Understanding the Curse of Horizon in Off-Policy Evaluation via Conditional Importance Sampling

Yao Liu [1] **Pierre-Luc Bacon** [2] **Emma Brunskill** [1]

Abstract

Off-policy policy estimators that use importance sampling (IS) can suffer from high variance in long-horizon domains, and there has been particular excitement over new IS methods that leverage the structure of Markov decision processes. We analyze the variance of the most popular approaches through the viewpoint of conditional Monte Carlo. Surprisingly, we find that in finite horizon MDPs there is no strict variance reduction of per-decision importance sampling or marginalized importance sampling, comparing with vanilla importance sampling. We then provide sufficient conditions under which the per-decision or marginalized estimators will provably reduce the variance over importance sampling with finite horizons. For the asymptotic (in terms of horizon T) case, we develop upper and lower bounds on the variance of those estimators which yields sufficient conditions under which there exists an exponential v.s. polynomial gap between the variance of importance sampling and that of the per-decision or stationary/marginalized estimators. These results help advance our understanding of if and when new types of IS estimators will improve the accuracy of off-policy estimation.

1. Introduction

Off-policy (Sutton & Barto, 2018) policy evaluation is the problem of estimating the expected return of a given *target* policy from the distribution of samples induced by a different policy. Due in part to the growing sources of data about past sequences of decisions and their outcomes – from marketing to energy management to healthcare – there is increasing interest in developing accurate and efficient algorithms for off-policy policy evaluation.

For Markov Decision Processes, this problem was addressed (Precup et al., 2000; Peshkin & Shelton, 2002) early on by importance sampling (IS)(Rubinstein, 1981), a method prone to large variance due to rare events (Glynn, 1994; L'Ecuyer et al., 2009). The *per-decision* importance sampling estimator of (Precup et al., 2000) tries to mitigate this problem by leveraging the temporal structure – earlier rewards cannot depend on later decisions – of the domain.

While neither importance sampling (IS) nor per-decision IS (PDIS) assumes the underlying domain is Markov, more recently, a new class of estimators (Hallak & Mannor, 2017; Liu et al., 2018; Gelada & Bellemare, 2019) has been proposed that leverages the Markovian structure. In particular, these approaches propose performing importance sampling over the stationary or marginalized state-action distributions induced by the corresponding Markov chain for a particular policy. By avoiding the explicit accumulation of likelihood ratios along the trajectories, it is hypothesized that such ratios of stationary/marginalized distributions could substantially reduce the variance of the resulting estimator, thereby overcoming the "curse of horizon" (Liu et al., 2018) plaguing off-policy evaluation. The recent flurry of empirical results shows significant performance improvements over the alternative methods on a variety of simulation domains. Yet so far there has not been a formal analysis of the accuracy of IS, PDIS, and marginalized state–action IS which will strengthen our understanding of their properties, benefits and limitations.

To formally understand the variance relationship between those unbiased estimators, we link this to a more general class of estimators: the *extended* (Bratley et al., 1987) form of the conditional Monte Carlo estimators (Hammersley, 1956; Dubi & Horowitz, 1979; Granovsky, 1981), and thus view those importance sampling estimators in a unified framework and referred to as conditional importance sampling, since they are computing weights as conditional expectation of likelihood ratio conditioning on different choice of statistics. Though the intuition from prior work suggests that marginalized importance sampling should have the best

[1]Department of Computer Science, Stanford University [2]Mila - University of Montreal. This research was conducted when Pierre-Luc Bacon was a post-doc at Stanford.. Correspondence to: Yao Liu <yaoliu@stanford.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

accuracy, followed by per-decision IS and then (worst) the crude IS estimator. Surprisingly, we show that this is not always the case. In particular, we construct short-horizon MDP examples in Figure 1 that demonstrate that the crude IS can have a lower variance estimate than per-decision IS or marginalized IS, and also show results for the other cases.

We then describe how this observation is quite natural when we note that all three estimators are instances of conditional expectation. If X and Y are two well-defined random variables on the same probability space such that $\theta = \mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$, then the conditional Monte Carlo estimator for θ is $\mathbb{E}[Y|x]$. By the law of total variance, the variance of the conditional Monte Carlo estimator cannot be larger than that of the crude Monte Carlo estimator y. However when X and Y are sequences of random variables, and we want to estimate $\mathbb{E}\left[\sum_{t=1}^{T} Y_t\right]$, the variance of the so-called *extended* conditional Monte Carlo estimator $\sum_{t=1}^{T} \mathbb{E}[Y_t|x_t]$ is not guaranteed to reduce variance due to covariance between the summands.

Building on these insights, we then provide a general variance analysis for conditional importance sampling estimators, as well as sufficient conditions for variance reduction in Section 5. In Section 6 we provide upper and lower bounds for the asymptotic variance of the crude, per-decision and marginalized estimators. These bounds show, under certain conditions, that the per-decision and marginalized importance sampling estimators can reduce the asymptotic variance to a polynomial function of the horizon compared to the exponential dependence of the per-decision estimator. Our proofs apply to general state spaces and use concentration inequalities for martingales. Importantly, these bounds characterize a set of common conditions under which the variance of marginalized importance sampling can be smaller than that of per decision importance sampling, which in turn, can have a smaller variance than the crude importance sampling estimator. In doing so, our results provide concrete theoretical foundations supporting recent empirical successes in long-horizon domains.

