Representation Balancing MDPs for Off-Policy Policy Evaluation

Yao Liu
Stanford University
yaoliu@stanford.edu

Omer Gottesman
Harvard University
gottesman@fas.harvard.edu

Aniruddh Raghu
Cambridge University
aniruddhraghu@gmail.com

Matthieu Komorowski
Imperial College London
matthieu.komorowski@gmail.com

Aldo Faisal
Imperial College London
a.faisal@imperial.ac.uk

Finale Doshi-Velez
Harvard University
finale@seas.harvard.edu

Emma Brunskill
Stanford University
ebrun@cs.stanford.edu

Abstract

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

1 Introduction

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

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

Approaches for OPPE must make a choice about whether and how to address this data distribution mismatch. Importance sampling (IS) based approaches [16, 23, 7, 5, 9, 22] are typically unbiased and strongly consistent, but despite recent progress tend to have high variance—especially if the evaluation policy is deterministic, as evaluating deterministic policies requires finding in the data

Preprint. Work in progress.

sequences where the actions exactly match the evaluation policy. However, in most real-world applications deterministic evaluation policies are more common—policies are typically to either amputate or not, rather than a policy that that flips a biased coin (to sample randomness) to decide whether to amputate. IS approaches also often rely on explicit knowledge of the behavior policy, which may not be feasible in situations such as medicine where the behaviors results from human actions. In contrast, some model based approaches ignore the data distribution mismatch, such as by fitting a maximum-likelihood model of the rewards and dynamics from the behavioral data, and then using that model to evaluate the desired evaluation policy. These methods may not converge to the true estimate of the evaluation policy's value, even in the limit of infinite data [15]. However, such model based approaches often achieve better empirical performance than the IS-based estimators [9].

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

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

Recent work [10, 19] on CATE[1] has obtained very promising results by learning a model to predict individual outcomes using a (model fitting) loss function that explicitly accounts for the data distribution shift between the treatment and control policies. We build on this work to introduce a new bound on the MSE for individual policy values, and a new loss function for fitting a model-based OPPE estimator. In contrast to most other OPPE theoretical analyses (e.g. [9, 5, 22]), we provide a finite sample generalization error instead of asymptotic consistency. In contrast to previous model value generalization bounds such as the Simulation Lemma [13], our bound accounts for the underlying data distribution shift if the data used to estimate the value of an evaluation policy were collected by following an alternate policy.

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

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

2 Related Work

Most prior work on OPPE in reinforcement learning falls into one of three approaches. The first, importance sampling (IS), reweights the trajectories to account for the data distribution shift. Under mild assumptions importance sampling estimators are guaranteed to be both unbiased and strongly consistent, and were first introduced to reinforcement learning OPPE by Precup et al. [16]. Despite

[1] Shalit 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.

recent progress (e.g.[23, 7]) IS-only estimators still often yield very high variance estimates, particularly when the decision horizon is large, and/or when the evaluation policy is deterministic. IS estimators also typically result in extremely noisy estimates for policy values of individual states. A second common approach is to estimate a dynamics and reward model, which can substantially reduce variance, but can be biased and inconsistent (as noted by [15]). The third approach, doubly robust estimators, originates from the statistics community [17]. Recently proposed doubly robust estimators for OPPE from the machine and reinforcement learning communities [5, 9, 22] have sometimes yielded orders of magnitude tighter estimates. However, most prior work that leverages 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 $\tau = (s_0, a_0, \ldots, s_H)$ 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 $p^\phi_{M,\mu}(\cdot)$ based on the representation space $\mathcal{Z}$. We focus on the undiscounted finite horizon case, using $V^\pi_{M,t}(s)$ 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: $\mathbb{E}_{s_0}[V^\pi_{\widehat{M}}(s_0) - V^\pi_M(s_0)]^2$. 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:*

$$\bar{\ell}_V(s, \widehat{M}, H-t) = \left(V^{\pi}_{\widehat{M},H-t}(s) - V^{\pi}_{M,H-t}(s)\right)^2 \qquad \bar{\ell}_r(s_t, a_t, \widehat{M}) = \left(\widehat{r}(s_t, a_t) - \bar{r}(s_t, a_t)\right)^2$$

$$\bar{\ell}_T(s_t, a_t, \widehat{M}) = \left(\int_{\mathcal{S}} \left(\widehat{T}(s'|s_t, a_t) - T(s'|s_t, a_t)\right) V^{\pi}_{\widehat{M},H-t-1}(s') ds'\right)^2 \tag{1}$$

Then the Simulation lemma ensures that

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned}\mathbb{E}_{s_0}\left[V^{\pi}_{\widehat{M}}(s_0) - V^{\pi}_{M}(s_0)\right]^2 \leq 2H \sum_{t=0}^{H-1} \Big[& B_{\phi,t} \mathit{IPM}_{G_t}\left(p^{\phi,F}_{M,\mu}(z_t), p^{\phi,CF}_{M,\mu}(z_t)\right) \\ & + \int_{\mathcal{S}} \frac{1}{p_{M,\mu}(a_{0:t} = \pi)} \left(\bar{\ell}_r(s_t, \pi(s_t), \widehat{M}) + \bar{\ell}_T(s_t, \pi(s_t), \widehat{M})\right) p_{M,\mu}(s_t, a_{0:t} = \pi) ds_t \Big]\end{aligned} \tag{3}$$

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

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:

$$\underbrace{\mathbb{E}_{s_0}\left[V^{\pi}_{\widehat{M}}(s_0) - V^{\pi}_{M}(s_0)\right]^2}_{\text{MSE}_\pi} + \underbrace{\mathbb{E}_{s_0}\left[V^{\mu}_{\widehat{M}}(s_0) - V^{\mu}_{M}(s_0)\right]^2}_{\text{MSE}_\mu}, \tag{4}$$

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

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:*

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

$$\ell_T(s_t, a_t, s'_t, \widehat{M}) = \left(\int_{\mathcal{S}} \widehat{T}(s'|s_t, a_t) V^{\pi}_{\widehat{M}, H-t-1}(s') ds' - V^{\pi}_{\widehat{M}, H-t-1}(s'_t) \right)^2 \tag{6}$$

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

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

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

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

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 $\widehat{M} \in \mathcal{M}_{\Phi}$:*

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

$m_{t,1}$ 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:

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

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:

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

then $\lim_{n\to\infty}\mathbb{E}_{s_0}[V^\pi_{\widehat{M}^*_{\phi^*}}(s_0)-V^\pi_M(s_0)]^2\to 0$ and estimator $V^\pi_{\widehat{M}^*_{\phi^*}}(s_0)$ 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. Künzel, 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,\phi-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.

