SOPE: Spectrum of Off-Policy Estimators

Anonymous Author(s)
Affiliation
Address
email

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 technique 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 stationary distributions (SIS) have recently been adopted. Unfortunately, while SIS often provides lower variance estimates, estimating the stationary 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, respectively. We then show that estimators in this spectrum allow us 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 a 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 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 of the most popular OPE techniques is trajectory-based (per-decision) 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 variance, 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 w between the stationary distribution induced by π_e and π_b. Unfortunately, not only can estimating this ratio accurately can require *unverifiable* strong assumptions on the parameters [Jiang and Huang, 2020], but also requires a non-trivial min-max saddle point optimization procedure [Yang et al., 2020]. Consequently, If the parameterization is not rich enough, then it may not be possible to represent the visitation ratios accurately, and when using rich non-linear function approximators (like neural networks) then the optimization procedure may get stuck in sub-optimal saddle points. In practice, these challenges can result in inaccurate estimation of the desired ratio w, thereby resulting in 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 low bias of IS and the low variance of SIS. Particularly, we show that:

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

- 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

The core idea to establish 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 for the first part and IS for the latter. using this insight, we explore various splitting criteria to obtain new estimators within this spectrum. Finally, we present empirical case studies to highlight the effectiveness of these new estimators.

2 Background

Notation: A Markov decision process (MDP) is a tuple $(\mathcal{S}, \mathcal{A}, r, T, \gamma, d_0)$, 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 parameter, and d_0 is the start-state distribution. Any policy π is a distribution over $\mathcal{A}$, conditioned on the state. Let $\boldsymbol{\tau} := (S_0, A_0, R_0, S_1, ..., S_L, A_L, R_L)$ be the random variable corresponding to a trajectory sampled from π, and let p_π denote the distribution of $\boldsymbol{\tau}$ under π. Let the performance of any policy π be $J(\pi) := \mathbb{E}_{\boldsymbol{\tau}\sim p_\pi}[\sum_{t=0}^{L} \gamma^t R_t]$, where L is the horizon length. Infinite horizon setting can be obtained when $L \to \infty$.

Problem Statement: We aim to estimate the performance $J(\pi_e)$ of an evaluation policy π_e using a historical data $D = \{\tau^{(i)}\}_{i=1}^m$ collected from a behavior policy π_b. However, 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: A popular way to estimate $J(\pi_e)$ is by using trajectory-based importance sampling. Let the single-step likelihood ratio be $\boldsymbol{\rho}_t := \frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)}$ and the likelihood ratio from steps j to k be $\boldsymbol{\rho}_{j:k} := \prod_{t=j}^{k} \boldsymbol{\rho}_t$. It was shown by Precup [2000] that under Assumption 1

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

are unbiased estimators of $J(\pi_e)$. That is, $J(\pi_e) = \mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}[\text{IS}(\boldsymbol{\tau})] = \mathbb{E}_{\boldsymbol{\tau}\sim \pi_b}[\text{PDIS}(\boldsymbol{\tau})]$. Unfortunately, however, both IS and PDIS directly depend on the product of importance ratios and thus suffer from variance exponential in the horizon length L [Jiang and Li, 2016, Guo et al., 2017].

Density-Based Importance Sampling: To mitigate variance's dependence on horizon length, recent works importance sample the state-action visitation distribution instead of individual trajectories. For any policy π, let d^π_t denote the average visitation of s_t at time step t. Let the average state-action visitation distribution be $d^\pi(s,a) := c^{-1}\sum_{t=0}^{L}\gamma^t d^\pi_t(s,a)$ where $c := \sum_{t=0}^{L}\gamma^t$. Let $(S,A) \sim d^\pi$ denote that (S,A) are drawn from the average visitation distribution $d^\pi(s,a)$. The performance of π_e can then be expressed as,

$$\begin{aligned}
J(\pi_e) &= \mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_e}}\left[\sum_{t=1}^{L}\gamma^{t-1}R_t\right] = \sum_{s,a}\sum_{t=1}^{L}\gamma^{t-1}\, d^{\pi_e}_t(s,a)r(s,a) = c\sum_{s,a} d^{\pi_e}(s,a)r(s,a) \\
&\overset{(a)}{=} c\sum_{s,a} d^{\pi_b}(s,a)\frac{d^{\pi_e}(s,a)}{d^{\pi_b}(s,a)}r(s,a) = \sum_{s,a}\sum_{t=1}^{L}\gamma^{t-1}d^{\pi_b}_t(s,a)\frac{d^{\pi_e}(s,a)}{d^{\pi_b}(s,a)}r(s,a), \\
&= \mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_b}}\left[\sum_{t=1}^{L}\gamma^{t-1}\frac{d^{\pi_e}(S_t,A_t)}{d^{\pi_b}(S_t,A_t)}R_t\right],
\end{aligned}$$

