

Representation Balancing MDPs for Off-Policy Policy Evaluation

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 We study the problem of off-policy policy evaluation (OPPE) in RL. In contrast
2 to prior work, we consider how to estimate both the individual policy value and
3 average policy value accurately. We draw inspiration from recent work in causal
4 reasoning, and propose a new finite sample generalization error bound for value
5 estimates from MDP models. Using this upper bound as an objective, we develop a
6 learning algorithm of an MDP model with a balanced representation, and show that
7 our approach can yield substantially lower MSE in a common synthetic domain
8 and on a challenging real-world sepsis management problem.

9 1 Introduction

10 In reinforcement learning, off-policy (batch) policy evaluation is the task of estimating the perfor-
11 mance of some *evaluation* policy given data gathered under a different *behavior* policy. Off-policy
12 policy evaluation (OPPE) is essential when deploying a new policy might be costly or risky, such as
13 in consumer marketing, healthcare, and education. Technically off-policy evaluation relates to other
14 fields that study counterfactual reasoning, including causal reasoning, statistics and economics.

15 Off-policy batch policy evaluation is challenging because the distribution of the data under the
16 behavior policy will in general be different than the distribution under the desired evaluation policy.
17 This difference in distributions comes from two sources. First, at a given state, the behavior policy
18 may select a different action than the one preferred by the evaluation policy—for example, a clinician
19 may chose to amputate a limb, whereas we may be interested in what might have happened if the
20 clinician had not. We never see the counterfactual outcome. Second, the distribution of future
21 states—not just the immediate outcomes—is also determined by the behavior policy. This challenge
22 is unique to sequential decision processes and is not covered by most causal reasoning work: for
23 example, the resulting series of a patient’s health states observed after amputating a patient’s limb is
24 likely to be significantly different than if the limb was not amputated.

25 Approaches for OPPE must make a choice about whether and how to address this data distribution
26 mismatch. Importance sampling (IS) based approaches [16, 23, 7, 5, 9, 22] are typically unbiased
27 and strongly consistent, but despite recent progress tend to have high variance—especially if the
28 evaluation policy is deterministic, as evaluating deterministic policies requires finding in the data
29 sequences where the actions exactly match the evaluation policy. However, in most real-world
30 applications deterministic evaluation policies are more common—policies are typically to either
31 amputate or not, rather than a policy that that flips a biased coin (to sample randomness) to decide
32 whether to amputate. IS approaches also often rely on explicit knowledge of the behavior policy,
33 which may not be feasible in situations such as medicine where the behaviors results from human
34 actions. In contrast, some model based approaches ignore the data distribution mismatch, such as by
35 fitting a maximum-likelihood model of the rewards and dynamics from the behavioral data, and then
36 using that model to evaluate the desired evaluation policy. These methods may not converge to the

37 true estimate of the evaluation policy’s value, even in the limit of infinite data [15]. However, such
38 model based approaches often achieve better empirical performance than the IS-based estimators [9].

39 In this work, we address the question of building model-based estimators for OPPE that both *do* have
40 theoretical guarantees and yield better empirical performance than model-based approaches that ignore
41 the data distribution mismatch. Typically we evaluate the quality of an OPPE estimate $\hat{V}^{\pi_e}(s_0)$, where
42 s_0 is an initial state, by evaluating its mean squared error (MSE). Most previous research (e.g. [9, 22])
43 evaluates their methods using MSE for the average policy value (APV): $[\mathbb{E}_{s_0} \hat{V}^{\pi_e}(s_0) - \mathbb{E}_{s_0} V^{\pi_e}(s_0)]^2$,
44 rather than the MSE for individual policy values (IPV): $\mathbb{E}_{s_0} [\hat{V}^{\pi_e}(s_0) - V^{\pi_e}(s_0)]^2$. This difference is
45 crucial for applications such as personalized healthcare since ultimately we may want to assess the
46 performance of a policy for an specific individual (patient) state.

47 Instead, in this paper we develop an upper bound of the MSE for individual policy value estimates.
48 Note that this bound is automatically an upper bound on the average treatment effect. Our work
49 is inspired by recent advances[19, 12, 10, 11] in estimating conditional averaged treatment effects
50 (CATE), also known as heterogeneous treatment effects (HTE), in the contextual bandit setting with
51 a single (typically binary) action choice. CATE research aims to obtain precise estimates in the
52 difference in outcomes for giving the treatment vs control intervention for an individual (state).

53 Recent work [10, 19] on CATE¹ has obtained very promising results by learning a model to predict
54 individual outcomes using a (model fitting) loss function that explicitly accounts for the data distribu-
55 tion shift between the treatment and control policies. We build on this work to introduce a new bound
56 on the MSE for individual policy values, and a new loss function for fitting a model-based OPPE
57 estimator. In contrast to most other OPPE theoretical analyses (e.g. [9, 5, 22]), we provide a finite
58 sample generalization error instead of asymptotic consistency. In contrast to previous model value
59 generalization bounds such as the Simulation Lemma [13], our bound accounts for the underlying
60 data distribution shift if the data used to estimate the value of an evaluation policy were collected by
61 following an alternate policy.

