Shaping Control Variates for Off-Policy Evaluation

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

Abstract

We study the problem of off-policy evaluation in RL settings where reward signals are sparse. We introduce a new model-based control variate for variance reduction in off-policy evaluation inspired by reward shaping, and provide conditions under which our these shaped control variates may be used to reduce the variance of standard importance sampling-based methods, whilst remaining unbiased. In dense settings, we show that our resulting estimator behaves like the class of doubly robust estimators, but can be easier to train, while under sparse reward signals, our new estimator leads to improved performance.

1 Introduction

Reliably estimating the performance of a proposed policy from observational data is essential for many domains. For instance, if a health system is using reinforcement learning (RL) to identify potentially useful policies from past ICU records, one would want to be as sure as possible about its benefits prior to a costly human inspection and validation and a prospective trial with real risks. To this end, the field of *off-policy evaluation* (OPE) aims to estimate the performance of an *evaluation policy* based on data sampled by following a different *behaviour policy*. It is well-studied across reinforcement learning (Precup, 2000; Thomas, 2015; Thomas and Brunskill, 2016; Jiang and Li, 2016), contextual bandits (Dudík et al., 2011; Wang et al., 2017), causal inference (Tennenholtz et al., 2019; Oberst and Sontag, 2019), as well as applied settings such as healthcare (Liao et al., 2019; Parbhoo et al., 2017, 2018; Gottesman et al., 2019), education (Mandel et al., 2014), and marketing (Silver et al., 2013).

A common approach to OPE is importance sampling (IS) (Precup, 2000), which computes the average of trajectories' outcomes in the data, properly weighted to account for the difference between the evaluation and behaviour policies. Unfortunately, when evaluating deterministic policies, the weights of a trajectory will be zero whenever an observed action in the data is different from the action the evaluation policy would have chosen, leading to very small effective sample-sizes (Gottesman et al., 2018). This problem can be overcome by per-decision variants of the IS estimator which can gain statistical power from only parts of the trajectories, even if the entire trajectory has weight zero. These estimators significantly reduce the estimation variance, but rely on frequent observations of the rewards and tend to perform poorly in settings where these observations are sparse. Another class of methods known as Doubly Robust (DR) algorithms (Jiang and Li, 2016; Thomas and Brunskill, 2016; Rotnitzky and Robins, 1995; Robins et al., 1994; Bang and Robins, 2005) improve standard IS methods by using model-based control variates to reduce variance.

In this work, we propose learning a new type of model-based control variate for sparse reward settings to overcome the high variance of IS. Specifically, we draw upon a popular approach for accelerating on-policy RL known as *reward-shaping*(Ng et al., 1999; Harutyunyan et al., 2015), to learn a new shaped control variate which can be integrated into existing OPE estimators to reduce their variance. The shaping control variate serves as a means of densifying the signal where rewards are sparse, thereby guiding the off-policy estimate. Unlike popular doubly robust methods such as (Dudík et al., 2011; Jiang and Li, 2016) that also use control variates for variance reduction, our approach has fewer

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

parameters (only one per state, rather than one per state-action); we find this results in it being much easier to optimise for variance reduction in practice.

Our specific contributions are as follows: (i) We introduce a new shaped control variate for variance reduction and integrate this into a family of traditional per-step IS estimators. (ii) We prove theoretically that using shaped control variates with per-step IS produces an unbiased, consistent estimator that only requires a model over states for variance reduction. (iii) We derive the variance of this estimator such that it can be minimised, and provide intuition for when this variance is expected to be lower than doubly robust methods in sparse reward settings, and when the estimator behaves similarly to doubly-robust methods. (iv) On benchmark problems, we demonstrate that integrating shaped control variates into per-step IS produces off-policy estimates that are more accurate than both IS and DR baselines.