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,

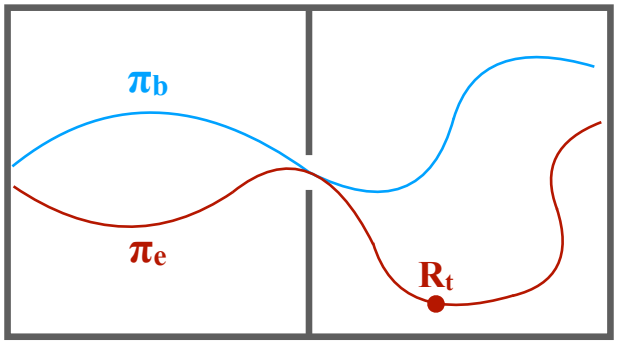


Figure 1: Illustration of two room domain. The domain consists of two rooms, the left room and the right room separated by a connecting door. π_b and π_e are two different policies that move from the left room to the right room. Although π_b and π_e have two different behaviors in the left room and right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

Jiang and Huang, 2020],

$$\text{SIS}(D) := \frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{L}\gamma^t w(s_t^i, a_t^i) r_t^i,$$

where $w(s,a) := d^{\pi_e}(s,a)/d^{\pi_b}(s,a)$. Notice that SIS($\tau$) has a single ratio term w instead of the product of importance ratios $\rho_{0:t}$, and thus helps mitigate variance's dependency 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., $\mathbb{E}_{\pi_b}[\text{SIS}(\tau)] = J(\pi_e)$. 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 can 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 density-based importance sampling methods are typically presented as alternative methods for off-policy evaluation. However, in this section we show that trajectory-based PDIS and the density-based SIS can actually be combined in a single importance sampling estimate. Furthermore, using this combination, we will be able to derive a spectrum of estimators that allows to interpolate between PDIS and IS and trade-off between the bias and variance of trajectory and density-based importance sampling methods.

PDIS and SIS provide two different ways of correcting the distribution mismatch between trajectories. PDIS corrects the distribution mismatch by how likely policies are to take the same actions and applies the action likelihood ratio as the correction term. SIS corrects the mismatch by how likely policies are to visit the same states—while remaining agnostic to *how* they arrived—and applies the density visitation ratio as the correction term. However, using the density ratio and likelihood ratio correction terms are not mutually exclusive, and we can draw on both types of correction terms to derive combined estimators.

To build up intuition to see how likelihood ratios and density ratios can be combined, we first consider the example of the two rooms domain shown in Figure 1. In this example, we have two policies π_b, π_e which have different strategies for navigating from the left room to the right 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. Conditioning on passing through the connecting door, all parts of the trajectory that occur in the right room are independent from what as occurred in the left room. Thus, when considering a reward R_t that occurs in the right room, we only need to consider the probability of reaching the door and then the probability of the behavior of the policies in the right room.

Now, we formalize this intuition and derive the a generic estimator that can be used to combine PDIS and SIS. To do this, we start with PDIS and use the law of total expectation to condition on (S_z, A_z), the state and action at time z. By conditioning on (S_z, A_z), the trajectory $\boldsymbol{\tau}$ can be separated into two conditionally independent partial trajectories $\boldsymbol{\tau}_{0:z}$ and $\boldsymbol{\tau}_{z+1,L}$. 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 $t \geq z$. We will see that the expectation of $\rho_{1:z}$ conditioned on (S_z, A_z) is exactly equal to the state-action density ratio for (S_z, A_z) in Lemma 1. Formally, let $(S_z, A_z) \sim d^{\pi_b}$, then,

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

The second to last line follows from Lemma 1 which states that the expected value of product likelihood ratios $\rho_{1:z}$ conditioned (S_z, A_z) equal to the visitation density ratio for (S_z, A_z). The proof can be found in the appendix.