62 We use this to derive a loss function that we can use to fit a model for OPPE for deterministic
63 evaluation policies. Conceptually, this process gives us a model that prioritizes fitting the trajectories
64 in the batch data that match the evaluation policy. Our current estimation procedure works for
65 deterministic evaluation policies which covers a wide range of scenarios in real-world applications
66 that are particularly hard for previous methods. Like recently proposed IS-based estimators [22, 9, 6],
67 and unlike the MLE model-based estimator that ignores the distribution shift [15], we prove that
68 our model-based estimator is asymptotically consistent, as long as the true MDP model is realizable
69 within our chosen model class; we use neural models to give our model class high expressivity.

70 We demonstrate that our resulting models can yield substantially lower mean squared error estimators
71 than prior model-based and IS-based estimators on a classic benchmark RL task (even when the
72 IS-based estimators are given access to the true behavior policy). We also demonstrate our approach
73 can yield improved results on a challenging sepsis treatment management dataset extracted from the
74 MIMIC-III database [12].

75 2 Related Work

76 Most prior work on OPPE in reinforcement learning falls into one of three approaches. The first,
77 importance sampling (IS), reweights the trajectories to account for the data distribution shift. Under
78 mild assumptions importance sampling estimators are guaranteed to be both unbiased and strongly
79 consistent, and were first introduced to reinforcement learning OPPE by Precup et al. [16]. Despite
80 recent progress (e.g.[23, 7]) IS-only estimators still often yield very high variance estimates, par-
81 ticularly when the decision horizon is large, and/or when the evaluation policy is deterministic. IS
82 estimators also typically result in extremely noisy estimates for policy values of individual states.
83 A second common approach is to estimate a dynamics and reward model, which can substantially
84 reduce variance, but can be biased and inconsistent (as noted by [15]). The third approach, doubly
85 robust estimators, originates from the statistics community [17]. Recently proposed doubly robust
86 estimators for OPPE from the machine and reinforcement learning communities [5, 9, 22] have
87 sometimes yielded orders of magnitude tighter estimates. However, most prior work that leverages

¹Shalit et al. [19] use the term individual treatment effect (ITE) to refer to a criterion which is actually defined as CATE in most causal inference literature. We discuss the confusion about the two terms in the appendix B.

an approximate model has largely ignored the choice of how to select and fit the model parameters. Recently, Farajtabar et al. [6] introduced more robust doubly robust (MRDR), which involves fitting a Q function for the model-value function part of the doubly robust estimator based on fitting a weighted return to minimize the variance of doubly robust. In contrast, our work learns a dynamics and reward model using a novel loss function, to estimate a model that yields accurate individual policy value estimates. While our method can be combined in doubly robust estimators, we will also see in our experimental results that directly estimating the performance of the model estimator can yield substantially benefits over estimating a Q function for use in doubly robust.

OPPE in contextual bandits and RL also has strong similarities with the treatment effect estimation problem common in causal inference and statistics. Recently, different kinds of machine learning models such as Gaussian Processes [1], random forests [24], and GANs [8] have been used to estimate heterogeneous treatment effects (HTE), in non-sequential settings. Schulam and Saria [18] study using Gaussian process models for treatment effect estimation in continuous time settings. Their setting differs from MDPs by not having sequential states. Most theoretical analysis of treatment effects focuses on asymptotic consistency rather than generalization error.

Our work is inspired by recent research that learns complicated outcome models (reward models in RL) to estimate HTE using new loss functions to account for covariate shift [10, 19, 2, 11]. In contrast to this prior work we consider the sequential state-action setting. In particular, Shalit et al. [19] provided an algorithm with a more general model class, and a corresponding generalization bound. We extend this idea from the binary treatment setting to sequential and multiple action settings.

3 Preliminaries: Notation and Setting

We consider undiscounted finite horizon MDPs, with finite horizon $H < \infty$, bounded state space $\mathcal{S} \subset \mathbb{R}^d$, and finite action space $\mathcal{A}$. Let $p_0(s)$ be the initial state distribution, and $T(s'|s, a)$ be the transition probability. Given a state action pair, the expectation of reward r is $\mathbb{E}[r|x, a] = \bar{r}(x, a)$. Given n trajectories collected from a stochastic behavior policy μ , our goal is to evaluate the policy value of $\pi(s)$. We assume the policy $\pi(s)$ is deterministic. We will learn a model of both reward and transition dynamics, $\widehat{M} = \langle \widehat{r}(s, a), \widehat{T}(s', s, a) \rangle$, based on a learned representation. The representation function $\phi : \mathcal{S} \mapsto \mathcal{Z}$ is a reversible and twice-differentiable function, where $\mathcal{Z}$ is the representation space. ψ is the reverse representation such that $\psi(\phi(s)) = s$. The specific form of our MDP model is: $\widehat{M}_\phi = \langle \widehat{r}(s, a), \widehat{T}(s', s, a) \rangle = \langle h_r(\phi(s), a), h_T(\phi(s'), \phi(s), a) \rangle$, where h_r and h_T is some function over space $\mathcal{Z}$. We will use the notation $\widehat{M}$ instead of $\widehat{M}_\phi$ later for simplicity.

