# Shaping Control Variates for Off-Policy Evaluation - Supplementary Material

**Sonali Parbhoo, Omer Gottesman, Finale Doshi-Velez**
Harvard University
sparbhoo@seas.harvard.edu

## A Existing Off-Policy Evaluation Methods

**Importance Sampling** A common approach to performing off-policy evaluation in RL uses IS or a family of IS-based estimators that assess the overall performance of the evaluation policy under consideration. The general IS estimate is given by

$$\hat{V}^{\pi_e}_{IS} := \frac{1}{n}\sum_{i=1}^{n}\omega^{(i)}_{0:T-1}\sum_{t=0}^{T-1}\gamma^t r^{(i)}_t, \tag{1}$$

where $\omega^{(i)}_{0:T-1} = \prod_{t=0}^{T-1}\frac{\pi_e(a^{(i)}_t|s^{(i)}_t)}{\pi_b(a^{(i)}_t|s^{(i)}_t)}$ and $r^{(i)}_t$ are the cumulative importance ratios and rewards reward at time step $t$ of trajectory $\tau^{(i)} \in \mathcal{D}$ respectively. Under Assumption 1 and the assumption that $\pi_b$ is known, it can be shown that the standard IS-estimator is unbiased, however suffers high variance that grows exponentially in size of the horizon. A variant of IS that often has less variance, while still being unbiased is Per-Step Importance Sampling (step-IS), where importance ratios are computed for *every* time step in a trajectory. That is,

$$\hat{V}^{\pi_e}_{stepIS} := \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{T-1}\gamma^t\omega^{(i)}_{0:t}r^{(i)}_t. \tag{2}$$

Note that when the behaviour policy $\pi_b$ is unknown and must be approximated, both IS and step-IS may not necessarily be unbiased.

**Weighted Importance Sampling** Having no bias is beneficial, however the high variance of standard IS estimators hinders their use in many applications, particularly in safety-critical domains. A variant of IS called *Weighted Importance Sampling (WIS)* and its per-decision counterpart (step-WIS), trade this bias for reduced variance, making them more suitable for practical applications overall. These are given by,

$$\hat{V}^{\pi_e}_{WIS} := \sum_{i=1}^{n}\frac{\omega^{(i)}_{0:T-1}}{\sum_{i=1}^{n}\omega^{(i)}_{0:T-1}}\sum_{t=0}^{T-1}\gamma^t r^{(i)}_t \tag{3}$$

$$\hat{V}^{\pi_e}_{stepWIS} := \sum_{i=1}^{n}\sum_{t=0}^{T-1}\gamma^t\frac{\omega^{(i)}_{0:t}r^{(i)}_t}{\sum_{i=1}^{n}\omega^{(i)}_{0:t}}. \tag{4}$$

**Doubly Robust Off-Policy Evaluation** DR estimators combine DM with IS and have been widely used in regression (Cassel et al., 1976), contextual bandits (Dudík et al., 2011), and RL (e.g. Thomas and Brunskill (2016); Jiang and Li (2016)). In RL, the DR estimate is given by,

$$\hat{V}^{\pi_e}_{DR}(\beta) := \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{T-1}\gamma^t\omega^{(i)}_{0:t}r^{(i)}_t - \gamma^t\left(\omega^{(i)}_{0:t}\hat{Q}^{\pi_e}(s^{(i)}_t, a^{(i)}_t;\beta) - \omega^{(i)}_{0:t-1}\hat{V}^{\pi_e}(s^{(i)}_t;\beta)\right) \tag{5}$$

Offline Reinforcement Learning Workshop at Neural Information Processing Systems, NeurIPS 2020.

The IS part of DR is based on step-IS while the model part relies on $\hat{Q}^{\pi_e}$ and $\hat{V}^{\pi_e}$ model estimates. Importantly, the bias of the DR estimator is a product of both the bias of DM and IS. As a result, DR is unbiased if either IS or DM is unbiased. When the behaviour policy $\pi_b$ is known as we assume in this paper, Eq. 5 is unbiased. The MRDR estimator modifies classic DR by learning the model parameter that minimises the variance of the DR estimator.