Lemma 1. $\mathbb{E}[\rho_{1:z}|S_z, A_z] = \frac{d^{\pi_e}(S_z,A_z)}{d^{\pi_b}(S_z,A_z)}$.

Using Eq (1), we can compute a generic estimator indexed by z that allows us to combine both action likelihood ratios and estimated density ratios,

$$
\hat{J}(\pi_e) = \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{t=1}^{z-1}\gamma^{t-1}\rho^i_{1:t}r^i_t + \sum_{t=z}^{L}\gamma^{t-1}w(s^i_z, a^i_z)\rho^i_{z+1:t}r^i_t\right). \qquad (2)
$$

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. Additionally note that this estimator can be used in both finite and infinite settings.*

Observe that Eq (1) is indexed by time z and for each time z we get a different estimator that combines SIS and PDIS in a different way. Intuitively, we can think of z as the "switch time", or the time such that we switch from using density ratio estimate to likelihood ratios. Specifically, we use density ratios to estimate the probability of visiting (s^i_z, a^i_z) and use likelihood ratios to correct for the probability of actions taken after time z.

We further observe that z does not have to be a fixed constant—we can have a random variable Z representing the random time of occurrence of any event. Additionally z can also be function of t, $z(t)$, so that each reward R_t can utilize a different switch time.

This estimator that combines PDIS and SIS is useful since neither trajectory-based nor density-based importance sampling methods are always strictly better in all scenarios, even when we have access to

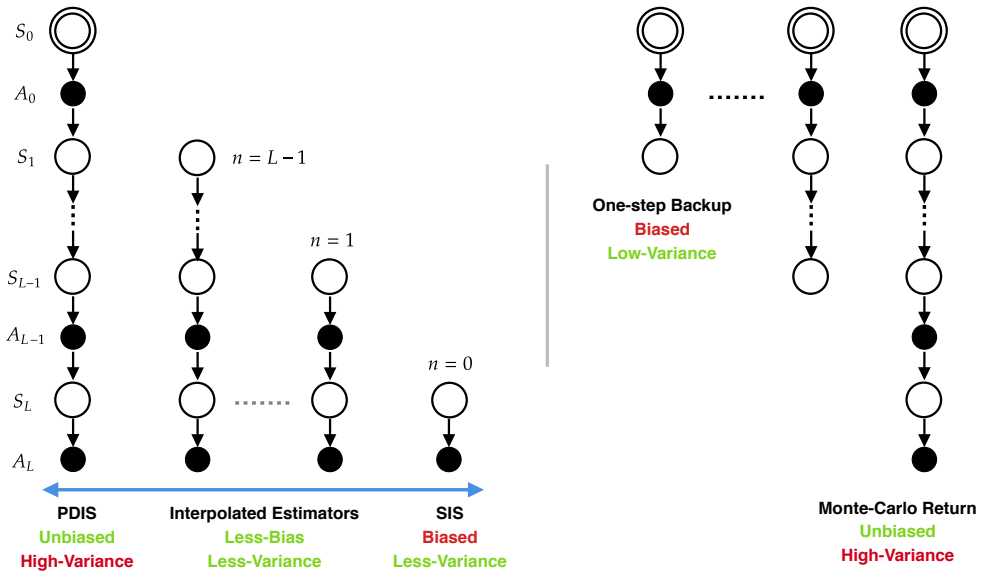


Figure 2: For the purpose of this illustration, consider that only at the last time step there is a non-zero reward R_L. We propose the NSTEP(n) estimator (left) that allows for control of how much an estimate depends on density ratios vs. likelihood ratios. Notice that NSTEP(0) results in SIS, NSTEP(L) 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 where SIS can have higher variance than PDIS [Liu et al., 2020, Metelli et al., 2020]).

