INFINITE-HORIZON OFF-POLICY POLICY EVALUATION WITH MULTIPLE BEHAVIOR POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider off-policy policy evaluation when the trajectory data are generated by multiple behavior policies. Recent work has shown the key role played by the state or state-action stationary distribution corrections in the infinite horizon context for off-policy policy evaluation. We propose estimated mixture policy (EMP), a novel class of partially policy-agnostic methods to accurately estimate those quantities. With careful analysis, we show that EMP gives rise to estimates with reduced variance for estimating the state stationary distribution correction while it also offers a useful induction bias for estimating the state-action stationary distribution correction. In extensive experiments with both continuous and discrete environments, we demonstrate that our algorithm offers significantly improved accuracy compared to the state-of-the-art methods.

1 INTRODUCTION

In many real-world decision-making scenarios, evaluating a novel policy by directly executing it in the environment is generally costly and can even be downright risky. Examples include evaluating a recommendation policy (Swaminathan et al., 2017; Zheng et al., 2018), a treatment policy (Hirano et al., 2003; Murphy et al., 2001), and a traffic light control policy (Van der Pol & Oliehoek, 2016). Off-policy policy evaluation methods (OPPE) utilize a set of previously-collected trajectories (for example, website interaction logs, patient trajectories, or robot trajectories) to estimate the value of a novel decision-making policy without interacting with the environment (Precup et al., 2001; Dudík et al., 2011). For many reinforcement learning applications, the value of the decision is defined in a long- or infinite-horizon, which makes OPPE more challenging.

The state-of-the-art methods for infinite-horizon off-policy policy evaluation rely on learning *(discounted) state stationary distribution corrections* or *ratios*. In particular, for each state in the environments, these methods estimate the likelihood ratio of the long-term probability measure for the state to be visited in a trajectory generated by the *target policy*, normalized by the probability measure generated by the *behavior policy*. This approach can effectively avoid the exponentially high variance compared to the more classic importance sampling (IS) estimation methods (Precup, 2000; Dudík et al., 2011; Hirano et al., 2003; Wang et al., 2017; Murphy et al., 2001), especially for infinite-horizon policy evaluation (Liu et al., 2018; Nachum et al., 2019; Hallak & Mannor, 2017). However, learning state stationary distribution requires detailed information on distributions of the behavior policy, and we call them *policy-aware* methods. As a consequence, policy-aware methods are difficult to apply when off-policy data are pre-generated by multiple behavior policies or when the behavior policy's form is unknown. To address this issue, Nachum et al. (2019) proposes a *policy-agnostic* method, DualDice, which learns the joint state-action stationary distribution correction that is much higher dimension, and therefore needs more model parameters than the state stationary distribution. Besides, there is no theoretic comparison between policy-aware and policy-agnostic methods.

In this paper, we propose a *partially policy-agnostic* method, EMP (estimated mixture policy) for infinite-horizon off-policy policy evaluation with multiple known or unknown behavior policies. EMP is partially policy-agnostic in the since that it does not necessarily require knowledge of the individual behavior policies. Instead, it involves a pre-estimation step to estimate a single mixed policy that will be defined formally later. Like the method in Liu et al. (2018), EMP also learns the state stationary distribution correction, so it remains computationally cheap and is scalable in terms of the number of behavior policies. Inspired by Hanna et al. (2019), we construct a theoretical bound for the mean square error (MSE) of the stationary distribution corrections learned by EMP. In particular, we show that in the single-behavior policy setting, EMP yields smaller MSE than the policy-aware method. On the other hand, compared to DualDice, EMP learns the state stationary distribution correction of smaller dimension, more importantly the estimation of the mixture policy can be considered as an inductive bias as far as the stationary distribution correction is concerned, and hence could achieve better performance when the pre-estimation is not expensive. In addition, we propose an ad-hoc improvement of EMP, whose theoretical analysis is left for future studies. EMP is compared with both policy-aware and policy-agnostic methods in a set of continuous and discrete control tasks and shows significant improvement.

2 Background and Related Work

2.1 Infinite-horizon Off-policy Policy Evaluation

We consider a Markov Decision Process (MDP) and our goal is to estimate the infinite-horizon *average reward*. The environment is specified by a tuple $\mathcal{M} = \langle S, A, R, T \rangle$, consisting of a state space, an action space, a reward function, and a transition probability function. A policy π interacts with the environment iteratively, starting with an initial state s_0. At step $n = 0, 1, \dots$, the policy produces a distribution $\pi(\cdot|s_n)$ over the actions A, from which an action a_n is sampled and applied to the environment. The environment stochastically produces a scalar reward $r(s_n, a_n)$ and a next state $s_{n+1} \sim T(\cdot|s_n, a_n)$. The infinite-horizon average reward under policy π is

$$R_\pi = \lim_{N\to\infty} \frac{1}{N+1} \sum_{n=0}^{N} \mathcal{E}\left[r(s_n, a_n)\right].$$

Without gathering new data, off-policy policy evaluation (OPPE) considers the problem of estimating the expected reward of a target policy π via a pre-collected state-action-reward tuples from policies that are different from π, which are called behavior policies. In our paper, we consider the general setting that the data are generated by multiple behavior policies $\pi_j (j = 0, 1, .., m)$. Most OPPE literature has focused on the single-behavior-policy case where $m = 1$. In this case, we denote the behavior policy by π_0. Roughly speaking, most OPPE methods can be grouped into two categories: importance-sampling(IS) based OPPE and stationary-distribution-correction based OPPE.

2.2 Importance Sampling Policy Evaluation Using Exact and Estimated Behavior Policy

As for short-horizon off-policy policy evaluation, importance sampling policy evaluation (IS) methods (Precup et al., 2001; Dudík et al., 2011; Swaminathan et al., 2017; Precup et al., 2000; Horvitz & Thompson, 1952) have shown promising empirical results. The main idea of importance sampling based OPPE is using importance weighting π/π_j to correct the mismatch between the target policy π and the behavior policy π_j that generates the trajectory.

Li et al. (2015) and Hanna et al. (2019) show that using estimated behavior policy in the importance weighting can obtain importance sampling estimation with smaller mean square error (MSE). EMP also uses estimated policy, but there are two key difference between EMP and the previous works: (1) EMP is not an IS-based method, it involves a min-max problem; (2) EMP focuses on multiple-behavior-policy setting while previous works have focused on single-behavior setting.

2.3 Policy Evaluation via Learning Stationary Distribution Correction

The state-of-the-art methods for long-horizon off-policy policy evaluation are stationary-distribution-correction based (Liu et al., 2018; Nachum et al., 2019; Hallak & Mannor, 2017). Let $d_{\pi_0}(s)$ and $d_\pi(s)$ be the stationary distribution of state s under the behavior policy π_0 and target policy π respectively. The main idea is directly applying importance weighting by $\omega = d_\pi/d_{\pi_0}$ on the stationary state-visitation distributions to avoid the exploding variance suffered from IS, and estimate the average reward as

$$R_\pi = \mathbb{E}_{(s,a)\sim d_\pi}[r(s,a)] = \mathbb{E}_{(s,a)\sim d_{\pi_0}}\left[\omega(s) \cdot \frac{\pi(a|s)}{\pi_0(a|s)} r(s,a)\right].$$

For example, Liu et al. (2018) uses min-max approach to estimate ω directly from the data. This class of methods require exact knowledge of behavior policy π_0 and are not straightforward to apply in multiple-behavior-policy setting. Recently, Nachum et al. (2019) proposes DualDice to overcome such limitation by learning the state-action stationary distribution correction $\omega(s,a) = d_\pi(s)\pi(a|s)/d_{\pi_0}(s)\pi_0(a|s)$.