2 Related Work

There is a vast literature on several methods and techniques for performing off-policy evaluation. Direct Methods (DM) try to build a model of the environment from a batch of data. The value of the evaluation policy can subsequently be computed by simulating trajectories using the model (e.g. Paduraru (2013); Chow et al. (2015); Hanna et al. (2017); Fonteneau et al. (2013); Liu et al. (2018)). In general, it can be difficult to choose the loss function for model learning without knowing the evaluation policy in advance. Moreover, it may not be straightforward to minimise the bias of these models because of the lack of counterfactual data, which may play an important role in learning the dynamics of the model under the evaluation policy (Shalit et al., 2017; Johansson et al., 2016).

As a result, a second class of OPE estimators exists based on Inverse Propensity Scores or using IS to correct the sampling bias in off-policy data, such that an unbiased estimator may be obtained (e.g. Precup (2000); Horvitz and Thompson (1952); Thomas and Brunskill (2016)). The value of the evaluation policy is approximated using a weighted average of the returns over trajectories, where the weights reflect the distributional mismatch between the evaluation and behaviour policies. Unfortunately, these methods can suffer from high variance, particularly for longer time horizons. Our approach falls into the category of approaches that seek to manage this variance via adding control variates e.g. Advantage Sum or DR (e.g. Jiang and Li (2016); Thomas and Brunskill (2016); Wang et al. (2017)). Closest is the More Robust Doubly Robust (MRDR) estimator (Farajtabar et al., 2018), which specifically attempts to optimise for the control variate in DR.

Our approach uses ideas from reward shaping to design a control variate specifically targeted to minimise the variance of OPE in sparse reward settings. In dense settings, our new estimator can be seen as a variant of MRDR, where we learn a control variate based only on the state information.

Finally, we note that there exist other ways to manage the variance of IS-based estimates in long horizon settings. For example, works such as DualDICE and others (Nachum et al., 2019; Zhang et al., 2020) apply IS-type weighting to the stationary distribution. Our approach may be viewed as complementary to these, applicable both in settings where the required assumptions of stationary distributions are appropriate and when more traditional IS is needed.

3 Preliminaries and Notation

3.1 Markov Decision Processes

A Markov decision process (MDP) is a tuple of the form $M = (\mathcal{S}, \mathcal{A}, P, \gamma, R)$, where $\mathcal{S}$ is the set of all possible states, $\mathcal{A}$ is the set of available actions, $P(s, a, s')$ is the distribution of transitions from a state s to a subsequent state s' when applying a particular action a, $R(s, a)$ defines the reward for performing action a in state s, and $\gamma < 1$ is a discount factor that trades off the relative importance of immediate and long-term rewards. A (stationary) policy $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a stochastic mapping from states to actions, with $\pi(a|s)$ being the probability of taking action a in state s.

We denote by $\tau = (s_0, a_0, r_0, \dots, s_T)$ a T-step trajectory generated by policy π, and by $R_{0:T-1}(\tau) = \sum_{t=0}^{T-1} \gamma^t r_t$ the return of trajectory τ, where state $s_{t+1} \sim P(\cdot|s_t, a_t)$, and action $a_t \sim \pi(\cdot|s_t)$ $\forall i = t, \dots T$. A policy π is evaluated by computing the expectation of the return of the T-step trajectories it generates, $V^\pi = \mathbb{E}_\tau[R_{0:T-1}(\tau)]$. Further, we denote the value and action-values of a policy π for a state s and state-action pair (s, a) as $V^\pi(s)$ and $Q^\pi(s, a)$ respectively. These are the expected returns of a T-step trajectory generated by starting at state s, or state s and taking a, and following policy π respectively.

3.2 Off-Policy Evaluation Task

The general off-policy evaluation problem is when we are given a set of T-step trajectories $\mathcal{D} = \{\tau^{(i)}\}_{i=1}^n$ each independently generated by *behaviour policy* π_b, while our goal is to have a good estimate of the performance of a different policy, π_e, known as the *evaluation policy*. In general, the estimator $\hat{V}^{\pi_e}$ is a good estimator if it produces a low mean square error (MSE),

$$MSE(V^{\pi_e}, \hat{V}^{\pi_e}) = \mathbb{E}_{P_\tau^{\pi_b}}[(V^{\pi_e} - \hat{V}^{\pi_e})^2], \tag{1}$$

where $P_\tau^{\pi_b}$ denotes the distribution of trajectory τ under behaviour policy π_b.