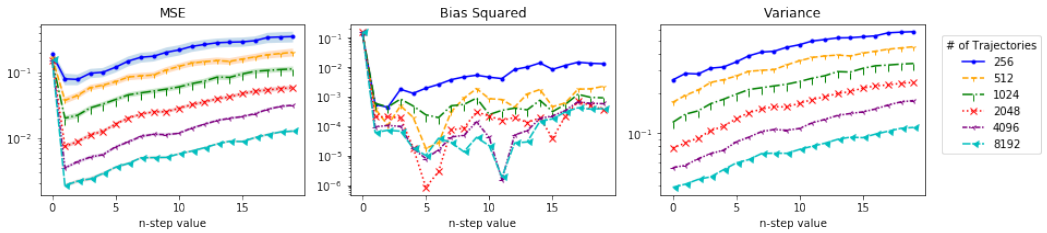
the true density ratios [Liu et al., 2020]. This suggests there may exist scenarios when interpolating estimators can outperform either extreme. Additionally, in practice we do not have access the oracle density ratio values w and must rely on an estimated $\hat{w}$. Then, the spectrum can be also be used to trade-off between a biased density ratio estimate and unbiased but high variance product of likelihood ratios. To best perform this bias and variance trade-off, we introduce and focus on a special case of this generic estimator, the NSTEP(n) estimator.

4 n-Step Interpolation Between PDIS and SIS for Bias-Variance Trade-off

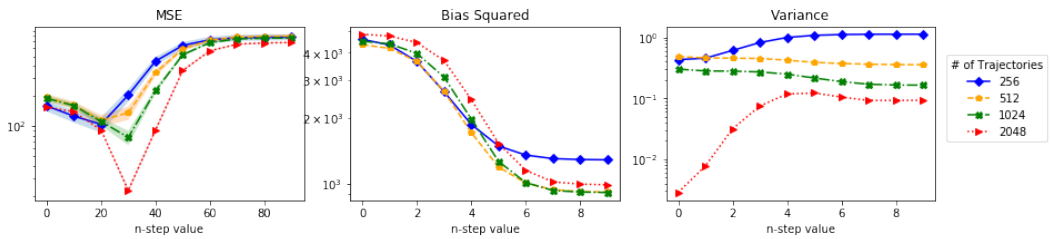
We can utilize the generic estimator that combines PDIS and SIS from Eq. (2) trade-off between the bias and variance of SIS and PDIS. Trajectory-based importance sampling methods like PDIS are unbiased but can suffer from variance that grows exponentially in the horizon length. On the other hand, SIS estimators eliminate variance's dependency on horizon length, but suffer from high-bias in practice due to model misspecification and estimation error of $\hat{w}$.

One simple method to to trade-off the bias and variance is to control the length of the product of the likelihood ratios. For each reward R_t, we can set our switch time z to give us an estimator that only considers the n most recent likelihood ratios $\rho_{t-n:t}$. To do this, we consider function $z(t) = \max\{t-n, n\}$. Using this switch time with Eq (2) gives us the n-step interpolation estimator,

$$\text{NSTEP}(n) = \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{t=1}^{n-1}\gamma^{t-1}\rho^i_{1:t}r^i_t + \sum_{t=n}^{L}\gamma^{t-1}\hat{w}(s^i_{t-n}, a^i_{t-n})\rho^i_{t-n+1:t}r^i_t\right).$$



(a) Graph Domain



(b) Toy Mountain Car Domain

Figure 3: Impact of varying size of historical dataset when computing WNSTEP(n) estimator on (a) graph domain (b) toy mountain car. Policies for experiments are $\pi_b(0) = 0.7, \pi_e(0) = 0.9$ for graph and $\pi_b(0) = 0.9, \pi_e(0) = 0.5$ for toy mountain car. For each domain, we report the estimated values for MSE, bias-squared, and variance of the WNSTEP(n) estimators for different values of n from 0 to L. The shaded regions denote 95% confidence regions on the mean of MSE. WNSTEP(0) gives us the weighted-SIS, and WNSTEP(L) gives us CWPDIS. WNSTEP(0) results in weighted-SIS, and WNSTEP(L) results in CWPDIS.