2. Notation and Problem Setting

We consider Markov Decision Processes (MDPs) with discounted or undiscounted rewards and under a fixed horizon T. An MDP is defined by a tuple $(\mathcal{S}, \mathcal{A}, P, p_1, r, \gamma, T)$, where $\mathcal{S} \subset \mathbb{R}^d$ is the state space and $\mathcal{A}$ is the action space, which we assume are both bounded compact Hausdorf spaces. We use the notation $P(S|s,a)$ to denote the transition probability kernel where $S \subset \mathcal{S}, s \in \mathcal{S}, a \in \mathcal{A}$ and $r_t(s,a) : \mathcal{S} \times \mathcal{A} \times [T] \mapsto [0,1]$ for the reward function. The symbol $\gamma \in [0,1]$ [1] refers to the discount factor. For simplicity we write the probability density function associated with P as $p(s'|s,a)$. Furthermore, our definition contains a probability density function of the initial state s_1 which we denote by p_1. We use $\pi(a_t|s_t)$ and $\mu(a_t|s_t)$ to denote the conditional probability density/mass functions associated with the policies π and μ. We call μ the *behavior* policy and π the *target* policy. We assume $\frac{\pi(a|s)}{\mu(a|s)} < \infty$ throughout this paper which is the necessary condition for the effectiveness of all importance sampling bassed methods. We are interested in estimating the value of π, defined as:

$$v^\pi = \mathbb{E}_\pi\left[\sum_{t=1}^{T} \gamma^{t-1} r_t\right] \ .$$

Furthermore, we use the notation $\tau_{1:T}$ to denote a T-step trajectory of the form: $\tau_{1:T} = \{(s_t, a_t, r_t)\}_{t=1}^{T}$. When appropriate, we use the subscript π or μ to specify if $\tau_{1:T}$ comes from the induced distribution of π or μ. We use the convention that the lack of subscript for $\mathbb{E}$ is equivalent to writing $\mathbb{E}_\mu$, but otherwise write $\mathbb{E}_\pi$ explicitly. We denote the 1-step likelihood ratio and the T-steps likelihood ratio respectively as:

$$\rho_t = \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}, \quad \rho_{1:T} = \prod_{t=1}^{T} \rho_t \ .$$

We define the T-step state distribution and marginalized state distribution under the behavior policy as:

$$d_t^\mu(s,a) = \Pr(s_t = s, a_t = a | s_1 \sim p_1, a_i \sim \mu(a_i|s_i))$$

$$d_{\gamma,1:T}^\mu(s,a) = \frac{\sum_{t=1}^{T} \gamma^t d_t^\mu(s,a)}{\sum_{t=1}^{T} \gamma^t}, \ d_\gamma^\mu = \lim_{T\to\infty} d_{\gamma,1:T}^\mu$$

For simplicity of notation, we drop the γ in $d_{\gamma,1:T}^\mu$ and d_γ^μ when $\gamma = 1$, and overload d^μ to denote the marginal state distribution as well: ie. $d^\mu(s) = \int_a d^\mu(s,a)\mathrm{d}a$ (and similarly for d^π). We use c to denote the KL divergence of μ and π and where the expectation is taken under d^μ over the states: $\mathbb{E}_{d^\mu}[D_{\mathrm{KL}}(\mu||\pi)]$. We assume $c > 0$, otherwise π and μ are identical and our problem reduces to on-policy policy evaluation.

In this paper, we define the estimator and discuss the variance over a single trajectory but of all our results carry to N trajectories by multiplying by a factor $1/N$. We define the *crude* importance sampling (IS) estimator and the per-decision importance sampling (PDIS) from (Precup et al., 2000) as:

$$\hat{v}_{\mathrm{IS}} = \rho_{1:T} \sum_{t=1}^{T} \gamma^{t-1} r_t, \ \ \hat{v}_{\mathrm{PDIS}} = \sum_{t=1}^{T} \gamma^{t-1} r_t \rho_{1:t} \ .$$

The marginalized importance sampling (MIS) estimator is defined as:

$$\hat{v}_{\mathrm{MIS}} = \sum_{t=1}^{T} \gamma^{t-1} r_t \frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)} \ .$$

[1] Our analysis works for both discounted and undiscounted reward.

All three estimators are unbiased. Our analysis mainly focuses on MIS (Xie et al., 2019) using the ratio between the marginalized state distributions (provided by an oracle) rather than the stationary distributions as in the prior work by (Liu et al., 2018; Hallak & Mannor, 2017). The usage of marginalized ratio is to tackle both the finite horizon and infinite horizon more easily under the same general framework by taking $T \to \infty$ when necessary, since the stationary ratio is undefined when $T < \infty$. However, the marginalized and stationary importance sampling estimators share the similar asymptotic properties because the ratio $\frac{d^\pi(s_t,a_t)}{d^\mu(s_t,a_t)}$ has the same asymptotic behavior as that of the stationary distribution ratio. Thus our analysis can be still viewed as over the group of marginalized and stationary importance sampling estimators. Surprisingly, we show that even under perfect knowledge of the marginalized ratio, it is generally non-trivial to guarantee a variance reduction for $\hat{v}_{\text{MIS}}$.

The next standard assumptions helps us analyze our estimators in the asymptotic regime by relying on a central limit property for general Markov chains.

Assumption 1 (Harris ergodic)**.** *The Markov chain of $\{s_t, a_t\}$ under μ is Harris ergodic. That is: the chain is aperiodic, ψ-irreducible, and positive Harris recurrent. See (Meyn & Tweedie, 2012) for more*

Assumption 2 (Drift property)**.** *There exist an everywhere-finite function $B : \mathcal{S} \times \mathcal{A} \mapsto [1, \infty)$, a constant $\lambda \in (0, 1)$, $b < \infty$ and a petite $K \subset \mathcal{S} \times \mathcal{A}$ such that:*

$$\mathbb{E}_{s',a'|s,a} B(s', a') \leq \lambda B(s, a) + b\mathbb{1}((s, a) \in K) \ .$$

These are standard assumptions to describe the ergodic and recurrent properties of general Markov chains. Assumption 1 is typically used to obtain the existence of a unique stationary distribution (Meyn & Tweedie, 2012) and assumption 2 is used to measure the concentration property (Meyn & Tweedie, 2012; Jones et al., 2004).

3. Counterexamples

It is tempting to presume that the root cause of the variance issues in importance sampling pertains entirely to the explicit *multiplicative* (Liu et al., 2018) accumulation of importance weights over long trajectories. The reasoning ensuing from this intuition is that the more terms one can drop off this product, the better the resulting estimator would be in terms of variance. In this section, we show that this intuition is misleading as we can construct small MDPs in which per-decision or marginalized importance sampling does not necessarily reduce the variance of the crude importance sampling. We then explain this phenomenon in section 4 using the extended conditional Monte Carlo method and point out that the lack of variance reduction is attributable to the interaction of some covariance terms across time steps.

However, all is not lost and section 6 shows that asymptotically ($T \to \infty$) marginalized importance sampling achieves much lower variance than crude importance sampling or the per-decision variant.