[23] P. S. Thomas, G. Theocharous, and M. Ghavamzadeh. High-confidence off-policy evaluation. In *AAAI*, 2015.

[24] S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, just-accepted, 2017.

A IS with approximate behavior policy

In this section, we include some theoretical and empirical results about the effect of using an estimated behavior policy in importance sampling, when the true behavior policy is not accessible.

Proposition 1. *Assume the reward is in $[0, R_{\max}]$. For any estimator $\widehat{\mu}(a|s)$ of the true behavior policy $\mu(a|s)$, let $V_{IS}(\hat{\mu})$ be the IS estimator using this estimated $\hat{\mu}(a|s)$ and $V_{IS}(\mu)$ be the IS estimator with true behavior policy. Both of the IS estimators are computed using n trajectories that are independent from the data used to estimate $\widehat{\mu}(a|s)$. If the relative error of $\widehat{\mu}(a|s)$ is bounded by δ: $\| \frac{\widehat{\mu}(a|s)-\mu(a|s)}{\mu(a|s)} \|_\infty \leq \delta$, then for any given dataset:*

$$|V_{IS}(\hat{\mu}) - V_{IS}(\mu)| \leq \max\left\{ \left(\frac{1}{1-\delta}\right)^H - 1, 1 - \left(\frac{1}{1+\delta}\right)^H \right\} V_{IS}(\mu)$$

The bias of $V_{IS}(\hat{\mu})$ is bounded by:

$$|\mathbb{E}V_{IS}(\hat{\mu}) - v| \leq \max\left\{ \left(\frac{1}{1-\delta}\right)^H - 1, 1 - \left(\frac{1}{1+\delta}\right)^H \right\} v,$$

where v is the true evaluation policy value.

Proof.

$$\begin{aligned}
|V_{IS}(\hat{\mu}) - V_{IS}(\mu)| &= \left| \frac{1}{n}\sum_{i=1}^{n}\prod_{t=0}^{H-1} \frac{\pi(a_t^{(i)}|s_t^{(i)})}{\hat{\mu}(a_t^{(i)}|s_t^{(i)})} R_{0:H-1}^{(i)} - \frac{1}{n}\sum_{i=1}^{n}\prod_{t=0}^{H-1} \frac{\pi(a_t^{(i)}|s_t^{(i)})}{\mu(a_t^{(i)}|s_t^{(i)})} R_{0:H-1}^{(i)} \right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \left| \prod_{t=0}^{H-1} \frac{\pi(a_t^{(i)}|s_t^{(i)})}{\hat{\mu}(a_t^{(i)}|s_t^{(i)})} R_{0:H-1}^{(i)} - \prod_{t=0}^{H-1} \frac{\pi(a_t^{(i)}|s_t^{(i)})}{\mu(a_t^{(i)}|s_t^{(i)})} R_{0:H-1}^{(i)} \right| \qquad (11) \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \left| \prod_{t=0}^{H-1} \frac{\mu(a_t^{(i)}|s_t^{(i)})}{\hat{\mu}(a_t^{(i)}|s_t^{(i)})} - 1 \right| \left| \prod_{t=0}^{H-1} \frac{\pi(a_t^{(i)}|s_t^{(i)})}{\mu(a_t^{(i)}|s_t^{(i)})} R_{0:H-1}^{(i)} \right| \qquad (12)
\end{aligned}$$

According to the condition, for any $a^{(i)}$ and $s^{(i)}$, $1-\delta \leq \frac{\widehat{\mu}(a_t^{(i)}|s_t^{(i)})}{\mu(a_t^{(i)}|s_t^{(i)})} \leq 1+\delta$. Then

$$\frac{1}{1+\delta} \leq \frac{\mu(a_t^{(i)}|s_t^{(i)})}{\hat{\mu}(a_t^{(i)}|s_t^{(i)})} \leq \frac{1}{1-\delta},$$

and:

$$\left(\frac{1}{1+\delta}\right)^H \leq \prod_{t=0}^{H-1} \frac{\mu(a_t^{(i)}|s_t^{(i)})}{\hat{\mu}(a_t^{(i)}|s_t^{(i)})} \leq \left(\frac{1}{1-\delta}\right)^H,$$

So:

$$\left| \prod_{t=0}^{H-1} \frac{\mu(a_t^{(i)}|s_t^{(i)})}{\hat{\mu}(a_t^{(i)}|s_t^{(i)})} - 1 \right| \leq \max\left\{ \left(\frac{1}{1-\delta}\right)^H - 1, 1 - \left(\frac{1}{1+\delta}\right)^H \right\}$$

Plug this into Equation 12:

$$\begin{aligned}
|V_{IS}(\hat{\mu}) - V_{IS}(\mu)| &\leq \max\left\{ \left(\frac{1}{1-\delta}\right)^H - 1, 1 - \left(\frac{1}{1+\delta}\right)^H \right\} \frac{1}{n}\sum_{i=1}^{n}\prod_{t=0}^{H-1} \frac{\pi(a_t^{(i)}|s_t^{(i)})}{\mu(a_t^{(i)}|s_t^{(i)})} R_{0:H-1}^{(i)} \\
&= \max\left\{ \left(\frac{1}{1-\delta}\right)^H - 1, 1 - \left(\frac{1}{1+\delta}\right)^H \right\} V_{IS}(\mu) \qquad (13)
\end{aligned}$$