Since variance can scale exponentially with the length of product of likelihood ratios n, varying n will allow us control variance's dependency on the product of likelihood ratios. Decreasing n will decrease the length of the product of likelihood ratios, which will often lead to a decrease in variance. However, as we decrease n, we will also depend more and more on $\hat{w}$ in our estimate, thus increasing the bias. Likewise as n increases, we rely less and less on the biased $\hat{w}$, but the length of the product of likelihood ratios increases which often leads to an increase in variance. Further note that that for the endpoint values of this spectrum, $n = 0$ and $n = L$, we get SIS and PDIS,

$$\text{NSTEP}(0) = \frac{1}{m}\sum_{i=1}^{m}\sum_{t=1}^{L}\gamma^{t-1}w(s_z^i, a_z^i)r_t^i = \text{SIS}(D),$$

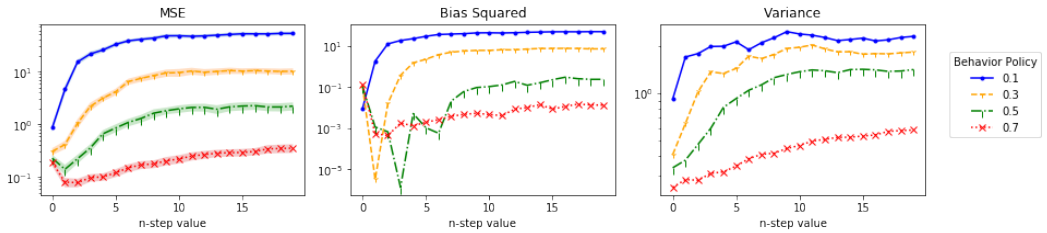
$$\text{NSTEP}(L) = \frac{1}{m}\sum_{i=1}^{m}\sum_{t=1}^{L}\gamma^{t-1}\rho_{1:t}^i r_t^i = \text{PDIS}(D).$$

Thus, by varying n, we can interpolate between PDIS and SIS and trade-off between bias and variance. An illustration of this idea can be seen in Figure 2.

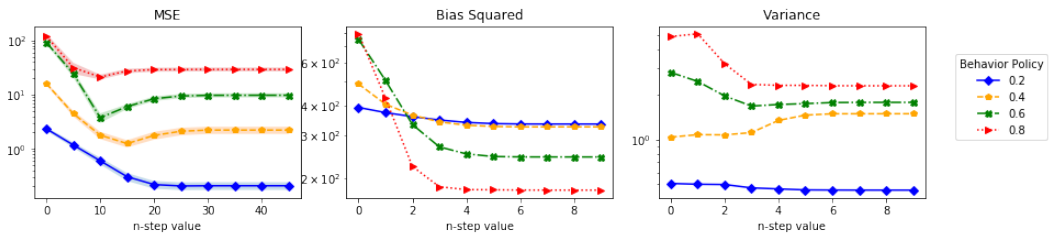
Additionally, we can create a weighted version of NSTEP(n) estimator that interpolates between a weighted-version of SIS and CWPDIS:

$$\text{WNSTEP}(n) = \sum_{t=0}^{n-1}\left(\gamma^{t-1}\sum_{i=1}^{m}\frac{\rho_{1:t}^i}{\sum_{i=1}^{m}\rho_{1:t}^i}r_t^i\right) + \sum_{t=n}^{M}\left(\gamma^{t-1}\sum_{i=1}^{m}\frac{w(s_{t-n}^i, a_{t-n}^i)\rho_{t-n:n}^i}{\sum_{i=1}^{m}w(s_{t-n}^i, a_{t-n}^i)\rho_{t-n:}^i}r_t^i\right) \tag{3}$$

For WNSTEP(n), since CWPDIS is a biased but consistent estimator, we do not get as clean of bias-variance trade-off as with the unweighted NSTEP(n) estimator. However, the amount amount of bias introduced by normalization in CWPDIS is in practice still significantly less than the bias due to estimation error in $\hat{w}$. Additionally, while CWPDIS is much lower variance than PDIS, its variance still depends on length of the product of likelihood ratios so it is generally still higher variance than SIS estimates. Thus, practically the WNSTEP(n) estimator still allows for a bias-variance trade-off. We show this experimentally in Section 5.



(a) Graph Domain



(b) Toy Mountain Car Domain

Figure 4: Impact of varying the closeness of the behavior and evaluation policies when using WNSTEP(n) estimator on (a) graph domain and (b) toy mountain car. The evaluation policies are $\pi_e(a_0) = 0.9$ and $\pi_e(a_0) = 0.1$ for graph and mountain car respectively. π_b is varied in both experiments. For each domain, we compute the estimated values of the MSE, bias-squared, and variance of the WNSTEP(n) estimators for different values of n from 0 to L. The shaded regions denote 95% confidence regions on the mean of MSE. WNSTEP(0) gives us the weighted-SIS, and WNSTEP(L) gives us CWPDIS.