In all examples, we use a two-steps MDP with deterministic transitions, undiscounted reward, and in which a uniform behavior policy is always initialized from the state s_1 (see figure 1). We then show that the ordering of the estimators based on their variance can vary by manipulating the target policy and the reward function so as to induce a different covariance structure between the reward and the likelihood ratio. We can then compute the exact variance (table 1) of each estimator manually (see appendix A). Example 1a shows that the per-decision estimator can have a larger variance than the crude estimator when marginalized estimator improves on per-decision estimator. Example 1b shows an instance where the marginalized estimator does not improve on the per-decision importance sampling, but per-decision importance sampling has a smaller variance than crude importance sampling. Finally, example 1c provides a negative example where the ordering goes against our intuition and shows that the marginalized estimator is worse than the per-decision estimator, which in turn has a larger variance than the crude estimator. Note that the lack of variance reduction for marginalized IS occurs even with perfect knowledge of the marginalized ratio. We show in section 4 that the problem comes from the covariance terms across time steps.

	IS		PDIS		MIS
Example 1a	0.12	$<$	0.2448	$>$	0.2
Example 1b	0.5424	$>$	0.4528	$<$	0.52
Example 1c	0.2304	$<$	0.2688	$<$	0.32

Table 1. Analytical variance of different estimators. See figure 1 for the problem structure.

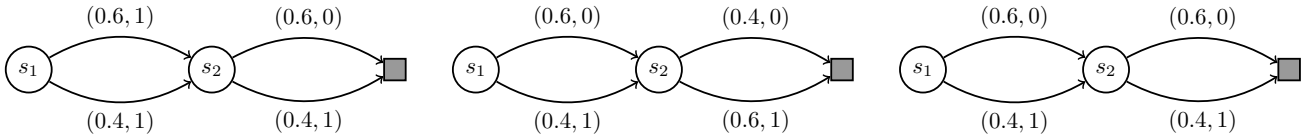
4. Conditional Importance Sampling

The unbiasedness of crude importance sampling (IS) estimator follows from the fact that:

$$\mathbb{E}\left[\rho_{1:T} \sum_{t=1}^{T} \gamma^{t-1} r_t\right] = \mathbb{E}_\pi\left[\sum_{t=1}^{T} \gamma^{t-1} r_t\right] = v^\pi \ ,$$

Let G_T be the total (discounted) return $\sum_{t=1}^{T} \gamma^{t-1} r_t$ and if ϕ_T is some statistics such that $\rho_{1:T}$ is conditionally independent with G_T given ϕ_T, then by the law of total expectation:

$$\begin{aligned}\mathbb{E}\left[\rho_{1:T} G_T\right] &= \mathbb{E}\left[\mathbb{E}\left[\rho_{1:T} G_T | \phi_T, G_T\right]\right] \\ &= \mathbb{E}\left[G_T \mathbb{E}\left[\rho_{1:T} | \phi_T, G_T\right]\right] \\ &= \mathbb{E}\left[G_T \mathbb{E}\left[\rho_{1:T} | \phi_T\right]\right] .\end{aligned}$$



(a) $\text{Var}(\hat{v}_{\text{IS}}) < \text{Var}(\hat{v}_{\text{MIS}}) < \text{Var}(\hat{v}_{\text{PDIS}})$ (b) $\text{Var}(\hat{v}_{\text{PDIS}}) < \text{Var}(\hat{v}_{\text{MIS}}) < \text{Var}(\hat{v}_{\text{IS}})$ (c) $\text{Var}(\hat{v}_{\text{IS}}) < \text{Var}(\hat{v}_{\text{PDIS}}) < \text{Var}(\hat{v}_{\text{MIS}})$

Figure 1. Counterexamples. The labels for each edge are of the form `(target policy probability, reward)` where the first component is the transition probability induced by the given target policy, and the second component is the reward function for this transition. All examples assume deterministic transitions and the same initial state s_1. The square symbol represents a terminating state.

Furthermore, by the law of total variance we have:

$$\begin{aligned}&\text{Var}\left(G_T\mathbb{E}\left[\rho_{1:T}|\phi_T\right]\right)\\&= \text{Var}\left(G_T\rho_{1:T}\right) - \mathbb{E}\left[\text{Var}(G_T\phi_{1:T}|\phi_T, G_T)\right]\\&= \text{Var}\left(G_T\rho_{1:T}\right) - \mathbb{E}\left[G_T^2\text{Var}(\rho_{1:T}|\phi_T)\right] \quad .\end{aligned}$$

Because the second term is always non-negative, it follows that $\text{Var}\left(G_T\mathbb{E}\left[\rho_{1:T}|\phi_T\right]\right) \leq \text{Var}\left(G_T\rho_{1:T}\right)$. This conditioning idea is the basis for the conditional Monte Carlo (CMC) as a variance reduction method.

If we now allow ourselves to conditions in a stage-dependent manner rather than with a fixed statistics ϕ_T, we obtain estimators belonging to the so-called *extended* conditional Monte Carlo methods (Bratley et al., 1987). Assuming that r_t is conditionally independent with $\rho_{1:T}$ given ϕ_t, then by the law of total expectation:

$$\begin{aligned}v^\pi = \mathbb{E}\left[G_T\rho_{1:T}\right] &= \sum_{t=1}^{T}\gamma^{t-1}\mathbb{E}\left[\mathbb{E}\left[r_t\rho_{1:T}|\phi_t, r_t\right]\right]\\&= \mathbb{E}\left[\sum_{t=1}^{T}\gamma^{t-1}r_t\mathbb{E}\left[\rho_{1:T}|\phi_t\right]\right] \quad .\end{aligned}$$

We refer to estimators in this form as "*extended* conditional importance sampling estimators": a family of estimators encompassing both the per-decision importance sampling (PDIS) estimator of (Precup et al., 2000) as well as the *stationary/marginalized* variants (Hallak & Mannor, 2017; Liu et al., 2018; Gelada & Bellemare, 2019). In this paper, we use "conditional importance sampling"[2] to refer to all variants of importance sampling based on the conditional Monte Carlo method, in its "extended" form or not. To obtain the per-decision estimator in our framework, it suffices to define the stage-dependent statistics ϕ_t to be the history $\tau_{1:t}$ up to time t:

$$v^\pi = \mathbb{E}\left[\sum_{t=1}^{T}\gamma^{t-1}r_t\mathbb{E}\left[\rho_{1:T}|\tau_{1:t}\right]\right] = \mathbb{E}\left[\sum_{t=1}^{T}\gamma^{t-1}r_t\rho_{1:t}\right] \quad .$$

[2](Bucklew, 2004; 2005) also uses this expression to describe the "g-method" of (Srinivasan, 1998), which is CMC applied to IS. Our work considers the extended form of the CMC method for Markov chains: a more general setting with very different variance properties.