3 Single Behavior Policy

We first consider the task of stationary distribution correction learning in the simple case where the data are generated by a single behavior policy as previous state stationary distribution correction methods. To explain the min-max problem formulation of the learning task, we first breifly review the method introduced by Liu et al. (2018) in Section 3.1, which we shall refer as the BCH method in the rest of the paper. In Section 3.2, we show that it is beneficial to replace the exact values of the behavior policy in the min-max problem by their estimated values in two folds. First, this extends the method to application setting where the behavior policy is unknown. Second, even when the behavior policy is known with exact values, we prove that the stationary distribution correction learned by the min-max problem with estimated behavior policy has smaller MSE. We will deal with multiple-behavior-policy cases in Section 4.

3.1 Learning Stationary Distribution Correction with Exact Behavior Policy

Assume the data, consisting of state-action-next-state tuples, are generated by a single behavior policy π_0, i.e. $\mathcal{D} = \{(s_n, a_n, s'_n) : n = 1, 2, ..., N\}$. Recall that d_{π_0} and d_π are the stationary state distribution under the behavior and target policy respectively, and $\omega = d_\pi/d_{\pi_0}$ is the *stationary distribution correction*. In the rest of Section 3, by slight notation abusion, we also denote $d_\pi(s,a) = d_\pi(s)\pi(a|s)$, $d_{\pi_0}(s,a) = d_{\pi_0}(s)\pi_0(a|s)$ and $d_{\pi_0}(s,a,s') = d_{\pi_0}(s)\pi_0(a|s)T(s'|a,s)$.

We briefly review the BCH method proposed by Liu et al. (2018). As $d_\pi(s)$ is the stationary distribution of s_n as $n \to \infty$ under policy π, it follows that:

$$d_\pi(s') = \sum_{s,a} d_\pi(s)\pi(a|s)P(s'|s,a) = \sum_{s,a} \omega(s)\frac{\pi(a|s)}{\pi_0(a|s)} d_{\pi_0}(s)\pi_0(a|s)T(s'|a,s), \quad \forall s'. \tag{1}$$

Therefore, for any function $f : S \to \mathbb{R}$,

$$\sum_{s'} \omega(s')d_{\pi_0}(s')f(s') = \sum_{s,a,s'} \omega(s)\frac{\pi(a|s)}{\pi_0(a|s)} d_{\pi_0}(s)\pi(a|s)T(s'|a,s)f(s').$$

Recall that $d_{\pi_0}(s,a,s') = d_{\pi_0}(s)\pi_0(a|s)T(s'|a,s)$, so ω and the data sample satisfy the following equation

$$\mathbb{E}_{(s,a,s')\sim d_{\pi_0}}\left[\left(\omega(s') - \omega(s)\frac{\pi(a|s)}{\pi_0(a|s)}\right) f(s')\right] = 0, \text{ for all } f.$$

BCH solves the above equation via the following min-max problem:

$$\min_\omega \max_f \mathbb{E}_{(s,a,s')\sim d_{\pi_0}}\left[\left(\omega(s') - \omega(s)\frac{\pi(a|s)}{\pi_0(a|s)}\right) f(s')\right]^2, \tag{2}$$

and use *kernel method* to solve ω. The derivation of kernel method are put in Appendix A.

3.2 Learning Stationary Distribution Correction with Estimated Behavior Policy

The objective function in the min-max problem (2), evaluated by data sample, can be viewed as a one-step importance sampling estimation. As shown in Hanna et al. (2019), importance sampling with estimated behavior policy has smaller MSE. Motivated by this fact and the heuristic that better objective function evaluation will lead to more accurate solution, we show that the BCH method can also be improved by using estimated behavior policy and obtain smaller asymptotic MSE. We will use this result to build theoretic guarantee for the performance of EMP method in Section 4.

To formally state the theoretic result, we need introduce more notation. Assume that we are given a class of stationary distribution correction $\Omega = \{\omega(\eta; s) : \eta \in \mathcal{E}_\eta\}$, and there exists $\eta_0 \in \mathcal{E}_\eta$ such that the true distribution correction $\omega(s) = \omega(\eta_0; s)$. Let $\omega(\tilde{\eta}; s)$ be the stationary distribution correction learned by the min-max problem (2) and $\omega(\hat{\eta}; s)$ be that learned by a min-max problem using estimated policy:

$$\min_{\omega} \max_{f} \; \mathbb{E}_{(s,a,s')\sim\mathcal{D}} \left[\left(\omega(s') - \omega(s) \frac{\pi(a|s)}{\hat{\pi}_0(a|s)} \right) f(s') \right]^2 . \tag{3}$$

The intuition is that the value of $\hat{\pi}_0$ is estimated from the data sample and appears in the denominator, as a result, it could cancel out a certain amount of random error in data sample. We use a maximum likelihood method to estimate the behavior policies for discrete and continuous control tasks. The details are in Appendix E.2. Based on the proof techniques in Henmi et al. (2007), we establish the following theoretic guarantee that using estimated behavior policy yields better estimates of the stationary distribution correction.

Theorem 1. *Under some mild conditions, we have, asymptotically*

$$E[(\hat{\eta} - \eta_0)^2] \leq E[(\tilde{\eta} - \eta_0)^2].$$

As a direct consequence, we derive the finite-sample error bound for $\hat{\eta}$.

Corollary 1. *(informal) Let N be the number of (s, a, s') tuples in the data,*

$$\mathbb{E}[(\hat{\eta} - \eta_0)^2] = O\left(\frac{1}{N}\right)$$

.

The precise conditions for Theorem 1 and Corollary 1 to hold and their proofs are in Appendix B.

4 EMP FOR MULTIPLE BEHAVIOR POLICIES

In this section, we shall propose our EMP method for off-policy policy evaluation with multiple known or unknown behavior policies and establish theoretic results on variance reduction of EMP.

Before that, we first give a detailed description on the data sample and its distribution. Assume the state-action-next-state tuples are generated by m different unknown behavior policies $\pi_j, j = 1, 2, ..., m$. Let $d_{\pi_j}(s)$ be the stationary state distribution and N_j be the number of state-action-next-state tuples by policy π_j, for $j = 1, 2, ..., m$. Let $N = \sum_j N_j$ and denote by $w_j = N_j/N$ the proportion of data generated by policy π_j. We use $\mathcal{D}$ to denote the data set and $\mathcal{D} = \{(s_{j,n_j}, a_{j,n_j}, s'_{j,n_j}) : j = 1, 2, .., m, n_j = 1, 2, ..., N_j\}$. Note that the policy label j in the subscript is only for notation clarity and it is not revealed in the data. Then, a single (s, a, s') tuple simply follows the marginal distribution $d_0(s, a, s') := \sum_j w_j d_{\pi_j}(s)\pi_j(a|s)T(s'|a, s)$. With slight notation abusion, we write $d_0(s, a) = \sum_j w_j d_{\pi_j}(s)\pi(a|s)$.

4.1 EMP METHOD

Now we derive the EMP method in the multiple-behavior-policy setting and explicitly explain what is the mixed policy to be estimated in EMP.

Let $d_0 := \sum_j w_j d_{\pi_j}$ be the mixture of stationary distributions of the behavior policies. For each state-action pair (a, s), define $\pi_0(a|s)$ as the weighted average of the behavior policies:

$$\pi_0(a|s) := \sum_j \frac{w_j d_{\pi_j}(s)}{d_{\pi_0}(s)} \pi_j(a|s), \forall\ (s, a). \tag{4}$$

It is easy to check that for each s, $\pi_0(\cdot|s)$ is a distribution on the action space and hence defines a policy by itself. We call π_0 the *mixed policy*. Let $\omega = d_\pi/d_0$, which is a state distribution ratio. Then, d_0, π_0 and ω satisfy the following relation with the average reward R_π.