Similarly, for the bias:

$$\begin{aligned}
|\mathbb{E}V_{IS}(\hat{\mu}) - v| &= |\mathbb{E}V_{IS}(\hat{\mu}) - \mathbb{E}V_{IS}(\mu)| && (14)\\
&= \left| \mathbb{E} \prod_{t=0}^{H-1} \frac{\pi(a_t^{(0)}|s_t^{(0)})}{\hat{\mu}(a_t^{(0)}|s_t^{(0)})} R_{0:H-1}^{(0)} - \mathbb{E} \prod_{t=0}^{H-1} \frac{\pi(a_t^{(0)}|s_t^{(0)})}{\mu(a_t^{(0)}|s_t^{(0)})} R_{0:H-1}^{(0)} \right| && (15)\\
&\leq \left| \prod_{t=0}^{H-1} \frac{\mu(a_t^{(0)}|s_t^{(0)})}{\hat{\mu}(a_t^{(0)}|s_t^{(0)})} - 1 \right| \left| \mathbb{E} \left[\prod_{t=0}^{H-1} \frac{\pi(a_t^{(0)}|s_t^{(0)})}{\mu(a_t^{(0)}|s_t^{(0)})} R_{0:H-1}^{(0)} \right] \right| && (16)\\
&\leq \max \left\{ \left(\frac{1}{1-\delta} \right)^H - 1, 1 - \left(\frac{1}{1+\delta} \right)^H \right\} v && (17)
\end{aligned}$$

□

We bound the error of IS estimates by the relative error of behavior policy estimates. Proposition 3 from Farajtabar et al. [6] gave an expression for the bias when using an empirical estimate of behavior policy in IS. The result in Farajtabar et al. [6] is similar to this proposition, but the authors did not explicitly bound the bias by the error of behavior policy. Note that this bound increases exponentially with the horizon H, which shows the accumulated error effect of behavior policy error.

By using the tree MDP example in Jiang and Li [9] we can show that the order of magnitude $O(\exp(H))$ is tight: there exists an MDP and a policy estimator $\hat{\mu}$ with $\|(\hat{\mu} - \mu)/\mu\|_\infty = \delta$ such that the bias of $V_{IS}(\hat{\mu})$ is $O(\exp(H))$. Define a binary discrete tree MDP [9] as following: At each node in a binary tree, we can take two actions $a = 0, a = 1$, leading to the two next nodes with observations $o = 0, o = 1$. The state of a node is defined by the whole path to the root: $o_0 a_0 o_1 a_i \ldots o_h$. That means each node in the tree will have a unique state. The depth of the tree, as well as the horizon of the trajectories, is H. Only the leftmost leaf node (by always taking $a = 0$) has non-zero reward $r = 1$. Denote this state as the target state. The evaluation policy always takes action $a = 0$ and the behavior policy μ is a uniform random policy. Let the estimated policy $\hat{\mu}$ differ from μ with $-\delta/2$ in all the state-action pairs on the path to target state. That means the action probability in $\hat{\mu}$ is $1/2 - \delta/2$ for all state-action pairs on the path to target state. The IS estimator with μ has expectation 1 since it is unbiased. It is easy to verify that the IS estimator using $\hat{\mu}$ has expectation $(1-\delta)^{-H}$. Thus the bias is $O(\exp(H))$.

This result represents the worst-case upper bound on the bias of IS when using an estimated behaviour policy; the fact it is exponential in the trajectory length illustrates the problem when using IS without knowing the true behaviour policy. To support this result with an empirical example illustrating the challenge of using IS with an unknown behaviour policy for a real data distribution, consider Figure 1 which represents the error in OPPE (found using Per-Decision WIS) as we vary the accuracy of the behaviour policy estimation. Two different behaviour policies are considered. The domain used in this example is a continuous 2D map ($s \in \mathbb{R}^2$) with a discrete action space, $\mathcal{A} = \{1, 2, 3, 4, 5\}$, with actions representing a movement of one unit in one of the four coordinate directions or staying in the current position. Gaussian noise of zero mean and specifiable variance is added onto state of the agent after each action, to provide environmental stochasticity. An agent starts in the top left corner of the domain and receives a positive reward within a given radius of the top right corner, and a negative reward within a given radius of the bottom left corner. The horizon is set to be 15. A k-Nearest Neighbours (kNN) model is used to estimate the behaviour policy distribution, given a set of training trajectories. The accuracy of the model is varied by changing the number of trajectories available and the number of neighbours used for behaviour policy estimation.

This plot shows how IS suffers from very poor estimates with even slight errors in the estimated behaviour policy – average absolute errors of as small as 0.06 can incur errors of up over 50% in OPPE. This provides additional motivation for our approach – we do not require the behaviour policy to be known for OPPE, avoiding the significant errors incurred by using incorrectly estimated behaviour policies.

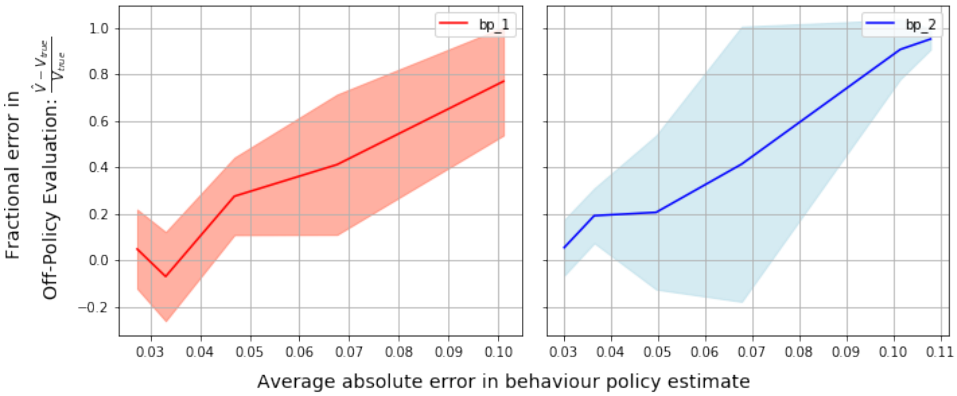

Figure 1: Plots showing the mean and standard deviation of the fractional error in OPE, $\frac{\hat{V}-V}{V}$, as a function of the average absolute error in behaviour policy estimation, $\frac{1}{n}\sum_{i=1}^{n}|\mu(a^{(i)}|s^{(i)}) - \hat{\mu}(a^{(i)}|s^{(i)})|$, for two different behaviour policies. The quality of OPE with IS has a very significant dependence on the accuracy of the behaviour policy estimation.