In this last expression, $\mathbb{E}[\rho_{1:T}|\tau_{1:t}] = \rho_{1:t}$ follows from the fact that the likelihood ratio is a martingale (L'Ecuyer & Tuffin, 2008). Similarly, the marginalized importance sampling (MIS) estimator can be derived by conditioning on the state and action at time t:

$$\begin{aligned}v^\pi &= \mathbb{E}\left[\sum_{t=1}^{T}\gamma^{t-1}r_t\mathbb{E}\left[\rho_{1:t}|s_t, a_t\right]\right]\\&= \mathbb{E}\left[\sum_{t=1}^{T}\gamma^{t-1}r_t\frac{d_t^\pi(s_t, a_t)}{d_t^\mu(s_t, a_t)}\right] .\end{aligned}$$

In this case, the connection between the expected importance sampling weights conditioned on (s_t, a_t) and the ratio of marginalized distributions warrants a lengthier justification which we formalize in the following lemma (proved in appendix).

Lemma 1. $\mathbb{E}(\rho_{1:t}|s_t, a_t) = \frac{d^\pi(s_t, a_t)}{d^\mu(s_t, a_t)}$.

Assuming that an unbiased estimator of the conditional weights $\mathbb{E}\left[\rho_{1:t}|\phi_t\right]$ is available, the conditional importance sampling estimators are also unbiased. However, the law of total variance no longer implies a variance reduction because the variance is now over a sum of random variables of the form:

$$\text{Var}\left(\sum_{t=1}^{T}r_tw_t\right) = \sum_{t=1}^{T}\text{Var}(r_tw_t) + \sum_{k\neq t}\text{Cov}(r_kw_k, r_tw_t) \quad .$$

where $w_t = \mathbb{E}\left[\gamma^{t-1}\rho_{1:t}|\phi_t\right]$. In general, there is no reason to believe that the sum of covariance terms interact in such a way as to provide a variance reduction. If stage-dependent conditioning of the importance weights need not reduce the variance in general, all we are left with is to "optimistically" (Bratley et al., 1987) suppose that the covariance structure plays in our favor. Over the next sections, we develop sufficient conditions for variance reduction in both the finite and infinite horizon setting. More specifically, theorem 1 provides sufficient conditions for a variance reduction with the per-decision estimator while theorem 2 applies to the marginalized importance sampling estimator. In section 6, we develop an asymptotic analysis of the variance when $T \to \infty$. We show that under some mild assumptions,

the variance of the crude importance sampling estimator is always exponentially large in the horizon T. Nevertheless, we show that there are cases where the per-decision or marginalized estimators can help reduce the variance to $O(T^2)$.

5. Finite-Horizon Analysis

While the counterexamples of section 3 show that there is no consistent order in general between the different IS estimators and their variance, we are still interested in characterizing when a variance reduction can occur. In this section, we provide theorems to answer when $\text{Var}(\hat{v}_{\text{PDIS}})$ is guaranteed to be smaller than $\text{Var}(\hat{v}_{\text{IS}})$ and when $\text{Var}(\hat{v}_{\text{MIS}})$ is guaranteed to be smaller than $\text{Var}(\hat{v}_{\text{PDIS}})$. We start by introducing a useful lemma to analyze the variance of the sum of conditional expectations.

Lemma 2. *Let X_t and Y_t be two sequences of random variables. Then*

$$\begin{aligned}&Var\left(\sum_t Y_t\right) - Var\left(\sum_t \mathbb{E}[Y_t|X_t]\right)\\ &\geq 2\sum_{t<k}\mathbb{E}[Y_tY_k] - 2\sum_{t<k}\mathbb{E}[\mathbb{E}[Y_t|X_t]\mathbb{E}[Y_k|X_k]]\ .\end{aligned}$$

This lemma states that the variance reduction of the stage-dependent conditional expectation depends on the difference between the covariance of the random variables and that of their conditional expectations. The variance reduction analysis of PDIS and MIS in theorems 1 and 2 can be viewed as a consequence of this result. We develop in those cases some sufficient conditions to guarantee that the difference between the covariance terms is positive.

Theorem 1 (Variance reduction of PDIS)**.** *If for any $1 \leq t \leq k \leq T$ and initial state s, $\rho_{0:k}(\tau)$ and $r_t(\tau)\rho_{0:k}(\tau)$ are positively correlated, $Var(\hat{v}_{PDIS}) \leq Var(\hat{v}_{IS})$.*

This theorem guarantees the variance reduction of PDIS given a positive correlation between the likelihood ratio and the importance-weighted reward. The random variables $\rho_{0:k}(\tau)$ and $r_t(\tau)\rho_{0:k}(\tau)$ are positively correlated when for a trajectory with large likelihood ratio, the importance-weighted reward (which is an unbiased estimator of reward under the target policy π) is also large. Intuitively, a positive correlation is to be expected if the target policy π is more likely to take a trajectory with a higher reward. We expect that this property may hold in applications where the target policy is near the optimal value for example.

Theorem 2 (Variance reduction of MIS)**.** *If for any fixed $0 \leq t \leq k < T$,*

$$Cov\left(\rho_{1:t}r_t, \rho_{1:k}r_k\right) \geq Cov\left(\frac{d_t^\pi(s,a)}{d_t^\mu(s,a)}r_t, \frac{d_k^\pi(s,a)}{d_k^\mu(s,a)}r_k\right)$$

then $Var(\hat{v}_{MIS}) \leq Var(\hat{v}_{PDIS})$

This theorem implies that the relative order of variance between MIS and PDIS depends on the ordering of the covariance terms between time-steps. In the case when T is very large, the covariance on the right is very close to zero, and if the covariance on the left is positive (which is true for many MDPs) the variance of MIS can be smaller than PDIS.