Proposition 1.

$$R_\pi = E_{(s,a)\sim d_0} \left[\omega(s) \frac{\pi(a|s)}{\pi_0(a|s)} r(s, a) \right] . \tag{5}$$

Besides, the state distribution ratio ω can be characterized by the stationary equation.

Proposition 2. *The function $\omega(s) = d_\pi(s)/d_{\pi_0}(s)$ (up to a constant) if and only if,*

$$\mathbb{E}_{(s,a,s')\sim d_0}\left[\left(\omega(s') - \omega(s)\frac{\pi(a|s)}{\pi_0(a|s)}\right) f(s')\right] = 0, \textit{ for all } f : S \to \mathbb{R}. \tag{6}$$

In the special case when $m = 1$, i.e. the data are generated by a single behavior policy, Proposition 2 reduces to Theorem 1 of Liu et al. (2018). The above two Propositions indicate that, to certain extend, the (s, a, s') tuples generated by multiple behavior policies can be pooled together and treated as if they are generated by a single behavior policy π_0.

Note that expression of π_0 (4) involves not only the behavior policies but also the state stationary distributions. In EMP method, we shall use a pre-estimation step to generate an estimate $\hat{\pi}_0$ from the data. Based on Proposition 2, the state distribution ratio ω can be estimated by the following min-max problem

$$\min_{\omega} \max_{f} \, \mathbb{E}_{(s,a,s')\sim\mathcal{D}}\left[\left(\omega(s') - \omega(s)\frac{\pi(a|s)}{\hat{\pi}_0(a|s)}\right) f(s')\right]^2. \tag{7}$$

Finally, EMP estimates the average reward according to (5) where d_0 is approximate by data $\mathcal{D}$.

Applying Theorem 1, we show that using the estimated $\hat{\pi}_0$ in EMP can actually reduce the MSE of the learned stationary distribution ratio ω.

Proposition 3. *Under the same conditions of Theorem 1, if $\omega(\tilde{\eta}; s)$ and $\omega(\hat{\eta}; s)$ are the stationary distribution correction learned from (7) and from the same min-max problem but with exact value of π_0, then, asymptotically*

$$E[(\hat{\eta} - \eta_0)^2] \leq E[(\tilde{\eta} - \eta_0)^2].$$

As a result, $E[(\hat{\eta} - \eta_0)^2] = O\left(\frac{1}{N}\right)$.

4.2 Why Pooling is Beneficial for EMP

One important feature of EMP is that it pools the data from different policy behaviors together and treat them as if they are from a single mixed policy. Of course, pooling makes EMP applicable to settings with minimal information on the behavior policies, for instance, EMP does not even require the knowledge on the number of behavior policies. In this part, we show that, the pooling feature of EMP is not just a compromise to the lack of behavior policy information, it also leads to variance reduction in an intrinsic manner.

If instead, the data can be classified according to the behavior policies and treated separately, we can still use EMP, which reduces to (3), or any other single-behavior-policy method, to obtain the stationary distribution correction $\omega_j = d_\pi/d_{\pi_j}$ for each behavior policy. Given ω_j, a common approach for variance reduction is to apply multiple importance sampling (MIS) (Tirinzoni et al., 2019; Veach & Guibas, 1995) technique and the average reward estimator is of the form

$$\hat{R}_{MIS} = \sum_{j=1}^{m} \frac{1}{N_j} \sum_{n=1}^{N_j} h_j(s_{j,n})\omega_j(s_{j,n})\pi(a_{j,n}|s_{j,n})r(s_{j,n}, a_{j,n}), \tag{8}$$

where the function h is often referred to as heuristics and must be a partition of unity, i.e., $\sum_j h_j(s) = 1$ for all $s \in S$. It has been proved by (Veach & Guibas, 1995) that MIS is unbiased, and, for given $w_j = N_j/N$, there is an optimal heuristic function to minimize the variance of $\hat{R}_{MIS}$.

Proposition 4. *For MIS with fixed values of $w_j, j = 1, 2, ..., m$, among all possible values of heuristics h, the balanced heuristic*

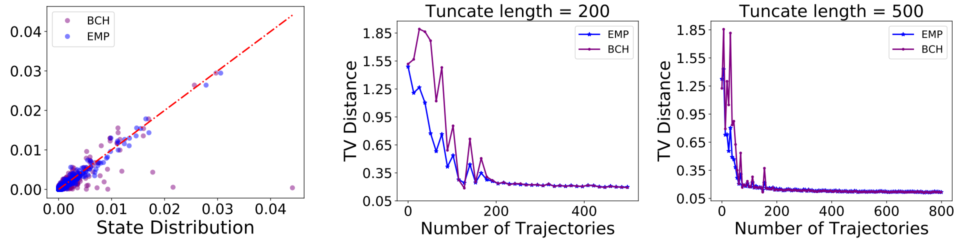
$$h_j(s) = \frac{w_j d_{\pi_j}(s)}{\sum_{j=1}^{m} w_j d_{\pi_j}(s)}, \; \forall j = 1, 2, ..., m \textit{ and } s \in S,$$

reaches the minimal variance.

Plug the optimal heuristic $h_j(s)$ into MIS estimator (8), and we will obtain that the optimal MIS estimator coincides with the EMP estimator (5), i.e.

$$\hat{R}_{MIS} = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\frac{d_\pi(s)}{d_0(s)}\pi(a|s)r(s,a)\right]. \tag{9}$$

In this light, by pooling the data together and directly learning ω. EMP also learns the optimal MIS weight inexplicitly.



(a) State Distribution Comparison

(b) TV Distance

Figure 1: (a) shows that scatter plot of pairs $(\hat{d}_{\pi_{\text{true}}}, d_\pi)$ and pairs $(\hat{d}_{\pi_{\text{esti}}}, d_\pi)$. The diagonal line indicates exact estimation. The default values of the number of trajectories is 200, and the length of horizon is 200. (b) shows the weighted total variation distance (TV distance) between $\hat{d}_{\pi_{\text{true}}}$ and d_π, $\hat{d}_{\pi_{\text{esti}}}$ and d_π respectively, along different number of trajectories and the length of horizons.

5 Experiment

In this section, we conduct experiments in three discrete-control tasks Taxi, Singlepath, Gridworld and one continuous-control task Pendulum (see Appendix E.1 for the details), with following purposes: (i) to compare the performance of distribution correction learning using policy-aware, policy-agnostic and partially agnostic-policy methods (in Sec 5.1); (ii) to compare the performance of the proposed EMP with existing OPPE methods (in Sec 5.1 and 5.2); (iii) to explore potential improvement of EMP methods (in Sec 5.2). We will release the codes with the publication of this paper for relevant study.

5.1 Results for Single Behavior Policy

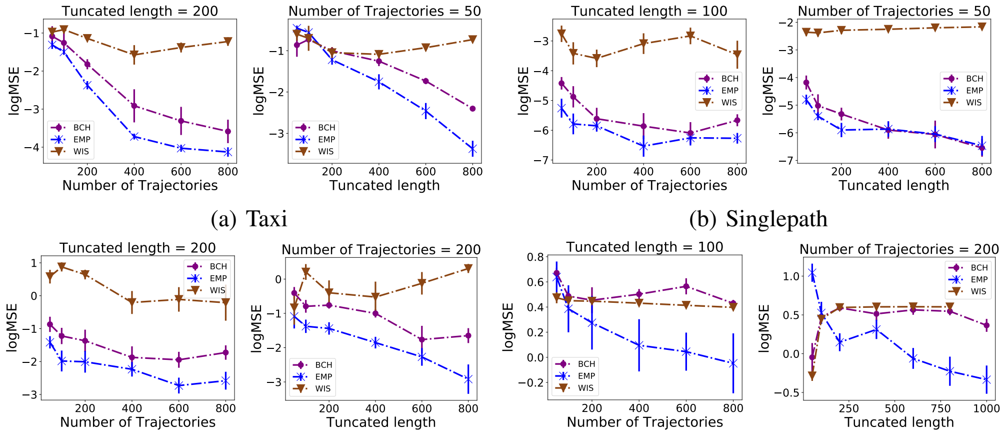
In this section, we compare the EMP method with the BCH method and step-wise importance sampling (IS) in the setting of single-behavior policy, i.e. the data is generated from a single behavior policy.