In what follows, we make the following regularity assumption. This is a standard assumption in OPE and typically prevents the use of deterministic behaviour policies.

Assumption 1. *(Absolute Continuity). For all state-action pairs $(s, a) \in \mathcal{S} \times \mathcal{A}$, if $\pi_b(a|s) = 0$ then $\pi_e(a|s) = 0$.*

Our second assumption states we are given a single behaviour policy.

Assumption 2. *(Single Behaviour Policy). For all $i, j \in 1, \ldots n$, $\pi_b^{\tau^{(i)}} = \pi_b^{\tau^{(j)}}$.*

3.3 Potential-based Reward Shaping (PBRS)

Reward shaping is a technique that is used to modify the original reward function using a reward-shaping function $F : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ to typically make RL methods converge faster with more instructive feedback. The original MDP $M = (\mathcal{S}, \mathcal{A}, P, \gamma, R)$ is transformed into a *shaped-MDP* $M' = \mathcal{S}, \mathcal{A}, P, \gamma, R' = R + F)$. Although it is possible to perform reward shaping with various kinds of reward-shaping functions, PBRS Ng et al. (1999) retains the optimal policy, as we summarise below.

Definition 1. *(Potential-based Reward Shaping Ng et al. (1999)). $F : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is a potential-based shaping function if there exists a real-valued function $\phi : \mathcal{S} \to \mathbb{R}$, such that $\forall (s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,*

$$F(s, a, s') = \gamma\phi(s') - \phi(s), \tag{2}$$

where $\phi(s)$ is known as a potential function.

Theorem 1. *(Policy Invariance under PBRS Ng (2003)). The condition that F is a potential-based shaping function is necessary and sufficient for it to guarantee consistency with the optimal policy. Formally, for $M = (\mathcal{S}, \mathcal{A}, P, \gamma, R)$ and $M' = (\mathcal{S}, \mathcal{A}, P, \gamma, R' = R + F)$, if $F(s, a, s') = \gamma\phi(s') - \phi(s)$ then $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$,*

$$\begin{aligned} V^\pi_{M'}(s) &= V^\pi_M(s) - \phi(s) \qquad (3) \\ Q^\pi_{M'}(s, a) &= Q^\pi_M(s, a) - \phi(s) \qquad (4) \\ V^*_{M'}(s) &= V^*_M(s) - \phi(s) \qquad (5) \\ Q^*_{M'}(s, a) &= Q^*_M(s, a) - \phi(s) \qquad (6) \end{aligned}$$

Hence the optimal policy derived from $Q^*_{M'}$ or $V^*_{M'}$ *remains the same* as the optimal policy derived from the original MDP M.

Another way of writing Eq. 4 is as the following Bellman Equation,

$$\begin{aligned} Q^\pi_{M'}(s, a) &= \mathbb{E}_{s'}[R(s, a, s') + F(s, a, s') + \gamma \max_{a' \in \mathcal{A}} Q^\pi_{M'}(s, a)] \\ &= \mathbb{E}_{s'}[R'(s, a, s') + \gamma \max_{a' \in \mathcal{A}} Q^\pi_{M'}(s, a)]. \qquad (7) \end{aligned}$$

4 Shaping Control Variates for Off-Policy Evaluation

In this section, we introduce a new shaped control variate for off-policy evaluation in sparse reward settings. We define a new estimator called Shaping Control for Off-Policy Evaluation (SCOPE) that integrates control variates based on reward shaping into per-step IS estimators for OPE. The main idea of SCOPE is to learn shaping parameters such that the variance of the estimator is minimised.

Definition 2. *(Shaping Control Variates for Off-Policy Evaluation). The SCOPE estimator for OPE is given by:*

$$\hat{V}^{\pi_e}_{SCOPE} := \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{T-1}\gamma^t\omega^{(i)}_{0:t}(r^{(i)}_t + \gamma\phi(s^{(i)}_{t+1};\beta) - \phi(s^{(i)}_t;\beta)), \tag{8}$$

where $\phi(s)$ is a potential function parameterised by β.