6. Asymptotic Analysis

We have seen in section 3 that a variance reduction cannot be guaranteed in the general case and we then proceeded to derive sufficient conditions. However, this section shows that the intuition behind per-decision and marginalized importance sampling does hold under some conditions and in the limit of the horizon $T \to \infty$. Under the light of these new results, we expect those estimators to compare favorably to crude importance sampling for very long horizons: an observation also implied by the sufficient conditions derived in the last section.

In the following discussion, we consider the asymptotic rate of the variance as a function when $T \to \infty$. We show that under some mild assumptions, the variance of crude importance sampling is exponential with respect to T and bounded from two sides. For the per-decision estimator, we provide conditions when the variance is at least exponential or at most polynomial with respect to T. Under some standard assumptions, we also show that the variance of marginalized importance sampling can be polynomial with respect to T, indicating an exponential variance reduction. As a starting point, we prove a result characterizing the asymptotic distribution of the importance-weighted return.

Theorem 3. *Under Assumption 1, if $\log(\frac{\pi(a|s)}{\mu(a|s)})$ is a continuous function of (s,a) in the support of μ then for $\pi \neq \mu$, $\lim_T (\rho_{1:T})^{1/T} = e^{-c}$, $\overline{\lim}_T |\hat{v}_{IS}|^{1/T} < e^{-c}$ a.s.*

Corollary 1. *Under the same condition as theorem 3, $\rho_{1:T} \to_{a.s.} 0$, $\rho_{1:T}\sum_{t=1}^T \gamma^{t-1} r_t \to_{a.s.} 0$*

Although crude importance sampling is unbiased, this result shows that it also converges to zero almost surely. Theorem 3 further proves that it converges to an exponentially small term $\exp(-cT)$. This indicates that in most cases the return is almost zero, leading to poor estimates of v^π, and under some rare events the return can be very large and the expectation is $v^\pi > 0$.

Equipped with these results, we can now show that the variance of the crude importance sampling estimator is exponential with respect to T. To quantitatively describe the variance, we need the following assumptions so that $\log \rho_t$ is bounded:

Assumption 3. $|\log \rho_t| < \infty$

This assumption entails that ρ_t is both upper-bounded (a

common assumption) and lower-bounded. We only need the assumption on the lower bound of ρ_t in the proof of a lower bound part in theorem 4 and 5. For the lower bound part, it essentially amounts to the event where all likelihood ratio terms on a trajectory are greater than zero. Then by the law of total variance, the original variance can only be larger than the variance of all returns conditioned on this event. Before we characterize the variance of the IS estimator, we first prove that the log-likelihood ratio is a martingale with bounded differences.

Lemma 3. *Under Assumption 1, 2 and 3, there exists a function $\hat{f} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ such that:*

1. $\forall (s,a)$, $|\hat{f}(s,a)| < c_1\sqrt{B(s,a)}$ *for constant* c_1.
2. *For any* $T > 0$, $\log \rho_{1:T} + Tc - \hat{f}(s_1,a_1) + \hat{f}(s_{T+1},a_{T+1})$ *is a mean-zero martingale with respect to the sequence* $\{s_t,a_t\}_{t=1}^T$ *with martingale differences bounded by* $2c_1\sqrt{||B||_\infty}$.

We are now ready to give both upper and lower bounds on the variance of the crude importance sampling estimator using an exponential function of T from both sides.

Theorem 4 (Variance of IS estimator)**.** *Under Assumption 1, 2 and 3, there exist $T_0 > 0$ such that for all $T > T_0$,*

$$Var(\hat{v}_{IS}) \geq \frac{(v^\pi)^2}{4}\exp\left(\frac{Tc^2}{8c_1^2\|B\|_\infty}\right) - (v^\pi)^2$$

where B is defined in Assumption 2, c_1 is some constant defined in lemma 3, $c = \mathbb{E}_{d^\mu}[D_{KL}(\mu||\pi)]$. If $\mathbb{E}_{a\sim\mu}\left[\frac{\pi(a|s)^2}{\mu(a|s)^2}\right] \leq M_\rho^2$ for any s, then $Var(\hat{v}_{IS}) \leq T^2M^{2T} - (v^\pi)^2$.

The lower bound part shows that the variance is at least an exponential function of the horizon T, and the rate depends on the distance between the behavior and target policies, as well as the recurrent property of the Markov chain associated with the behavior policy. This result differs from that of (Xie et al., 2019), which is based on the CLT for i.i.d sequences, since our analysis considers more broadly a distribution of samples from a Markov chain.

Proof Sketch. Let Y be the IS estimator and Z be indicator function $\mathbb{1}(Y > v^\pi/2)$. By the law of total variance, $\text{Var}(Y) \geq \text{Var}(\mathbb{E}(Y|Z))$. Since the expectation of $\mathbb{E}(Y|Z)$ is a constant, we only need to show that the second moment of $\mathbb{E}(Y|Z)$ is asymptotically exponential. To achieve this, we observe that $\mathbb{E}[(\mathbb{E}[Y|Z])^2] \geq \Pr(Y > v^\pi/2)(\mathbb{E}[Y|Y > v^\pi/2])^2$. We can then establish that $\mathbb{E}[Y|Y > v^\pi/2]$ is $\Omega(1/\Pr(Y > v^\pi/2))$ using the fact that the expectation of Y is a constant. It follows that we can upper bound $\Pr(Y > v^\pi/2)$ by an exponentially small term. This can be done by a concentration inequality for martingales. The upper bound part is proved by bounding the absolute range of each variable. □

Now we prove upper and lower bounds for the variance of the per-decision estimator as a function of γ, the expected reward at time t, $\mathbb{E}_\pi[r_t]$, and other properties of MDP. We then give a sufficient condition for the variance of PDIS to have an exponential lower bound, and when it is at most polynomial.