Let $\tau = (s_0, a_0, \dots, s_H)$ be a trajectory of $H + 1$ states and actions, sampled from the joint distribution of MDP M and a policy μ . The joint distributions of τ are: $p_{M,\mu}(\tau) = p_0(s_0) \prod_{t=0}^{H-1} [T(s_{t+1}|s_t, a_t)\mu(a_t|s_t)]$. Given the joint distribution, we denote the associated marginal and conditional distributions as $p_{M,\mu}(s_0), p_{M,\mu}(s_0, a_0), p_{M,\mu}(s_0|a_0)$ etc. We also have the joint, marginal and conditional, distributions $p_{M,\mu}^\phi(\cdot)$ based on the representation space $\mathcal{Z}$. We focus on the undiscounted finite horizon case, using $V_{M,t}^\pi(s)$ to denote the t -step value function of policy π .

4 Generalization Error Bound for MDP based OPPE estimator

Our goal is to learn a MDP model $\widehat{M}$ that directly minimizes a good upper bound of the MSE for the individual evaluation policy π values: $\mathbb{E}_{s_0} [V_{\widehat{M}}^\pi(s_0) - V_M^\pi(s_0)]^2$. This model can provide value function estimates of the policy π and be used as part of doubly robust methods.

In the on-policy case, the Simulation Lemma ([13] and repeated for completeness in Lemma 1) shows that MSE of a policy value estimate can be upper bounded by a function of the reward and transition prediction losses. Before we state this result, we first define some useful notation.

Definition 1. The square error loss function of value function, reward, transition are:

$$\begin{aligned} \bar{\ell}_V(s, \widehat{M}, H-t) &= \left(V_{\widehat{M}, H-t}^\pi(s) - V_{M, H-t}^\pi(s) \right)^2 & \bar{\ell}_r(s_t, a_t, \widehat{M}) &= (\widehat{r}(s_t, a_t) - \bar{r}(s_t, a_t))^2 \\ \bar{\ell}_T(s_t, a_t, \widehat{M}) &= \left(\int_{\mathcal{S}} \left(\widehat{T}(s'|s_t, a_t) - T(s'|s_t, a_t) \right) V_{\widehat{M}, H-t-1}^\pi(s') ds' \right)^2 \end{aligned} \quad (1)$$

133 Then the Simulation lemma ensures that

$$\mathbb{E}_{s_0} \left[V_{\widehat{M}}^{\pi}(s_0) - V_M^{\pi}(s_0) \right]^2 \leq 2H \sum_{t=0}^{H-1} \mathbb{E}_{s_t, a_t \sim p_{M, \pi}} \left[\bar{\ell}_r(s_t, a_t, \widehat{M}) + \bar{\ell}_T(s_t, a_t, \widehat{M}) \right], \quad (2)$$

134 The right hand side can be used to formulate an objective to fit a model for policy evaluation. In
 135 off-policy case our data is from a different policy μ , and one can get unbiased estimation of the RHS
 136 of Equation 2 by importance sampling. However, this will provide an objective function with high
 137 variance, especially for a long horizon MDP or a deterministic evaluation policy due to the product of
 138 IS weights. An alternative is to learn an MDP model by directly optimizing the prediction loss over
 139 our observational data, ignoring the covariate shift. From the Simulation Lemma this minimizes an
 140 upper bound of MSE of behavior policy value, but the resulting model may not be a good one for
 141 estimating the evaluation policy value. In this paper we propose a new upper bound on the MSE of
 142 the individual evaluation policy values inspired by recent work in treatment effect estimation, and use
 143 this as a loss function for fitting models.

144 Before proceeding we first state our assumptions, which are common in most OPPE algorithms:

- 145 1. Support of behavior policy covers the evaluation policy: for any state s and action a ,
 146 $\mu(a|s) = 0$ only if $\pi(a|s) = 0$.
- 147 2. Strong ignorability: there are no hidden confounders that influence the choice of actions
 148 other than the current observed state.

149 Denote a *factual* sequence to be a trajectory that matches the evaluation policy, $a_0 =$
 150 $\pi(s_0), \dots, a_{t-1} = \pi(s_{t-1})$ as $a_{0:t-1} = \pi$. Let a *counterfactual* action sequence $a_{0:t-1} \neq \pi$
 151 be an action sequence with at least one action that does not match $\pi(s)$. $p_{M, \mu}(\cdot)$ is the distribution
 152 over trajectories under M and policy μ . We define the $H - t$ step value error with respect to the state
 153 distribution given the factual action sequence.

154 **Definition 2.** $H - t$ step value error is: $\epsilon_V(\widehat{M}, H - t) = \int_S \bar{\ell}_V(s_t, H - t) p_{M, \mu}(s_t | a_{0:t-1} = \pi) ds_t$

155 We use the idea of bounding the distance between representations given factual and counterfactual
 156 action sequences to adjust the distribution mismatch. Here the distance between representation
 157 distributions is formalized by Integral Probability Metric (IPM).