B Clarification of CATE/HTE and ITE

In the causal inference literature[14], for an single unit i with covariate (state) x_i, we observe $Y_i(1)$ if we give the unit treatment and $Y_i(0)$ if not. The Individual Treatment Effect (ITE) is defined as:

$$D_i = Y_i(1) - Y_i(0), \tag{18}$$

for this particular set of observations Y_i, x_i. However $Y_i(1)$ and $Y_i(0)$ cannot be observed at the same time, which makes ITE unidentifiable without strong additional assumptions. Thus the conditional average treatment effect (CATE), also known as heterogeneous treatment effect (HTE) is defined as:

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|x] \tag{19}$$

which is a function of x and is identifiable. Shalit et al. [19] defined ITE as $\tau(x)$, which is actually named as CATE or HTE in most causal reasoning literature. So we use the name CATE/HTE to refer to this quantity and it is inconsistent with Shalit et al. 's work. We clarify it here so that it does confuse the reader.

C Proofs of Section 4

C.1 Proofs of Theorem 1 and Corollary 1

Before we prove Lemma 4 and Theorem 1, we need some useful lemmas and assumptions. We restate a well-known variant of Simulation Lemma [13] in finite horizon case here:

Lemma 1. *(Simulation Lemma with finite horizon case) Define that $V^{\pi}_{\widehat{M},0}(s) = V^{\pi}_{M,0}(s) = 0$. For any approximate MDP model $\widehat{M}$, any policy π, and $t = 0, \ldots, H-1$:*

$$V^{\pi}_{\widehat{M},H-t}(s_t) - V^{\pi}_{M,H-t}(s_t) = \mathbb{E}_{a_t\sim\pi}\Big[\widehat{r}(s_t,a_t) - \bar{r}(s_t,a_t) + \int_{\mathcal{S}}\Big(\widehat{T}(s'|s_t,a_t) - T(s'|s_t,a_t)\Big) V^{\pi}_{\widehat{M},H-t-1}(s')ds' + \int_{\mathcal{S}} T(s'|s_t,a_t)\Big(V^{\pi}_{\widehat{M},H-t-1}(s') - V^{\pi}_{M,H-t-1}(s')\Big)ds'\Big] \tag{20}$$

Then:

$$V^{\pi}_{\widehat{M},H}(s) - V^{\pi}_{M,H}(s) = \mathbb{E}_{\pi,M}\sum_{t=0}^{H-1}\Big[\widehat{r}(s_t,a_t) - \bar{r}(s_t,a_t) + \int_{\mathcal{S}}\Big(\widehat{T}(s'|s_t,a_t) - T(s'|s_t,a_t)\Big) V^{\pi}_{\widehat{M},H-t-1}(s')ds'|s_0 = s\Big] \tag{21}$$

Lemma 2. *Let $J_\psi(z)$ be the absolute of the determinant of the Jacobian of $\psi(z)$. Then for any $z_t = \phi(s_t)$ and any sequence of actions $a_{0:t} = a_0, \dots, a_t$:*

$$p^\phi_{M,\mu}(z_t|a_{0:t}) = J_\psi(z_t) p_{M,\mu}(\psi(z_t)|a_{0:t})$$

Proof. By the change of variable formula in a probability density function, we have:

$$p^\phi_{M,\mu}(z_t|a_{0:t}) = \frac{p^\phi_{M,\mu}(z_t, a_{0:t})}{p^\phi_{M,\mu}(a_{0:t})} = \frac{p_{M,\mu}(\psi(z_t), a_{0:t}) J_\psi(z_t)}{p_{M,\mu}(a_{0:t})} = J_\psi(z_t) p_{M,\mu}(\psi(z_t)|a_{0:t})$$

□

Lemma 3. *Let p and q be two distributions over the state space with the form of $p_{M,\mu}(s_t|a_{0:t})$(the action sequence $a_{0:t}$ might be different for p and q), and p^ϕ and q^ϕ be the corresponding distributions over the representation space. For any real valued function over the state space f, if there exists a constant $B_\phi > 0$ and a function class G such that: $\frac{1}{B_\phi} f(\psi(z)) \in G$ then we have that*

$$\int_{\mathcal{S}} f(s)p(s)ds - \int_{\mathcal{S}} f(s)q(s)ds \leq B_\phi \textit{IPM}_G(p^\phi, q^\phi)$$

Proof.

$$\begin{aligned}
\int_{\mathcal{S}} f(s)p(s)ds - \int_{\mathcal{S}} f(s)q(s)ds &= \int_{\mathcal{S}} f(s)(p(s) - q(s))ds && (22)\\
&= \int_{\mathcal{Z}} f(\psi(z))(p(\psi(z)) - q(\psi(z)))J_\psi(z)dz && (23)\\
&= \int_{\mathcal{Z}} f(\psi(z))(p^\phi(z) - q^\phi(z))dz && (24)\\
&= B_\phi \int_{\mathcal{Z}} \frac{1}{B_\phi} f(\psi(z))(p^\phi(z) - q^\phi(z))dz && (25)\\
&\leq B_\phi \left| \int_{\mathcal{Z}} \frac{1}{B_\phi} f(\psi(z))(p^\phi(z) - q^\phi(z))dz \right| && (26)\\
&\leq B_\phi \sup_{g \in G} \left| \int_{\mathcal{Z}} g(z)(p^\phi(z) - q^\phi(z))dz \right| && (27)\\
&= B_\phi \text{IPM}_G(p^\phi, q^\phi) && (28)
\end{aligned}$$

□

The following lemma recursively bounds $\epsilon_V(\widehat{M}, H - t)$ by $\epsilon_V(\widehat{M}, H - t - 1)$, whose result allows us to bound $\text{MSE}_\pi = \epsilon_V(\widehat{M}, H)$.