Theorem 5 (Variance of the PDIS estimator)**.** *Under Assumption 1, 2 and 3, $\exists T_0 > 0$ s.t. $\forall T > T_0$,*

$$Var(\hat{v}_{PDIS}) \geq \sum_{t=T_0}^{T} \frac{\gamma^{2t-2}(\mathbb{E}_\pi(r_t))^2}{4}\exp\left(\frac{tc^2}{8c_1^2\|B\|_\infty}\right) - (v^\pi)^2$$

where B c_1 and c are same constants in theorem 4, and C is some constant. For the upper bound:

1. *If $\mathbb{E}_{a\sim\mu}\left[\frac{\pi(a|s)^2}{\mu(a|s)^2}\right] \leq M_\rho^2$ for any s, $Var(\hat{v}_{PDIS}) \leq T\sum_{t=1}^T M_\rho^{2t}\gamma^{2t-2} - (v^\pi)^2$.*
2. *Let $U_\rho = \sup_{s,a}\frac{\pi(a|s)}{\mu(a|s)} < \infty$, $Var(\hat{v}_{PDIS}) \leq T\sum_{t=1}^T U_\rho^{2t}\gamma^{2t-2}\mathbb{E}_\mu[r_t^2] - (v^\pi)^2$.*

Proof Sketch. The proof of the lower bound part is similar to the proof of the last theorem where we first lower bound the square of the sum by a sum of squares. We then apply the proof techniques of theorem 4 for the time-dependent terms. The proof for the upper bound relies the Cauchy-Schwartz inequality on the square of sum and then upper bound each term directly. □

Using theorem 5, we can now give sufficient conditions for the variance of the PDIS estimator to be at least exponential or at most polynomial.

Corollary 2. *With theorem 5 holds, $Var(\hat{v}_{PDIS}) = \Omega(\exp(\epsilon T))$ if the following conditions hold: 1) $\gamma \geq \exp\left(\frac{-c^2}{16c_1^2\|B\|_\infty}\right)$; 2) There exist a $\epsilon > 0$ such that*

$$\mathbb{E}_\pi(r_t) = \Omega\left(\exp\left(-t\left(\frac{c^2}{16c_1^2\|B\|_\infty} + \log\gamma - \epsilon/2\right)\right)\right)$$

This corollary says that if γ is close enough to 1 and the expected reward under the target policy is larger than an exponentially decaying function, then the variance of $\hat{v}_{\text{PDIS}}$is still at least exponentially large. We note that the second condition is satisfied if $r_t(s,a)$ is a function that does not depends on t and $\mathbb{E}_{d^\pi}(r(s,a)) > 0$. This is due to the fact that $\mathbb{E}_\pi(r_t) \to \mathbb{E}_{d^\pi}(r(s,a))$ as $t \to \infty$ and we obtain a constant which is larger than any exponentially decaying function.

Corollary 3. *Let* $U_\rho = \sup_{s,a} \frac{\pi(a|s)}{\mu(a|s)}$. *If* $U_\rho \gamma \leq 1$ *or* $U_\rho \gamma \lim_T \left(\mathbb{E}_\pi[r_T]\right)^{1/T} < 1$, *Var*$(\hat{v}_{PDIS}) = O(T^2)$.

This corollary says that when γ and the reward $\mathbb{E}_\pi(r_t)$ decreases fast enough, the variance of PDIS is polynomial in T, indicating an exponential improvement over crude importance sampling for long horizons. We can now prove an upper bound on the variance of marginalized importance sampling.

Theorem 6 (Variance of the MIS estimator)**.**

$$Var\left(\hat{v}_{MIS}\right) \leq T \sum_{t=1}^{T} \gamma^{t-1} \left(\mathbb{E}\left[\left(\frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)} \right)^2 \right] - 1 \right)$$

The proof uses Cauchy-Schwartz to bound each covariance term. In this theorem, the left hand side, is very close to $O(T^2)$ but $\mathbb{E}\left[\left(\frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)} \right)^2 \right]$ still depends on t. Intuitively, the assumption that the ratio of marginalized distributions is bounded is enough for this to hold since d_t^μ and d_t^π is close to d^μ and d^π for large t. We formally show this idea in the next corollary. However, we first need to introduce a continuity definition for function sequences.

Definition 1 (asymptotically equi-continuous)**.** *A function sequence* $f_t : \mathbb{R}^d \mapsto \mathbb{R}$ *is asymptotically equi-continuous if for ant* $\epsilon > 0$ *there exist* $n, \delta > 0$ *such that for all* $t > n$ *and dist*$(x_1, x_2) \leq \delta$, $|f_t(x_1) - f_t(x_2)| \leq \epsilon$

Corollary 4. *If* $d_t^\mu(s_t)$ *and* $d_t^\pi(s_t)$ *are asymptotically equi-continuous,* $\frac{d^\pi(s)}{d^\mu(s)} \leq U_s$, *and* $\frac{\pi(a|s)}{\mu(a|s)} \leq U_\rho$, *then Var* $\left(\hat{v}_{MIS}\right) = O(T^2)$

This corollary implies that as long as the stationary ratio and one step ratios are bounded, the variance of marginalized IS is $O(T^2)$. This result is predicated on having access to an oracle of d_t^π / d_t^μ because our results characterizes the variance reduction due to *conditioning* irrespective of the choice of estimators for d_t^π / d_t^μ. For general case of approximating choice of the ratio d_t^π / d_t^μ by a function $w_t(s_t, a_t)$, we could show an plug-in type estimator and a variance upper bound based on the Theorem above.

Now we consider approximate MIS estimators, which approximate density ratio $\frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)}$ by $w_t(s_t, a_t)$ and plug it into the MIS estimator. More specifically,

$$\hat{v}_{\text{AMIS}} = \sum_{t=1}^{T} \gamma^{t-1} w_t(s,a) r_t \tag{1}$$

This approximate MIS estimator is often biased based on the choice of $w_t(s, a)$, so we consider the upper bound of their mean square error with respect to T and the error of the ratio estimator. We can show the following bound under the same condition as the oracle ratio case:

Corollary 5. *Under the same condition of Corollary 4,* $\hat{v}_{AMIS}$ *with* w_t *such that where* $\mathbb{E}_\mu \left(w_t(s_t,a_t) - \frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)} \right)^2 \leq \epsilon_w$ *has a MSE of* $O(T^2(1+\epsilon_w))$

Different accuracy bound of w_t result in estimators with different variance. Previous work (Xie et al., 2019) shows the existence of w_t estimator with polynomial MSE. This bound match the dependency on the horizon $O(T^3)$ from (Xie et al., 2019) for an $O(T)$ accurate w_t, with our proof considers general spaces with samples coming from a Markov chain, and potentially works for more general choice of w_t. This result, along with the lower bound for variance of PDIS and IS, suggests that for long-horizon problems MIS reduces the variance significantly, from $\exp(T)$ to $O(T^2)$. Only when corollary 3 holds, which requires a much stronger assumption than this, PDIS yields $O(T^2)$ variance.

