SOPE: Spectrum of Off-Policy Estimators

Christina J. Yuan
University of Texas at Austin
cjyuan@cs.utexas.edu

Yash Chandak
University of Massachusetts
ychandak@cs.umass.edu

Stephen Giguere
University of Texas at Austin
sgiguere@cs.utexas.edu

Philip S. Thomas
University of Massachusetts
pthomas@cs.umass.edu

Scott Niekum
University of Texas at Austin
sniekum@cs.utexas.edu

Abstract

Many sequential decision making problems are high-stakes and require off-policy evaluation (OPE) of a new policy using historical data collected using some other policy. One of the most common OPE techniques that provides unbiased estimates is trajectory based importance sampling (IS). However, due to the high variance of trajectory IS estimates, importance sampling methods based on state-action visitation distributions (SIS) have recently been adopted. Unfortunately, while SIS often provides lower variance estimates for long horizons, estimating the state-action distribution ratios can be challenging and lead to biased estimates. In this paper, we present a new perspective on this bias-variance trade-off and show the existence of a spectrum of estimators whose endpoints are SIS and IS. Additionally, we also establish a spectrum for doubly-robust and weighted version of these estimators. We provide empirical evidence that estimators in this spectrum can be used to trade-off between the bias and variance of IS and SIS and can achieve lower mean-squared error than both IS and SIS.

1 Introduction

Many sequential decision making problems, such as automated health-care, robotics, and online recommendations are high-stakes in terms of health, safety, or finance [Liao et al., 2020, Brown et al., 2020, Theocharous et al., 2020]. For such problems, collecting new data to evaluate the performance of a new decision rule, called an evaluation policy π_e, may be expensive or even dangerous if π_e results in undesired outcomes. Therefore, one of the most important challenges in such problems is the estimation of the performance $J(\pi_e)$ of the policy π_e *before its deployment*.

Many off-policy evaluation (OPE) methods enable estimation of $J(\pi_e)$ with historical data collected using an existing decision rule, called a behavior policy π_b. One popular OPE technique is trajectory-based importance sampling (IS) [Precup, 2000]. While this method is both non-parametric and provides unbiased estimates of $J(\pi_e)$, it suffers from the *curse of horizon* and can have variance exponential in the horizon length [Jiang and Li, 2016, Guo et al., 2017]. To mitigate this problem, recent methods use stationary distribution importance sampling (SIS) to adjust the *stationary distribution* of the Markov chain induced by the policies, instead of the individual trajectories [Liu et al., 2018, Gelada and Bellemare, 2019, Nachum and Dai, 2020]. This requires (parametric) estimation of the ratio between the stationary distribution induced by π_e and π_b. Unfortunately, estimating this ratio accurately can require *unverifiable* strong assumptions on the parameters [Jiang and Huang, 2020], and often requires solving non-trivial min-max saddle point optimization problems [Yang et al., 2020]. Consequently, if the parameterization is not rich enough, then it may not be possible to represent the distribution ratios accurately, and when using rich function approximators (such as neural networks) then the optimization procedure may get stuck in sub-optimal saddle points.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

In practice, these challenges can introduce error when estimating the distribution ratio, potentially leading to arbitrarily biased estimates of $J(\pi_e)$, even when an infinite amount of data is available.

In this work, we present a new perspective on the bias-variance trade-off for OPE that bridges the unbiasedness of IS and the often lower variance of SIS. Particularly, we show that

- There exists a *spectrum* of OPE estimators whose end-points are IS and SIS, respectively.
- Estimators in this spectrum can have lower mean-squared error than both IS and SIS.
- This spectrum can also be established for doubly-robust and weighted version of IS and SIS.

In Sections 3 and 4 we show how trajectory-based and distribution-based methods can be combined. The core idea establishing the existence of this spectrum relies upon first splitting individual trajectories into two parts and then computing the probability of the first part using SIS and IS for the latter. In Section 5, we introduce weighted and doubly-robust extensions of the spectrum. Finally, in Section 6, we present empirical case studies to highlight the effectiveness of these new estimators.

2 Background

Notation: A Markov decision process (MDP) is a tuple $(\mathcal{S}, \mathcal{A}, r, T, \gamma, d_1)$, where $\mathcal{S}$ is the state set, $\mathcal{A}$ is the action set, r is the reward function, T is the transition function, γ is the discounting factor, and d_1 is the initial state distribution. Although our results extend to the continuous setting, for simplicity of notation we assume that $\mathcal{S}$ and $\mathcal{A}$ are finite. A policy π is a distribution over $\mathcal{A}$, conditioned on the state. Starting from initial state $S_1 \sim d_1$, policy π interacts with the environment iteratively by sampling action A_t at every time step t from $\pi(\cdot|S_t)$. The environment then produces reward R_t with the expected value $r(S_t, A_t)$, and transitions to the next state S_{t+1} according to $T(\cdot|S_t, A_t)$. Let $\boldsymbol{\tau} := (S_1, A_1, R_1, S_2, ..., S_L, A_L, R_L)$ be the sequence of random variables corresponding to a trajectory sampled from π, where L is the horizon length. Let p_π denote the distribution of $\boldsymbol{\tau}$ under π.

Problem Statement: The performance of any policy π is given by its value defined by the expected discounted sum of rewards $J(\pi) := \mathbf{E}_{\boldsymbol{\tau} \sim p_\pi}[\sum_{t=1}^{L} \gamma^{t-1} R_t]$. The infinite horizon setting can be obtained by letting $L \to \infty$. In general, for any random variable, we use the superscript of i to denote the trajectory associated with it. The goal of the off-policy policy evaluation (OPE) problem is to estimate the performance $J(\pi_e)$ of an evaluation policy π_e using only a batch of historical trajectories $D := \{\tau^i\}_{i=1}^m$ collected from a different behavior policy π_b. This problem is challenging because $J(\pi_e)$ must be estimated using only observational, off-policy data from the deployment of a different behavior policy π_b. Additionally, this problem might not be feasible if the data collected using π_b is not informative about the outcomes possible under π_e. Therefore, to make the problem tractable, we make the following standard support assumption, which implies that any outcome possible under π_e also has non-zero probability of occurring under π_b.

Assumption 1. *For all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, the ratio $\frac{\pi_e(a|s)}{\pi_b(a|s)} < \infty$.*

Trajectory-Based Importance Sampling: One of the earliest methods for estimating $J(\pi_e)$ is trajectory-based importance sampling. This method corrects the difference in distribution of π_b and π_e by re-weighting the trajectories from π_b in D by the probability ratio of the trajectory under π_e and π_b, i.e. $\frac{p_{\pi_e}(\tau)}{p_{\pi_b}(\tau)} = \prod_{t=1}^{L} \frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)}$. Let the single-step action likelihood ratio be denoted $\boldsymbol{\rho}_t := \frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)}$ and the likelihood ratio from steps j to k be denoted $\boldsymbol{\rho}_{j:k} := \prod_{t=j}^{k} \boldsymbol{\rho}_t$. The full-trajectory importance sampling (IS) estimator and the per-decision importance sampling (PDIS) estimator [Precup, 2000] can then be defined as:

$$\text{IS}(D) := \frac{1}{m} \sum_{i=1}^{m} \rho_{1:L}^{i} \sum_{t=1}^{L} \gamma^{t-1} R_t^i, \qquad \text{PDIS}(D) := \frac{1}{m} \sum_{i=1}^{m} \sum_{t=1}^{L} \gamma^{t-1} \rho_{1:t}^{i} R_t^i,$$

It was shown by Precup [2000] that under Assumption 1, IS(D) and PDIS(D) are unbiased estimators of $J(\pi_e)$. That is, $J(\pi_e) = \mathbf{E}_{\boldsymbol{\tau} \sim p_{\pi_b}}[\text{IS}(\boldsymbol{\tau})] = \mathbf{E}_{\boldsymbol{\tau} \sim \pi_b}[\text{PDIS}(\boldsymbol{\tau})]$. Unfortunately, however, both IS and PDIS directly depend on the product of importance ratios and thus can often suffer from exponentially high-variance in the horizon length L, known as the "curse of horizon" [Jiang and Li, 2016, Guo et al., 2017, Liu et al., 2018].

Distribution-Based Importance Sampling: To eliminate the dependency on trajectory length, recent works apply importance sampling over the state-action space rather than the trajectory space. For any policy π, let d_t^π denote the induced state-action distribution at time step t, i.e. $d_t^\pi(s,a) = p_\pi(S_t = a, A_t = a)$. Let the average state-action distribution be $d^\pi(s,a) := (\sum_{t=1}^L \gamma^{t-1} d_t^\pi(s,a))/(\sum_{t=1}^L \gamma^{t-1})$. This gives the likelihood of encountering (s,a) when following policy π and averaging over time with γ-discounting. Let $(S,A) \sim d^\pi$ and $(S,A) \sim d_t^\pi$ denote that (S,A) are sampled from d^π and d_t^π respectively. The performance of π_e can be expressed as,

$$J(\pi_e) = \mathbf{E}_{\boldsymbol{\tau} \sim p_{\pi_e}}\left[\sum_{t=1}^L \gamma^{t-1} R_t\right] = \sum_{s,a}\sum_{t=1}^L \gamma^{t-1}\, d_t^{\pi_e}(s,a) r(s,a) = \left(\sum_{t=1}^L \gamma^{t-1}\right)\sum_{s,a} d^{\pi_e}(s,a) r(s,a)$$

$$\overset{(a)}{=} \left(\sum_{t=1}^L \gamma^{t-1}\right)\sum_{s,a} d^{\pi_b}(s,a)\frac{d^{\pi_e}(s,a)}{d^{\pi_b}(s,a)} r(s,a) = \sum_{s,a}\sum_{t=1}^L \gamma^{t-1} d_t^{\pi_b}(s,a)\frac{d^{\pi_e}(s,a)}{d^{\pi_b}(s,a)} r(s,a),$$

$$= \mathbf{E}_{\boldsymbol{\tau} \sim p_{\pi_b}}\left[\sum_{t=1}^L \gamma^{t-1}\frac{d^{\pi_e}(S_t,A_t)}{d^{\pi_b}(S_t,A_t)} R_t\right],$$

where (a) is possible due to Assumption 1. Using this observation, recent works have considered the following stationary-distribution importance sampling estimator [Liu et al., 2018, Yang et al., 2020, Jiang and Huang, 2020],

$$\text{SIS}(D) := \frac{1}{m}\sum_{i=1}^m \sum_{t=1}^L \gamma^{t-1} w(S_t^i, A_t^i) R_t^i,$$

where $w(s,a) := \frac{d^{\pi_e}(s,a)}{d^{\pi_b}(s,a)}$ is the distribution correction ratio. Notice that SIS(τ) marginalizes over the product of importance ratios $\rho_{1:t}$, and thus can help in mitigating variance's dependence on horizon length for PDIS and IS estimators. When an unbiased estimate of w is available, then SIS(τ) is also an unbiased estimator, i.e., $\mathbf{E}_{\boldsymbol{\tau} \sim \pi_b}[\text{SIS}(\tau)] = J(\pi_e)$. Unfortunately, such an estimate of w is often not available. For large-scale problems, parametric estimation w is required in practice and we replace the true density ratios w with an estimate $\hat{w}$. However, estimating w accurately may require both a non-verifiable strong assumption on the parametric function class, and global solution to a non-trivial min-max optimization problem [Jiang and Huang, 2020, Yang et al., 2020]. When these conditions are not met, SIS estimates can be arbitrarily biased, even when an infinite amount of data is available.