# B Preliminaries

**Corollary 1.** *(Khintchine's Strong Law of Large Numbers). Let $\{X_i\}_{i=1}^{\infty}$ be independent and identically distributed random variables. Then $(\frac{1}{n}\sum_{i=1}^{n} X_i)_{n=1}^{\infty}$ is a sequence of random variables that converges almost surely to $\mathbb{E}[X_1]$. Proof: See Sen and Singer (2017).*

**Corollary 2.** *(Kolmogorov's Strong Law of Large Numbers). Let $\{X_i\}_{i=1}^{\infty}$ be independent but not necessarily identically distributed random variables. If all $X_i$ have the same mean and bounded variance, then $(\frac{1}{n}\sum_{i=1}^{n} X_i)_{n=1}^{\infty}$ is a sequence of random variables that converges almost surely to $\mathbb{E}[X_1]$. Proof: See Sen and Singer (2017).*

# C Shaping Control Variates for Off-Policy Evaluation

**Lemma 1.** *1 The SCOPE estimator for stochastic evaluation policy $\pi_e$ is an unbiased estimator of the shaped value function $V_{M'}^{\pi_e}$.*

*Proof:*

$$\begin{aligned} Bias(\hat{\rho}_{SCOPE}^{\pi_e}) &= |V^{\pi_e} + \gamma^T\Phi(s_T) - \Phi(s_0) - \mathbb{E}_{\pi_b}[\hat{\rho}_{SCOPE}^{\pi_e}]| \\ &= \left| V^{\pi_e} + \gamma^T\Phi(s_T) - \Phi(s_0) - \mathbb{E}_{\pi_b}\left[\sum_{t=0}^{T-1}\gamma^t\omega_{0:t}r_t\right] - \mathbb{E}_{\pi_b}\left[\sum_{t=0}^{T-1}\gamma^t\omega_{0:t}(\gamma\Phi(s_{t+1}) - \Phi(s_t))\right]\right| \\ &= \left|V^{\pi_e} + \gamma^T\Phi(s_T) - \Phi(s_0) - V^{\pi_e} - \gamma^T\Phi(s_T) + \Phi(s_0)\right| = 0. \end{aligned} \tag{6}$$

*Hence given $\pi_b$, the estimator $\hat{\rho}_{SCOPE}^{\pi_e}$ is unbiased for any choice of potential function $\Phi$.*

**Lemma 2.** *(Consistency) SCOPE is a strongly consistent estimator of $V_{M'}^{\pi_e}$ i.e. $\lim_{n\to\infty}\hat{V}_{SCOPE}^{\pi_e} = V_{M'}^{\pi_e}$ almost surely. As a result, this implies that estimators with shaped control variates are well-posed.*

*Proof: Since we know the estimate is unbiased from Lemma 1 and our data set $D$ consists of $n$ independent and identically distributed samples, if Assumptions 1 and 2 hold, we can infer from Khintchine's Strong Law of Large Numbers that $\lim_{n\to\infty}\hat{V}_{SCOPE}^{\pi_e} = V_{M'}^{\pi_e}$.*

**Corollary 3.** *(Generalisation to multiple behaviour policies) If we assume that there is a constant $\beta < \infty$ such that $\forall(t,i) \in \mathbb{N}_{\geq 0} \times 1,\ldots,n, \omega_t \geq \beta$ i.e. that our importance weights are bounded, then SCOPE is a strongly consistent estimator of $V_{M'}^{\pi_e}$.*

*Proof: We know from Lemma 1 that the SCOPE estimator is unbiased for $i \in \{1,\ldots,n\}$. However, when there are multiple behaviour policies, the SCOPE estimate is computed over a set of $n$ independent but not necessarily identically distributed random variables and we cannot apply Khintchine's Strong Law of Large Numbers. Instead we apply Kolmogorov's Strong Law of Large Numbers, which requires each random variable to be bounded. As a result, $\lim_{n\to\infty}\hat{V}_{SCOPE}^{\pi_e} = V_{M'}^{\pi_e}$.*