Experiment Set-up. A single behavior policy which is learned by a certain reinforcement learning algorithm [1] for evaluating BCH and IS. This single behavior policy then generates a set of trajectories consisting of s-a-s-r tuples. These tuples are used to estimate the behaviour policy for EMP methods as well as estimating the stationary distribution corrections for estimating the average step reward of the target policy.

Stationary Distribution Learning Performance. We choose the Taxi domain as an example to compare the stationary distribution $\hat{d}_{\pi_{\text{true}}}$ and $\hat{d}_{\pi_{\text{esti}}}$ learned by BCH and EMP. Figure 1(a) shows the scatter pairs $(\hat{d}_{\pi_{\text{true}}}, d_\pi)$ and $(\hat{d}_{\pi_{\text{esti}}}, d_\pi)$ estimated by 200 trajectories of 200 steps. It shows that $\hat{d}_{\pi_{\text{esti}}}$ approximate d_π better than $\hat{d}_{\pi_{\text{true}}}$. Figure 1(b) and Figure 1(b) compare the TV distance from $\hat{d}_{\pi_{\text{true}}}$ and $\hat{d}_{\pi_{\text{esti}}}$ to d_π under different data sample sizes. The results indicate that both $\hat{d}_{\pi_{\text{true}}}$ and $\hat{d}_{\pi_{\text{esti}}}$ converge, while $\hat{d}_{\pi_{\text{esti}}}$ converges faster and is significantly closer to d_π when the data size is small. These observations are well consistent with Theorem 1.

Policy Evaluation Performance. Figure 2 reports the MSE of policy evaluation by EMP, BCH and IS methods for the 4 different environments. We observe that, (i) EMP consistently obtains smaller

[1] We use Q-learning in discrete control tasks and Actor Critic in continuous control tasks.



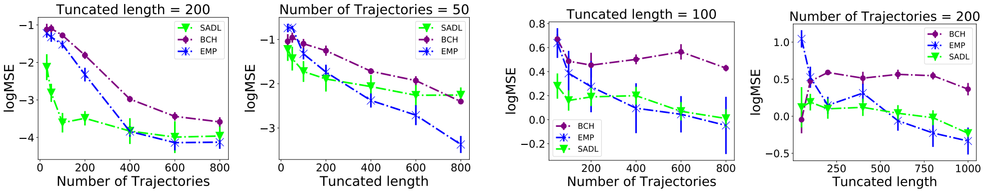
(a) Taxi

(b) Singlepath

(c) Gridworld

(d) Pendulum

Figure 2: Single-behavior-policy results of BCH, EMP and WIS across continuous and discrete environments with average reward. Each node indicates the mean value and the bars represents the standard error of the mean.



(a) Taxi

(b) Pendulum

Figure 3: Comparison results among policy-aware (BCH), partially policy-agnostic (EMP) and policy-agnostic (SADL) on continuous and discrete control tasks.

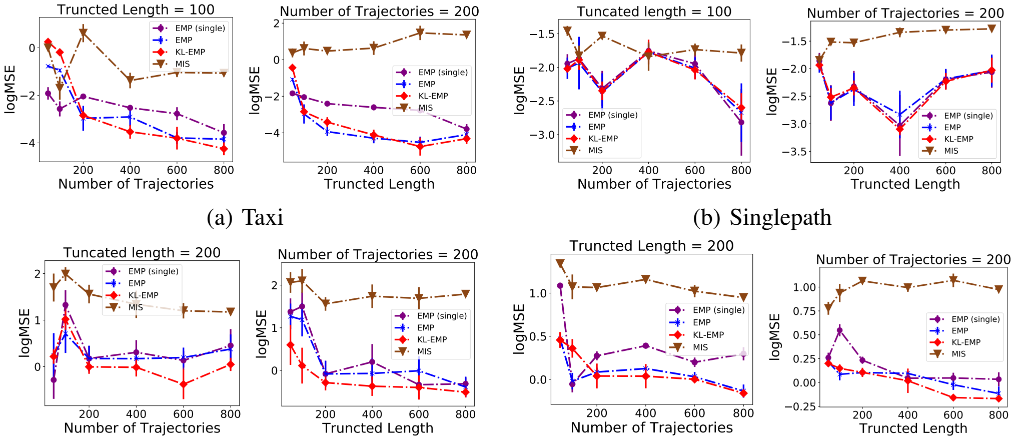
MSE than the other two methods for different sample scales and different environments. (ii) The performance of EMP and BCH improves as the number of trajectories and length of horizons increase, while the IS method suffers from growing variance. our method correctly estimates the true density ratio over the state space.

Partially Policy-agnostic versus Policy-agnostic OPPE. Figure 3 reports the comparison results for the *policy-aware* BCH, *partially policy-agnostic* EMP and a *policy-agnostic* method, which we call it state-action distribution learning (SADL) and whose formal formulation is given in Appendix D. The results show that all three methods obtain improvement as the number of length of trajectories increase. Roughly speaking, both EMP and SADL outperform BCH. The policy-agnostic SADL is better than EMP in the cases of small sample size. But when the sample size increases so that the estimated behavior policy is more accurate, EMP gradually exceeds SADL.

Remark: In our implementation of SADL, we use the same min-max formulation and optimization solver as EMP so that the comparison could shed more lights on the impact of behavior policy information on the performance of off-policy policy evaluation. We will report the comparison result between EMP and DualDice once the code is released.

5.2 Results for Multiple Unknown Behavior Policies

As for multiple behavior policies, we conduct experiments in policy-aware and partially policy-agnostic settings. We report the results of partially policy-agnostic setting in this section and the policy-aware setting is described in Appendix E.4. Because partially policy-agnostic version consistently achieves better performance.



(a) Taxi

(b) Singlepath

(c) Gridworld

(d) Pendulum

Figure 4: Multiple-behavior-policy results of EMP (single), EMP, KL-EMP and MIS across continuous and discrete environments with average reward.

Experiment Set-up. We implement the following 4 methods: (1) the proposed EMP; (2) the multiple importance sampling (MIS) method as in Tirinzoni et al. (2019) using balanced heuristics; (3) EMP (single), in which we apply EMP for each subgroup of samples generated by a single behavior policy to obtain one OPPE value and finally output their average; (4) KL-EMP, which is an ad-hoc improvement of EMP using more information on the behavior policies and whose implementation details are given in Appendix E.3.

Policy Evaluation Performance. Figure 4 reports the log MSE of the 4 methods in different environments with different sample scales. It shows that the proposed EMP outperforms both MIS and EMP (single). It is interesting to note that in EMP (single), actually more information on the behavior policies is learned than in EMP, but the learned stationary distribution corrections are mixed with naively equal weights. So, the advantage of EMP over EMP (single) can be probably attributed to (1) the robustness due to less required information on behavior policies; (2) a near-optimal weighted average that is automatically learned by pooling together the samples from different behavior policies.

On the other hand, we see that the performance of KL-EMP has greater improvement with the increase of sample size and eventually outperform EMP in cases of large sample size. This is because, KL-EMP replaces the fixed sample proportion (i.e. w_j as defined in Section 4.2) with a KL divergence-based proportion, which is better estimated with more data sample.