158 **Definition 3.** Let p, q be two distributions and let G be a family of real-valued functions defined over
 159 the same space. The integral probability metric is: $IPM_G(p, q) = \sup_{g \in G} \left| \int g(x)(p(x) - q(x)) dx \right|$

160 Some important instances of IPM include the Wasserstein metric where G is 1-Lipschitz continuous
 161 function class, and Maximum Mean Discrepancy where G is norm-1 function class in RKHS.

162 Let $p_{M, \mu}^{\phi, F}(z_t) = p_{M, \mu}^{\phi}(z_t | a_{0:t} = \pi)$ and $p_{M, \mu}^{\phi, CF}(z_t) = p_{M, \mu}^{\phi}(z_t | a_t \neq \pi, a_{0:t-1} = \pi)$, where F and
 163 CF denote factual and counterfactual. We first give an upper bound of MSE in terms of an expected
 164 loss term and then develop a finite sample bound which can be used as a learning objective.

165 **Theorem 1.** For any MDP M , approximate MDP model $\widehat{M}$, behavior policy μ and deterministic
 166 evaluation policy π , let $B_{\phi, t}$ and G_t be a real number and function family that satisfy the condition
 167 in Lemma 4. Then:

$$\begin{aligned} \mathbb{E}_{s_0} \left[V_{\widehat{M}}^{\pi}(s_0) - V_M^{\pi}(s_0) \right]^2 &\leq 2H \sum_{t=0}^{H-1} \left[B_{\phi, t} IPM_{G_t} \left(p_{M, \mu}^{\phi, F}(z_t), p_{M, \mu}^{\phi, CF}(z_t) \right) \right. \\ &\quad \left. + \int_S \frac{1}{p_{M, \mu}(a_{0:t} = \pi)} \left(\bar{\ell}_r(s_t, \pi(s_t), \widehat{M}) + \bar{\ell}_T(s_t, \pi(s_t), \widehat{M}) \right) p_{M, \mu}(s_t, a_{0:t} = \pi) ds_t \right] \quad (3) \end{aligned}$$

168 **(Proof Sketch)** The key idea is to use Equation 20 in Lemma 1 to view each step as a contextual
 169 bandit problem, and bound $\epsilon_V(\widehat{M}, H)$ recursively. We decompose the value function error into a one
 170 step reward loss, a transition loss and a next step value loss, with respect to the on-policy distribution.
 171 We can treat this as a contextual bandit problem, and we build on the method in Shalit et al.'s work
 172 [19] about binary action bandits to bound the distribution mismatch by a representation distance
 173 penalty term; however, additional care is required due to the sequential setting since the next states
 174 are also influenced by the policy. By adjusting the distribution for the next step value loss, we reduce
 175 it into $\epsilon_V(\widehat{M}, H - t - 1)$, allowing us recursively repeat this process for H steps. $\square$

This theorem bounds the MSE for the individual evaluation policy value by a loss on the distribution of the behavior policy, with the cost of an additional representation distribution metric. The first IPM term measures how different the state representations are conditional on factual and counterfactual action history. Intuitively, a balanced representation can generalize better from the observational data distribution to the data distribution under the evaluation policy, but we also need to consider the prediction ability of the representation on the observational data distribution. This bound quantitatively describes those two effects about MSE by the IPM term and the loss terms. The re-weighted expected loss terms over the *observational* data distribution is weighted by the marginal action probabilities ratio instead of the conditional action probability ratio, which is used in importance sampling. The marginal probabilities ratio has lower variance than the importance sampling weights in practice, especially when the sample size is limited.

One natural approach might be to use the right hand side of Equation 3 as a loss, and try to directly optimize a representation and model that minimizes this upper bound on the mean squared error in the individual value estimates. Unfortunately, doing so can suffer from two important issues. (1) The subset of the data that matches the evaluation policy can be very sparse for large t , and though the above bound re-weights data, fitting a model to it can be challenging due to the limited data size. (2) Unfortunately this approach ignores all the other data present that do not match the evaluation policy. If we are also learning a representation of the domain in order to scale up to very large problems, we suspect that we may benefit from framing the problem as related to transfer or multitask learning.