**Theorem 4.** *The variance of the SCOPE estimator for stochastic evaluation policy $\pi_e$ is given by,*

$$\begin{aligned} \mathbb{V}_{P_\tau^{\pi_b}}(\hat{V}_{SCOPE}^{\pi_e}) &= \mathbb{V}_{P_\tau^{\pi_b}}[\hat{V}_{stepIS}^{\pi_e}] + 2\mathbb{E}_{P_\tau^{\pi_b}}[R_t(\delta + \gamma^T\omega_{0:T-1}\phi(s_T;\beta) - \phi(s_0;\beta))] \\ &- 2V_{M'}^{\pi_e}\mathbb{E}_{P_\tau^{\pi_b}}[\delta + \gamma^T\omega_{0:T-1}\phi(s_T;\beta) - \phi(s_0;\beta)] + \mathbb{V}_{P_\tau^{\pi_b}}[\delta]^2 \\ &+ \mathbb{V}_{P_\tau^{\pi_b}}\left[\left(\gamma^T\omega_{0:T-1}\phi(s_T;\beta) - \phi(s_0;\beta)\right)\right] \\ &- 2\mathbb{E}_{P_\tau^{\pi_b}}\left[\delta\left(\gamma^T\omega_{0:T-1}\phi(s_T;\beta) - \phi(s_0;\beta)\right)\right] \\ &+ 2\mathbb{E}_{P_\tau^{\pi_b}}[\delta]\mathbb{E}_{P_\tau^{\pi_b}}\left[\gamma^T\omega_{0:T-1}\phi(s_T;\beta) - \phi(s_0;\beta)\right] \end{aligned} \tag{7}$$

*where* $\delta = \sum_{t=1}^{T-1} \gamma^t \phi(s_t;\beta)(\omega_{0:t-1} - \omega_{0:t})$.
*Proof:*
*For ease of notation we drop the parameter* $\beta$ *in each of the shaping terms in the derivation below, however* $\phi$ *remains parameterised by* $\beta$.

$$
\begin{aligned}
\mathbb{E}_{P_\tau^{\pi_b}}\left[{\hat{V}^{\pi_e}_{SCOPE}}^2\right] &= \mathbb{E}_{P_\tau^{\pi_b}}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^t\omega^n_{0:t}(r^n_t + \gamma\phi(s^n_{t+1}) - \phi(s^n_t))\right)^2\right] \\
&= \mathbb{E}_{P_\tau^{\pi_b}}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^t\omega^n_{0:t}r^n_t + \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^{t+1}\omega^n_{0:t}\phi(s^n_{t+1}) - \gamma^t\phi(s^n_t))\right)^2\right] \\
&= \mathbb{E}_{P_\tau^{\pi_b}}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^t\omega^n_{0:t}r^n_t + \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T-1}\gamma^t\phi(s^n_t)(\omega^n_{0:t-1} - \omega^n_{0:t}) - \phi(s^n_0)\right)^2\right] \\
&= \mathbb{E}_{P_\tau^{\pi_b}}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^t\omega^n_{0:t}r^n_t\right)^2\right. \\
&+ 2\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^t\omega^n_{0:t}r^n_t\right)\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T-1}\gamma^t\phi(s^n_t)(\omega^n_{0:t-1} - \omega^n_{0:t}) - \phi(s^n_0)\right) \\
&+ \left.\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T-1}\gamma^t\phi(s^n_t)(\omega^n_{0:t-1} - \omega^n_{0:t}) - \phi(s^n_0)\right)^2\right] \\
&= \mathbb{E}_{P_\tau^{\pi_b}}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^t\omega^n_{0:t}r^n_t\right)^2\right] \\
&+ 2\mathbb{E}_{P_\tau^{\pi_b}}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^t\omega^n_{0:t}r^n_t\right)\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T-1}\gamma^t\phi(s^n_t)(\omega^n_{0:t-1} - \omega^n_{0:t}) - \phi(s^n_0)\right)\right] \\
&+ \mathbb{E}_{P_\tau^{\pi_b}}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T-1}\gamma^t\phi(s^n_t)(\omega^n_{0:t-1} - \omega^n_{0:t}) - \phi(s^n_0)\right)^2\right]
\end{aligned}
\tag{8}
$$

$$
\begin{aligned}
\mathbb{E}\left[{\hat{V}^{\pi_e}_{SCOPE}}^2\right] &= \mathbb{E}_{P_\tau^{\pi_b}}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^t\omega^n_{0:t}r^n_t\right)^2\right] \\
&+ 2\mathbb{E}_{P_\tau^{\pi_b}}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^t\omega^n_{0:t}r^n_t\right)\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T-1}\gamma^t\phi(s^n_t)(\omega^n_{0:t-1} - \omega^n_{0:t}) - \phi(s^n_0)\right)\right] \\
&+ \mathbb{E}_{P_\tau^{\pi_b}}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T-1}\gamma^t\phi(s^n_t)(\omega^n_{0:t-1} - \omega^n_{0:t})\right)^2\right] \\
&- 2\mathbb{E}_{P_\tau^{\pi_b}}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T-1}\gamma^t\phi(s^n_t)(\omega^n_{0:t-1} - \omega^n_{0:t})\ \phi(s^n_0)\right] + \mathbb{E}_{P_\tau^{\pi_b}}[\phi(s_0)^2]\right.
\end{aligned}
\tag{9}
$$