In this form, the SCOPE estimator may be viewed as a variant of step-IS[1] where we introduce an importance weighted shaping term based on PBRS to the step-IS estimate. Note however that the idea of using control variates based on shaping is general enough to be integrated into various other OPE estimators such as weighted-IS variants, as well as DR and DualDICE to produce a family of shaped OPE estimators. Next, we state the bias and consistency of the SCOPE estimator given Assumptions 1 and 2.

Lemma 1. *The SCOPE estimator for stochastic evaluation policy π_e is an unbiased estimator of the shaped value function $V^{\pi_e}_{M'}$. Proof: See appendix for details.*

Lemma 2. *SCOPE is a strongly consistent estimator of $V^{\pi_e}_{M'}$ i.e. $\lim_{n\to\infty}\hat{V}^{\pi_e}_{SCOPE} = V^{\pi_e}_{M'}$ almost surely. As a result, this implies that SCOPE estimators are well-posed.*
Proof Sketch: Since we know the estimate is unbiased from Lemma 1 and our data set D consists of n independent and identically distributed samples, we can infer from Khintchine's Strong Law of Large Numbers that $\lim_{n\to\infty}\hat{V}^{\pi_e}_{SCOPE} = V^{\pi_e}_{M'}$. See appendix for details.

Since SCOPE is an unbiased estimate of i.i.d. trajectories, standard concentration inequalities for uncertainty in our off-policy estimates such as Hoeffding's Inequality and others Thomas (2015) apply.

4.1 Variance Analysis

Just because an estimator is unbiased and consistent does not mean that it is useful; it may still have high variance given a finite number of samples. In this section, we present the variance of the SCOPE estimator for any choice of shaping control variate; next we will describe how the choice of variate can be optimised to minimise the variance of the estimator.

Lemma 3. *The variance of the SCOPE estimator in Equation 8 for stochastic evaluation policy π_e is given by,*

$$\begin{aligned}
\mathbb{V}_{P^{\pi_b}_\tau}(\hat{V}^{\pi_e}_{SCOPE}) &= \mathbb{V}_{P^{\pi_b}_\tau}[\hat{V}^{\pi_e}_{stepIS}] + 2\mathbb{E}_{P^{\pi_b}_\tau}[R_t(\delta + \gamma^T\omega_{0:T-1}\phi(s_T;\beta) - \phi(s_0;\beta))] \\
&- 2V^{\pi_e}_{M'}\mathbb{E}_{P^{\pi_b}_\tau}[\delta + \gamma^T\omega_{0:T-1}\phi(s_T;\beta) - \phi(s_0;\beta)] + \mathbb{V}_{P^{\pi_b}_\tau}[\delta]^2 \\
&+ \mathbb{V}_{P^{\pi_b}_\tau}\left[\left(\gamma^T\omega_{0:T-1}\phi(s_T;\beta) - \phi(s_0;\beta)\right)\right] \\
&- 2\mathbb{E}_{P^{\pi_b}_\tau}\left[\delta\left(\gamma^T\omega_{0:T-1}\phi(s_T;\beta) - \phi(s_0;\beta)\right)\right] \\
&+ 2\mathbb{E}_{P^{\pi_b}_\tau}[\delta]\mathbb{E}_{P^{\pi_b}_\tau}\left[\gamma^T\omega_{0:T-1}\phi(s_T;\beta) - \phi(s_0;\beta)\right]
\end{aligned} \tag{9}$$

where $\delta = \sum_{t=1}^{T-1}\gamma^t\phi(s_t;\beta)(\omega_{0:t-1} - \omega_{0:t})$.
Proof: See appendix for details.

The variance formulation in Equation 9 holds for any parameterisation of ϕ and can be minimised in general as we discuss Section 4.2. Like Thomas and Brunskill (2016), we note that it is also possible to derive the weighted SCOPE estimator where the per-step IS weights in SCOPE are replaced by per-step weighted IS (step-WIS). This results in the introduction of some bias, but can potentially reduce the overall variance and thus the MSE.