Motivated by viewing off-policy policy evaluation as a transfer learning task, we can view the source task as the evaluating the behavior policy, for which we have on-policy data, and view the target task as evaluating the evaluation policy, for which we have the high-variance re-weighted data from importance sampling. This is similar to transfer learning where we only have a few, potentially noisy, data points for the target task. Thus we can take the idea of co-learning a source task and a target task at the same time as a sort of regularization given limited data. More precisely, we now bound the OPPE error by an upper bound of the sum of two terms:

$$\underbrace{\mathbb{E}_{s_0} [V_{\widehat{M}}^{\pi}(s_0) - V_M^{\pi}(s_0)]^2}_{\text{MSE}_{\pi}} + \underbrace{\mathbb{E}_{s_0} [V_{\widehat{M}}^{\mu}(s_0) - V_M^{\mu}(s_0)]^2}_{\text{MSE}_{\mu}}, \quad (4)$$

where we bound the former part using Theorem 1. Thus our upper bound of this objective can address the issues with separately using MSE_{π} and MSE_{μ} as objective: compared with IS estimation of MSE_{π} , the "marginal" action probability ratio has lower variance. The representation distribution distance term regularizes the representation layer such that the learned representation would not vary significantly between the state distribution under the evaluation policy and the state distribution under the behavior policy. That reduces the concern that using MSE_{μ} as an objective will force our model to evaluate the behavior policy, rather than the evaluation policy, more effectively.

Our work is also inspired by treatment effect estimation in the casual inference literature, where we estimate the difference between the treated and control groups. An analogue in RL would be estimating the difference between the target policy value and the behavior policy value, by minimizing the MSE of policy difference estimation. The objective above is an upper bound of the MSE of policy

difference estimator: $\frac{1}{2} \mathbb{E}_{s_0} \left[\left(V_{\widehat{M}}^{\pi}(s_0) - V_M^{\mu}(s_0) \right) - \left(V_M^{\pi}(s_0) - V_M^{\mu}(s_0) \right) \right]^2 \leq \text{MSE}_{\pi} + \text{MSE}_{\mu}$

We now bound Equation 4 further by finite sample terms. For the finite sample generalization bound, we first introduce a minor variant of the loss functions, with respect to the sample set.

Definition 4. Let r_t and s'_t be an observation of reward and next step given state action pair s_t, a_t . Define the loss functions as:

$$\ell_r(s_t, a_t, r_t, \widehat{M}) = (\widehat{r}(s_t, a_t) - r_t)^2 \quad (5)$$

$$\ell_T(s_t, a_t, s'_t, \widehat{M}) = \left(\int_S \widehat{T}(s'|s_t, a_t) V_{\widehat{M}, H-t-1}^{\pi}(s') ds' - V_{\widehat{M}, H-t-1}^{\pi}(s'_t) \right)^2 \quad (6)$$

Definition 5. Define the empirical risk over the behavior distribution and weighted distribution as:

$$\widehat{R}_{\mu}(\widehat{M}) = \frac{1}{n} \sum_{i=1}^n \sum_{t=0}^{H-1} \ell_r(s_t^{(i)}, a_t^{(i)}, r_t^{(i)}, \widehat{M}) + \ell_T(s_t^{(i)}, a_t^{(i)}, s_t'^{(i)}, \widehat{M}) \quad (7)$$

$$\widehat{R}_{\pi, u}(\widehat{M}) = \frac{1}{n} \sum_{i=1}^n \sum_{t=0}^{H-1} \frac{\mathbb{1}(a_{0:t}^{(i)} = \pi)}{\widehat{u}_{0:t}} \left[\ell_r(s_t^{(i)}, a_t^{(i)}, r_t^{(i)}, \widehat{M}) + \ell_T(s_t^{(i)}, a_t^{(i)}, s_t'^{(i)}, \widehat{M}) \right], \quad (8)$$

where n is the dataset size, $s_t^{(i)}$ is the state of the t^{th} step in the i^{th} trajectory, and $\hat{u}_{0:t} = \sum_{i=1}^n \frac{\mathbb{1}(a_{0:t}^{(i)} = \pi)}{n}$.

Theorem 2. Suppose $\mathcal{M}_\Phi$ is a model class of MDP models based on representation ϕ . For n trajectories sampled by μ , let $\ell_t(s_t, a_t, \widehat{M}_\phi) = \ell_r(s_t, a_t, r_t, \widehat{M}) + \ell_T(s_t, a_t, s'_t, \widehat{M})$, and d_t be the pseudo-dimension of function class $\{\ell_t(s_t, a_t, \widehat{M}_\phi), \widehat{M}_\phi \in \mathcal{M}_\Phi\}$. Suppose $\mathcal{H}$ is the reproducing kernel Hilbert space induced by k , and $\mathcal{F}$ is the unit ball in it. Assume there exists a constant $B_{\phi,t}$ such that $\frac{1}{B_{\phi,t}}\ell_t(\psi(z), \pi(\psi(z)), \widehat{M}_\phi) \in \mathcal{F}$. With probability $1 - 3\delta$, for any $\widehat{M} \in \mathcal{M}_\Phi$:

$$\begin{aligned} \mathbb{E}_{s_0} \left[V_M^\pi(s_0) - V_M^\pi(s_0) | \widehat{M} \right]^2 &\leq \text{MSE}_\pi + \text{MSE}_\mu \leq 2H\widehat{R}_\mu(\widehat{M}) + 2H\widehat{R}_{\pi,u}(\widehat{M}) \\ &+ 2H \sum_{t=0}^{H-1} B_{\phi,t} \left(\text{IPM}_{\mathcal{F}} \left(\widehat{p}_{M,\mu}^{\phi,F}(z_t), \widehat{p}_{M,\mu}^{\phi,CF}(z_t) \right) + \min \left\{ \mathcal{D}_\delta^{\mathcal{F}} \left(\frac{1}{\sqrt{m_{t,1}}} + \frac{1}{\sqrt{m_{t,2}}} \right), 2\nu \right\} \right) \\ &+ 2H \sum_{t=0}^{H-1} \frac{\mathcal{C}_{n,\delta,t}^{\mathcal{M}}}{n^{3/8}} \left(\mathbb{V} \left[\frac{\mathbb{1}(a_{0:t} = \pi)}{\widehat{u}_{0:t}}, \ell_t \right] + \mathbb{V}[1, \ell_t] + \ell_{t,\max} \mathbb{V} \left[\frac{\mathbb{1}(a_{0:t} = \pi)}{u_{0:t}}, 1 \right] \right) \quad (9) \end{aligned}$$

$m_{t,1}$ and $m_{t,2}$ are the number of samples used to estimate $\widehat{p}_{M,\mu}^{\phi,F}(z_t)$ and $\widehat{p}_{M,\mu}^{\phi,CF}(z_t)$ respectively. $\mathcal{D}_\delta^{\mathcal{F}}$ is a function of the kernel k . $\mathcal{C}_{n,\delta,t}^{\mathcal{M}}$ is a function of d_t . $\mathbb{V}[w, \ell_t] = \max\{\sqrt{\mathbb{E}_{p_{M,\mu}}[w^2 \ell_t^2]}, \sqrt{\mathbb{E}_{\widehat{p}_{M,\mu}}[w^2 \ell_t^2]}\}$, $\ell_{t,\max} = \max_{s_t, a_t} |\ell_t(s_t, a_t)|$.

The first term is the empirical loss over the observational data distribution. The second term is a re-weighted empirical loss, which is an empirical version of the first term in Theorem 1. As said previously, this has less variance than importance sampling ratios in practice, especially when the sample size is limited. Our bound is based on the empirical estimate of the marginal probability $u_{0:t}$ and we are not required to know the behavior policy. Our method’s independence of the behavior policy is a significant advantage over IS methods which are very susceptible to errors its estimation, as we discuss in appendix A. In practice, this marginal probability $u_{0:t}$ is easier to estimate than μ when μ is unknown. The third term is an empirical estimate of IPM, which we described in Theorem 1. We use norm-1 RKHS functions and MMD distance in this theorem and our algorithm. There are similar but worse results for Wasserstein distance and total variation distance [20]. $\mathcal{D}_\delta^{\mathcal{F}}$ measures how complex $\mathcal{F}$ is. It is obtained from concentration measures about empirical IPM estimators [20]. The constant $\mathcal{C}_{n,\delta,t}^{\mathcal{M}}$ measures how complex the model class is and it is derived from traditional learning theory results [4].

5 Algorithm for Representation Balancing MDPs

Based on our generalization bound above, we propose an algorithm to learn an MDP model for OPPE, minimizing the following objective function:

$$\mathcal{L}(\widehat{M}_\phi; \alpha_t) = \widehat{R}_\mu(\widehat{M}_\phi) + \widehat{R}_{\pi,u}(\widehat{M}_\phi) + \sum_{t=0}^{H-1} \alpha_t \text{IPM}_{\mathcal{F}} \left(\widehat{p}_{M,\mu}^{\phi,F}(z_t), \widehat{p}_{M,\mu}^{\phi,CF}(z_t) \right) + \frac{\mathfrak{R}(\widehat{M}_\phi)}{n^{3/8}} \quad (10)$$

This objective is based on Equation 9 in Theorem 2. We minimize the terms in this upper bound that are related to the model $\widehat{M}_\phi$. Note that since $B_{\phi,t}$ depends on the loss function, we cannot know $B_{\phi,t}$ in practice. We therefore use a tunable factor α in our algorithm. $\mathfrak{R}(\widehat{M}_\phi)$ here is some kind of bounded regularization term of model that one could choose, corresponding to the model class complexity term in Equation 9. This objective function matches our intuition about using lower-variance weights for the re-weighting component and using IPM of the representation to avoid fitting the behavior data distribution.

In this work, $\phi(s)$ and $\widehat{M}_\phi$ are parameterized by neural networks, due to their strong ability to learn representations. We use an estimator of IPM term from Sriperumbudur et al. [21]. All terms in the objective function are differentiable, allowing us to train them jointly by minimizing the objective by a gradient based optimization algorithm.