*Similarly for* $(\mathbb{E}_{P_\tau^{\pi_b}}[\hat{V}^{\pi_e}_{SCOPE}])^2$ *we have,*

$$
\begin{aligned}
\mathbb{E}_{P_\tau^{\pi_b}}\left[\hat{V}_{SCOPE}^{\pi_e}\right]^2 &= \mathbb{E}_{P_\tau^{\pi_b}}\left[\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^t\omega_{0:t}^n(r_t^n+\gamma\phi(s_{t+1}^n)-\phi(s_t^n))\right]^2 \\
&= \mathbb{E}_{P_\tau^{\pi_b}}\left[\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^t\omega_{0:t}^n r_t^n+\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^{t+1}\omega_{0:t}^n\phi(s_{t+1}^n)-\gamma^t\phi(s_t^n)\right]^2 \\
&= \mathbb{E}_{P_\tau^{\pi_b}}\left[\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^t\omega_{0:t}^n r_t^n+\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T-1}\gamma^t\phi(s_t^n)(\omega_{0:t-1}^n-\omega_{0:t}^n)-\phi(s_0^n)\right]^2 \\
&= \mathbb{E}_{P_\tau^{\pi_b}}\left[\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^t\omega_{0:t}^n r_t^n\right]^2 \\
&+ 2\mathbb{E}_{P_\tau^{\pi_b}}\left[\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^t\omega_{0:t}^n r_t^n\right]\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T-1}\gamma^t\phi(s_t^n)(\omega_{0:t-1}^n-\omega_{0:t}^n)-\phi(s_0^n)\right] \\
&+ \mathbb{E}_{P_\tau^{\pi_b}}\left[\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T-1}\gamma^t\phi(s_t^n)(\omega_{0:t-1}^n-\omega_{0:t}^n)-\phi(s_0^n)\right]^2 \qquad (10)
\end{aligned}
$$

*Now subtracting Eq. 10 from Eq. 9 yields,*

$$
\begin{aligned}
\mathbb{V}_{P_\tau^{\pi_b}}(\hat{V}^{\pi_e}_{SCOPE}) &= \mathbb{E}_{P_\tau^{\pi_b}}\left[{\hat{V}^{\pi_e}_{SCOPE}}^2\right] - \mathbb{E}_{P_\tau^{\pi_b}}\left[\hat{V}^{\pi_e}_{SCOPE}\right]^2 \\
&= \mathbb{E}_{P_\tau^{\pi_b}}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^t\omega^n_{0:t}r^n_t\right)^2\right] - \mathbb{E}_{P_\tau^{\pi_b}}\left[\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^t\omega^n_{0:t}r^n_t\right]^2 \\
&+ 2\mathbb{E}_{P_\tau^{\pi_b}}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^t\omega^n_{0:t}r^n_t\right)\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T-1}\gamma^t\phi(s^n_t)(\omega^n_{0:t-1}-\omega^n_{0:t})-\phi(s^n_0)\right)\right] \\
&- 2\mathbb{E}_{P_\tau^{\pi_b}}\left[\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^t\omega^n_{0:t}r^n_t\right]\mathbb{E}_{P_\tau^{\pi_b}}\left[\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T-1}\gamma^t\phi(s^n_t)(\omega^n_{0:t-1}-\omega^n_{0:t})-\phi(s^n_0)\right] \\
&+ \mathbb{E}_{P_\tau^{\pi_b}}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T-1}\gamma^t\phi(s^n_t)(\omega^n_{0:t-1}-\omega^n_{0:t})\right)^2\right] \\
&- \mathbb{E}_{P_\tau^{\pi_b}}\left[\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T-1}\gamma^t\phi(s^n_t)(\omega^n_{0:t-1}-\omega^n_{0:t})\right]^2 \\
&- 2\mathbb{E}_{P_\tau^{\pi_b}}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T-1}\gamma^t\phi(s^n_t)(\omega^n_{0:t-1}-\omega^n_{0:t})\ \phi(s^n_0)\right)\right] \\
&+ 2\mathbb{E}_{P_\tau^{\pi_b}}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T-1}\gamma^t\phi(s^n_t)(\omega^n_{0:t-1}-\omega^n_{0:t})\right)\mathbb{E}_{P_\tau^{\pi_b}}\left[\phi(s^n_0)\right]\right] \\
&+ \mathbb{E}_{P_\tau^{\pi_b}}\left[\phi(s^n_0)^2\right] - \mathbb{E}_{P_\tau^{\pi_b}}\left[\phi(s^n_0)\right]^2 \\
&= \mathbb{E}_{P_\tau^{\pi_b}}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^t\omega^n_{0:t}r^n_t\right)^2\right] - \mathbb{E}_{P_\tau^{\pi_b}}\left[\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^t\omega^n_{0:t}r^n_t\right]^2 \\
&+ 2\mathbb{E}_{P_\tau^{\pi_b}}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^t\omega^n_{0:t}r^n_t\right)\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T-1}\gamma^t\phi(s^n_t)(\omega^n_{0:t-1}-\omega^n_{0:t})-\phi(s^n_0)\right)\right] \\
&- 2\mathbb{E}_{P_\tau^{\pi_b}}\left[\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\gamma^t\omega^n_{0:t}r^n_t\right]\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T-1}\gamma^t\phi(s^n_t)(\omega^n_{0:t-1}-\omega^n_{0:t})-\phi(s^n_0)\right] \\
&+ \mathbb{E}_{P_\tau^{\pi_b}}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T-1}\gamma^t\phi(s^n_t)(\omega^n_{0:t-1}-\omega^n_{0:t})\right)^2\right] \\
&- \mathbb{E}_{P_\tau^{\pi_b}}\left[\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T-1}\gamma^t\phi(s^n_t)(\omega^n_{0:t-1}-\omega^n_{0:t})\right]^2,
\end{aligned}
\tag{11}
$$