The main idea to prove this is using Equation 20 from simulation lemma to decompose the loss of value functions into a one step reward loss, a transition loss and a next step value loss, with respect to the on-policy distribution. We can treat this as a contextual bandit problem, with the right side of Equation 20 as the loss function. For the distribution mismatch term, we follow the method in Shalit et al.'s work [19] about binary action bandits to bound the distribution mismatch by a representation distance penalty term. By converting the next step value error in the right side of Equation 20 into $\epsilon_V(\widehat{M}, H - t - 1)$, we can repeat this process recursively to bound the value error for H steps.

Lemma 4. *For any MDP M, approximate MDP model $\widehat{M}$, behavior policy μ and deterministic evaluation policy π, let $B_{\phi,t}$ and G_t be a scalar and a function family such that:*

$$\begin{aligned}
&\frac{1}{B_{\phi,t}} \Big[\bar{\ell}_r(\psi(z_t), \pi(\psi(z_t)), \widehat{M}) + \bar{\ell}_T(\psi(z), \pi(\psi(z_t)), \widehat{M}) \\
&\qquad + \frac{1}{2(H - t - 1)} \int_{\mathcal{Z}} h_T(z'|\psi(z_t), \pi(\psi(z_t))) \bar{\ell}_V(\psi(z'), \widehat{M}, H - t - 1)dz' \Big] \in G_t \qquad (29)
\end{aligned}$$

Then for any $t \le H-1$:

$$\epsilon_V(\widehat{M}, H-t) \le (H-t)\left[2\int_{\mathcal{S}}\left[\bar{\ell}_r(s_t,\pi(s_t),\widehat{M}) + \bar{\ell}_T(s_t,\pi(s_t),\widehat{M})\right] p_{M,\mu}(s_t|a_{0:t}=\pi)ds_t \right.$$
$$\left. + \frac{\epsilon_V(\widehat{M}, H-t-1)}{H-t-1} + 2B_{\phi,t}\mathit{IPM}_{G_t}\left(p^{\phi}_{M,\mu}(z_t|a_{0:t}=\pi), p^{\phi}_{M,\mu}(z_t|a_t \neq \pi, a_{0:t-1}=\pi)\right)\right]$$

Proof. Note the recursive form in Lemma 1. We could treat RL as dealing with a contextual bandit problem at each step. Here we view the right side of recursive result in simulation lemma (restated here)

$$\begin{aligned} V^{\pi}_{\widehat{M},H-t}(s_t) - V^{\pi}_{M,H-t}(s_t) = [&\widehat{r}(s_t,\pi(s_t)) - \bar{r}(s_t,\pi(s_t)) \\ &+ \int_{\mathcal{S}}\left(\widehat{T}(s'|s_t,\pi(s_t)) - T(s'|s_t,\pi(s_t))\right) V^{\pi}_{\widehat{M},H-t-1}(s')ds' \\ &+ \int_{\mathcal{S}} T(s'|s_t,\pi(s_t))\left(V^{\pi}_{\widehat{M},H-t-1}(s_t) - V^{\pi}_{M,H-t-1}(s_t)\right)ds'\Big] \end{aligned} \quad (30)$$

as a kind of square loss for a one-step prediction problem, and we bound the whole loss by recursively bounding these one-step losses. The key here is to find the recursive form of this square loss.

Recall that the definition of $\bar{\ell}_r, \bar{\ell}_T$ We will apply Cauchy-Schwarz inequality to bound Equation 30. Note that if $X_n = X_{n-1} + a_n + b_n$. Then $X_n^2 = (\frac{X_{n-1}}{\sqrt{n-1}}\sqrt{n-1} + \sqrt{2}a_n\frac{1}{\sqrt{2}} + \sqrt{2}b_n\frac{1}{\sqrt{2}})^2 \le (\frac{X_{n-1}^2}{n-1} + 2a_n^2 + 2b_n^2)n$. By applying this to Equation 30, we have that

$$\begin{aligned} \epsilon_V(\widehat{M}, H-t) &= \int_{\mathcal{S}} \bar{\ell}_V(s_t, H-t) p_{M,\mu}(s_t|a_{0:t-1}=\pi)ds_t & (31)\\ &\le (H-t)\int_{\mathcal{S}}\Big[2\bar{\ell}_r(s_t,\pi(s_t),\widehat{M}) + 2\bar{\ell}_T(s_t,\pi(s_t),\widehat{M}) & (32)\\ &\quad + \frac{1}{H-t-1}\int_{\mathcal{S}} T(s'|s_t,\pi(s_t))\bar{\ell}_V(s,\widehat{M},H-t-1)ds'\Big] p_{M,\mu}(s_t|a_{0:t-1}=\pi)ds_t \end{aligned}$$

Note that:

$$\begin{aligned} p_{M,\mu}(s_t|a_{0:t-1}=\pi) &= p_{M,\mu}(s_t, a_t=\pi|a_{0:t-1}=\pi) + p_{M,\mu}(s_t, a_t\neq\pi|a_{0:t-1}=\pi) & (33)\\ &= p_{M,\mu}(s_t|a_{0:t}=\pi)p(a_t=\pi|a_{0:t-1}=\pi) & (34)\\ &\quad + p_{M,\mu}(s_t|a_t\neq\pi, a_{0:t-1}=\pi)p(a_t\neq\pi|a_{0:t-1}=\pi) & (35) \end{aligned}$$