[1] Note that in general, the notion of PBRS would not be expected to help in the standard (non-step) version of IS because if the weight of a full trajectory is zero, then no control variate—SCOPE or otherwise—will help.

4.2 Optimising the variance of SCOPE

Equation 9 provides a closed form expression for the variance of the SCOPE estimator. Hence we can compute all the expectations with respect to $P_\tau^{\pi_b}$ using Monte Carlo gradients to determine the optimal parameters β that minimise the variance for arbitrary ϕ. We examine the case where the potential function $\phi(s;\beta) = \beta^\top \psi(s)$ for some for some K-dimensional vector of feature functions $\psi(s)$. Here, the variance formulation in Equation 9 is a quadratic convex function that is smooth in β. In this case, optimising over β yields,

$$\begin{aligned}\nabla_\beta \mathbb{V}(\hat{V}^{\pi_e}_{SCOPE}) &= \nabla_\beta \mathbb{E}_{P_\tau^{\pi_b}}[\hat{V}^{\pi_e^2}_{SCOPE}] - 2\mathbb{E}_{P_\tau^{\pi_b}}[\hat{V}^{\pi_e}_{SCOPE}]\nabla_\beta \mathbb{E}_{P_\tau^{\pi_b}}[\hat{V}^{\pi_e}_{SCOPE}] \\ &= \nabla_\beta[\hat{V}^{\pi_e^2}_{stepIS} + 2\beta^\top \Delta \hat{V}^{\pi_e}_{stepIS} + \beta^\top \Delta\Delta^\top \beta] - 2V^{\pi_e}_{M'}\mathbb{E}_{P_\tau^{\pi_b}}[\Delta], \quad (10)\end{aligned}$$

since the estimator is unbiased and $\Delta = \sum_{t=0}^{T-1} \gamma^t \omega_{0:t}(\gamma\psi(s_{t+1}) - \psi(s_t))$. Thus if we minimise the variance by setting Equation 10 to 0 we get,

$$\beta = \mathbb{E}_{P_\tau^{\pi_b}}[\Delta\Delta^\top]^{-1}[V^{\pi_e}_{M'}\mathbb{E}_{P_\tau^{\pi_b}}[\Delta] - \mathbb{E}_{P_\tau^{\pi_b}}[\Delta \hat{V}^{\pi_e}_{stepIS}]] \quad (11)$$

When the potential function Φ is constant, this is analogous to having a constant DM component in DR which produces some improvements over classic IS (see for instance, Voloshin et al. (2019)).

Some intuition can also be gained from contrasting SCOPE with MRDR—both algorithms learn a control variate whose sole purpose is to minimise variance, relying on the unbiasedness of DR to take care of the overall bias. However, while MRDR learns a state-action dependent control variate, SCOPE learns a control variate which only depends on the state. This property makes it easier to estimate and optimise, and less susceptible to noise due to action stochasticity. For domains where some states can be thought of as having intrinsic values (for example, being physically close to a goal in a navigational domain or having a low number of diseased cells in a healthcare setting), this simplification can reap the benefit of ease of optimisation, without sacrificing important aspects of the dynamics. Moreover, in sparse reward settings using a poor estimate of the true value function as a control variate in DR may result in high variance. In these instances, it can be easier to estimate a shaping control variate dependent only on the state for OPE to still produce a lower variance estimate overall.