There might also be question on if MIS estimators can potentially achieve a better error bound than $O(T^2)$. We demonstrate an example to show that for any T even with an oracle of the ratio $\frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)}$ or $\frac{d^\pi(s)}{d^\mu(s)}$, there exist an MDP such that $\text{Var}(\hat{v}_{\text{MIS}})$ is at least $\Theta(T^2)$.

Definition 2. *Given any* $T > 3$, *define an MDP and off-policy evaluation problem in the following way:*

1. *There are two actions* a_1 *and* a_2 *in initial state* s_0, *leading to* s_1 *and* s_2 *separately. After* s_1, *the agent will go through* $s_3, s_5, \ldots s_{2n-1}$ *sequentially, no matter which action was taken in* $s_1, s_3, s_5, \ldots s_{2T-1}$. *Similarly,* s_2 *leads to a chain of* $s_4, s_6, \ldots s_{2T}$.
2. *The reward for* s_0, a_1 *and any action on* $s_1, s_3, s_5, \ldots s_{2n-1}$ *is one, and zero otherwise.*
3. *Behavior policy is an uniformly random policy, gives* 0.5 *probability to go through each of the chains. Evaluation policy will always choose* a_1 *which leads to the chain of* s_1.

In this example, it is easy to verify that the distribution of MIS given oracle $\frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)}$ is a uniform distribution over $\{0, 2T\}$, and the variance is T^2.

7. Related Work

This idea of substituting the importance ratios for their conditional expectations can be found in the thesis of (Hesterberg, 1988) under the name *conditional weights* and is presented as an instance of the conditional Monte Carlo method. Here instead, we consider the class of importance sampling estimators arising from the extended conditional Monte Carlo method and under a more general conditional independence assumption than that of (Hesterberg, 1988, p.48). The "conditional" form of the per-decision and stationary estimators are also discussed in appendix A of (Liu et al.,

2018) where the authors hypothesize a potential connection to the more stringent concept of Rao-Blackwellization; our work shows that PDIS and MIS belong to the extended conditional Monte Carlo method and on which our conditional importance sampling framework is built.

The extended conditional Monte Carlo method is often attributed to (Bratley et al., 1987). (Glasserman, 1993) studies the extended conditional Monte Carlo more generally under the name *filtered Monte Carlo*. The sufficient condition for variance reduction in section 5 is closely related to theorem 3.8 of (Glasserman, 1993), theorem 12 of (Glynn & Iglehart, 1988), the main theorem of (Ross, 1988) on page 310 and exercise 2.6.3 of (Bratley et al., 1987). Our results in section 6 use elements of the proof techniques of (Glynn et al., 1996; Glynn & Olvera-Cravioto, 2019) but in the context of importance sampling for per-decision and marginalized methods rather than for derivative estimation. The multiplicative structure of the importance sampling ratio in our setting renders impossible a direct application of those previous results to our setting.

Prior work has shown worst-case exponential lower bounds on the variance of IS and weighted IS (Jiang & Li, 2015; Guo et al., 2017). On the upper bound side Metelli et al. (2018, Lemma 1) provides similar upper bound as our Theorem 4, but with one difference: our bound only use the second moment of one step ratio while theirs is the ratio of the whole trajectory. Additionally, we focus on the order of variance terms and derive lower bound and upper bound for the different estimators.. However, these results are derived with respect to specific MDPs while our Theorem 4 provides general variance bounds. The recent work on stationary importance sampling (Hallak & Mannor, 2017; Gelada & Bellemare, 2019; Liu et al., 2018) has prompted multiple further investigations. First, (Xie et al., 2019) introduces a tabular marginalized importance sampling estimator to refer to the specific use of a marginalized importance sampling estimator in conjunction with an estimate of an MDP model. This idea is related to both model-based reinforcement learning and the *control variates* method for variance reduction (Bratley et al., 1987; L'Ecuyer, 1994); our work takes a different angle based on the extended conditional Monte Carlo. Our Corollary 4 about the variance of the marginalized estimator matches their $O(T^2)$ dependency on the horizon but our result holds for general spaces and does not rely on having an estimate of the reward function.

Voloshin et al. (2019) also observed empirically that stationary/marginalized importance sampling can yield a less accurate estimate than the crude importance sampling estimator or PDIS. Our analysis also considers how IS and PDIS might also vary in their accuracy, but focuses more broadly on building a theoretical understanding of those estimators and provide new variance bounds. Finally, parallel work by Kallus & Uehara (2019b) studies and analyzes incorporating control variates with marginalized importance sampling by leveraging ideas of "double" machine learning (Kallus & Uehara, 2019a; Chernozhukov et al., 2016) from semi-parametric inference. In contrast to that work, we provide a formal characterization of the variance of important sampling without control variates, and our results do not make the assumptions of a consistent value function estimator which is necessary for analysis by Kallus & Uehara (2019b).

8. Discussion

Our analysis sheds new light on the commonly held belief that the stationary/marginalized importance sampling estimators necessarily improve on their per-decision counterparts. As we show in section 3, in short-horizon settings, there exist MDPs in which the marginalized importance sampling estimator is provably worse than the per-decision one and both are worse than the crude importance sampling estimator. Furthermore, this increase in the variance occurs even if the marginalized importance sampling ratio is given as oracle. To better understand this phenomenon, we establish a new connection between the per-decision and marginalized estimators to the extended conditional Monte Carlo method. From this perspective, the potential lack of variance reduction is no longer surprising once we extend previous theoretical results from the simulation community to what we call "conditional importance sampling". This formalization help us derive sufficient conditions for variance reduction in theorems 1 and 2 for the per-decision and marginalized settings respectively.