Let $c_t = p(a_t=\pi|a_{0:t-1}=\pi)$ then $1-c_t = p(a_t\neq\pi|a_{0:t-1}=\pi)$. Then

$$\begin{aligned} &\epsilon_V(\widehat{M}, H-t) \\ \le\ & c_t(H-t)\int_{\mathcal{S}}\Big[2\bar{\ell}_r(s_t,\pi(s_t),\widehat{M}) + 2\bar{\ell}_T(s_t,\pi(s_t),\widehat{M}) \\ & + \frac{1}{H-t-1}\int_{\mathcal{S}} T(s'|s_t,\pi(s_t))\bar{\ell}_V(s,\widehat{M},H-t-1)ds'\Big] p_{M,\mu}(s_t|a_{0:t}=\pi)ds_t \\ & + (1-c_t)(H-t)\int_{\mathcal{S}}\Big[2\bar{\ell}_r(s_t,\pi(s_t),\widehat{M}) + 2\bar{\ell}_T(s_t,\pi(s_t),\widehat{M}) + \frac{1}{H-t-1} \\ & \int_{\mathcal{S}} T(s'|s_t,\pi(s_t))\bar{\ell}_V(s,\widehat{M},H-t-1)ds'\Big] p_{M,\mu}(s_t|a_t\neq\pi, a_{0:t-1}=\pi)ds_t \end{aligned} \quad (36)$$

Let

$$f(s_t) = 2\bar{\ell}_r(s_t,\pi(s_t),\widehat{M}) + 2\bar{\ell}_T(s_t,\pi(s_t),\widehat{M}) + \frac{1}{H-t-1}\int_{\mathcal{S}} T(s'|s_t,\pi(s_t))\bar{\ell}_V(s,\widehat{M},H-t-1)ds'$$

We are going use Lemma 3 to bound the difference between $\int_S f(s_t)p_{M,\mu}(s_t|a_{0:t}=\pi)ds_t$ and $\int_S f(s_t)p_{M,\mu}(s_t|a_t\neq\pi, a_{0:t-1}=\pi)ds_t$. Let G_t and $B_{\phi,t} > 0$ be a function class and a constant satisfying that:

$$\frac{1}{B_{\phi,t}}\Big[\bar{\ell}_r(\psi(z_t),\pi(\psi(z_t)),\widehat{M}) + \bar{\ell}_T(\psi(z),\pi(\psi(z_t)),\widehat{M}) + \frac{1}{2(H-t-1)}\int_{\mathcal{Z}} h_T(z'|\psi(z_t),\pi(\psi(z_t)))\bar{\ell}_V(\psi(z'),\widehat{M},H-t-1)dz'\Big] \in G_t \tag{37}$$

Then following Lemma 3 we have that:

$$\begin{aligned}
&\int_S f(s_t)p_{M,\mu}(s_t|a_t \neq \pi, a_{0:t-1}=\pi)ds_t \\
\leq &\int_S f(s_t)p_{M,\mu}(s_t|a_{0:t}=\pi)ds_t + 2B_{\phi,t}\text{IPM}_{G_t}\left(p^{\phi}_{M,\mu}(z_t|a_{0:t}=\pi), p^{\phi}_{M,\mu}(z_t|a_t\neq\pi, a_{0:t-1}=\pi)\right) \\
&= \int_S f(s_t)p_{M,\mu}(s_t|a_{0:t}=\pi)ds_t + 2B_{\phi,t}\text{IPM}_{G_t}\left(p^{\phi,F}_{M,\mu}(z_t), p^{\phi,CF}_{M,\mu}(z_t)\right) \qquad (38)
\end{aligned}$$

Substituting this into Equation 36 we have that

$$\begin{aligned}
&\epsilon_V(\widehat{M},H-t) \\
\leq\ & (H-t)\int_S f(s_t)p_{M,\mu}(s_t|a_{0:t}=\pi)ds_t + 2(1-c_t)(H-t)B_{\phi,t}\text{IPM}_{G_t}\left(p^{\phi,F}_{M,\mu}(z_t), p^{\phi,CF}_{M,\mu}(z_t)\right) \qquad (39) \\
\leq\ & (H-t)\int_{\mathcal{S}}\left[2\bar{\ell}_r(s_t,\pi(s_t),\widehat{M}) + 2\bar{\ell}_T(s_t,\pi(s_t),\widehat{M}) + \frac{\int_{\mathcal{S}}\widehat{T}(s'|s_t,\pi(s_t))\bar{\ell}_V(s,\widehat{M},H-t-1)ds'}{H-t-1}\right] \\
& p_{M,\mu}(s_t|a_{0:t}=\pi)ds_t + 2(1-c_t)(H-t)B_{\phi,t}\text{IPM}_{G_t}\left(p^{\phi,F}_{M,\mu}(z_t), p^{\phi,CF}_{M,\mu}(z_t)\right) \qquad (40) \\
\leq\ & 2(H-t)\int_{\mathcal{S}}\left[\bar{\ell}_r(s_t,\pi(s_t),\widehat{M}) + \bar{\ell}_T(s_t,\pi(s_t),\widehat{M})\right]p_{M,\mu}(s_t|a_{0:t}=\pi)ds_t \\
& + \frac{H-t}{H-t-1}\epsilon_V(\widehat{M},H-t-1) + 2(H-t)B_{\phi,t}\text{IPM}_{G_t}\left(p^{\phi,F}_{M,\mu}(z_t), p^{\phi,CF}_{M,\mu}(z_t)\right) \qquad (41)
\end{aligned}$$

Thus we finish the proof. □

Iteratively applying this result for $t = 0, 1, \ldots, H$ we will have Theorem 1.