*Finally, grouping terms and considering the telescopic series yields the result.*

# D Empirical Studies

In this section we provide more details of experiments comparing the SCOPE estimator to various IS estimators as well as DR. We illustrate the performance of SCOPE across 3 different domains namely, GridWorld, HIV and Cancer Simulators.

## D.1 GridWorld

In addition to the experimental results shown in the paper, we compared the performance of SCOPE to several other baselines such as MAGIC and WDR for the GridWorld domain. These results are

shown in Figures 2, 1 and 3. The same general trends are visible. Though WDR tends to outperform DR, it does not perform as well as SCOPE or MRDR in both sparse and dense reward settings. MAGIC tends to perform similarly to WDR in dense reward settings.

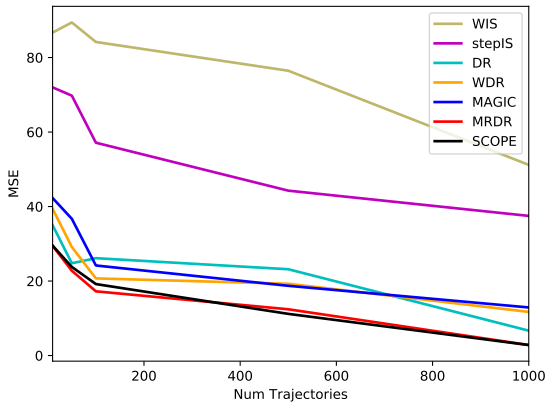



Figure 1: MSE of SCOPE vs baselines under dense rewards for Gridworld. SCOPE and MRDR perform similarly.

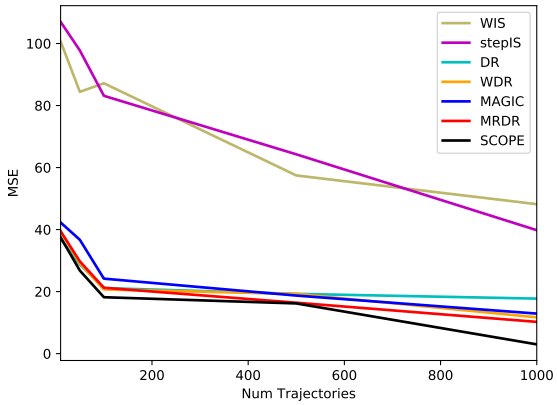



Figure 2: MSE of SCOPE vs baselines under sparse rewards for Gridworld. SCOPE outperforms the baselines.