3 Combining Trajectory-Based and Density-Based Importance Sampling

Trajectory-based and distribution-based importance sampling methods are typically presented as alternative methods of applying importance sampling for off-policy evaluation. However, in this section we show that the choice of estimator is not binary, and these two styles of computing importance weights can actually be combined into a single importance sampling estimate. Furthermore, using this combination, in the next section, we will derive a spectrum of estimators that allows interpolation between the trajectory-based PDIS and distribution-based SIS, which will often allow us trade-off between the strengths and weaknesses of these methods.

Intuitively, trajectory-based and distribution-based importance sampling provide two different ways of correcting the distribution mismatch under the evaluation and behavior policies. Trajectory-based importance sampling corrects the distribution mismatch by examining how likely policies are to take the same sequence of actions and thus applies the action likelihood ratio as the correction term. Distribution-based importance sampling corrects the mismatch by how likely policies are to visit the same state and action pairs—while remaining agnostic to *how* they arrived—and applies the distribution ratio as the importance weight. However, using distribution ratio and action likelihood ratio correction terms are not mutually exclusive, and one can draw on both types of correction terms to derive combined estimators.

To build intuition for why likelihood ratios and distribution ratios can naturally be combined, we consider the two rooms domain shown in Figure 3. In this example, there are two policies π_b, π_e

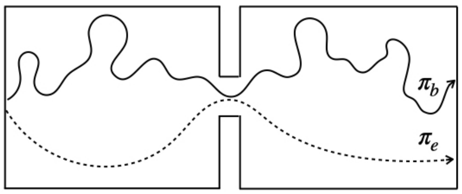


Figure 1: Illustration of two room domain. The domain consists of two rooms, the left room and the right room separated by a connecting door. π_b and π_e are two different policies that move from the left room to the right room. Note that, although π_b and π_e have two different behaviors in the left room and right room, both pass through the connecting door.

which have different strategies for navigating from the first room to the second room. Note that while the behavior of the two policies are very different in the left room, both policies must pass through the connecting door to get to the right room at some point in time. Conditioning on having passed through the connecting door at a point in time, all parts of the trajectory that occur in the right room are independent from what has occurred in the left room by the Markov property. Thus, when considering a reward R_t that occurs in the right room, it is natural to consider the probability of reaching the door and then the probability of the action sequence policy in the right room under each policy.

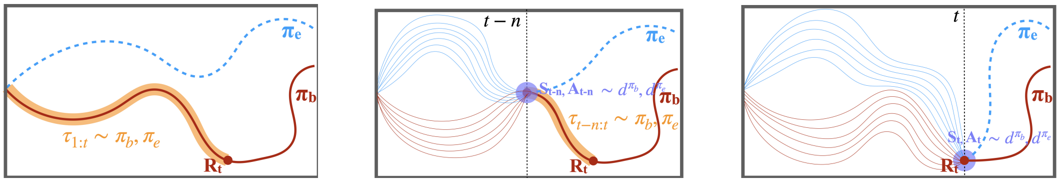
Now, we formalize this intuition and show how trajectory-based and density-based importance sampling can be combined in the same estimator. Given a trajectory $\boldsymbol{\tau}$, we can consider (S_z, A_z), the state and action at time z in the trajectory. By conditioning on (S_z, A_z), trajectory $\boldsymbol{\tau}$ can be separated into two conditionally independent partial trajectories $\boldsymbol{\tau}_{0:z}$ and $\boldsymbol{\tau}_{z+1,L}$ by the Markov property. Since the segments of $\boldsymbol{\tau}$ before and after time z are conditionally independent, then $\rho_{1:z}$, the likelihood ratio for the trajectory before time z, is conditionally independent from $\rho_{z+1:L}$ and from R_t for all $t \geq z$. Formally, let $(S_z, A_z) \sim d_z^{\pi_b}$, then,

$$
\begin{aligned}
J(\pi_e) &= \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}[\text{PDIS}(\boldsymbol{\tau})] = \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\sum_{t=1}^{L}\gamma^{t-1}\rho_{1:t}R_t\right] \\
&= \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\sum_{t=1}^{z}\gamma^{t-1}\rho_{1:t}R_t\right] + \mathbf{E}_{\substack{(S_z,A_z)\\ \sim d_z^{\pi_b}}}\left[\mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\sum_{t=z+1}^{L}\gamma^{t-1}\rho_{1:z}\rho_{z+1:t}R_t\middle| S_z, A_z\right]\right] \\
&= \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\sum_{t=1}^{z}\gamma^{t-1}\rho_{1:t}R_t\right] + \mathbf{E}_{\substack{(S_z,A_z)\\ \sim d_z^{\pi_b}}}\left[\sum_{t=z+1}^{L}\gamma^{t-1}\mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\rho_{1:z}|S_z, A_z\right]\mathbf{E}_{\boldsymbol{\tau}\sim \pi_b}\left[\rho_{z+1:t}R_t|S_z, A_z\right]\right] \\
&\overset{(a)}{=} \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\sum_{t=1}^{z}\gamma^{t-1}\rho_{1:t}R_t\right] + \mathbf{E}_{\substack{(S_z,A_z)\\ \sim d_z^{\pi_b}}}\left[\sum_{t=z+1}^{L}\gamma^{t-1}\frac{d_z^{\pi_e}(S_z, A_z)}{d_z^{\pi_b}(S_z, A_z)}\mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\rho_{z+1:t}R_t\middle| S_z, A_z\right]\right] \\
&= \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\sum_{t=1}^{z}\gamma^{t-1}\rho_{1:t}R_t + \sum_{t=z+1}^{L}\gamma^{t-1}\frac{d_z^{\pi_e}(S_z, A_z)}{d_z^{\pi_b}(S_z, A_z)}\rho_{z+1:t}R_t\right]. \qquad (1)
\end{aligned}
$$

where (a) follows from the following Property 1, which states that the expected value of product likelihood ratios $\rho_{1:z}$ conditioned on (S_z, A_z) is equal to the time-dependent state-action distribution ratio for (S_z, A_z). We provide a detailed proof of Property 1 in Appendix A.

Property 1 ([Liu et al., 2018])**.** *Under Assumption 1,* $\mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}[\rho_{1:t}|S_t = s, A_t = a] = \frac{d_t^{\pi_e}(s,a)}{d_t^{\pi_b}(s,a)}$.

Observe that Eq (1) is indexed by time z. Intuitively, z can be thought of as the time to switch from using distribution ratios to action likelihood ratios in the importance weight. Specifically, the distribution ratios are used to estimate the probability of being in state S_z and taking action A_z at time z and action likelihood ratios are used to correct for the probability of actions taken after time z. Further observe that z does not have to be a fixed constant—$z(t)$ can be a function of t so that each reward in the trajectory R_t can utilize a different switching time. In the next section, we show that by using a function $z(t)$ that allows the switching time to be time-dependent, we are able to further marginalize over time and create an estimator that interpolates between *average* state-action distribution ratios $w(s,a) = \frac{d^{\pi_e}(s,a)}{d^{\pi_b}(s,a)}$, rather than time-dependent distribution ratios $\frac{d_t^{\pi_e}(s,a)}{d_t^{\pi_b}(s,a)}$.


π_e
π_b
τ_{1:t} ~ π_b, π_e
R_t
t − n
S_{t-n}, A_{t-n} ~ d^{π_b}, d^{π_e}
τ_{t−n:t} ~ π_b, π_e
R_t
t
S_t, A_t ~ d^{π_b}, d^{π_e}
R_t

(a) PDIS

(b) SOPE$_n$

(c) SIS

Figure 2: Illustrations of the PDIS, SOPE$_n$ and SIS estimators. The dotted blue line represents an example trajectory drawn from π_e, and the solid red line represents an example trajectory from π_b. All three importance sampling methods work by re-weighting each reward R_t in the trajectory from π_b. (a) Trajectory-based PDIS works by re-weighting each reward by $\frac{p_{\pi_e}(\tau_{1:t})}{p_{\pi_b}(\tau_{1:t})}$, the probability ratio of the sub-trajectory leading up to R_t under the π_b and π_e, respectively. This factors into $\rho_{1:t}$, the product of t action likelihood ratios. (c) Distribution-based SIS considers the probability of encountering (S_t, A_t) under π_e and π_b, and re-weights R_t by $\frac{d^{\pi_e}(S_t,A_t)}{d^{\pi_b}(S_t,A_t)}$, (b) SOPE$_n$ combines trajectory and distribution importance sampling weights by considering the probability of each policy visiting (S_{t-n}, A_{t-n}), the state-action pair n steps in the past, and additionally the probability of the sub-trajectory $\tau_{t-n+1:t}$ from n steps in the past to t. Thus, SOPE$_n$ re-weights R_t by $\frac{d^{\pi_e}(S_{t-n},A_{t-n})}{d^{\pi_b}(S_{t-n},A_{t-n})}\rho_{t-n+1:t}$.