6 Conclusion

In this paper, we advocate the viewpoint of partial policy-awareness and the benefits of estimating a mixture policy for off-policy policy evaluation. The theoretical results of reduced variance coupled with experimental results illustrate the power of this class of methods. One key question that still remains is the following: if we are willing to estimate the individual behavior policies, can we further improve EMP by developing an efficient algorithm to compute the optimal weights? One other question is a direct comparison of DualDice and EMP when the code of DualDice is released, this will allow us to see the props and cons of inductive bias offered by the Bellman equation used by DualDice and direct estimation of the mixture policy used by EMP.

REFERENCES

Dietterich and Thomas G. Hierarchical reinforcement learning with the maxq value function decomposition. *Journal of artificial intelligence research*, 13:227–303, 2000.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *ICML*, pp. 1097–1104, 2011.

Assaf Hallak and Shie Mannor. Consistent on-line off-policy evaluation. In *ICML*, pp. 1372–1383, 2017.

Josiah Hanna, Scott Niekum, and Peter Stone. Importance sampling policy evaluation with an estimated behavior policy. In *ICML*, pp. 2605–2613, 2019.

Masayuki Henmi, Ryo Yoshida, and Shinto Eguchi. Importance sampling via the estiamted sampler. *Biometrika*, (4):985–991, 2007.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, pp. 1161–1189, 2003.

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.

Lihong Li, Rémi Munos, and Csaba Szepesvári. Toward minimax off-policy value estimation. *JMLR*, 2015.

Qaing Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *NeurIPS*, 2018.

Susan A Murphy, Mark J van der Laan, James M Robins, and Conduct Problems Prevention Research Group. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, pp. 1410–1423, 2001.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *NeurIPS*, 2019.

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

Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In *ICML*, pp. 759–766, 2000.

Doina Precup, Richard S. Sutton, and Sanjoy Dasgupta. Off-policy temporal-difference learning with function approximation. In *ICML*, pp. 417–424, 2001.

Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miroslav Dudík, John Langford, Damien Jose, and Imed Zitouni. Off-policy evaluation for slate recommendation. In *NeurIPS*, pp. 3632–3642, 2017.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *ICML*, pp. 2139–2148, 2016.

Andrea Tirinzoni, Mattia Salvini, and Marcello Restelli. Transfer of samples in policy search via multiple importance sampling. In *ICML*, pp. 6264–6274, 2019.

Elise Van der Pol and Frans A Oliehoek. Coordinated deep reinforcement learners for traffic light control. In *NeurIPS*, 2016.

Eric Veach and Leonidas J Guibas. Optiamlly combining sampling techniques for Monte Carlo rendering. In *SIGGRAPH*, pp. 419–428, 1995.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudik. Optimal and adaptive off-policy evaluation in contextual bandits. In *ICML*, pp. 3589–3597, 2017.

Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. Drn: A deep reinforcement learning framework for news recommendation. In *WWW*, pp. 167–176, 2018.

A Kernel Method

We use the reproducing kernel Hilbert space to solve the mini-max problem of BCH (Liu et al. (2018)). The key property of RKHS we leveraged is called **reproducing property**. The reproducing property claims, for any function $f \in \mathcal{H}$ ($\mathcal{H}$ is a RKHS), the evaluation of f at point x equals its inner product with another function in RKHS: $f(s) = \langle f, k(s, \cdot)\rangle_{\mathcal{H}}$.

Given the objective function of BCH $L(w, f) = \mathbb{E}_{(s,a,s')\sim d_{\pi_0}}[(\omega(s)\frac{\pi(a|s}{\pi_0(a|s)-\omega})f(s')]$. We use the reproducing property to obtain the closed form representation of $\max_{f\in\mathcal{F}} L(w, f)^2$, which is shown as follows:

$$\max_{f\in\mathcal{F}} L(w, f)^2 = \mathbb{E}_{(s,a,s')\sim d_{\pi_0},(\bar{s},\bar{a},\bar{s}')\sim d_{\pi_0}}\left[\Delta\left(\omega; s, a, s'\right)\Delta\left(w; \overline{s}, \overline{a}, \overline{s}'\right)k\left(s', \overline{s}'\right)\right]$$

.

This equation has been proved in BCH Liu et al. (2018).

B Proof of Theorem 1

B.1 Assumptions

In this appendix, we provide the mathematical details and proof of Theorem 1. We first introduce some notations and assumptions.

We assume the behavior policy $\pi_0(a|s)$ belongs to a class of policies $\Pi = \{\pi(\theta; a, s) : \theta \in \mathcal{E}_\theta\}$, where $\mathcal{F}_\theta$ is the parameter space, i.e. there exists $\theta_0 \in E_1$ such that $\pi_0(a|s) = \pi(\theta_0; a, s)$. The estimated behavior policy $\hat{\pi}_0 = \pi_{\hat{\theta}}$ is obtained via maximum likelihood method, i.e.

$$\hat{\theta} = \arg\max \sum_{n=0}^{N-1} \log(\pi(\theta; s_n, a_n)).$$

We assume central limit theorem holds for $\hat{\theta}$. Recall that we have assumed in Section 3.2 that the true stationary distribution correction $\omega(s) = \omega(\eta_0; s)$. Using the kernel method introduced in Appendix A, our estimation $\hat{\omega}(s) = \omega(\hat{\eta}; s)$ is obtained via

$$\min_{\eta} \sum_{0\leq i,j\leq N-1} G(\eta, \hat{\theta}; x_i, x_j),$$

with $x_i = (s_i, a_i, s_i')$ and

$$G(\eta, \hat{\theta}; (x_i, x_j)) = \left(\omega(\eta; s_i)\frac{\pi(a_i|s_i)}{\pi(\hat{\theta}, a_i, s_i)} - \omega(\eta, s_i')\right)\left(\omega(\eta; s_j)\frac{\pi(a_j|s_j)}{\pi(\hat{\theta}, a_j, s_j)} - \omega(\eta, s_j')\right)k(s_i', s_j').$$

Assumption 1. *We assume the following regularity conditions on G:*

1. *G is second order differentiable.*
2. $\mathbb{E}[\partial_\eta\partial_\theta G(\eta_0, \theta_0; x_i, x_j)]$ *is finite.*
3. $\mathbb{E}[\partial_\eta^2 G(\eta_0, \theta_0; x_i, x_j)]$ *is finite and non-zero.*
4. $\mathbb{E}[\partial_\eta G(\eta_0, \theta_0; x_i, x_j)^2]$ *is finite.*

Here we simply write $\mathbb{E}_{x_i\sim d_{\pi_0},x_j\sim d_{\pi_0}}$ as $\mathbb{E}$ for the simplicity of notation.

B.2 Proof of Theorem 1

Proof. Following the kernel method,

$$\hat{\eta} = \arg\min_{\eta} = \arg\min_{\eta} \sum_{0\leq i,j\leq N-1} G(\eta, \hat{\theta}; (x_i, x_j)) \text{ with } x_i = (s_i, a_i, s_i') \text{ and } s_i' \triangleq s_{i+1}.$$

Then, $\sum_{1\le i,j\le N} \partial_\eta G(\hat{\eta}, \hat{\theta}; (x_i, x_j)) = 0$, we have

$$\begin{aligned}0 =& \frac{1}{N\sqrt{N}} \sum_{0\le i,j\le N-1} \partial_\eta G(\eta_0, \theta_0; (x_i, x_j)) + \sqrt{N}(\hat{\eta} - \eta_0) \frac{1}{N^2} \sum_{0\le i,j\le N-1} \partial_\eta^2 G(\eta_0, \theta_0; (x_i, x_j)) \\ &+ \sqrt{N}(\hat{\theta} - \theta) \frac{1}{N^2} \sum_{0\le i,j\le N-1} \partial_\theta \partial_\eta G(\eta_0, \theta_0; (x_i, x_j)) \\ =& \frac{1}{N\sqrt{N}} \sum_{0\le i,j\le N-1} \partial_\eta G(\eta_0, \theta_0; (x_i, x_j)) + \sqrt{N}(\hat{\eta} - \eta_0) \mathbb{E}\left[\partial_\eta^2 G(\eta_0, \theta_0; (x_1, x_2))\right] \\ &+ \sqrt{N}(\hat{\theta} - \theta) \mathbb{E}\left[\partial_\theta \partial_\eta G(\eta_0, \theta_0; (x_1, x_2))\right] + o_p(1).\end{aligned}$$