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

$$\mathbb{E}_{s_0}\left[V^{\pi}_{\widehat{M}}(s_0) - V^{\pi}_{M}(s_0)\right]^2 \leq 2H\sum_{t=0}^{H-1}\Big[B_{\phi,t}\textit{IPM}_{G_t}\left(p^{\phi,F}_{M,\mu}(z_t), p^{\phi,CF}_{M,\mu}(z_t)\right) + \int_{\mathcal{S}}\left(\bar{\ell}_r(s_t,\pi(s_t),\widehat{M}) + \bar{\ell}_T(s_t,\pi(s_t),\widehat{M})\right)p_{M,\mu}(s_t|a_{0:t}=\pi)ds_t\Big]$$

For MSE_μ, we can apply the simulation lemma to bound it by the reward and transition losses since the data distribution matches the policy μ. Then we combine it with the theorem above. Note that $\text{MSE}_\pi \leq \text{MSE}_\pi + \text{MSE}_\mu$.

Corollary 1. *For any MDP M, approximate MDP model $\widehat{M}$, behavior policy μ and deterministic evaluation policy π, let $B_{\phi,t}$ and G_t be a real number and function family that satisfies the condition in Lemma 4. Let $u_{0:t} = p_{\mu,M}(a_{0:t}=\pi)$. Then:*

$$\begin{aligned}
&\textit{MSE}_\pi \leq \textit{MSE}_\pi + \textit{MSE}_\mu \\
&\leq 2H\sum_{t=0}^{H-1}\Big[\frac{1}{u_{0:t}}\int_{\mathcal{S}}\left(\bar{\ell}_r(s_t,\pi(s_t),\widehat{M}) + \bar{\ell}_T(s_t,\pi(s_t),\widehat{M})\right)p_{M,\mu}(s_t,a_{0:t}=\pi)ds_t + \\
&\qquad \int_{\mathcal{S}}\sum_{a_t\in\mathcal{A}}\left(\bar{\ell}_r(s_t,a_t,\widehat{M}) + \bar{\ell}_T(s_t,a_t,\widehat{M})\right)p_{\mu,M}(s_t,a_t)ds_t \\
&\qquad\qquad + B_{\phi,t}\textit{IPM}_{G_t}\left(p^{\phi,F}_{M,\mu}(z_t), p^{\phi,CF}_{M,\mu}(z_t)\right)\Big]
\end{aligned}$$

Proof. According to Lemma 1, the mean square error of estimating the behavior policy value could be written as:

$$\begin{align*}
& \left[\mathbb{E}_{s_0} V^{\mu}_{\widehat{M}}(s_0) - \mathbb{E}_{s_0} V^{\mu}_{M}(s_0)\right]^2 \\
= \ & \left[\mathbb{E}_{\mu,M}\left(\sum_{t=0}^{H-1} \widehat{r}(s_t,a_t) - \bar{r}(s_t,a_t) + \int_{\mathcal{S}} \left(\widehat{T}(s'|s_t,a_t) - T(s'|s_t,a_t)\right) V^{\pi}_{\widehat{M},H-t-1}(s')ds' \mid s_0\right)\right]^2 \\
\leq \ & \mathbb{E}_{\mu,M}\left[\left(\sum_{t=0}^{H-1} \widehat{r}(s_t,a_t) - \bar{r}(s_t,a_t) + \int_{\mathcal{S}} \left(\widehat{T}(s'|s_t,a_t) - T(s'|s_t,a_t)\right) V^{\pi}_{\widehat{M},H-t-1}(s')ds'\right)^2\right] \tag{42} \\
\leq \ & \mathbb{E}_{\mu,M}\left[2H\sum_{t=0}^{H-1} \bar{\ell}_r(s_t,a_t,\widehat{M}) + \bar{\ell}_T(s_t,a_t,\widehat{M})\right] \tag{43} \\
= \ & 2H\sum_{t=0}^{H-1}\int_{\mathcal{S}}\sum_{a_t}\left(\bar{\ell}_r(s_t,a_t,\widehat{M}) + \bar{\ell}_T(s_t,a_t,\widehat{M})\right) p_{\mu,M}(s_t,a_t)ds_t \tag{44}
\end{align*}$$

The first step follows from Lemma 1. The second step follows from Jensen's inequality, and the third step follows from Cauchy-Schwarz inequality. By combining the results above with Theorem 1, we have that:

$$\begin{align*}
& \mathbb{E}_{s_0}\left[V^{\pi}_{\widehat{M}}(s_0) - V^{\pi}_{M}(s_0)\right]^2 \\
\leq \ & \mathbb{E}_{s_0}\left[V^{\pi}_{\widehat{M}}(s_0) - V^{\pi}_{M}(s_0)\right]^2 + \mathbb{E}_{s_0}\left[V^{\mu}_{\widehat{M}}(s_0) - V^{\mu}_{M}(s_0)\right]^2 \tag{45} \\
\leq \ & 2H\sum_{t=0}^{H-1}\left[\int_{\mathcal{S}}\left(\bar{\ell}_r(s_t,\pi(s_t),\widehat{M}) + \bar{\ell}_T(s_t,\pi(s_t),\widehat{M})\right) p_{M,\mu}(s_t|a_{0:t}=\pi)ds_t\right. \\
& + \int_{\mathcal{S}}\sum_{a_t\in\mathcal{A}}\left(\bar{\ell}_r(s_t,a_t,\widehat{M}) + \bar{\ell}_T(s_t,a_t,\widehat{M})\right) p_{\mu,M}(s_t,a_t)ds_t \\
& \left. + B_{\phi,t}\text{IPM}_{G_t}\left(p^{\phi,F}_{M,\mu}(z_t), p^{\phi,CF}_{M,\mu}(z_t)\right)\right] \tag{46} \\
= \ & 2H\sum_{t=0}^{H-1}\left[\frac{1}{u_{0:t}}\int_{\mathcal{S}}\left(\bar{\ell}_r(s_t,\pi(s_t),\widehat{M}) + \bar{\ell}_T(s_t,\pi(s_t),\widehat{M})\right) p_{M,\mu}(s_t,a_{0:t}=\pi)ds_t\right. \\
& \left. + \int_{\mathcal{S}}\sum_{a_t\in\mathcal{A}}\left(\bar{\ell}_r(s_t,a_t,\widehat{M}) + \bar{\ell}_T(s_t,a_t,\widehat{M})\right) p_{\mu,M}(s_t,a_t)ds_t\right] \\
& + 2H\sum_{t=0}^{H-1} B_{\phi,t}\text{IPM}_{G_t}\left(p^{\phi,F}_{M,\mu}(z_t), p^{\phi,CF}_{M,\mu}(z_t)\right) \tag{47}
\end{align*}$$