4 Bias-Variance Trade-off using n-step Interpolation Between PDIS and SIS

We now build upon the ideas from Section 3 to derive a spectrum of off-policy estimators that allows for interpolation between the trajectory-based PDIS and distribution-based SIS estimators. This spectrum contains PDIS and SIS at the endpoints and allows for smooth interpolation between them to obtain new estimators that can often trade-off the strengths and weaknesses of PDIS and SIS. An illustration of the key idea can be found in Figure 2.

One simple way to perform this trade-off is to control the number of terms in the product in the action likelihood ratio for each reward R_t. Specifically, for any reward R_t, we propose including only the n most recent action likelihood ratios $\boldsymbol{\rho}_{t-n+1:t}$ in the importance weight, rather than $\boldsymbol{\rho}_{1:t}$. Thus, the overall importance weight becomes the re-weighted probability of visiting (S_{t-n}, A_{t-n}), followed by the re-weighted probability of taking the last n actions leading up to reward R_t. This reduces the exponential impact that horizon length L has on the variance of PDIS, and provides control over this reduction via the parameter n. To get an estimator to perform this trade-off, we start with the derivation in (1) with $z(t) = t - n$, then accumulate the time-dependent state-action distributions d_t over time. The final expression for the finite horizon setting requires some additional constructs and is thus presented along with its derivations and additional discussion in Appendix B. In the following we present the result for the infinite horizon setting.

$$J(\pi_e) = \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\sum_{t=1}^{n}\gamma^{t-1}\rho_{1:t}R_t + \sum_{t=n+1}^{\infty}\gamma^{t-1}\frac{d^{\pi_e}(S_{t-n}, A_{t-n})}{d^{\pi_b}(S_{t-n}, A_{t-n})}\rho_{t-n+1:t}R_t\right]. \tag{2}$$

Using the sample estimate of (2), we obtain the Spectrum of Off-Policy Estimators (SOPE$_n$),

$$\text{SOPE}_n(D) = \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{t=1}^{n}\gamma^{t-1}\rho^i_{1:t}R^i_t + \sum_{t=n+1}^{\infty}\gamma^{t-1}\hat{w}(S^i_{t-n}, A^i_{t-n})\rho^i_{t-n+1:t}R^i_t\right).$$

Remark 1. *Note that since we generally do not have access to the true density ratios, in practice we substitute w with the estimated density ratios $\hat{w}$ similarly as in SIS. Since SOPE$_n$ is agnostic to how $\hat{w}$ is estimated, it can readily leverage existing and new methods for estimating $\hat{w}$.*

Observe that SOPE$_n$ doesn't just give a single estimator, but a spectrum of off-policy estimators indexed by n. An illustration of this spectrum can be seen in Figure 3. As n decreases, the number of terms in the action likelihood ratio decreases, and SOPE$_n$ depends more on the distribution correction ratio and is more like SIS. Likewise as n increases, the number of terms in the action likelihood ratio increases, and SOPE$_n$ is closer to PDIS. Further note that that for the endpoint values of this

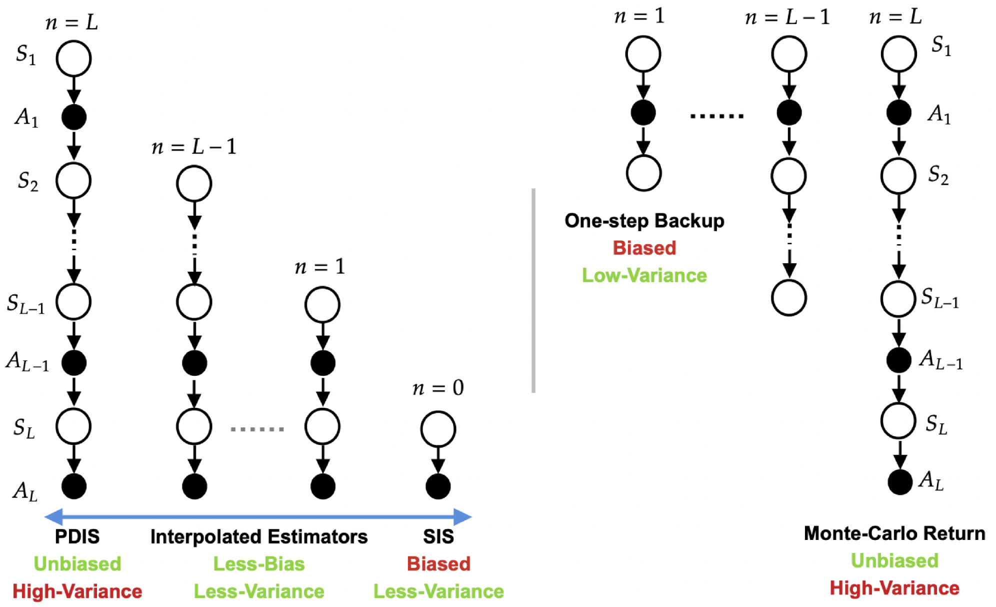


Figure 3: On the left side of the figure, we show an illustration the SOPE$_n$ spectrum of estimators. For the purpose of this illustration, consider that only at the last time step there is a non-zero reward R_L. The SOPE$_n$ spectrum allows for control of how much an estimate depends on distribution ratios vs action likelihood ratios. Notice that SOPE$_0$ results in SIS, SOPE$_L$ results in PDIS estimator, and other values of n result in new interpolated estimators. As an analogy, consider the backup-diagram [Sutton and Barto, 2018] for the n-step q-estimate as illustrated on the right-hand side of the solid vertical line. Notice that in the n-step q-estimate, returns are backed up from possible *future outcomes*, whereas in the n-step interpolation estimators the probabilities are 'backed-up' from the possible *histories*. (In the diagram, bias-variance characterization of PDIS and SIS is based on typical practical observations [Voloshin et al., 2019, Fu et al., 2021], however it is worth noting that SIS is not biased when oracle density ratios are available, and there are also edge cases, particularly for short horizon problems, where SIS can have higher variance than PDIS [Liu et al., 2020, Metelli et al., 2020]).

spectrum, $n = 0$ and $n = L$, SOPE$_n$ gives the SIS and PDIS estimators exactly (for PDIS, horizon length needs to be L instead of ∞ for the estimator to be well defined),

$$\text{SOPE}_0(D) = \frac{1}{m}\sum_{i=1}^{m}\sum_{t=1}^{L}\gamma^{t-1}w(S_t^i, A_t^i)R_t^i = \text{SIS}(D),$$

$$\text{SOPE}_L(D) = \frac{1}{m}\sum_{i=1}^{m}\sum_{t=1}^{L}\gamma^{t-1}\rho_{1:t}^i R_t^i = \text{PDIS}(D).$$

5 Doubly-Robust and Weighted IS Extensions to SOPE$_n$

An additional advantage of SOPE$_n$ is that it can be readily extended to obtain a spectrum for other estimators. For instance, to mitigate variance further a popular technique is to leverage domain knowledge from (imperfect) models using doubly-robust estimators [Jiang and Li, 2016, Jiang and Huang, 2020]. In the following we can create a doubly robust version of the SOPE$_n$ estimator.

Before moving further, we introduce some additional notation. Let,

$$w(t,n) := \begin{cases} \frac{d^{\pi_e}(S_{t-n},A_{t-n})}{d^{\pi_b}(S_{t-n},A_{t-n})}\left(\prod_{j=0}^{n-1}\frac{\pi_e(A_{t-j}|S_{t-j})}{\pi_b(A_{t-j}|S_{t-j})}\right) & \text{if } t > n \\ \prod_{j=1}^{t}\frac{\pi_e(A_j|S_j)}{\pi_b(A_j|S_j)} & 1 \le t \le n \\ 1 & \text{otherwise} \end{cases}$$

Let q be an estimate for the q-value function for π_e, computed using the (imperfect) model. For brevity, we make the random variable $\boldsymbol{\tau} \sim p_{\pi_b}$ implicit for the expectations in this section. For a given value of n, performance (2) of π_e can then be expressed as,

$$J(\pi_e) = \mathbf{E}\left[\sum_{t=1}^{\infty} w(t,n)\gamma^{t-1}R_t\right].$$

We now use this form to create a spectrum of doubly-robust estimators,

$$\begin{aligned}
J(\pi_e) &= \mathbf{E}\left[\sum_{t=1}^{\infty} w(t,n)\gamma^{t-1}R_t\right] + \underbrace{\mathbf{E}\left[\sum_{t=1}^{\infty} w(t,n)\gamma^{t-1}q(S_t,A_t)\right] - \mathbf{E}\left[\sum_{t=1}^{\infty} w(t,n)\gamma^{t-1}q(S_t,A_t)\right]}_{=0} \\
&\overset{(a)}{=} \mathbf{E}\left[\sum_{t=1}^{\infty} w(t,n)\gamma^{t-1}R_t\right] + \mathbf{E}\left[\sum_{t=1}^{\infty} w(t-1,n)\gamma^{t-1}q(S_t,A_t^{\pi_e})\right] - \mathbf{E}\left[\sum_{t=1}^{\infty} w(t,n)\gamma^{t-1}q(S_t,A_t)\right] \\
&= \mathbf{E}\left[w(0,n)\gamma^0 q(S_1,A_1^{\pi_e})\right] + \mathbf{E}\left[\sum_{t=1}^{\infty} w(t,n)\gamma^{t-1}\Big(R_t + \gamma q(S_{t+1},A_{t+1}^{\pi_e}) - q(S_t,A_t)\Big)\right] \\
&= \mathbf{E}\Big[q(S_1,A_1^{\pi_e})\Big] + \mathbf{E}\left[\sum_{t=1}^{\infty} w(t,n)\gamma^{t-1}\Big(R_t + \gamma q(S_{t+1},A_{t+1}^{\pi_e}) - q(S_t,A_t)\Big)\right], \qquad (3)
\end{aligned}$$

where in (a) we used the notation $A_t^{\pi_e}$ to indicate the $A_t \sim \pi_e(\cdot|S_t)$. Using $A_t^{\pi_e}$ eliminates the need for correcting A_t sampled under π_b. We define DR-SOPE$_n(D)$ to be the sample estimate of (3), i.e., a doubly-robust form for the SOPE$_n(D)$ estimator. It can now be observed that existing doubly-robust estimators are end-points of DR-SOPE$_n(D)$ (for trajectory-wise settings, horizon length needs to be L instead of ∞ for the estimator to be well defined),

DR-SOPE$_L(D)$ = Trajectory-wise DR [Jiang and Li, 2016, Thomas and Brunskill, 2016],
DR-SOPE$_0(D)$ = State-action distribution DR [Jiang and Huang, 2020, Kallus and Uehara, 2020].