Similarly, we have

$$\tilde{\eta} = \arg\max_\eta \sum_{1\le i,j\le N} G(\eta, \theta_0; (x_i, x_j)),$$

and $0 = \frac{1}{N\sqrt{N}} \sum_{0\le i,j\le N-1} \partial_\eta G(\eta_0, \theta_0; (x_i, x_j)) + \sqrt{N}(\tilde{\eta} - \eta_0) \mathbb{E}\left[\partial_\eta^2 G(\eta_0, \theta_0; (x_1, x_2))\right] + o_p(1)$. Define S$(\theta; (x_i, x_j)) = \log(\pi(\theta; s_i, a_i)) + \log(\pi(\theta; s_j, a_j))$. According to our estimation method,

$$\hat{\theta} = \arg\max_\theta \sum_{0\le i,j\le N-1} S(\theta; (x_i, x_j)).$$

Therefore, $0 = \frac{1}{N\sqrt{N}} \sum_{0\le i,j\le N-1} \partial_\theta S(\theta_0; (x_i, x_j)) + \sqrt{N}(\hat{\theta} - \theta_0) \mathbb{E}\left[\partial_\theta^2 S(\theta_0; (x_1, x_2))\right] + o_p(1)$. Following the proof of Theorem 1 of (Henmin et al. 2007), it suffices to prove that

$$\mathbb{E}\left[\partial_\theta \partial_\eta G(\eta_0, \theta_0; (x_1, x_2))\right] = \mathbb{E}\left[-\partial_\eta G(\eta_0, \theta_0; (x_1, x_2)) \partial_\theta S(\theta_0; (x_1, x_2))\right]. \tag{10}$$

One can check

$$\begin{aligned}&\mathbb{E}\left[\partial_\eta G(\eta_0, \theta_0; (x_1, x_2))\right] \\ =&\mathbb{E}\left[k(s_1', s_2') \left[\left(\partial_\eta \omega(\eta_0; s_1) \frac{\pi(a_1|s_1)}{\pi(\theta_0; a_1, s_1)} - \partial_\eta \omega(\eta_0; s_1')\right) \left(\omega(\eta_0; s_2) \frac{\pi(a_2|s_2)}{\pi(\theta_0; a_2, s_2)} - \omega(\eta_0; s_2')\right)\right.\right. \\ &+ \left.\left.\left(\partial_\eta \omega(\eta; s_2) \frac{\pi(a_2|s_2)}{\pi(\theta_0; a_2, s_2)} - \partial_\eta \omega(\eta_0; s_2')\right) \left(\omega(\eta_0; s_1) \frac{\pi(a_1|s_1)}{\pi(\theta_0; a_1, s_1)} - \omega(\eta_0; s_1')\right)\right]\right] \\ =&\mathbb{E}\left[(k(s_1', s_2') + k(s_2', s_1')) \left(\partial_\eta \omega(\eta_0; s_1) \frac{\pi(a_1|s_1)}{\pi(\theta_0; a_1, s_1)} - \partial_\eta \omega(\eta_0; s_1')\right) \left(\omega(\eta_0; s_2) \frac{\pi(a_2|s_2)}{\pi(\theta_0; a_2, s_2)} - \omega(\eta_0; s_2')\right)\right].\end{aligned}$$

The last equality holds because $(x_1, x_2) \sim d_{\pi_0}(s_1)\pi(x_1; \eta_0) \otimes d_{\pi_0}(s_2)\pi(x_2; \eta_0)$. Besides, we have

$$\partial_\theta S(\theta; (x_1, x_2)) = \frac{\partial_\theta \pi(\theta; a_1, s_1)}{\pi(\theta; a_1, s_1)^2} + \frac{\partial_\theta \pi(\theta; a_2, s_2)}{\pi(\theta; a_2, s_2)^2}.$$

Then, we derive

$$\begin{aligned}&\mathbb{E}\left[\partial_\theta \partial_\eta G(\eta_0, \theta; (x_1, x_2))\right] \\ =&\mathbb{E}\left[(k(s_1', s_2') + k(s_2', s_1')) \left[-\partial_\eta \omega(\eta_0; s_1) \frac{\pi'(a_1|s_1)}{\pi(\theta_0; a_1, s_1)^2} \left(\omega(\eta_0; s_2) \frac{\pi(a_2|s_2)}{\pi(\theta_0; a_2, s_2)} - \omega(\eta_0; s_2')\right)\right.\right. \\ &\left.\left.-\omega(\eta_0; s_2) \frac{\pi'(a_2|s_2)}{\pi(\theta_0, a_2, s_2)^2} \left(\partial_\eta \omega(\eta_0; s_1) \frac{\pi(a_1|s_1)}{\pi(\theta_0; a_1, s_1)} - \partial_\eta \omega(\eta_0; s_1')\right)\right]\right] \\ =&\mathbb{E}\left[(k(s_1', s_2') + k(s_2', s_1')) \left[-\left(\partial_\eta \omega(\eta_0; s_1) \frac{\pi'(a_1|s_1)}{\pi(\theta_0; a_1, s_1)^2} - \frac{\partial_\eta \omega(\eta_0; s_1')}{\pi(\theta_0; a_1, s_1)}\right) \left(\omega(\eta_0; s_2) \frac{\pi(a_2|s_2)}{\pi(\theta_0; a_2, s_2)} - \omega(\eta_0; s_2')\right)\right.\right. \\ &\left.\left.- \left(\omega(\eta_0; s_2) \frac{\pi'(a_2|s_2)}{\pi(\theta_0; a_2, s_2)^2} - \frac{\omega(\eta_0; s_j')}{\pi(\theta_0; a_2, s_2)}\right) \left(\partial_\eta \omega(\eta_0; s_1) \frac{\pi(a_1|s_1)}{\pi(\theta_0; a_1, s_1)} - \partial_\eta \omega(\eta_0; s_1')\right)\right]\right] \\ &- \mathbb{E}\left[(k(s_1', s_2') + k(s_2', s_1')) \left[\frac{\partial_\eta \omega(\eta_0; s_1')}{\pi(\theta_0; a_1, s_1)} \left(\omega(\eta_0; s_2) \frac{\pi(a_1|s_1)}{\pi(\theta_0; a_2, s_2)} - \omega(\eta_0; s_2')\right)\right.\right. \\ &\left.\left.+\frac{\omega(\eta_0; s_2')}{\pi(\theta_0; a_2, s_2)} \left(\partial_\eta \omega(\eta_0; s_1) \frac{\pi(a_1|s_1)}{\pi(\theta_0; a_1, s_1)} - \partial_\eta \omega(\eta_0; s_1')\right)\right]\right] \\ \triangleq&\mathbb{E}\left[-\partial_\eta G(\eta_0, \theta_0; (x_1, x_2)) \partial_\theta S(\theta_0; (x_1, x_2))\right] + \mathbb{E}\left[H(\eta_0, \theta_0; (x_1, x_2))\right].\end{aligned}$$