We then reconcile our theory with the known empirical success of marginalized importance sampling through the theorems of section 6. We show that under some assumptions, the intuition regarding PDIS and MIS does hold asymptotically and their variance can be polynomial in the horizon (corollary 3 and 4 respectively) rather than exponential for the crude importance sampling estimator (theorem 4). Furthermore, we show through corollary 2 and corollary 4 that there exist conditions under which the variance of the marginalized estimator is provably lower than the variance of the per-decision estimator.

A natural next direction is exploring other statistics that better leverage the specific structure of an MDP, such as rewards, state abstractions, and find better conditional importance sampling estimator. Concurrent work (Rowland et al., 2020) shows an interesting application of the reward conditioned estimator in online TD learning. However, in the batch setting, we prove that a reward conditioned estimator with a linear regression estimator yields an estimator that is equivalent to the vanilla IS estimator (See Appendix E). This interesting result highlights the subtle differences

between online and batch settings. Exploring other statistics or lower bounds on this class of estimator is an interesting future direction.

In summary, the proposed framework of conditional importance sampling estimator both helps us understand existing estimators for batch off-policy policy evaluation and may lead to interesting future work by conditioning on different statistics.

9. Acknowledgements

The authors would like to thank Peter Glynn and Pierre L'Écuyer for the useful discussions on variance reduction techniques and the conditional Monte Carlo method. This work was supported in part by an NSF CAREER award and an ONR Young Investigator Award.

References

Bratley, P., Fox, B. L., and Schrage, L. E. *A Guide to Simulation (2Nd Ed.)*. Springer-Verlag, Berlin, Heidelberg, 1987. ISBN 0-387-96467-3.

Bucklew, J. A. *Introduction to Rare Event Simulation*. Springer New York, 2004.

Bucklew, J. A. Conditional importance sampling estimators. *IEEE transactions on information theory*, 51(1):143–153, 2005.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. K. Double machine learning for treatment and causal parameters. Technical report, cemmap working paper, 2016.

Dubi, A. and Horowitz, Y. S. The interpretation of conditional monte carlo as a form of importance sampling. *SIAM Journal on Applied Mathematics*, 36(1):115–122, 1979.

Gelada, C. and Bellemare, M. G. Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019.*, pp. 3647–3655, 2019.

Glasserman, P. Filtered monte carlo. *Mathematics of Operations Research*, 18(3):610–634, August 1993.

Glynn, P. W. Importance sampling for markov chains: asymptotics for the variance. *Communications in Statistics. Stochastic Models*, 10(4):701–717, January 1994.

Glynn, P. W. and Iglehart, D. L. Simulation methods for queues: An overview. *Queueing Systems*, 3(3):221–255, Sep 1988.

Glynn, P. W. and Olvera-Cravioto, M. Likelihood ratio gradient estimation for steady-state parameters. *Stochastic Systems*, 9(2):83–100, June 2019.

Glynn, P. W., Meyn, S. P., et al. A liapounov bound for solutions of the poisson equation. *The Annals of Probability*, 24(2):916–931, 1996.

Granovsky, B. L. Optimal formulae of the conditional monte carlo. *SIAM Journal on Algebraic Discrete Methods*, 2 (3):289–294, September 1981.

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

Hallak, A. and Mannor, S. Consistent on-line off-policy evaluation. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 1372–1383, 2017.

Hammersley, J. M. Conditional monte carlo. *J. ACM*, 3(2): 73–76, April 1956. ISSN 0004-5411.

Hesterberg, T. C. *Advances in Importance Sampling*. PhD thesis, Stanford University, August 1988.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. *arXiv preprint arXiv:1511.03722*, 2015.

Jones, G. L. et al. On the markov chain central limit theorem. *Probability surveys*, 1(299-320):5–1, 2004.

Kallus, N. and Uehara, M. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *arXiv preprint arXiv:1908.08526*, 2019a.

Kallus, N. and Uehara, M. Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes. *arXiv preprint arXiv:1909.05850*, 2019b.

L'Ecuyer, P. Efficiency improvement and variance reduction. In *Proceedings of the 26th Conference on Winter Simulation*, WSC '94, pp. 122–132, San Diego, CA, USA, 1994. Society for Computer Simulation International. ISBN 0-7803-2109-X.

L'Ecuyer, P. and Tuffin, B. Approximate zero-variance simulation. In *2008 Winter Simulation Conference*. IEEE, December 2008.

L'Ecuyer, P., Mandjes, M., and Tuffin, B. Importance sampling in rare event simulation. In *Rare Event Simulation using Monte Carlo Methods*, pp. 17–38. John Wiley & Sons, Ltd, March 2009.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pp. 5361–5371, 2018.

Metelli, A. M., Papini, M., Faccio, F., and Restelli, M. Policy optimization via importance sampling. In *Advances in Neural Information Processing Systems*, pp. 5442–5454, 2018.

Meyn, S. P. and Tweedie, R. L. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

Peshkin, L. and Shelton, C. R. Learning from scarce experience. *arXiv preprint cs/0204043*, 2002.

Precup, D., Sutton, R. S., and Singh, S. P. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pp. 759–766, 2000.

Ross, S. M. Simulating average delay–variance reduction by conditioning. *Probability in the Engineering and Informational Sciences*, 2(3):309–312, July 1988.

Rowland, M., Harutyunyan, A., van Hasselt, H., Borsa, D., Schaul, T., Munos, R., and Dabney, W. Conditional importance sampling for off-policy learning. *AISTATS*, 2020.

Rubinstein, R. Y. *Simulation and the Monte Carlo Method*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1981. ISBN 0471089176.

Srinivasan, R. Some results in importance sampling and an application to detection. *Signal Processing*, 65(1):73–88, February 1998.

Sutton, R. and Barto, A. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning series. MIT Press, 2018. ISBN 9780262039246.

Voloshin, C., Le, H. M., and Yue, Y. Empirical analysis of off-policy policy evaluation for reinforcement learning. *Real-world Sequential Decision Making Workshop at ICML 2019*, 2019.

Xie, T., Ma, Y., and Wang, Y. Optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, 2019.