A variation of PDIS that can often also help in mitigating the variance of PDIS method is the Consistent Weighted Per-Decision Importance Sampling estimator (CWPDIS) [Thomas, 2015]. CWDPIS renormalizes the importance ratio at each time with the sum of importance weights, which causes CWPDIS to be biased (but consistent) and often have lower variance than PDIS.

$$\text{CWPDIS}(D) := \sum_{t=1}^{L} \gamma^{t-1} \frac{\sum_{i=1}^{m} \rho_{1:t}^i R_t^i}{\sum_{i=1}^{m} \rho_{1:t}^i}.$$

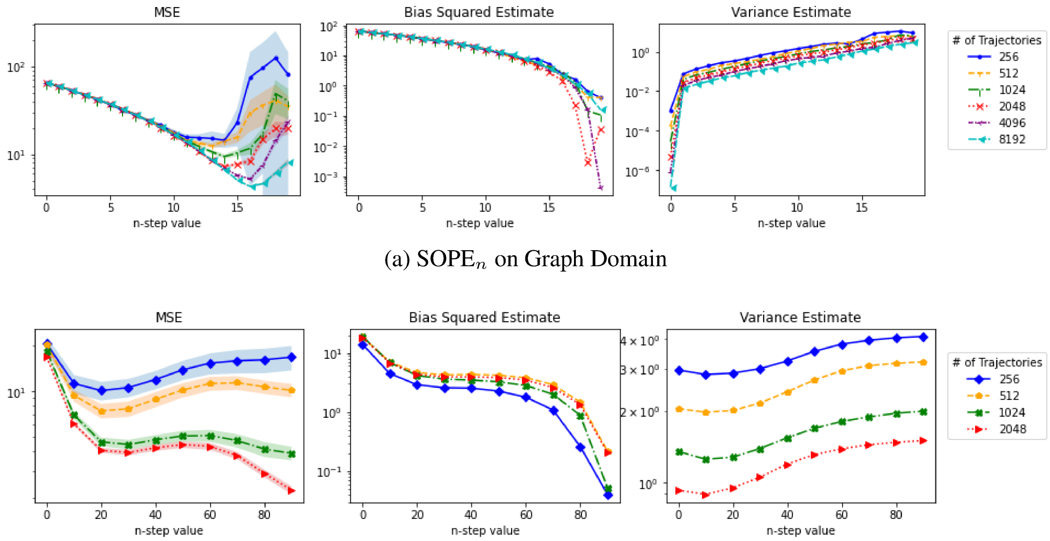
Similar DR-SOPE$_n$, we can create a weighted version of SOPE$_n$ estimator that interpolates between a weighted-version of SIS and CWPDIS:

$$\text{W-SOPE}_n(D) := \sum_{t=1}^{n}\left(\gamma^{t-1}\sum_{i=1}^{m}\frac{\rho_{1:t}^i}{\sum_{i=1}^{m}\rho_{1:t}^i}R_t^i\right) + \sum_{t=n+1}^{\infty}\left(\gamma^{t-1}\sum_{i=1}^{m}\frac{w(S_{t-n}^i, A_{t-n}^i)\rho_{t-n+1:t}^i}{\sum_{i=1}^{m} w(S_{t-n}^i, A_{t-n}^i)\rho_{t-n+1:t}^i}R_t^i\right).$$

Since, unlike PDIS, CWPDIS is a biased (but consistent) estimator, W-SOPE$_n$ interpolates between two biased estimators as endpoints. Nonetheless, we show experimentally in Section 6 that in practice even W-SOPE$_n$ can allow for bias-variance trade-off.

6 Experimental Results

In this section, we present experimental results showing that interpolated estimators within the SOPE$_n$ and W-SOPE$_n$ spectrums can outperform the SIS/weighted-SIS and PDIS/CWPDIS endpoints. In each experiment, we evaluate SOPE$_n$ and W-SOPE$_n$ for different values of n ranging from 0 to L. This allows us to compare the different estimators we get for each n and see trends of how the



(a) SOPE$_n$ on Graph Domain

(b) SOPE$_n$ on Toy Mountain Car Domain

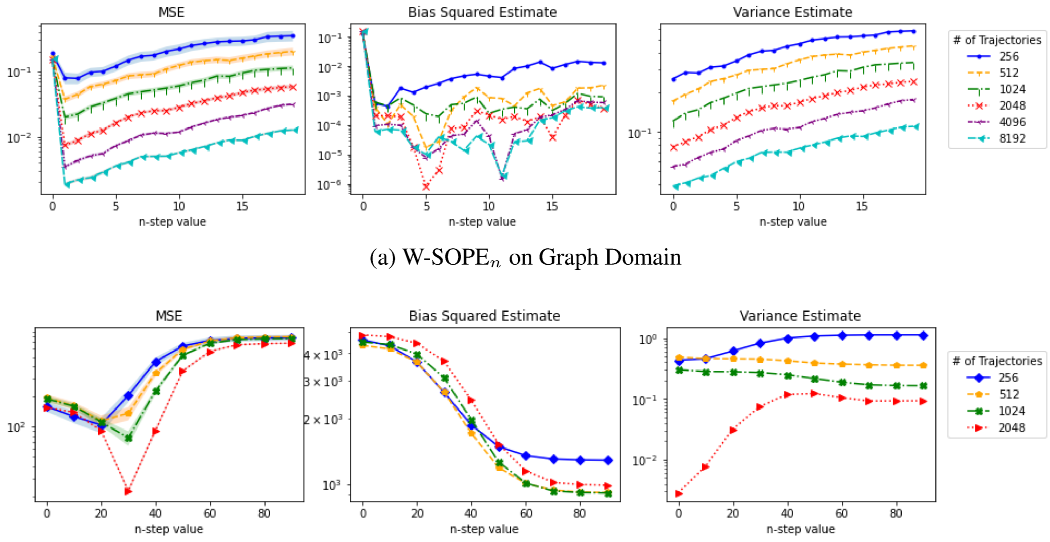
Figure 4: Experimental results from evaluating the SOPE$_n$ estimator on the Graph and Toy Mountain Car domains. The x-axis for each plot indicates the value of n in the SOPE$_n$ estimate. The shaded regions denote 95% confidence regions on the mean of MSE. Recall that SOPE$_0$ gives SIS and SOPE$_L$ gives PDIS. The evaluation and behavior policies are $\pi_e(a = 0) = 0.9$ and $\pi_b(a = 0) = 0.5$ for the experiments on the Graph Domain and and $\pi_e(a = 0) = 0.5$ and $\pi_b(a = 0) = 0.6$ for the Toy Mountain Car domain. In both these domains, we can see that there exist interpolating estimators in the SOPE$_n$ spectrum that outperform SIS and PDIS, and that the SOPE$_n$ spectrum empirically performs a bias-variance trade-off.

performance changes as n varies. Additionally, we plot estimates of the bias and the variance for the different values of n to further investigate the properties of estimators in this spectrum.

For our experiments, we utilize the environments and implementations of baseline estimators in the Caltech OPE Benchmarking Suite (COBS) [Voloshin et al., 2019]. In this section, we present results on the Graph and Toy Mountain Car environments. To obtain an estimate of the density ratios $\hat{w}$, we use COBS's implementation of infinite horizon methods from [Liu et al., 2018]. Full experimental details and additional experimental results can be found in Appendix D. Additional experiments include an investigation on the impact on the degree of π_e and π_b mismatch on SOPE$_n$ and W-SOPE$_n$, as well as additional experiments on the Mountain Car domain.

The experimental results for the SOPE$_n$ and W-SOPE$_n$ estimators can be seen in Figures 4 and 5 respectively. We observe that for both SOPE$_n$ and W-SOPE$_n$, the plots of mean-squared error (MSE) have a U-shape indicating that there exist interpolated estimators within the spectrum with lower MSE than the endpoints. Additionally, from the bias and variance plots, we can see that SOPE$_n$ and W-SOPE$_n$ perform a bias-variance trade-off in these experiments. We observe that as n increases and the estimators become closer to PDIS/CWPDIS, the bias decreases but the variance increases. Likewise, as n decreases and the estimators become closer to SIS/weighted-SIS, the variance decreases but the bias increases. This bias-variance trade-off trend is more notable for the unweighted SOPE$_n$ which trades-off between biased SIS and unbiased PDIS endpoints. However, we still can see this trend even with the W-SOPE$_n$ estimator, although the trade-off is not as clean because W-SOPE interpolates between biased SIS and the also biased (but consistent) CWPDIS.

Finally, note that our plots also show the results for different batch sizes of historical data. In our plots, as batch size increases, for some domains the PDIS/CWPDIS endpoints eventually outperform the SIS/weighted-SIS endpoints. However, even in this case, there still exist interpolated estimators that outperform both endpoints.



(a) W-SOPE$_n$ on Graph Domain

(b) W-SOPE$_n$ on Toy Mountain Car Domain

Figure 5: Experimental results from evaluating the W-SOPE$_n$ estimator on the Graph and Toy Mountain Car domains. The x-axis for each plot indicates the value of n in the SOPE$_n$ estimate. The shaded regions denote 95% confidence regions on the mean of MSE. Recall that W-SOPE$_0$ gives weighted-SIS and W-SOPE$_L$ gives CWPDIS. The evaluation and behavior policies are $\pi_e(a=0) = 0.9$ and $\pi_b(a=0) = 0.7$ for the experiments on the Graph Domain and and $\pi_e(a=0) = 0.5$ and $\pi_b(a=0) = 0.9$ for the Toy Mountain Car domain. In both these domains, we can see that there exist interpolating estimators in the W-SOPE$_n$ spectrum that outperform SIS and PDIS. Similar to Figure 4, we see that W-SOPE also performs a bias-variance trade-off, however, is not as clean since the CWPDIS endpoint of the spectrum is biased (but consistent).