## E Experiments with Simulators

### HIV Simulator

**Dynamics** The immune response to Human Immunodeficiency Virus (HIV) and antiretroviral therapy has been frequently studied using RL approaches in the past (Ernst et al., 2006; Parbhoo et al., 2017; Killian et al., 2017). For this task, we consider a dynamical system formulation for this problem from (Adams et al., 2004; Ernst et al., 2006). Here, a patient's response to therapy is modelled in terms of 6 parameters that describe their state in terms of $CD4^+$, $CD8^+$ and macrophage counts. There are 4 possible actions available at each step depending on whether a particular class of antiretrovirals is administered or not. The reward function is based on the immune response of a patient after a period of treatment as in Ernst et al. (2006). The complete dynamics of the model are described by the set of Equations 15. Here, $T_1$ ($T_1^*$) denotes the number of non-infected (respectively

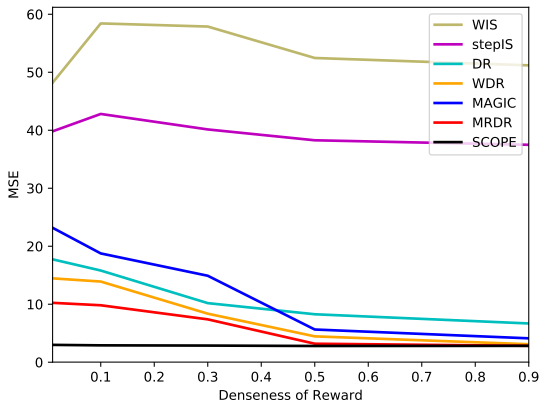


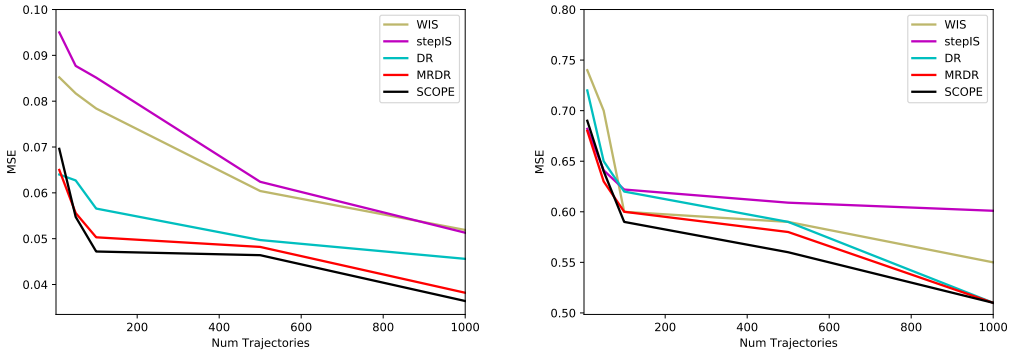

Figure 3: Influence of the denseness of reward on MSE. SCOPE outperforms baselines in sparse reward settings.

infected) CD4$^+$ T-lymphocytes (in cells/$ml$), $T_2$ ($T_2^*$) the number of non-infected (respectively infected) macrophages (in cells/$ml$), $V$ the number of free HI viruses (in copies/$ml$) and $E$ the number of cytotoxic T-lymphocytes (in cells/$ml$). The values of the various parameters of the model are taken directly are listed in Table 1.

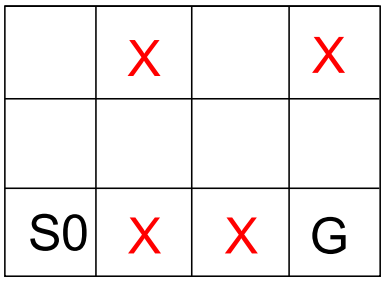
The model consists of two populations of target cells representing CD4$^+$ T-cells and macrophages respectively (Adams et al., 2004). In particular, the classes of RTIs and PIs are modelled. The model includes parameters for the drug efficacy of each class of antiretroviral under consideration. These parameters are $\epsilon_1$ and $\epsilon_2$. They describe how effective the RTI and PI classes of drugs are in reducing infection respectively (Adams et al., 2004). The model assumes that the RTI class of drugs is more effective in the CD4$^+$ population of cells than in macrophages where the efficacy is reduced by a factor of $f$, $f \in [0, 1]$. The PI class of drugs is only included in equations describing the change in the viral load under antiretroviral therapy, since these drugs operate by directly interfering with the formation of viral proteins that are necessary for viral production. The model assumes both T-cells and macrophages have the same death rates, $d_1 = d_2$. Immune effector cells are produced in response to the presence of infected cells and existing immune effectors. The action of these immune effector cells triggers lysing of infected T-cells and macrophages which results in their removal from the system of equations at the rates of $m_1$ and $m_2$ respectively. The rate at which the virus infects both types of cells is assumed to be different and is given by the parameters $k_1$ and $k_2$ respectively. Free virus particles are produced by both infected macrophages and infected T-cells; the model assumes these are produced at the same rate. Adams et al. (2004) demonstrate that when both $\epsilon_1$ and $\epsilon_2$ are zero, the dynamic model has three physical equilibrium points where all the variables are non-negative. These equilibria are:

$$(T_1, T_2, T_1^*, T_2^*, V, E) = (10^6, 3\,198, 0, 0, 0, 10), \tag{12}$$
$$(T_1, T_2, T_1^*, T_2^*, V, E) = (967\,839, 621, 76, 6, 415, 353\,108), \tag{13}$$
$$(T_1, T_2, T_1^*, T_2^*, V, E) = (163\,573, 5, 11\,945, 46, 63\,919, 24). \tag{14}$$