After we learn an MDP by minimizing the objective above, we use Monte-Carlo estimates or value iteration to get the value for any initial state s_0 as an estimator of policy value for that state. We show that if there exists an MDP and representation model in our model class that could achieve:

$$\min_{\widehat{M}_\phi} \left(R_\mu(\widehat{M}_\phi) + R_{\pi,u}(\widehat{M}_\phi) + \sum_{t=0}^{H-1} \alpha_t \text{IPM}_{\mathcal{F}} \left(p_{M,\mu}^{\phi,F}(z_t), p_{M,\mu}^{\phi,CF}(z_t) \right) \right) = 0,$$

then $\lim_{n \rightarrow \infty} \mathbb{E}_{s_0} [V_{\widehat{M}_{\phi^*}^\pi}^\pi(s_0) - V_M^\pi(s_0)]^2 \rightarrow 0$ and estimator $V_{\widehat{M}_{\phi^*}^\pi}^\pi(s_0)$ is a consistent estimator for any s_0 . See Corollary 2 in Appendix for detail.

We can use our model in any OPPE estimators that leverage model-based estimators, such as doubly robust [9] and MAGIC [22], though our generalization MSE bound is just for the model value.

6 Experiments

6.1 Synthetic domain: Cart Pole

We test our algorithm on Cart Pole, a continuous-state benchmark domain. We use a greedy policy from a learned Q function as the evaluation policy, and an ϵ -greedy policy with $\epsilon = 0.2$ as the behavior policy. We collect 1024 trajectories for OPPE, whose average length is around 190. This setting is challenging for IS-based OPPE estimators due to the deterministic evaluation policy and long horizon, which will give the IS weights high variance. However determinism and long horizons can be very common in real-world domains, and they are key limitations of current OPPE algorithms.

We compare our method RepBM, with a baseline approximate model (AM), doubly robust (DR), more robust doubly robust (MRDR), and importance sampling (IS). The baseline approximate model is an MDP model-based estimator trained by minimizing the empirical risk, using the same model class as RepBM. DR(RepBM) is doubly robust estimator using our model and DR(AM) is a doubly robust estimator using the baseline model. MRDR [6] is a recent method that trains a Q function as the model-based part in DR to minimize the resulting variance. We include their Q function estimator (MRDR Q), the doubly robust estimator that combines this Q function with IS (MRDR).

The reported results are square root of the average MSE over 100 runs. α is set to 0.01 for RepBM. We report mean and individual MSEs, corresponding to MSEs of average policy value and individual policy value, $[\mathbb{E}_{s_0} \widehat{V}(s_0) - \mathbb{E}_{s_0} V(s_0)]^2$ and $\mathbb{E}_{s_0} [\widehat{V}(s_0) - V(s_0)]^2$ respectively. IS and DR methods re-weight samples, so their estimates for single initial states are not applicable, especially in continuous state space. A comparison across more methods is included in the appendix.

Table 1: Root MSE for Cart Pole

Long Horizon	RepBM	DR(RepBM)	AM	DR(AM)	MRDR Q	MRDR	IS
Mean	0.4121	1.359	0.7535	1.786	151.1	202	194.5
Individual	1.033	-	1.313	-	151.9	-	-
Short Horizon	RepBM	DR(RepBM)	AM	DR(AM)	MRDR Q	MRDR	IS
Mean	0.07836	0.02081	0.1254	0.0235	3.013	0.258	2.86
Individual	0.4811	-	0.5506	-	3.823	-	-

Representation Balancing MDPs outperform baselines for long time horizons. We observe that MRDR variants and IS methods have high MSE in the long horizon setting. The reason is that the IS weights for 200-step trajectories are extremely high-variance, and MRDR whose objective depends on the square of IS weights, also fails. Compared with the baseline model, we can see that our method is better than AM for both the pure model case and when used in doubly robust. We also observe that the IS part in doubly robust actually hurts the estimates, for both RepBM and AM.

Representation Balancing MDPs outperform baselines in deterministic settings. We also include results on Cart Pole with a shorter horizon, by using weaker evaluation and behavior policies. The average length of trajectories is about 23 in this setting. Here, we observe that RepBM is still better than other model-based estimators, and doubly robust that uses RepBM is still better than other doubly robust methods. Though MRDR produces substantially lower MSE than IS, which matches the report in Farajtabar et al. [6], it still has higher MSE than RepBM and AM, due to the high variance of its learning objective when the evaluation policy is deterministic.

Representation Balancing MDPs produce accurate estimates even when the behavior policy is unknown. For both horizon cases, we observe that RepBM learned with no knowledge of the behavior policy is better than methods such as MRDR and IS that use the true behavior policy.

6.2 Real-world domain: Sepsis

We demonstrate our method on a real world example of evaluating policies for sepsis treatment in medical intensive care units (ICU). We extract time series data for 9983 sepsis patients from the MIMIC III data set [12]. The data for each patient consists of 47 features, recorded at 4 hour intervals. We define the patient state as a concatenation of the last 4 observations with 188 features. The actions we consider are the prescription of IV fluids and vasopressors. Each of the two treatments is binned into 5 discrete actions according to the dosage amounts, resulting in 25 possible actions. The rewards are defined as the change of an acuity measure, the SOFA score, following an action. We treat the data as a task with a fixed time horizon by only observing the first 10 transitions of every patient.