7 Related Work

Off-policy evaluation (also related to counterfactual inference in the causality literature [Pearl, 2009]) is one the most crucial aspects of RL, and importance sampling [Metropolis and Ulam, 1949, Horvitz and Thompson, 1952] plays a central role in it. Precup [2000] first introduced IS, PDIS, and WIS estimates for OPE. Since then there has been a flurry of research in this direction: using partial-models to develop doubly robust estimators [Jiang and Li, 2016, Thomas and Brunskill, 2016], using multi-importance sampling [Papini et al., 2019, Metelli et al., 2020], estimating the behavior policy [Hanna et al., 2019], clipping importance ratios [Bottou et al., 2013, Thomas et al., 2015, Munos et al., 2016, Schulman et al., 2017], dropping importance ratios [Guo et al., 2017], importance sampling the entire return distribution [Chandak et al., 2021], importance resampling of trajectories [Schlegel et al., 2019], emphatic weighting of TD methods [Mahmood et al., 2015, Hallak et al., 2016, Patterson et al., 2021], and estimating state-action distributions [Hallak and Mannor, 2017, Liu et al., 2018, Gelada and Bellemare, 2019, Xie et al., 2019, Nachum and Dai, 2020, Yang et al., 2020, Zhang et al., 2020, Jiang and Huang, 2020, Uehara et al., 2020].

Perhaps the most relevant to our work are the recent works by Liu et al. [2020] and Rowland et al. [2020] that use the conditional IS (CIS) framework to show how IS, PDIS, and SIS are special instances of CIS. Similarly, our proposed method for combining trajectory and density-based importance sampling also falls under the CIS framework. Liu et al. [2020] also showed that in the finite horizon setting, none of IS, PDIS, or SIS has variance *always* lesser than the other. Similarly, Rowland et al. [2020] used sufficient conditional functions to create new off-policy estimators and showed that return conditioned estimates (RCIS) can provide optimal variance reduction. However, using RCIS requires a challenging task of estimating *density ratios for returns* (not state-action pair) and Liu et al. [2020] established a negative result that estimating these ratios using linear regression may result in the IS estimate itself.

Our analysis complements these recent works by showing that there exists interpolated estimators that can provide lower variance estimates than any of IS, PDIS, or SIS. Our proposed estimator SOPE$_n$

provides a natural interpolation technique to trade-off between the strengths and weaknesses of these trajectory and density based methods. Additionally, while it is known that $q^\pi(s,a)$ and $d^\pi(s,a)$ have a primal-dual connection [Wang et al., 2007], our time-based interpolation technique also sheds new light on connections between their n-step generalizations.

8 Conclusions

We present a new perspective in off-policy evaluation connecting two popular estimators, PDIS and SIS, and show that PDIS and SIS lie as endpoints on the Spectrum of Off-Policy Estimators SOPE$_n$ which interpolates between them. Additionally, we also derive a weighted and doubly robust version of this spectrum of estimators. With our experimental results, we illustrate that estimators that lie on the interior of the SOPE$_n$ and W-SOPE$_n$ spectrums can be used outperform their endpoints SIS/weighted-SIS and PDIS/CWPDIS.

While we are able to show there exist SOPE$_n$ estimators that are able to outperform PDIS and SIS, it remains as future work to devise strategies to automatically select n to trade-off bias and variance. Future directions may include developing methods to select n or combine all estimators for all n using λ-trace methods [Sutton and Barto, 2018] to best trade-off bias and variance.

Finally, like all off-policy evaluation methods, our approach carries risks if used inappropriately. When using OPE for sensitive or safety-critical applications such as medical domains, caution should be taken to carefully consider the variance and bias of the estimator that is used. In these cases, high-confidence OPE methods [Thomas et al., 2015] may be more appropriate.

9 Acknowledgement

We thank members of the Personal Autonomous Robotics Lab (PeARL) at the University of Texas at Austin for discussion and feedback on early stages of this work. We especially thank Jordan Schneider, Harshit Sikchi, and Prasoon Goyal for reading and giving suggestions on early drafts. We additionally thank Ziyang Tang for suggesting an additional marginalization step in the main proof that helped us unify the results for finite and infinite horizon setting. The work also benefited from feedback by Nan Jiang during initial stages of this work. We would also like to thank the anonymous reviewers for their suggestions which helped improve the paper.

This work has taken place in part in the Personal Autonomous Robotics Lab (PeARL) at The University of Texas at Austin. PeARL research is supported in part by the NSF (IIS-1724157, IIS-1638107, IIS-1749204, IIS-1925082), ONR (N00014-18-2243), AFOSR (FA9550-20-1-0077), and ARO (78372-CS). This research was also sponsored by the Army Research Office under Cooperative Agreement Number W911NF-19-2-0333, a gift from Adobe, NSF award #2018372, and the DEVCOM Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196 (ARL IoBT CRA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Daniel Brown, Russell Coleman, Ravi Srinivasan, and Scott Niekum. Safe imitation learning via fast bayesian reward inference from preferences. In *International Conference on Machine Learning*, pages 1165–1177. PMLR, 2020.

Yash Chandak, Scott Niekum, Bruno Castro da Silva, Erik Learned-Miller, Emma Brunskill, and Philip S Thomas. Universal off-policy evaluation. *arXiv preprint arXiv:2104.12820*, 2021.

Justin Fu, Mohammad Norouzi, Ofir Nachum, George Tucker, Ziyu Wang, Alexander Novikov, Mengjiao Yang, Michael R Zhang, Yutian Chen, Aviral Kumar, et al. Benchmarks for deep off-policy evaluation. *arXiv preprint arXiv:2103.16596*, 2021.

Carles Gelada and Marc G Bellemare. Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3647–3655, 2019.

Zhaohan Daniel Guo, Philip S Thomas, and Emma Brunskill. Using options and covariance testing for long horizon off-policy policy evaluation. *arXiv preprint arXiv:1703.03453*, 2017.

Assaf Hallak and Shie Mannor. Consistent on-line off-policy evaluation. In *International Conference on Machine Learning*, pages 1372–1383. PMLR, 2017.

Assaf Hallak, Aviv Tamar, Rémi Munos, and Shie Mannor. Generalized emphatic temporal difference learning: Bias-variance analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Josiah Hanna, Scott Niekum, and Peter Stone. Importance sampling policy evaluation with an estimated behavior policy. In *International Conference on Machine Learning*, pages 2605–2613. PMLR, 2019.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Nan Jiang and Jiawei Huang. Minimax confidence interval for off-policy evaluation and policy optimization. *arXiv preprint arXiv:2002.02081*, 2020.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR, 2016.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020.

Peng Liao, Predrag Klasnja, and Susan Murphy. Off-policy estimation of long-term average outcomes with applications to mobile health. *Journal of the American Statistical Association*, pages 1–10, 2020.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *arXiv preprint arXiv:1810.12429*, 2018.

Yao Liu, Pierre-Luc Bacon, and Emma Brunskill. Understanding the curse of horizon in off-policy evaluation via conditional importance sampling. In *International Conference on Machine Learning*, pages 6184–6193. PMLR, 2020.

A Rupam Mahmood, Huizhen Yu, Martha White, and Richard S Sutton. Emphatic temporal-difference learning. *arXiv preprint arXiv:1507.01569*, 2015.

Alberto Maria Metelli, Matteo Papini, Nico Montali, and Marcello Restelli. Importance sampling techniques for policy optimization. *Journal of Machine Learning Research*, 21(141):1–75, 2020.

Nicholas Metropolis and Stanislaw Ulam. The monte carlo method. *Journal of the American statistical association*, 44(247):335–341, 1949.

Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc G Bellemare. Safe and efficient off-policy reinforcement learning. *arXiv preprint arXiv:1606.02647*, 2016.

Ofir Nachum and Bo Dai. Reinforcement learning via Fenchel-Rockafellar duality. *arXiv preprint arXiv:2001.01866*, 2020.

Matteo Papini, Alberto Maria Metelli, Lorenzo Lupo, and Marcello Restelli. Optimistic policy optimization via multiple importance sampling. In *International Conference on Machine Learning*, pages 4989–4999. PMLR, 2019.

Andrew Patterson, Sina Ghiassian, D Gupta, A White, and M White. Investigating objectives for off-policy value estimation in reinforcement learning, 2021.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

Mark Rowland, Anna Harutyunyan, Hado Hasselt, Diana Borsa, Tom Schaul, Rémi Munos, and Will Dabney. Conditional importance sampling for off-policy learning. In *International Conference on Artificial Intelligence and Statistics*, pages 45–55. PMLR, 2020.

Matthew Schlegel, Wesley Chung, Daniel Graves, Jian Qian, and Martha White. Importance resampling for off-policy prediction. *arXiv preprint arXiv:1906.04328*, 2019.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT Press, Cambridge, MA, 2 edition, 2018.

Georgios Theocharous, Yash Chandak, Philip S Thomas, and Frits de Nijs. Reinforcement learning for strategic recommendations. *arXiv preprint arXiv:2009.07346*, 2020.

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

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Philip S Thomas. *Safe reinforcement learning*. PhD thesis, University of Massachusetts Libraries, 2015.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR, 2020.

Cameron Voloshin, Hoang M Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. *arXiv preprint arXiv:1911.06854*, 2019.

Tao Wang, Michael Bowling, and Dale Schuurmans. Dual representations for dynamic programming and reinforcement learning. In *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, pages 44–51. IEEE, 2007.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *arXiv preprint arXiv:1906.03393*, 2019.

Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-policy evaluation via the regularized lagrangian. *arXiv preprint arXiv:2007.03438*, 2020.

Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. Gendice: Generalized offline estimation of stationary values. *arXiv preprint arXiv:2002.09072*, 2020.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 8
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 8
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Appendix
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proof of Lemma 1

Liu et al. [2018] first showed that stationary importance sampling methods can be viewed as Rao-Blackwellization of IS estimator, and claimed that the expectation of the likelihood-ratios conditioned on state and action is equal to the distribution ratio, as stated in Property 1. For completeness, we present a proof of Property 1. Recall that $d_t^\pi(s,a) = p_\pi(S_t = s, A_t = a)$.

$$
\begin{aligned}
&\mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\rho_{1:t} | S_t = s, A_t = a\right] \\
&= \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\frac{p_{\pi_e}(\boldsymbol{\tau}_{1:t})}{p_{\pi_b}(\boldsymbol{\tau}_{1:t})}\middle| S_t = s, A_t = a\right] \\
&= \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\frac{p_{\pi_e}(S_1, A_1, \dots, S_t, A_t)}{p_{\pi_b}(S_1, A_1, \dots, S_t, A_t)}\middle| S_t = s, A_t = a\right] \\
&= \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\frac{p_{\pi_e}(S_1, A_1, \dots, S_t, A_t)}{p_{\pi_e}(S_t, A_t)}\frac{p_{\pi_b}(S_t, A_t)}{p_{\pi_b}(S_1, A_1, \dots, S_t, A_t)}\frac{p_{\pi_e}(S_t, A_t)}{p_{\pi_b}(S_t, A_t)}\middle| S_t = s, A_t = a\right] \\
&= \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\frac{p_{\pi_e}(\boldsymbol{\tau}_{1:t} | S_t, A_t)}{p_{\pi_b}(\boldsymbol{\tau}_{1:t} | S_t, A_t)}\middle| S_t = s, A_t = a\right]\frac{p_{\pi_e}(S_t = s, A_t = a)}{p_{\pi_b}(S_t = s, A_t = a)} \\
&\overset{(a)}{=} \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\frac{p_{\pi_e}(\boldsymbol{\tau}_{1:t} | S_t, A_t)}{p_{\pi_b}(\boldsymbol{\tau}_{1:t} | S_t, A_t)}\middle| S_t = s, A_t = a\right]\frac{d_t^{\pi_e}(s,a)}{d_t^{\pi_b}(s,a)} \\
&= \left(\sum_\tau \frac{p_{\pi_e}(\tau_{1:t} | S_t = s, A_t = a)}{p_{\pi_b}(\tau_{1:t} | S_t = s, A_t = a)} p_{\pi_b}(\tau | S_t = s, A_t = a)\right)\frac{d_t^{\pi_e}(s,a)}{d_t^{\pi_b}(s,a)} \\
&\overset{(b)}{=} \left(\sum_\tau \frac{p_{\pi_e}(\tau_{1:t} | S_t = s, A_t = a)}{p_{\pi_b}(\tau_{1:t} | S_t = s, A_t = a)} p_{\pi_b}(\tau_{1:t} | S_t = s, A_t = a) p_{\pi_b}(\tau_{t+1:L} | S_t = s, A_t = a)\right)\frac{d_t^{\pi_e}(s,a)}{d_t^{\pi_b}(s,a)} \\
&\overset{(c)}{=} \left(\sum_{\tau_{1:t}} p_{\pi_e}(\tau_{1:t} | S_t = s, A_t = a) \sum_{\tau_{t+1:L}} p_{\pi_b}(\tau_{t+1:L} | S_t = s, A_t = a)\right)\frac{d_t^{\pi_e}(s,a)}{d_t^{\pi_b}(s,a)} \\
&= \frac{d_t^{\pi_e}(s,a)}{d_t^{\pi_b}(s,a)}.
\end{aligned}
$$

Line (a) follows from $d_t^\pi(s,a) = p_\pi(S_t = s, A_t = a)$. In line (b), we use the Markov property which gives that $\boldsymbol{\tau}_{1:t}$ and $\boldsymbol{\tau}_{t+1:L}$ are independent conditioned on $(S_t = s, A_t = a)$. Line (c) follows from splitting the summation over τ into to summations over $\tau_{1:t}$ and $\tau_{t+1:L}$.

B Full Derivation of SOPE$_n$ Estimator

To derive the SOPE$_n$ estimator, we repeat the derivation of (1) with z being a function of time, $z(t) = \max\{t-n, 0\}$. This gives us the expression

$$
J(\pi_e) = \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\sum_{t=1}^{n}\gamma^{t-1}\rho_{1:t}R_t + \sum_{t=n+1}^{L}\gamma^{t-1}\frac{d_{t-n}^{\pi_e}(S_{t-n}, A_{t-n})}{d_{t-n}^{\pi_b}(S_{t-n}, A_{t-n})}\rho_{t-n+1:t}R_t\right]. \tag{4}
$$

Since $z(t)$ is function of t, we can accumulate the d_t^π across time so that we can write the interpolating expression using *average* state-action distribution ratios, rather than time-dependent ones. This additional marginalization step over time allows us to consider time-independent distribution ratios. Notation-wise, let $d_{1:T}^\pi := (\sum_{t=1}^T \gamma^{t-1} d_t^\pi(s,a))/(\sum_{t=1}^T \gamma^{t-1})$ for any time T. $d_{1:T}$ can be thought of as at the average state-action visitation over the first T time-steps. Note that $d^\pi = \lim_{T\to\infty} d_{1:T}^\pi$ where d^π is the average state-action distribution. Then, using the law of total expectation, we can write the expectation of the second sum in (4) as:

$$
\begin{aligned}
&\mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\sum_{t=n+1}^{L}\gamma^{t-1}\frac{d_{t-n}^{\pi_e}(S_{t-n}, A_{t-n})}{d_{t-n}^{\pi_b}(S_{t-n}, A_{t-n})}\rho_{t-n+1:t}R_t\right] \\
&= \sum_{t=n+1}^{L}\gamma^{t-1}\mathbf{E}_{\substack{(S_{t-n}, A_{t-n}) \\ \sim d_{t-n}^{\pi_b}}}\left[\mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\frac{d_{t-n}^{\pi_e}(S_{t-n}, A_{t-n})}{d_{t-n}^{\pi_b}(S_{t-n}, A_{t-n})}\rho_{t-n+1:t}R_t\middle| S_{t-n}, A_{t-n}\right]\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{t=n+1}^{L} \gamma^{t-1} \mathbf{E}_{\substack{(S_{t-n},A_{t-n})\\ \sim d^{\pi_b}_{t-n}}} \left[\frac{d^{\pi_e}_{t-n}(S_{t-n},A_{t-n})}{d^{\pi_b}_{t-n}(S_{t-n},A_{t-n})} \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\rho_{t-n+1:t}R_t | S_{t-n},A_{t-n}\right]\right] \\
&= \sum_{t=n+1}^{L} \gamma^{t-1} \sum_{s,a} d^{\pi_b}_{t-n}(s,a) \frac{d^{\pi_e}_{t-n}(s,a)}{d^{\pi_b}_{t-n}(s,a)} \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\rho_{t-n+1:t}R_t | S_{t-n}=s,A_{t-n}=a\right] \\
&= \sum_{t=n+1}^{L} \gamma^{t-1} \sum_{s,a} d^{\pi_e}_{t-n}(s,a) \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\rho_{t-n+1:t}R_t | S_{t-n}=s,A_{t-n}=a\right] \\
&\overset{(a)}{=} \sum_{s,a} \left(\sum_{t=n+1}^{L} \gamma^{t-1} d^{\pi_e}_{t-n}(s,a) \right) \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\rho_{1:n}R_n | S_1=s,A_1=a\right] \\
&= \sum_{s,a} \left(\sum_{t=1}^{L-n} \gamma^{t-1} d^{\pi_e}_{t}(s,a) \right) \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\rho_{1:n}R_n | S_1=s,A_1=a\right] \\
&\overset{(b)}{=} \sum_{s,a} \left(\sum_{t=1}^{L-n} \gamma^{t-1} \right) d^{\pi_b}_{1:L-n}(s,a) \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\rho_{1:n}R_n | S_1=s,A_1=a\right] \\
&\overset{(c)}{=} \sum_{s,a} \left(\sum_{t=1}^{L-n} \gamma^{t-1} \right) d^{\pi_b}_{1:L-n}(s,a) \frac{d^{\pi_e}_{1:L-n}(s,a)}{d^{\pi_b}_{1:L-n}(s,a)} \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\rho_{1:n}R_n | S_1=s,A_1=a\right] \\
&\overset{(d)}{=} \sum_{s,a} \left(\sum_{t=1}^{L-n} \gamma^{t-1} d^{\pi_b}_{t}(s,a) \right) \frac{d^{\pi_e}_{1:L-n}(s,a)}{d^{\pi_b}_{1:L-n}(s,a)} \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\rho_{1:n}R_n | S_1=s,A_1=a\right] \\
&= \sum_{s,a} \left(\sum_{t=n+1}^{L} \gamma^{t-1} d^{\pi_b}_{t-n}(s,a) \right) \frac{d^{\pi_e}_{1:L-n}(s,a)}{d^{\pi_b}_{1:L-n}(s,a)} \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\rho_{1:n}R_n | S_1=s,A_1=a\right] \\
&= \sum_{t=n+1}^{L} \gamma^{t-1} \sum_{s,a} d^{\pi_b}_{t-n}(s,a) \frac{d^{\pi_e}_{1:L-n}(s,a)}{d^{\pi_b}_{1:L-n}(s,a)} \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\rho_{t-n+1:t}R_t | S_{t-n}=s,A_{t-n}=a\right] \\
&= \sum_{t=n+1}^{L} \gamma^{t-1} \mathbf{E}_{\substack{(S_{t-n},A_{t-n})\\ \sim d^{\pi_b}_{t-n}}} \left[\mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\frac{d^{\pi_e}_{1:L-n}(S_{t-n},A_{t-n})}{d^{\pi_b}_{1:L-n}(S_{t-n},A_{t-n})} \rho_{t-n+1:t}R_t \middle| S_{t-n},A_{t-n}\right]\right] \\
&= \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\sum_{t=n+1}^{L} \gamma^{t-1} \frac{d^{\pi_e}_{1:L-n}(S_{t-n},A_{t-n})}{d^{\pi_b}_{1:L-n}(S_{t-n},A_{t-n})} \rho_{t-n+1:t}R_t \right].
\end{aligned}
\tag{5}
$$

In line (a), we use $\mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}[\rho_{t-n+1:t}R_t|S_{t-n}=s,A_{t-n}=a] = \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}[\rho_{1:n}R_n|S_1=s,A_1=a]$ which follows from noting that conditioning on S_{t-n}, A_{t-n} and considering the n time steps after is equivalent to conditioning on S_1, A_1 and considering the n time steps after that. Lines (b) and (d) follow from $d^{\pi}_{1:L-n} = \left(\sum_{t=1}^{L-n} \gamma^{t-1} d^{\pi}_t(s,a)\right) / \left(\sum_{t=1}^{L-n}\gamma^{t-1}\right)$. Line (c) is possible due to Assumption 1. Plugging in the final expression from (5) back into (4) gives us

$$
J(\pi_e) = \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\sum_{t=1}^{n} \gamma^{t-1}\rho_{1:t}R_t + \sum_{t=n+1}^{L} \gamma^{t-1} \frac{d^{\pi_e}_{1:L-n}(S_{t-n},A_{t-n})}{d^{\pi_b}_{1:L-n}(S_{t-n},A_{t-n})} \rho_{t-n+1:t}R_t \right]. \tag{6}
$$

Note that $\frac{d^{\pi_e}_{1:L-n}(s,a)}{d^{\pi_b}_{1:L-n}(s,a)}$ is the state-action distribution ratio over the first $L-n$ time-steps. In practice, to estimate this ratio, one can discard the data from time-step $L-n$ to L, and use the same min-max optimization procedures used to estimate $\frac{d^{\pi_e}_{1:L}(s,a)}{d^{\pi_b}_{1:L}(s,a)}$ on the remaining data to estimate this ratio.

