2 d^{\pi}_t(s_t) \\ \leq &\frac{4\tau_s\tau_a}{n} \sum_{t=1}^{h} \sum_{s_t} \mathbb{P}^{\pi}_{h,t}(s_h|s_t) d^{\pi}_t(s_t) \\ = &\frac{4h\tau_s\tau_a}{n} d^{\pi}_h(s_h)\end{aligned}$$

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.4. We start by applying the law of total variance to obtain the following recursive equation

$$\begin{aligned}\mathrm{Cov}[\widetilde{d}^{\pi}_h] &= \mathbb{E}\left[\mathrm{Cov}\left[\widetilde{\mathbb{P}}^{\pi}_{h,h-1}\widetilde{d}^{\pi}_{h-1}\middle|\mathrm{Data}_{h-1}\right]\right] + \mathrm{Cov}\left[\mathbb{E}\left[\widetilde{\mathbb{P}}^{\pi}_{h,h-1}\widetilde{d}^{\pi}_{h-1}\middle|\mathrm{Data}_{h-1}\right]\right] \\ &= \mathbb{E}\left[\mathrm{Cov}\left[\sum_{s_{h-1}} \widetilde{\mathbb{P}}^{\pi}_{h,h-1}(\cdot|s_{h-1})\widetilde{d}^{\pi}_{h-1}(s_{h-1})\middle|\mathrm{Data}_{h-1}\right]\right] + \mathrm{Cov}\left[\mathbb{E}\left[\widetilde{\mathbb{P}}^{\pi}_{h,h-1}\widetilde{d}^{\pi}_{h-1}\middle|\mathrm{Data}_{h-1}\right]\right] \\ &= \underbrace{\mathbb{E}\left[\sum_{s_{h-1}} \mathrm{Cov}\left[\widetilde{\mathbb{P}}^{\pi}_{h,h-1}(\cdot|s_{h-1})\middle|\mathrm{Data}_{h-1}\right] \widetilde{d}^{\pi}_{h-1}(s_{h-1})^2\right]}_{(***)} + \mathbb{P}^{\pi}_{h,h-1}\mathrm{Cov}[\widetilde{d}^{\pi}_{h-1}][\mathbb{P}^{\pi}_{h,h-1}]^T.\end{aligned} \tag{B.6}$$

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}.

$$
\begin{aligned}
(***) =& \mathbb{E}\left[\sum_{s_{h-1}} \text{Cov}\left[\frac{1}{n_{s_{h-1}}}\sum_{i=1}^{n}\frac{\pi(a_{h-1}^{(i)}|s_{h-1}^{(i)})}{\mu(a_{h-1}^{(i)}|s_{h-1}^{(i)})}\mathbf{1}_{\{s_{h-1}^{(i)}=s_{h-1}\}}\mathbf{e}_{s_h^{(i)}}\middle|\text{Data}_{h-1}\right]\mathbf{1}_{\{n_{s_{h-1}}\geq\frac{nd_{h-1}^{\mu}(s_{h-1})}{2}\}}\widetilde{d}_{h-1}^{\pi}(s_{h-1})^2\right] \\
=& \mathbb{E}\left[\sum_{s_{h-1}}\frac{1}{n_{s_{h-1}}}\text{Cov}\left[\frac{\pi(a_{h-1}^{(1)}|s_{h-1})}{\mu(a_{h-1}^{(1)}|s_{h-1})}\mathbf{e}_{s_h^{(1)}}\middle|s_{h-1}^{(1)}=s_{h-1}\right]\mathbf{1}_{\{n_{s_{h-1}}\geq\frac{nd_{h-1}^{\mu}(s_{h-1})}{2}\}}\widetilde{d}_{h-1}^{\pi}(s_{h-1})^2\right] \\
=& \sum_{s_{h-1}}\left\{\mathbb{E}\left[\frac{1}{n_{s_{h-1}}}\mathbf{1}_{\{n_{s_{h-1}}\geq\frac{nd_{h-1}^{\mu}(s_{h-1})}{2}\}}\widetilde{d}_{h-1}^{\pi}(s_{h-1})^2\right]\left(\sum_{a_{h-1}}\frac{\pi(a_{h-1}|s_{h-1})^2}{\mu(a_{h-1}|s_{h-1})}\text{diag}[\mathbb{P}_{h,h-1}(\cdot|s_{h-1},a_{h-1})]\right.\right. \\
& \left.\left. -\mathbb{P}_{h,h-1}^{\pi}(\cdot|s_{h-1})[\mathbb{P}_{h,h-1}^{\pi}(\cdot|s_{h-1})]^T\right)\right\} \\
\prec& \sum_{s_{h-1}}\left\{\frac{2d_{h-1}^{\pi}(s_{h-1})^2+2\text{Var}[\widetilde{d}_{h-1}^{\pi}(s_{h-1})]}{nd_{h-1}^{\mu}(s_{h-1})}\sum_{a_{h-1}}\frac{\pi(a_{h-1}|s_{h-1})^2}{\mu(a_{h-1}|s_{h-1})}\text{diag}[\mathbb{P}_{h,h-1}(\cdot|s_{h-1},a_{h-1})]\right\}
\end{aligned}
\tag{B.7}
$$