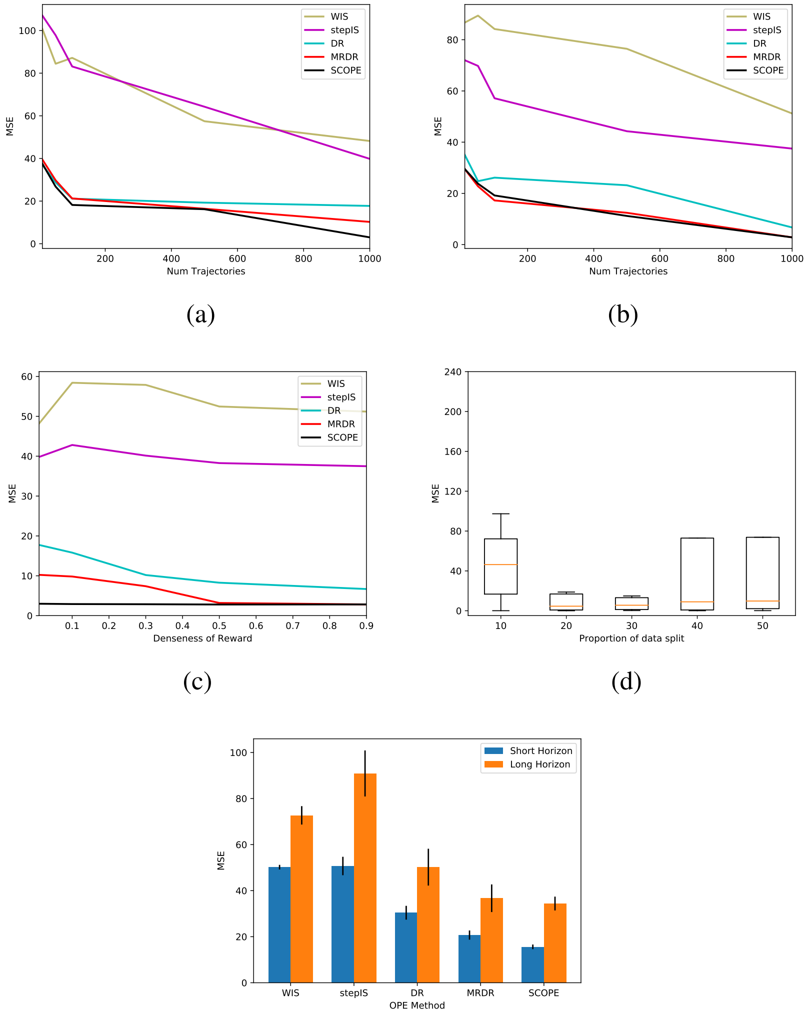
Learning an appropriate ϕ is key to the performance of SCOPE. Though in practice it may be easier to obtain a reasonable estimate of ϕ than learning an approximate model of the value function, like DR-based methods, SCOPE requires us to subset our data to empirically estimate β, while using the remaining samples to perform OPE. For this purpose, we use bootstrapping over different splits of the data. The optimal data split minimises the MSE. This procedure is summarised in Algorithm 1.

Algorithm 1 Off-Policy evaluation using SCOPE

```
Input: $\pi_e$, $\pi_b$, $\mathcal{D}$, confidence $c$ and number of bootstrap estimates $B$
Output: $\hat{V}^{\pi_e}_{SCOPE}$
$j = 1$
While $j \leq B${
Partition $\mathcal{D}$ into $\mathcal{D}\backslash\mathcal{D}_j$ and $\mathcal{D}_j = \{H_1^j, \dots, H_k^j\}$ of histories $H_k^j$
Estimate $\beta$ using $\mathcal{D}_j$ from Eqn. 10 and obtain $\phi$
$\hat{V}^{\pi_e}_{SCOPE_j}$ = SCOPE($\pi_e, \pi_b, \phi, \mathcal{D}\backslash\mathcal{D}_j$)
}
Sort $\hat{V}^{\pi_e}_{SCOPE_j}$
$l$ = $cB$
Return $\hat{V}^{\pi_e}_{SCOPE_l}$
```

5 Gridworld Demonstration

Domain Description and Baselines: We consider the 2D rectangular gridworld environment of size 3 x 4, where the agent tries to move a start state s_0 in the bottom left to a goal state in the bottom right. We assume that the initial state s_0 is fixed. We train an ϵ-greedy policy with $\epsilon = 0.3$ with an



(a)

(b)

(c)

(d)

(e)