The main difficulty in performing the real data experiment is that we do not have access to the true value of any policy for computing the MSE on the value estimates. We address this issue by artificially splitting the data into two subsets, with each subset representing a different distribution of actions. As a control group, we select one sub-group containing half of the patients who were never treated with vasopressors (chosen randomly from all such patients), and assign the rest of the patients to the other group which we treat as our training dataset. This is equivalent to observing two separate datasets collected under different behavior policies. We approximate the evaluation and behavior policies by a nearest-neighbors classifier, discussed in detail in Appendix E.2. We compute the value of the policy which never uses vasopressors by taking the empirical average of the value from the control group, and treat this average as the ground truth policy value. We then try to estimate the value of this policy by performing off-policy policy evaluation with the data from the training dataset. We ran the experiments $N = 10$ times, where for each experiment we repeat the splitting process. The results for the RMSE of the average policy value are reported in Table 2. For this real-world dataset, the lack of access to the true value of a policy makes computation of the RMSE for individual patients very noisy, and we therefore focus on population averages.

Table 2: Root MSE for Sepsis

	RepBM	DR(RepBM)	AM	DR(AM)	IS	WDR(RepBM)	WDR(AM)
Mean	0.09	$3 \cdot 10^5$	0.23	$4 \cdot 10^5$	$5 \cdot 10^5$	0.8	0.73

On a real-world domain, RepBM outperforms all other OPE methods, and is the only one which can confidently differentiate between the behavior and evaluation policies. We observe that RepBM has the lowest root MSE on estimating the value of the evaluation policy. The average expected return computed using on-policy estimation is -0.07 for the no vasopressor policy and -0.17 for the policy including vasopressors (more negative is better). As the domain prior implies, the policy including the administration of vasopressors does better in terms of SOFA score reduction, as it employs a wider range of treatment options. RepBM is the only method for which the RMSE is smaller than the difference in expected return between the two policies, and is thus the only method which can confidently differentiate between them.

7 Discussion and Conclusion

One interesting issue for our method is the effect of the hyper-parameter α on the quality of estimator. In the appendix, we include the results of RepBM across different values of α . We find that our method outperforms prior work for a large range of alphas, for both domains. In both domains we observe that the effect of IPM adjustment (non-zero α) is less than the effect of "marginal" IS re-weighting, which matches the results in Shalit et al.'s work in the binary action bandit case [19].

To conclude, in this work we give an MDP model learning method for the individual OPPE problem in RL, based on a new finite sample generalization bound of MSE for the model value estimator. We show our method results in substantially smaller MSE estimates compared to state-of-the-art baselines in a common benchmark simulator and on a challenging real-world dataset on sepsis management.

References

- [1] A. M. Alaa and M. van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *arXiv preprint arXiv:1704.02801*, 2017.
- [2] O. Atan, W. R. Zame, and M. van der Schaar. Learning optimal policies from observational data. *arXiv preprint arXiv:1802.08679*, 2018.
- [3] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.
- [4] C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In *Advances in neural information processing systems*, pages 442–450, 2010.
- [5] M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.
- [6] M. Farajtabar, Y. Chow, and M. Ghavamzadeh. More robust doubly robust off-policy evaluation. *arXiv preprint arXiv:1802.03493*, 2018.
- [7] Z. D. Guo, P. S. Thomas, and E. Brunskill. Using options for long-horizon off-policy evaluation. *arXiv preprint arXiv:1703.03453*, 2017.
- [8] J. J. J. Yoon and M. van der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. *ICLR*, 2018.
- [9] N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. *arXiv preprint arXiv:1511.03722*, 2015.
- [10] F. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pages 3020–3029, 2016.
- [11] F. D. Johansson, N. Kallus, U. Shalit, and D. Sontag. Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598*, 2018.
- [12] A. E. Johnson, D. J. Stone, L. A. Celi, and T. J. Pollard. The mimic code repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, page ocx084, 2017.
- [13] M. Kearns and S. Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2-3):209–232, 2002.
- [14] S. Künnel, J. Sekhon, P. Bickel, and B. Yu. Meta-learners for estimating heterogeneous treatment effects using machine learning. *arXiv preprint arXiv:1706.03461*, 2017.
- [15] T. Mandel, Y.-E. Liu, S. Levine, E. Brunskill, and Z. 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.
- [16] D. Precup, R. S. Sutton, and S. P. Singh. Eligibility traces for off-policy policy evaluation. In *ICML*, pages 759–766. Citeseer, 2000.
- [17] J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.
- [18] P. Schulam and S. Saria. Reliable decision support using counterfactual models. In *Advances in Neural Information Processing Systems*, pages 1697–1708, 2017.
- [19] U. Shalit, F. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. *arXiv preprint arXiv:1606.03976*, 2016.
- [20] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. Lanckriet. On integral probability metrics, ϕ -divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.
- [21] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, G. R. Lanckriet, et al. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.
- [22] P. Thomas and E. Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148, 2016.

- 389 [23] P. S. Thomas, G. Theocharous, and M. Ghavamzadeh. High-confidence off-policy evaluation. In *AAAI*,
390 2015.
- 391 [24] S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests.
392 *Journal of the American Statistical Association*, just-accepted, 2017.