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.7) 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.6) and (B.7) and by recursively apply them, we get the stated results. □

Combine Lemma B.1, (B.5), Theorem B.1, 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:*

$$r_h(s_h, a_h) := \mathbb{E}_\pi \left[r_h^{(1)} \middle| s_h^{(1)} = s_h, a_h^{(1)} = a_h \right] \in [0, R_{\max}]$$

$$\sigma_h(s_h, a_h) := \mathrm{Var}_\pi \left[r_h^{(1)} \middle| s_h^{(1)} = s_h, a_h^{(1)} = a_h \right]^{1/2} \leq \sigma$$

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

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*

$$n \geq \max \left\{ \frac{4t\tau_a\tau_s}{\min_{t,s_t} \max\{d_t^\pi(s_t), d_t^\mu(s_t)\}}, \frac{4\log_{e/2}(nS)}{\min_{t,s_t} d_t^\mu(s_t)} \right\}$$

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

$$\mathbb{E}[(\mathcal{P}\widehat{v}^\pi - v^\pi)^2] \leq \frac{4}{n} \sum_{h=0}^{H} \sum_{s_h} \frac{d_h^\pi(s_h)^2}{d_h^\mu(s_h)} \sum_{a_h} \frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)} (\|V_{h+1}^\pi\|^2_{P_{h+1,h}(\cdot|s_h,a_h)} + \sigma_h^2(s_h,a_h) + r_h(s_h,a_h)^2) + \frac{19\tau_a^2\tau_s^2 SH^2(\sigma^2 + R_{\max}^2 + V_{\max}^2)}{n^2},$$

Corollary 3. *In the familiar setting when* $V_{\max} = HR_{\max}$*, then the same conditions in the above theorem implies that:*

$$\mathbb{E}[(\mathcal{P}\widehat{v}^\pi - v^\pi)^2] \leq \frac{8}{n}\tau_a\tau_s(H\sigma^2 + H^3R_{\max}^2).$$

Proof of Theorem 4.1. Lemma B.2, Lemma 4.1 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 the stated results.

$$\begin{aligned}
&\mathbb{E}[(\mathcal{P}\widetilde{v}^\pi - v^\pi)^2] \\
\leq & \frac{\|V_1^\pi\|^2_{d_1(\cdot)}}{n} + \frac{4}{n} \sum_{h=1}^{H} \sum_{s_h} \frac{d_h^\pi(s_h)^2}{d_h^\mu(s_h)} \sum_{a_h} \frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)} (\|V_{h+1}^\pi\|^2_{P_{h+1,h}(\cdot|s_h,a_h)} + \sigma^2 + R_{\max}^2) \\
&+ \frac{4}{n} \sum_{h=1}^{H} \sum_{s_h} \frac{4h\tau_a\tau_s}{n} \frac{d_h^\pi(s_h)}{d_h^\mu(s_h)} \sum_{a_h} \frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)} (\|V_{h+1}^\pi\|^2_{P_{h+1,h}(\cdot|s_h,a_h)} + \sigma^2 + R_{\max}^2) \\
&+ 3H^3SR_{\max}^2 \left(\frac{2}{e}\right)^{\frac{n \min_{t,s_t} d_t^\mu(s_t)}{2}}.
\end{aligned}$$

The first set of assumption on n ensures that we can apply Theorem B.1, our second assumption on n, the logarithm term is smaller than $3n^{-2}H^3SR_{\max}^2$, which is combined with a simple upper bound of the $O(1/n^2)$ term.

To obtain the "Or simply" part, 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. *The result implies a sample complexity (in terms of the number of episodes) of* H^3SA/ϵ^2*, which matches the information-theoretic lower bound in the PAC RL setting [Dann and Brunskill, 2015]*[4]*, and the regret lower bound in an online learning setting[see, e.g., Jin et al., 2018, Theorem*

[4] Careful readers may notice that the sample complexity lower bound of [Dann and Brunskill, 2015] is H^2SA/ϵ^2 for a stationary transition kernel, in our setting a factor of H is there to account for the unknown time-varying transition probabilities.