Note that in the infinite horizon setting where $L \to \infty$ and for finite n, (6) becomes

$$
J(\pi_e) = \mathbf{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\sum_{t=1}^{n} \gamma^{t-1}\rho_{1:t}R_t + \sum_{t=n+1}^{\infty} \gamma^{t-1} \frac{d^{\pi_e}(S_{t-n},A_{t-n})}{d^{\pi_b}(S_{t-n},A_{t-n})} \rho_{t-n+1:t}R_t \right].
$$

In this case, the typical optimization procedures for estimating $\frac{d^{\pi_e}(s,a)}{d^{\pi_b}(s,a)}$ in the infinite horizon setting can be used to estimate the distribution ratios.

Additionally, note that specifically for the infinite horizon setting, we can alternatively derive the SOPE$_n$ estimator using the Bellman equations for the average state-action distribution d^π. This alternative derivation can be found in Appendix C.

C Bellman Recursion Derivation of SOPE$_n$

We present an alternative derivation of the SOPE$_n$ estimator for the infinite horizon setting using the Bellman equations for the average state-action distribution d^π, which is:

$$\begin{aligned} d^\pi(s,a) &\coloneqq (1-\gamma)\sum_{t=1}^{\infty}\gamma^{t-1}\Pr(S_t = s, A_t = a\,;\pi) \\ &= (1-\gamma)d_1(s)\pi(a|s) + \gamma \sum_{s'\in\mathcal{S},a'\in\mathcal{A}} \Pr(s,a|s',a'\,;\pi)d^\pi(s',a'). \qquad (7) \end{aligned}$$

Now using (7) we can expand $J(\pi_e)$ and unroll d^{π_e} once to obtain

$$\begin{aligned} J(\pi_e) &= (1-\gamma)^{-1}\sum_{s\in\mathcal{S},a\in\mathcal{A}} r(s,a)d^{\pi_e}(s,a) \\ &= (1-\gamma)^{-1}\sum_{s\in\mathcal{S},a\in\mathcal{A}} r(s,a)\left[(1-\gamma)d_1(s)\pi_e(a|s) + \gamma\sum_{s'\in\mathcal{S},a'\in\mathcal{A}}\Pr(s,a|s',a'\,;\pi_e)d^{\pi_e}(s',a')\right] \\ &= \sum_{s\in\mathcal{S},a\in\mathcal{A}} r(s,a)d_1(s)\pi_e(a|s) + \gamma(1-\gamma)^{-1}\sum_{s\in\mathcal{S},a\in\mathcal{A}}\sum_{s'\in\mathcal{S},a'\in\mathcal{A}}\Pr(s,a|s',a'\,;\pi_e)d^{\pi_e}(s',a')r(s,a) \\ &\overset{(a)}{=} \sum_{s\in\mathcal{S},a\in\mathcal{A}} r(s,a)d_1(s)\pi_e(a|s) + \gamma(1-\gamma)^{-1}\sum_{s\in\mathcal{S},a\in\mathcal{A}}\sum_{s'\in\mathcal{S},a'\in\mathcal{A}}\Pr(s',a'|s,a\,;\pi_e)d^{\pi_e}(s,a)r(s',a') \\ &= \sum_{s\in\mathcal{S},a\in\mathcal{A}} \pi_b(a|s)r(s,a)d_1(s)\frac{\pi_e(a|s)}{\pi_b(a|s)} \\ &\quad + \gamma(1-\gamma)^{-1}\sum_{s\in\mathcal{S},a\in\mathcal{A}} d^{\pi_b}(s,a)\sum_{s'\in\mathcal{S},a'\in\mathcal{A}}\pi_b(a'|s')\Pr(s'|s,a)\frac{\pi_e(a'|s')}{\pi_b(a'|s')}\frac{d^{\pi_e}(s,a)}{d^{\pi_b}(s,a)}r(s',a') \\ &= \sum_{s\in\mathcal{S},a\in\mathcal{A}} \pi_b(a|s)r(s,a)d_1(s)\frac{\pi_e(a|s)}{\pi_b(a|s)} \\ &\quad + \gamma\sum_{s\in\mathcal{S},a\in\mathcal{A}}\sum_{t=1}^{\infty}\gamma^{t-1}\Pr(S_t = s, A_t = a\,;\pi_b)\sum_{s'\in\mathcal{S},a'\in\mathcal{A}}\pi_b(a'|s')\Pr(s'|s,a)\frac{\pi_e(a'|s')}{\pi_b(a'|s')}\frac{d^{\pi_e}(s,a)}{d^{\pi_b}(s,a)}r(s',a') \\ &= \mathbf{E}_{\boldsymbol{\tau}\sim\pi_b}\left[\frac{\pi_e(A_1|S_1)}{\pi_b(A_1|S_1)}r(S_1,A_1) + \sum_{t=1}^{\infty}\gamma^t\frac{d^{\pi_e}(S_t,A_t)}{d^{\pi_b}(S_t,A_t)}\frac{\pi_e(A_{t+1}|S_{t+1})}{\pi_b(A_{t+1}|S_{t+1})}r(S_{t+1},A_{t+1})\right] \\ &= \mathbf{E}_{\boldsymbol{\tau}\sim\pi_b}\left[\frac{\pi_e(A_1|S_1)}{\pi_b(A_1|S_1)}r(S_1,A_1) + \sum_{t=2}^{\infty}\gamma^{t-1}\frac{d^{\pi}(S_{t-1},A_{t-1})}{d^{\pi_b}(S_{t-1},A_{t-1})}\frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)}r(S_t,A_t)\right]. \qquad (8) \end{aligned}$$

where (a) follows by relabelling in the common notation such that (s,a) and (s',a') are consecutive state-action pairs. Notice that SOPE$_1(D)$ is the sample estimate of (8). Similarly, on unrolling d^{π_b} twice using (7),

$$
\begin{aligned}
J(\pi_e) &= (1-\gamma)^{-1} \sum_{s \in \mathcal{S}, a \in \mathcal{A}} r(s,a) d^{\pi_e}(s,a) \\
&= (1-\gamma)^{-1} \sum_{s \in \mathcal{S}, a \in \mathcal{A}} r(s,a) \Bigg[(1-\gamma) d_1(s) \pi_e(a|s) \\
&\quad + \gamma \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} \Pr(s,a|s',a';\pi_e) \Bigg[(1-\gamma) d_1(s') \pi_e(a'|s') + \gamma \sum_{s'' \in \mathcal{S}, a'' \in \mathcal{A}} \Pr(s',a'|s'',a'';\pi_e) d^{\pi_e}(s'',a'') \Bigg]\Bigg] \\
&= \sum_{s \in \mathcal{S}, a \in \mathcal{A}} r(s,a) d_1(s) \pi_e(a|s) + \gamma \sum_{s \in \mathcal{S}, a \in \mathcal{A}} r(s,a) \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} \Pr(s,a|s',a';\pi_e) d_1(s') \pi_e(a'|s') \\
&\quad + \gamma^2 (1-\gamma)^{-1} \sum_{s \in \mathcal{S}, a \in \mathcal{A}} r(s,a) \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} \Pr(s,a|s',a';\pi_e) \sum_{s'' \in \mathcal{S}, a'' \in \mathcal{A}} \Pr(s',a'|s'',a'';\pi_e) d^{\pi_e}(s'',a'') \\
&= \sum_{s \in \mathcal{S}, a \in \mathcal{A}} r(s,a) d_1(s) \pi_e(a|s) + \gamma \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} r(s',a') \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \Pr(s',a'|s,a;\pi_e) d_1(s) \pi_e(a|s) \\
&\quad + \gamma^2 (1-\gamma)^{-1} \sum_{s'' \in \mathcal{S}, a'' \in \mathcal{A}} r(s'',a'') \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} \Pr(s'',a''|s',a';\pi_e) \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \Pr(s',a'|s,a;\pi_e) d^{\pi_e}(s,a),
\end{aligned}
$$

Where the last line follows by relabelling the state-action pairs such that they match the common notation where (s, a), (s', a') and (s'', a'') are the state action tuples for three consecutive time-steps. Now changing the sampling distribution as earlier,

$$
\begin{aligned}
J(\pi_e) = \mathbf{E}_{\boldsymbol{\tau} \sim \pi_b} \Bigg[& \frac{\pi_e(A_1|S_1)}{\pi_b(A_1|S_1)} r(S_1, A_1) + \gamma \frac{\pi_e(A_1|S_1)}{\pi_b(A_1|S_1)} \frac{\pi_e(A_2|S_2)}{\pi_b(A_2|S_2)} r(S_2, A_2) \\
& + \sum_{t=3}^{\infty} \gamma^{t-1} \frac{d^{\pi}_e(S_{t-2}, A_{t-2})}{d^{\pi_b}(S_{t-2}, A_{t-2})} \frac{\pi_e(A_{t-1}|S_{t-1})}{\pi_b(A_{t-1}|S_{t-1})} \frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)} r(S_t, A_t) \Bigg]
\end{aligned} \tag{9}
$$