5 Experiments

In this section, we conduct experiments to show how our interpolated estimators compare to existing estimators. In particular, we focus on the WNSTEP(n) estimator in Eq. (3), and show that WNSTEP(n) estimators can achieve lower mean-squared error than both weighted-SIS and CWPDIS. We also perform ablation studies to demonstrate how WNSTEP(n) is trading off bias and variance. Additionally, the studies allow us to understand in which scenarios an interpolated estimator can outperform CWPDIS and weighted-SIS.

For our experiments, we utilized the environments and implementations of baseline estimators in the OPE-tools library [Voloshin et al., 2019]. To obtain an estimate of the density ratios $\hat{w}$, we used their implementation of infinite horizon methods from [Liu et al., 2018]. For our ablation studies, we chose to focus on tabular domains with finite and discrete state spaces, since in this setting, there will be no model misspecification when estimating $\hat{w}$. Specifically, our ablation studies are run on the Graph and Toy Mountain Car domains. Brief descriptions of the environments are given below. Further details and figures can be found in the appendix. Results of experiments in larger continuous state spaces can be found in the appendix.

Graph Environment The Graph environment is a two-chain environment with $2H$ states and 2 actions. The ends of the chain are starting state $x_0 = 0$ and absorbing state $x_{abs} = 2H$. In between x_0 and x_{abs}, the remaining states form two chains of length $H - 1$ each. The states on the top chain are labeled $1, 3, \ldots, 2H - 3$ and the states on the bottom chain are labeled $2, 4, \ldots, 2H - 2$. For each $t < H$, 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 $H = 20$ and $\gamma = 0.98$.

Toy Mountain Car Environment The Toy-MC environment [Voloshin et al., 2019] is a tabular simplification of mountain car. 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 $H = 100$ and $\gamma = 0.99$.

Impact on Amount of Historical Data on WNSTEP(n) In the first experiment, we examine the impact that the size of the historical dataset has has on the WNSTEP(n) estimator. The results of this experiment can be found in Figure 3. For each environment, we compute the WNSTEP(n) for several values of n that vary from 0 to L. Additionally, we can see that the interpolating estimators in the spectrum are able to trade-off between bias and variance. In both domains and for all batch sizes, there exist estimators in the middle of the spectrum that have lower MSE than the endpoints. Furthermore, observe that for large batch sizes, CWPDIS eventually outperforms weighted-IS since it is consistent. However, even in this setting, there still exists a WNSTEP(n) estimator that outperforms CWPDIS.

Impact of Policy Mismatch Between π_b and π_e on WNSTEP(n) In the second experiment, we examine the impact of the similartity between the evaluation and behavior policies on the performance of the WNSTEP(n) estimator. The results of this experiment can be found in Figure 4. Similarly to the previous environment, we compute the WNSTEP(n) estimator for several values of n that vary from 0 to to L. 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 true for the performance of the WNSTEP(n) estimator for all values of n. Additionally, we also observe that for all values of policy mismatch, there exists an interpolated WNSTEP(n) estimator that was able to achieve lower MSE than the endpoints of PDIS and SIS. Even for cases where the evaluation and behavior policies are close and so there is only a small mismatch, an interpolated WNSTEP(n) estimator is able to outperform the endpoints.

6 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. 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) accurately and Liu et al. [2020] established a negative result illustrating 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, even when oracle estimates of density ratios w are available. Further, the proposed work provides a natural interpolation technique to balance the bias-variance trade-off for mitigating the estimation bias of w in practice. Finally, 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.

7 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 a spectrum of estimators that interpolate between them. Using this spectrum, we derive the NSTEP(n) estimator that can be used to perform bias-variance trade-off between low variance but biased SIS and high variance but unbiased PDIS. With our experimental results, we illustrate that estimators that lie on the interior of the NSTEP(n) spectrum are able to trade-off bias and variance and outperform their endpoints PDIS and SIS.

While we are able to show there exist NSTEP(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.

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.

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.

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.

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 7
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 7
   (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]