Note that Equation 12 is an unstable equilibrium point representing an uninfected individual; Equations 13 and 14 are the stable equilibrium points representing an infected individual. Specifically,




(a) (b)




(c)

Figure 4: (a) Performance comparison of SCOPE on HIV Simulator where $\zeta = 0.3$ for varying $n$. (b) Performance comparison of SCOPE on Cancer simulator where $\zeta = 0.3$ for varying $n$. (c) Illustration of Gridworld domain where $s_0$ is the start state, $G$ is the goal and $X$ are pit states in the dense reward setting.

Equation 13 represents an individual with good immune control over the virus.

$$
\begin{aligned}
\frac{\mathrm{d}T_1}{\mathrm{d}t} &= \lambda_1 - d_1 T_1 - (1-\epsilon_1) k_1 V T_1 \\
\frac{\mathrm{d}T_2}{\mathrm{d}t} &= \lambda_2 - d_2 T_2 - (1-f\epsilon_1) k_2 V T_2 \\
\frac{\mathrm{d}T_1^*}{\mathrm{d}t} &= (1-\epsilon_1) k_1 V T_1 - \delta T_1^* - m_1 E T_1^* \\
\frac{\mathrm{d}T_2^*}{\mathrm{d}t} &= (1-f\epsilon_1) k_2 V T_2 - \delta T_2^* - m_2 E T_2^* \\
\frac{\mathrm{d}V}{\mathrm{d}t} &= (1-\epsilon_2) N_T \delta (T_1^* + T_2^*) - cV - \left[(1-\epsilon_1)\rho_1 k_1 T_1 + (1-f\epsilon_1)\rho_2 k_2 T_2\right] V \\
\frac{\mathrm{d}E}{\mathrm{d}t} &= \lambda_E + \frac{b_E (T_1^* + T_2^*)}{(T_1^* + T_2^*) + K_b} E - \frac{d_E (T_1^* + T_2^*)}{(T_1^* + T_2^*) + K_d} E - \delta_E E
\end{aligned}
\tag{15}
$$

The reward function is based on the patient's immune response after treatment as in Ernst et al. (2006).

**Experimental Setup** Our data set consists of 1000 trajectories each of approximately 40 time steps. To demonstrate the effect of sparsity on performance, we sparsify the rewards for both domains by randomly setting a proportion $\zeta = 0.3$ of the rewards in each trajectory to 0. Our evaluation policy $\pi_e$ is an $\epsilon$-greedy policy with $\epsilon = 0.4$ and $\gamma = 0.9$. For our behaviour policy we use a policy that prefers to using both RTIs and PIs when a patient is far from the steady state and switch to the $\epsilon$-greedy policy near the steady state. This is akin to real situations where a clinician might know how to treat ill patients, but be less certain about how to keep them stable in the long run when their condition is

more stable.We compared the performance of SCOPE to IS, PDIS, WIS, DR and MRDR. For both DR and MRDR we train a linear parametric model for our control variates using the last layer of a feed-forward neural network.

**Cancer Simulator** The dynamics of the cancer domain follow a set of ODEs described in Ribba et al. (2012) to model the response of cancer cells to treatment. Here, the state space consists of 4 features for the respective cell counts and medication concentrations. The time steps correspond to the months in which a clinician chooses to administer a type of therapy or not administer any therapy at all. Our data set consists of 1000 trajectories each of approximately 40 time steps. The reward at each time step is given by the total change in the diameter of cancerous cells.To demonstrate the effect of sparsity on performance, we sparsify the rewards for both domains by randomly setting a proportion $\zeta = 0.3$ of the rewards in each trajectory to 0. Our evaluation policy $\pi_e$ is an epsilon greedy policy with $\epsilon = 0.4$ and $\gamma = 0.9$. We then compared the performance of baselines from the previous section to SCOPE across both of these simulators in Figures 2(a) and 2(b) respectively. For both of these examples our evaluation policy $\pi_e$ is an epsilon greedy policy with $\epsilon = 0.4$ and $\gamma = 0.9$. Across both domains, the optimal data split uses 35% of the data for training $\phi$ and the remaining data for OPE.