It can be now observed that $\text{SOPE}_2(D)$ is the sample estimate of (9). Similarly, by generalizing this pattern it can be observed that on unrolling n times, we will get,

$$
\begin{aligned}
J(\pi_e) =& \mathbf{E}_{\boldsymbol{\tau} \sim \pi_b} \Bigg[\sum_{t=1}^{n} \left(\prod_{j=1}^{t} \frac{\pi_e(A_j|S_j)}{\pi_b(A_j|S_j)} \right) \gamma^{t-1} r(S_t, A_t) + \\
& \sum_{t=n+1}^{\infty} \gamma^{t-1} \frac{d^{\pi_e}(S_{t-n}, A_{t-n})}{d^{\pi_b}(S_{t-n}, A_{t-n})} \left(\prod_{j=0}^{n-1} \frac{\pi_e(A_{t-j}|S_{t-j})}{\pi_b(A_{t-j}|S_{t-j})} \right) r(S_t, A_t) \Bigg] \\
=& \mathbf{E}_{\boldsymbol{\tau} \sim p_{\pi_b}} \left[\sum_{t=1}^{n} \gamma^{t-1} \rho_{1:t} R_t + \sum_{t=n+1}^{\infty} \gamma^{t-1} \frac{d^{\pi_e}(S_{t-n}, A_{t-n})}{d^{\pi_b}(S_{t-n}, A_{t-n})} \rho_{t-n+1:t} R_t \right].
\end{aligned} \tag{10}
$$

Finally, it can be observed that that $\text{SOPE}_n(D)$ is the sample estimate of (10).

D Additional Experimental Details

For all experiments, we utilize the domains and algorithm implementations from Caltech OPE Benchmarking Suite (COBS) library by Voloshin et al. [2019]. Our code can be found at https://github.com/Pearl-UTexas/SOPE, and our experiments ran on 32 Intel Xeon cores.

D.1 Experimental Set-Up

For our experiments, we used the Graph, Toy Mountain Car, and standard Mountain Car [Brockman et al., 2016] domains provided in the COBS library. We include a brief description of each of these domains below, and a full description of each can be found in the work by Voloshin et al. [2019].

Graph Environment The Graph environment is a two-chain environment with $2L$ states and 2 actions. The ends of the chain are starting state $x_0 = 0$ and absorbing state $x_{abs} = 2L$. In between x_0 and x_{abs}, the remaining states form two chains of length $L - 1$ each. The states on the top chain are labeled $1, 3, \ldots, 2L - 3$ and the states on the bottom chain are labeled $2, 4, \ldots, 2L - 2$. For each $t < L$, taking action $a = 0$, the agent will try to enter the next state on the top chain $x_{t+1} = 2t + 1$, and taking action $a = 1$, the agent will try to enter the next state on the bottom chain $x_{t+1} = 2t + 2$. Since the environment is stochastic, the agent will succeed with probability 0.75 and slip into the wrong row with probability 0.25. The reward is +1 if the agent transitions to a state on the top chain and -1 otherwise. For our experiments, we set $L = 20$ and $\gamma = 0.98$.

Toy Mountain Car Environment The Toy-MC environment [Voloshin et al., 2019] is a tabular simplification of the classic Mountain Car domain. There are a total of 21 states: $x_0 = 0$ the starting point in the valley, 10 states to the left, and 10 states to the right. The right-most state is a terminal absorbing state. Taking action $a = 0$ moves the agent to the right and taking action $a = 1$ moves the agent to the left. The agent receives reward of $r = -1$ each time step, and the reward becomes 0 when the agent reaches the terminal absorbing states. For our experiments, we use random restart where start in a random state in the domain and set $L = 100$ and $\gamma = 0.99$.

Mountain Car Environment We use the Mountain Car environment from OpenAI gym with the simplifying modifications applied in Voloshin et al. [2019]. In particular, the car agent starts in a valley and needs to move back and forth in order to gain moment to reach the goal of getting to the top of the mountain. The state space is the position and velocity of the car. At each time step, the car agent can either accelerate move forward, move backwards, or do nothing. Additionally, at each time, the agent receives a reward of $r = -1$ until it reaches the goal. The environment is modified in the COBS library to decrease the effective trajectory length by applying each action a_t five times before observing x_{t+1}. Additionally, the initial start location is modified from being uniformly chosen between $[-.6, -.4]$ to be randomly chosen from $\{-.6, -.5, -.4\}$ with no velocity.

Policies For the tabular environments Graph and Toy Mountain Car, we utilize static policies that take action $a = 0$ with probability p and action $a = 1$ with probability $1 - p$. For the Mountain Car environment, we utilize an ϵ-greedy policies with the provided DDQN trained policy in the COBS library.

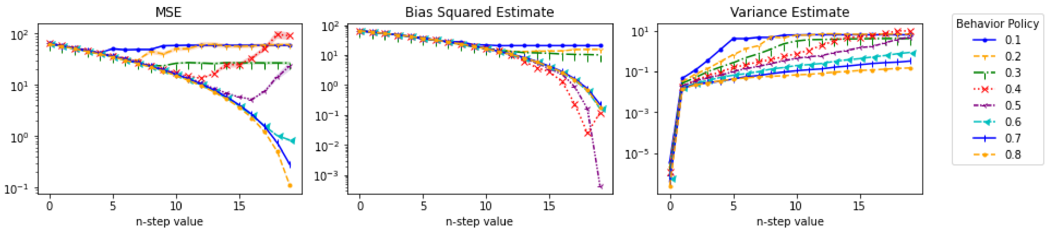
Methods For our experiments, we evaluate the performance of our proposed SOPE$_n$ and W-SOPE$_n$ estimators. To estimate the average state-action visitation ratios $\frac{d^{\pi_e}(s,a)}{d^{\pi_b}(s,a)}$, we utilize the implementation of methods from Liu et al. [2018] provided in the COBS library. For the Mountain Car experiments, we utilize the radial-basis function for the kernel estimate and a linear function class for the density estimate. Specific hyper-parameters can be found below.

Parameter	Graph	Toy-MC	Mountain Car
Quad. prog. regular.	1e-3	1e-3	-
NN Fit Epochs	-	-	1000
NN Batchsize	-	-	1024

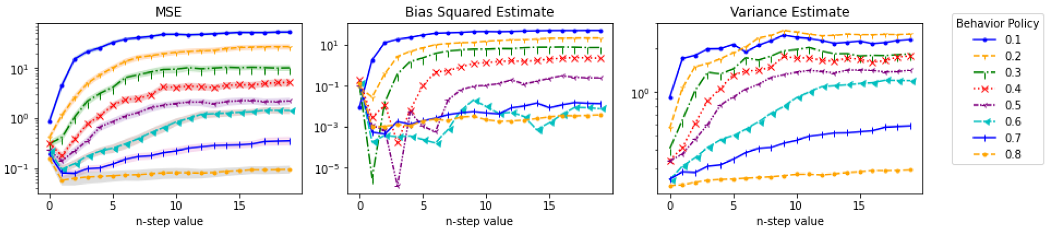
D.2 Impact of Policy Mismatch Between π_b and π_e on SOPE$_n$ and W-SOPE$_n$

We examine the impact of the policy mismatch between the behavior and evaluation policies on the performance of the SOPE$_n$ and W-SOPE$_n$ estimators. In this experiment, the evaluation policy takes action $a = 0$ with probability 0.9, and we vary the probability that the behavior policy takes $a = 0$ from 0.1 to 0.8 by increments of 0.1. We examine the performance of the SOPE$_n$ and W-SOPE$_n$ estimators across values of n for the different behavior policies. Results can be seen in the plots below.

The performance of PDIS and SIS has been known to be negatively correlated with the degree of policy mismatch [Voloshin et al., 2019]. We also find this to be generally true for the performance of the SOPE$_n$ and W-SOPE$_n$ estimators. Additionally, we observe that the degree of mismatch between



(a) SOPE$_n$ on Graph Domain

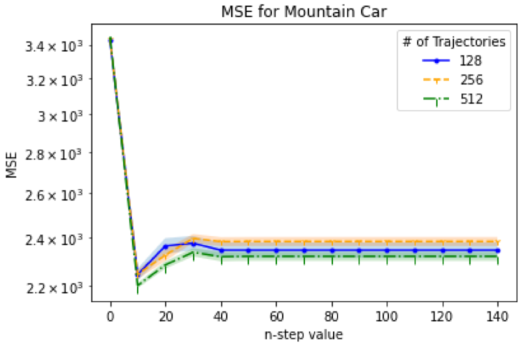


(b) W-SOPE$_n$ on Graph Domain

the evaluation and behavior policies has an impact on the existence of an interpolating estimator that is able to achieve lower MSE than the endpoints. For both SOPE$_n$ and W-SOPE$_n$, when the π_b is extremely different π_e, there are instances when the best estimate is SIS or weighted-SIS. In cases when the π_b is extremely close to the π_e, particularly for unweighted SOPE$_n$, there are cases when the trajectory-based importance sampling endpoint gives the lowest MSE. We do note that in cases when the difference between evaluation and behavior policies moderate but not extreme, there exists interpolating estimators that outperform the endpoints. This experiment helps to shed light on the possible conditions on the evaluation and behavior policies that allow for an interpolating estimator to have the lowest MSE.

D.3 Mountain Car Experimental Results

In addition to the experiments contained in the main paper, we also examine the performance of W-SOPE$_n$ on the Mountain Car domain. For these experiments, we used a provided DDQN trained policy as the base policy, and use ϵ-greedy versions of this policy as our behavior and evaluation policies. Specific information about this policy can be found in [Voloshin et al., 2019]. For our behavior policy, we use $\epsilon = 0.05$ and for our evaluation policy, we use $\epsilon = 0.9$. We average over 10 trials with $128, 256$ and 512 trajectories each. The results of this experiment can be found in the figure below.



We observe that this setting is an extremely challenging one for both trajectory-based and density-based importance sampling since the behavior and evaluation policies are so far apart. However, even in this extremely difficult setting, the there exists an interpolating estimator within the W-SOPE$_n$ spectrum that is able to have better performance than either of the endpoints.