Figure 1: (a) Comparison of MSE in sparse setting. The SCOPE estimator outperforms other estimators in sparse reward settings. Moreover, the SCOPE estimator is empirically consistent. (b) Comparison of MSE in dense reward setting. SCOPE and MRDR exhibit similar performance. (c) Comparison of MSE over different degrees of sparsity. SCOPE outperforms baselines in sparse reward settings. (d) MSE of SCOPE according to the proportion of data used for training potential functions. The optimal data split where the MSE is minimised uses 30% of the data for training ϕ. (e) Performance over varying horizon lengths. SCOPE is less sensitive to longer horizons.

average time of 100 time steps to complete the task. The true on-policy value estimate V^{π_e} is the Monte Carlo estimate via 1000 rollouts of π_e. We compared the performance of SCOPE to step-IS, WIS, DR and MRDR trained with a linear model for $\hat{Q}$ as baselines.[2]

Improvement in performance under sparse rewards: To assess the influence of varying sparsity on our estimates, we introduce a metric ζ to measure the reward sparsity per trajectory as the ratio of the number of non-zero rewards per trajectory over the trajectory length. We compared SCOPE to each of the baselines under (i) a sparse setting (ζ = 0.01) where the reward is 50 for entering the

[2] See appendix for further experiments and details on setup.

goal state and 0 for all other steps in a trajectory, and (ii) a dense reward setting (ζ = 0.9) where intermediate rewards of -5 are given for entering pit cells, and rewards of +1 are provided for 90% of the other steps. The results are shown in Figures 1(a) and 1(b). Overall, both sets of results serve as an empirical check for consistency of SCOPE. As the number of observed trajectories grows, the MSE decreases. In the dense setting, the performance of MRDR and SCOPE estimators is similar as both shaped and MRDR control variates operate similarly. In the sparse setting however, SCOPE outperforms MRDR. This is a result of SCOPE's ability to compensate for the lack of reward signal using the shaping control variate. We also show the performance of SCOPE for varying degrees of reward sparsity in Figure 1(c) for n = 1000 trajectories. SCOPE performs significantly better than the other baselines when rewards sparser.

Efficiency of learning shaping control variates ϕ: We used Algorithm 1 to determine the influence of the proportion of data used to determine ϕ. Figure 1(d) shows the bootstrapped variance over different splits of the data. The MSE is minimised when 30% of the data is used for shaping. Algorithm 1 enables us to compute the optimal data split such that we can minimise the variance. Importantly, the proportion of data required to learn the optimal shaping parameters is less than the amount of data required to learn a reasonable Q-function estimate in DR.

Robustness against varying horizon lengths: Traditional IS methods tend to be very sensitive to horizon length since π_e and π_b may have little overlap over longer horizons. We assess the performance of our estimator over horizon lengths of 100 and 500 respectively in Figure 1(e). SCOPE is less sensitive to increases in horizon. As we get closer to the goal, the effect of shaping is more pronounced since states near the goal state carry more information about the goal. The shaping control variate allows us to exploit this information for variance reduction.

6 Discussion and Conclusion

A lack of intermediate reward signal often increases the variance of standard OPE estimates. Control variates can help by introducing intermediate rewards in a principled way, and here we presented a new method for OPE based on reward shaping particularly well-suited to the sparse reward setting. Our approach achieves state-of-the-art performance whilst being straightforward to optimise. Specifically, we find

Shaped control variates outperform existing methods for OPE under sparse reward settings. SCOPE achieves optimal performance in comparison to existing baselines in sparse settings. Reward shaping serves as a special type of control variate that can be used instead of the standard value function for sparse problems. Here, the introduction of shaping terms densifies rewards to learn an improved control variate, where standard value function estimates would otherwise be poor as a result of reward sparsity.

For dense reward settings, linear SCOPE has performance equivalent to MRDR. In dense settings, the additional shaping terms in SCOPE in Eqn. 8 reduce to value function estimates. Learning the optimal shaping parameter is then equivalent to learning the model parameter that minimises the variance of DR estimators as in the MRDR approach. As a result, the performance of SCOPE is equivalent to that of MRDR.