4][5]. In fact, asymptotically, our bound also matches the Cramer-Rao lower bound for the discrete DAG-MDP model Jiang and Li [2016, Theorem 3] up to a universal constant of 4. *To the best of our knowledge, there has not been an analysis that achieves the optimal sample complexity for off-policy evaluation in the model-free setting. The only two known instances where correct dependence on* H *(or* $(1-\gamma)^{-1}$ *in infinite horizon settings) for tabular MDPs are the model-based approach [Azar et al., 2017] and under the additional assumption of a generative model [Sidford et al., 2018].*

Remark 3. *Note that the bound is not tight when* $\pi = \mu$. *The simple averages will achieve a variance bound of* $(H^2 R_{\max}^2 + H\sigma^2)/n$, *while Corollary 2 says* H^3. *This should not be alarming, as we commented earlier the bound in Corollary 2 cannot be improved in general. There is a more refined version of Theorem 4.1 that smoothly interpolates between the regime above to the regime of* $\pi = \mu$ *when the tight bound is on the order of* $(H^2 R_{\max}^2 + H\sigma^2)/n$. *The idea is to use the exact variance calculation in place of the inequality here (B.2), which will lead to slightly more complicated (but straightforward) calculations that knocks out a factor of* H *when* π *and* μ *are almost identical to each other.*

C Application to Other IS-Based Estimators

In this section, we unify some popular IS-based estimators using generic framework IS-based estimators. Then we show that our marginalized approach can be applied directly to these IS-based estimators.

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

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

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:

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

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:

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

[5]Their cumulative regret bound is $\sqrt{H^2 SAT}$ but T is the total number of steps we can take $T = nH$ and recover that one additional factor of $\sqrt{H}$.

and the WDR estimator:

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

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:

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

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 $w_t(s)$. We will discuss this property in detail in the next section.

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

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

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

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

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

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

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

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

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

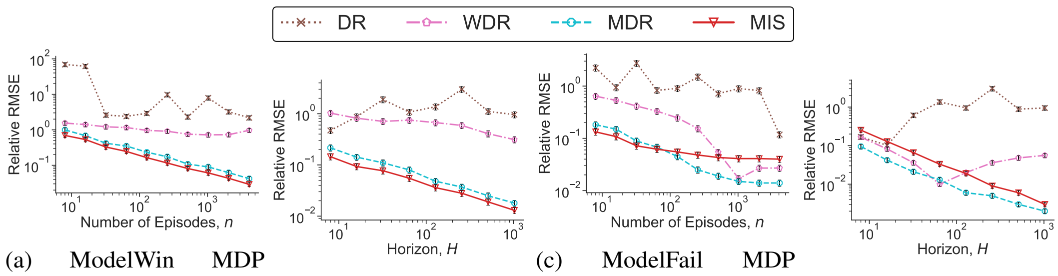
$$\widehat{v}_{BM}(\pi)=\frac{1}{n}\sum_{i=1}^{n}g(s_0^i)+\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{H}\widehat{w}_t(s_t^i)\rho_t^i\gamma^t(\widehat{r}(s_t^i,a_t^i)+f_t(s_t^i,a_t^i,s_{t+1}^i)). \tag{C.7}$$

Remark 4. *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.

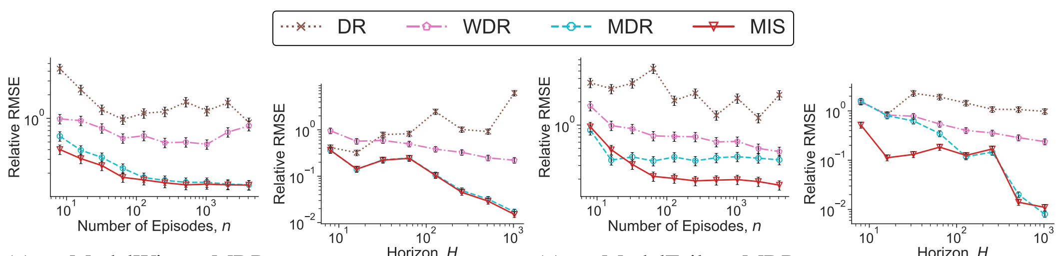
D Extended Experimental Studies

In this section, we 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.



(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.



(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 6: Results on time-varying MDPs.

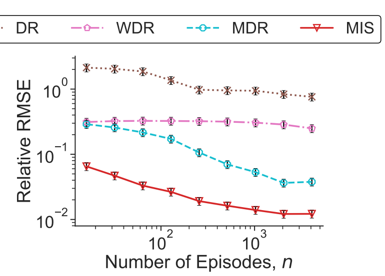


Figure 7: Mountain Car with different number of episodes.

The results are in Figure 5, Figure 6, and Figure 7. 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 projecting (D.1) into the probability simplex for better performance.

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

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

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

7: **end for**

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