Here, we define

$$H(\eta_0, \theta_0, (x_1, x_2)) = \frac{\partial_\eta \omega(\eta_0; s_1')}{\pi(\theta_0; a_1, s_1)} \left(\omega(\eta_0; s_2) \frac{\pi(a_2|s_2)}{\pi(\theta_0; a_2, s_2)} - \omega(\eta_0, s_2') \right) + \frac{\omega(\eta_0; s_2')}{\pi(\theta_0; a_2, s_2)} \left(\partial_\eta \omega(\eta_0; s_i) \frac{\pi(a_1|s_1)}{\pi(\theta_0; a_1, s_1)} - \partial_\eta \omega(\eta_0; s_1') \right).$$

Note that

$$\mathbb{E}\left[\left(\omega(\eta_0; s_2) \frac{\pi(a_2|s_2)}{\pi(\theta_0, a_2, s_2)} - \omega(\eta_0, s_2') \right) |a_1, s_1, s_1', s_2' \right] = 0,$$
$$\mathbb{E}\left[\left(\partial_\eta \omega(\eta_0; s_1) \frac{\pi(a_1|s_1)}{\pi(\theta_0; a_1, s_1)} - \partial_\eta \omega(\eta_0; s_1') \right) |a_2, s_2, s_2' \right] = 0.$$

Therefore $\mathbb{E}[H(\eta_0, \theta_0; (x_1, x_2))] = 0$. So we obtain (10). □

B.3 Proof of Corollary 1

Proof. In the prove of Theorem 1, we see that

$$\frac{1}{N^2} \sum_{0 \le i,j \le N-1} \partial_\eta G(\eta_0, \theta_0; (x_i, x_j)) + K_1(\hat{\eta} - \eta_0) + K_2(\hat{\theta} - \theta) = o_p(1).$$

with $K_1 = \mathbb{E}\left[\partial_\eta^2 G(\eta_0, \theta_0; (x_1, x_2))\right]$ and $K_2 = \mathbb{E}\left[\partial_\theta \partial_\eta G(\eta_0, \theta_0; (x_1, x_2))\right]$. Therefore,

$$\mathbb{E}[(\hat{\eta} - \eta_0)^2] \le 2K_1^{-2} \left(K_2^2 \mathbb{E}[(\hat{\theta} - \theta_0)^2] + \mathbb{E}\left[\left(\frac{1}{N^2} \sum_{0 \le i,j \le N-1} \partial_\eta G(\eta_0, \theta_0; (x_i, x_j)) \right)^2 \right] \right).$$

We assume that CLT holds for the maximum likelihood estimator $\hat{\theta}$, i.e. $\mathbb{E}[(\hat{\theta} - \theta_0)^2] = O(1/N)$. Besides, as $\mathbb{E}[\partial_\eta G(\eta_0, \theta_0; (x_i, x_j))] = 0$, under Condition 4 of Assumption 1, , we can apply the central limit theorem (for stationary Markov chain) and have

$$\mathbb{E}\left[\left(\frac{1}{N\sqrt{N}} \sum_{0 \le i,j \le N-1} \partial_\eta G(\eta_0, \theta_0; (x_i, x_j)) \right)^2 \right] = O(1).$$

Therefore,

$$\mathbb{E}[(\hat{\eta} - \eta_0)^2] = O(1/N).$$

□

C Proofs of Propositions for EMP

Proof of Proposition 1.

$$\mathbb{E}_{(s,a)\sim d_0}\left[\omega(s) \frac{\pi(a|s)}{\pi_0(a|s)} r(s,a) \right] = \sum_{s,a} \omega(s) \frac{\pi(a|s)}{\pi_0(a|s)} r(s,a) \sum_j w_j d_{\pi_j}(s) \pi_j(a|s)$$
$$= \sum_{s,a} \omega(s) \frac{\pi(a|s)}{\pi_0(a|s)} r(s,a) d_0(s) \pi_0(s) = \sum_{s,a} d_0(s) \omega(s) \pi(a|s) r(s,a) = R_\pi.$$

□

Proof of Proposition 2. If $\omega = d_\pi / d_0$, based on the stationary equation

$$d_\pi(s') = \sum_{s,a} d_\pi(s) \pi(a|s) T(s'|a,s), \text{ for any } s' \in S,$$

we have

$$\begin{aligned}\sum_{s'}\omega(s')d_0(s')f(s') &= \sum_{s,a}\omega(s)d_0(s)\pi(a|s)T(s'|a,s)f(s') \\ &= \sum_{j}\sum_{s,a}\omega(s)\frac{\pi(a|s)}{\sum_j w_j d_\pi(s)\pi_j(a|s)/d_0} w_j d_{\pi_j}(s)\pi_j(a|s)T(s'|a,s)f(s') \\ &= \mathbb{E}_{(s,a,s')\sim d_0}\left[\omega(s)\frac{\pi(a|s)}{p_0(a|s)}f(s')\right].\end{aligned}$$

Therefore,

$$\mathbb{E}_{(s,a,s')\sim d_0}\left[\left(\omega(s') - \omega(s)\frac{\pi(a|s)}{p_0(a|s)}\right)f(s')\right] = 0.$$

On the opposite way, if

$$\mathbb{E}_{(s,a,s')\sim d_0}\left[\left(\omega(s') - \omega(s)\frac{\pi(a|s)}{p_0(a|s)}\right)f(s')\right] = 0, \text{ for all function } f : S \to \mathbb{R},$$

we should have

$$d_0(s')\omega(s') = \sum_{s,a} d_0(s)\omega(s)\pi(a|s)T(s'|a,s).$$

Therefore, $d_0\omega$ satisfy the stationary equation and must equal to d_π (up to a constant). □

Proof of Proposition 3. The proof follows immediately from that of Theorem 1. In particular, assume $\pi_0 \in \{\pi(\theta; a, s) : \theta \in \mathcal{E}_\theta\}$ and the estimated $\hat{\pi}_0 = \pi(\hat{\theta}; \cdot)$ is obtained via

$$\hat{\theta} = \arg\max_{\theta}\sum_j\sum_{n=0}^{N_j-1}\log(\pi(\theta; s_{j,n}, a_{j,n})).$$

The rest part of the proof follows the same argument in the proof of Theorem 1. □

D State-Action Distribution Learning

Here we propose a behavior-agnostic approach that evaluates the target policy through learning occupation distribution correction instead of stationary distribution correction. Recall that the occupation distribution the stationary distribution of the state-action pair $d_\pi(s)\pi(a|s)$. We define occupation distribution correction as $d_\pi(s)\pi(a|s)/(d_{\pi_0}(s)\pi_0(a|s))$. Since $\pi(a|s)$ is known, we denote $u(a,s) = d_\pi(s)/(d_{\pi_0}(s)\pi_0(a|s))$ and formulate a min-max problem to learn u.

Following this notation, Equation (1) can be written as:

$$u(s',a')d_{\pi_0}(s',a') = \sum_{s,a}u(s,a)\pi(a|s)d_{\pi_0}(s,a)P(s'|a,s), \quad \forall s', \forall a'$$

For any test function $f(s',a') : S \times R \to \mathbb{R}$ and any probability density function $\omega(a')$ with respect to $a' \in A$, the equation above implies that:

$$\sum_{s',a'}u(s',a')\omega(a')f(s',a')d_{\pi_0}(s',a') = \sum_{s'}\sum_{a'}f(s',a')\omega(a')\sum_{s,a}u(s,a)\pi(a|s)d_{\pi_0}(s,a)P(s'|a,s),$$

which is equivalent to,

$$\mathbb{E}_{(s,a,s')\sim d_{\pi_0}}[u(s,a)\omega(a)f(s,a) - u(s,a)\pi(a|s)\mathbb{E}_{a'\sim\omega}[f(s',a')]] = 0, \quad \forall f.$$

This suggest the following mini-max problem can be used to estimate $u(s,a)$

$$\min_u\{D(u) := \max_{f\in\mathcal{F}} L(u,f)^2\},$$
$$\text{where } L(u,f) := \mathbb{E}_{(s,a,s')\sim d_{\pi_0}}\left[u(s,a)\omega(a)f(s,a) - u(s,a)\pi(a|s)\mathbb{E}_{a'\sim\omega}[f(s',a')]\right].$$

We simplify this mini-max problem into a minimization problem using the kernel method. Theorem 2 gives the closed form representation of $D(u)$ when $\mathcal{F}$ is a unit ball in a RKHS with kernel k.