Unlike MRDR, shaping control variates with more expressive function classes still allows for easy optimisation. When we perform SCOPE with higher order function classes that are not necessarily linear, the performance of SCOPE significantly improves regardless of whether we are in a sparse or dense reward setting. For instance, we see reductions in MSE to +/- 1.26 for the Gridworld task. Regardless of the function class, SCOPE can still be optimised easily.

Optimising SCOPE requires splitting data appropriately, but has a closed-form. An essential design consideration for the performance of SCOPE is how to divide the training data set appropriately to estimate the shaping parameters. In our experiments, this is done by bootstrapping over different data splits. Since we can in general, write out a closed form expression for the variance of SCOPE, optimisation is fairly straightforward in comparison to existing techniques.

References

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Yinlam Chow, Marek Petrik, and Mohammad Ghavamzadeh. Robust policy optimization with baseline guarantees. *arXiv preprint arXiv:1506.04514*, 2015.

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

Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pages 1446–1455, 2018.

Raphael Fonteneau, Susan A Murphy, Louis Wehenkel, and Damien Ernst. Batch mode reinforcement learning based on the synthesis of artificial trajectories. *Annals of operations research*, 208(1):383–416, 2013.

Omer Gottesman, Fredrik Johansson, Joshua Meier, Jack Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xuefeng Peng, et al. Evaluating reinforcement learning algorithms in observational health settings. *arXiv preprint arXiv:1805.12298*, 2018.

Omer Gottesman, Yao Liu, Scott Sussex, Emma Brunskill, and Finale Doshi-Velez. Combining parametric and nonparametric models for off-policy evaluation. *arXiv preprint arXiv:1905.05787*, 2019.

Josiah P Hanna, Peter Stone, and Scott Niekum. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 538–546. International Foundation for Autonomous Agents and Multiagent Systems, 2017.

Anna Harutyunyan, Sam Devlin, Peter Vrancx, and Ann Nowé. Expressing arbitrary reward functions as potential-based advice. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

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

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029, 2016.

Peng Liao, Predrag Klasnja, and Susan Murphy. Off-policy estimation of long-term average outcomes with applications to mobile health. *arXiv preprint arXiv:1912.13088*, 2019.

Yao Liu, Omer Gottesman, Aniruddh Raghu, Matthieu Komorowski, Aldo A Faisal, Finale Doshi-Velez, and Emma Brunskill. Representation balancing mdps for off-policy policy evaluation. In *Advances in Neural Information Processing Systems*, pages 2644–2653, 2018.

Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1077–1084. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, pages 2318–2328, 2019.

Andrew Y Ng. *Shaping and policy search in reinforcement learning*. PhD thesis, University of California, Berkeley, 2003.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *International Conference on Machine Learning*, 1999.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. *arXiv preprint arXiv:1905.05824*, 2019.

Cosmin Paduraru. *Off-policy evaluation in Markov decision processes*. PhD thesis, 2013.

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.

Sonali Parbhoo, Omer Gottesman, Andrew Slavin Ross, Matthieu Komorowski, Aldo Faisal, Isabella Bon, Volker Roth, and Finale Doshi-Velez. Improving counterfactual reasoning with kernelised dynamic mixing models. *PloS one*, 13(11):e0205839, 2018.

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

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

Andrea Rotnitzky and James M Robins. Semi-parametric estimation of models for means and covariances in the presence of missing data. *Scandinavian Journal of Statistics*, pages 323–333, 1995.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org, 2017.

David Silver, Leonard Newnham, David Barker, Suzanne Weller, and Jason McFall. Concurrent reinforcement learning from customer interactions. In *International Conference on Machine Learning*, pages 924–932, 2013.

Guy Tennenholtz, Shie Mannor, and Uri Shalit. Off-policy evaluation in partially observable environments. *arXiv preprint arXiv:1909.03739*, 2019.

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

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.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudik. Optimal and adaptive off-policy evaluation in contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3589–3597. JMLR. org, 2017.

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