## References


Brian Michael Adams, Harvey Thomas Banks, Hee-Dae Kwon, and Hien T Tran. Dynamic multidrug therapies for hiv: Optimal and sti control approaches. Technical report, North Carolina State University. Center for Research in Scientific Computation, 2004.

Claes M Cassel, Carl E Särndal, and Jan H Wretman. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620, 1976.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

Damien Ernst, Guy-Bart Stan, Jorge Goncalves, and Louis Wehenkel. Clinical data based optimal sti strategies for hiv: a reinforcement learning approach. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pages 667–672. IEEE, 2006.

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

Taylor W Killian, Samuel Daulton, George Konidaris, and Finale Doshi-Velez. Robust and efficient transfer learning with hidden parameter markov decision processes. In *Advances in neural information processing systems*, pages 6250–6261, 2017.

Sonali Parbhoo, Jasmina Bogojeska, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. Combining kernel and model based learning for hiv therapy selection. *AMIA Summits on Translational Science Proceedings*, 2017:239, 2017.

Benjamin Ribba, Gentian Kaloshi, Mathieu Peyre, Damien Ricard, Vincent Calvez, Michel Tod, Branka Čajavec-Bernard, Ahmed Idbaih, Dimitri Psimaras, Linda Dainese, et al. A tumor growth inhibition model for low-grade glioma treated with chemotherapy or radiotherapy. *Clinical Cancer Research*, 18(18):5071–5080, 2012.

Pranab K Sen and Julio M Singer. *Large Sample Methods in Statistics (1994): An Introduction with Applications*. CRC Press, 2017.

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

| Parameters | Value | Units | Description |
|---|---|---|---|
| $\lambda_1$ | 10 000 | $\frac{cells}{ml.day}$ | production rate of CD4$^+$ cells |
| $d_1$ | 0.01 | $\frac{1}{day}$ | death rate of CD4$^+$ cells |
| $\epsilon_1$ | $\in [0, 1)$ | - | efficacy of RTI |
| $\epsilon_2$ | $\in [0, 1)$ | - | efficacy of PI |
| $k_1$ | $8.0 \times 10^{-7}$ | $\frac{ml}{virions.day}$ | infection rate of CD4$^+$ cells |
| $\lambda_2$ | 31.98 | $\frac{cells}{ml.day}$ | production rate of macrophages |
| $d_2$ | 0.01 | $\frac{1}{day}$ | death rate of macrophages |
| $f$ | 0.34 | - | reduction of treatment efficacy for macrophages |
| $k_2$ | $1.0 \times 10^{-4}$ | $\frac{ml}{virions.day}$ | infection rate of macrophages |
| $\delta$ | 0.7 | $\frac{1}{day}$ | death rate of infected cell |
| $m_1$ | $1.0 \times 10^{-5}$ | $\frac{ml}{cells.day}$ | immune-induced clearance rate for CD4$^+$ cells |
| $m_2$ | $1.0 \times 10^{-5}$ | $\frac{ml}{cells.day}$ | immune-induced clearance rate for macrophages |
| $N_T$ | 100 | $\frac{virions}{cell}$ | virions produced per infected cell |
| $c$ | 13 | $\frac{1}{day}$ | natural death rate of virus |
| $\rho_1$ | 1 | $\frac{virions}{cell}$ | average number of virions infecting a CD4$^+$ cell |
| $\rho_2$ | 1 | $\frac{virions}{cell}$ | average number of virions infecting a macrophage |
| $\lambda_E$ | 1 | $\frac{cells}{ml.day}$ | production rate of immune effector/cytotoxic T-cell |
| $b_E$ | 0.3 | $\frac{1}{day}$ | maximum birth rate for cytotoxic T-cell |
| $K_b$ | 100 | $\frac{cells}{ml}$ | saturation constant for cytotoxic T-cell birth |
| $d_E$ | 0.25 | $\frac{1}{day}$ | maximum death rate for cytotoxic T-cell |
| $K_d$ | 500 | $\frac{cells}{ml}$ | saturation constant for cytotoxic T-cell death |
| $\delta_E$ | 0.1 | $\frac{1}{day}$ | natural death rate of cytotoxic T-cells |

Table 1: Parameters used in Equations 15.