Theorem 2. *Assume $\mathcal{H}$ is a RKHS of functions $f(s,a)$ with a positive definite kernel $k((s,a),(s',a'))$, and define $\mathcal{F} := \{f \in \mathcal{H} : ||f||_{\mathcal{H}} \leq 1\}$ to be the unit ball of $\mathcal{H}$. We have $\max_{f\in\mathcal{F}} L(u,f)^2 =$*

$$\begin{aligned}
&= \mathbb{E}_{(s,a,s')\sim d_{\pi_0}, a'\sim\omega, (\bar{s},\bar{a},\bar{s}')\sim d_{\pi_0}, \bar{a}'\sim\omega}[u(s,a)u(\bar{s},\bar{a})\omega(a)\omega(\bar{a})k((s,a),(\bar{s},\bar{a})) \\
&+ u(s,a)u(\bar{s},\bar{a})\pi(a|s)\pi(\bar{a}|\bar{s})k((s',a'),(\bar{s}',\bar{a}')) \\
&- u(s,a)u(\bar{s},\bar{a})\omega(a)\pi(\bar{a}|\bar{s})k((s,a),(\bar{s}',\bar{a}')) - u(s,a)u(\bar{s},\bar{a})\omega(\bar{a})\pi(a|s)k((\bar{s},\bar{a}),(s',a'))].
\end{aligned}$$

E Experimental Details

E.1 Environment Description

Taxi Taxi Dietterich & G (2000) is a 5 × 5 grid world simulating a taxi movement. Six actions are contained in Taxi: moves North, East, South, West, pick up and drop off a passenger. A reward of 20 is received when it picks up a passenger or drops her/he off at the right place, and a reward of -1 for each time step. The passengers are allow to randomly appear and disappear at every corner of the map at each time step. The 5 × 5 grid size yields 2000 states in total (25 × 24 × 5, corresponding to 25 taxi locations, 24 passenger appearance status and 5 taxi status (empty or with one of 4 destinations)).

Gridworld Gridworld Thomas & Brunskill (2016) is a 4 × 4 grid world which including one reward state, one terminate state and one fire state and thirteen normal state. Four action can be taken in this environment: up, down, left and right. A reward of -1 will be received while the agent in normal states, 1 reward is obtained in reward state, 100 reward is got in terminate state and -11 reward will got in fire state.

SinglePath This environment has 5 states, 2 actions. The agent begins in state 0 and both actions either take the agent from state n to state n + 1 or cause the agent to remain in state n. If the agent arrives at a new state, it will receive a +1 reward, otherwise it will get a 1 reward.

Pendulum Pendulum has a continuous state space of $\mathcal{R}^3$ which describes the triangle of and a action space of $[-2, 2]$.

E.2 Behavior Policy Estimation

EMP employs the maximum likelihood method as in Tirinzoni et al. (2019) to estimate the mixed policy $\hat{\pi}_0$ as

$$\hat{\pi}_0 = \arg\max_{\pi\in\Pi} \sum_{j=1}^{m} \sum_{n=1}^{N_j} \log \pi(a_n|s_n) \tag{11}$$

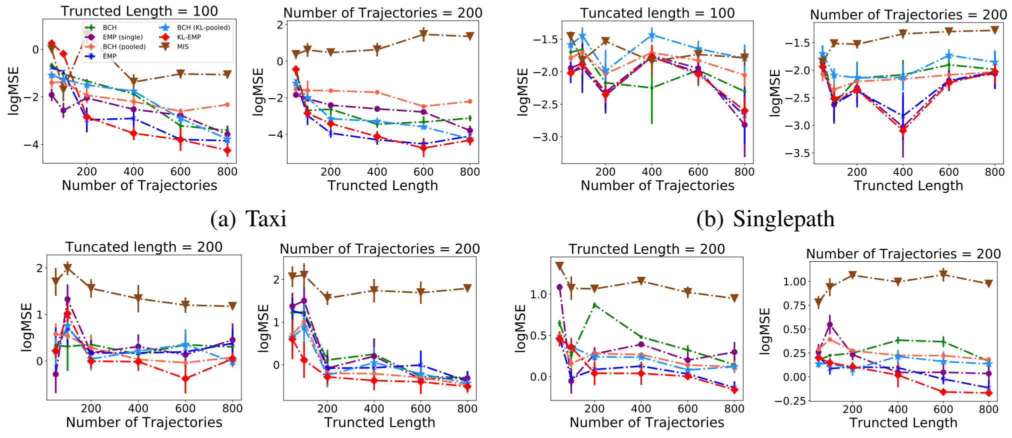
As for discrete control tasks, the optimal $\hat{\pi}_0$ coincides with the count-frequency.

E.3 Computation of KL-Weights

In an ad-hoc way, we optimize the weights w_j according to the KL-divergence between the behavior policies and the target policy, which can be estimated directly from the data. For finite-state space, we propose to choose

$$\begin{aligned}
w_j^{KL} &= \frac{\sum_{s\in S} \mathbf{1}(j = \arg\min_k D_{KL}(\pi(\cdot|s)||\pi_k(\cdot|s)))}{\sum_{i=1}^{m}\sum_{s\in S} \mathbf{1}(i = \arg\min_k D_{KL}(\pi(\cdot|s)||\pi_k(\cdot|s)))} \\
&= \frac{\sum_{s\in S} \mathbf{1}(j = \arg\min_k D_{KL}(\pi(\cdot|s)||\pi_k(\cdot|s)))}{|S|}
\end{aligned} \tag{12}$$

To implement this method for infinite- or continuous-state space in the numerical experiments, we replace the set of all possible states S in (12) with the set of all states that has been visited in the data buffer. The numerical results show that using the KL weights $\{w_j^{KL}\}$ could achieve smaller MSE compared to using $\{w_j\}$ as given by the data sample. We believe this approach deserves more careful analysis in future research studies.



(a) Taxi

(b) Singlepath

(c) Gridworld

(d) Pendulum

Figure 5: Results of policy-aware OPPE methods (BCH, BCH (pooled) and BCH (KL-pooled)) and their corresponding partially policy-agnostic version (EMP (single), EMP and KL-EMP) across continuous and discrete environments with average reward.

E.4 Additional Experiment Results

BCH, EMP and KL-EMP have both policy-aware and partially policy-agnostic versions in multiple behavior policies. In policy-aware BCH we first apply BCH for each subgroup of samples generated by each behavior policy followed by output their average value. As for partially policy-agnostic BCH, it is equal to EMP (single).

The policy-aware version of EMP is named as BCH (pooled). In BCH (pooled), the corresponding min-max problem formation is

$$\min_{\omega} \max_{f} \mathbb{E}_{(j,s,a,s')\sim\mathcal{D}}\left[\left(\omega(s') - \omega(s)\frac{\pi(a|s)}{\pi_j(a|s)}\right)f(s')\right]. \tag{13}$$

They both pool the data from different behavior policies together and the main difference is that BCH (pooled) uses the exact behavior policies.

The policy-aware version of KL-EMP is called BCH (KL-polled). The main difference between BCH (KL-polled) and BCH (polled) is that BCH (KL-polled) utilizes KL-divergence to calculate the weights.

The results of the two versions are shown in Figure 5. We observe that the partially policy-agnostic version OPPE consistently outperform the policy-aware version OPPE.