□

C.2 Proof of Theorem 2

We showed in Theorem 1 that we can bound MSE by expected losses under the behavior policy distribution and an IPM term. In this section, we are going to further bound this by empirical losses and a generalization gap. We will firstly define some loss terms that are based on observations, instead of losses that are on expected values, $\bar{r}$ and $T(\cdot|s,a)$. Then we will introduce some lemmas that allow us to bound the generalization gap of weighted losses and IPM terms from previous works. Finally we will prove the finite sample MSE bound by putting these generalization gaps together.

Definition 4. *(Restated) 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 function as:*

$$\begin{aligned}
\ell_r(s_t, a_t, r_t, \widehat{M}) &= \left(\widehat{r}(s_t, a_t) - r_t\right)^2 \\
\ell_T(s_t, a_t, s'_t, \widehat{M}) &= \left(\int_{\mathcal{S}} \left(\widehat{T}(s'|s_t, a_t) - \delta(s' - s'_t)\right) V^{\pi}_{\widehat{M},H-t-1}(s')ds'\right)^2 \\
&= \left(\int_{\mathcal{S}} \widehat{T}(s'|s_t, a_t) V^{\pi}_{\widehat{M},H-t-1}(s')ds' - V^{\pi}_{\widehat{M},H-t-1}(s'_t)\right)^2
\end{aligned}$$

where δ is the Dirac delta function.

Actually the difference between ℓ and $\bar{\ell}$ can be captured by the variance of the reward and transition dynamics, which only depend on the underlying dynamics. The following definition and lemmas show that.

Definition 6. *Define the variance of t-th step reward and transition with respect to the state-action distribution $q(s_t, a_t)$ as:*

$$\sigma_{q,t} = \sigma_q(r) + \sigma_q(T) \qquad \sigma_{q,t}(r) = \int_{\mathcal{S}} \sum_{a_t} \int_{\mathcal{R}} (r - \bar{r}(s_t, a_t))^2 p_M(r|s_t, a_t) q(s_t, a_t) dr ds_t$$

$$\sigma_{q,t}(T) = \int_{\mathcal{S}} \sum_{a_t} \int_{\mathcal{S}} \left(\int_{\mathcal{S}} T(s'|s_t, a_t) V^{\pi}_{\widehat{M},H-t-1}(s')ds' - V^{\pi}_{\widehat{M},H-t-1}(s'_t)\right)^2 p_M(s'_t|s_t, a_t) q(s_t, a_t) ds'_t ds_t$$

Lemma 5. *(Variance decomposition)*

$$\begin{aligned}
\int_{\mathcal{S}} \sum_{a_t} \left(\bar{\ell}_r(s_t, a_t, \widehat{M}) + \bar{\ell}_T(s_t, a_t, \widehat{M})\right) q(s_t, a_t) ds_t = \int_{\mathcal{S}} \sum_{a_t} \bigg(\int_{\mathcal{R}} \ell_r(s_t, a_t, r, \widehat{M}) p(r|s_t, a_t) dr \\
+ \int_{\mathcal{S}} \ell_T(s_t, a_t, s'_t, \widehat{M}) p(s'_t|s_t, a_t, s'_t) ds'_t\bigg) q(s_t, a_t) ds_t - \sigma_{q,t}
\end{aligned}$$

Proof. Let's start with the ℓ_r and ℓ_T terms:

$$\begin{aligned}
\ell_r(s_t, a_t, r_t, \widehat{M}) = \left(\widehat{r}(s_t, a_t) - r_t\right)^2 = \left(\widehat{r}(s_t, a_t) - \bar{r}(s_t, a_t)\right)^2 + \left(\bar{r}(s_t, a_t) - r_t\right)^2 \\
+ 2\left(\widehat{r}(s_t, a_t) - \bar{r}(s_t, a_t)\right)\left(\bar{r}(s_t, a_t) - r_t\right)
\end{aligned} \tag{48}$$

Note that $\mathbb{E}\left[\bar{r}(s_t, a_t) - r_t\right] = 0$ so the last term will be zero after we apply the integral. Then:

$$\int_{\mathcal{S}} \sum_{a_t} \int_{\mathcal{R}} \ell_r(s_t, a_t, r, \widehat{M}) p(r|s_t, a_t) q(s_t, a_t) dr ds_t = \int_{\mathcal{S}} \sum_{a_t} \bar{\ell}_r(s_t, a_t, \widehat{M}) q(s_t, a_t) ds_t + \sigma_{q,t}(r)$$

Similarly, for ℓ_T we have that:

$$\begin{aligned}
& \ell_T(s_t, a_t, s'_t, \widehat{M}) && (49) \\
= & \left(\int_{\mathcal{S}} \widehat{T}(s'|s_t, a_t) V^{\pi}_{\widehat{M},H-t-1}(s')ds' - V^{\pi}_{\widehat{M},H-t-1}(s'_t)\right)^2 && (50) \\
= & \left(\mathbb{E}_{s'\sim\widehat{T}} V^{\pi}_{\widehat{M},H-t-1}(s') - V^{\pi}_{\widehat{M},H-t-1}(s'_t)\right)^2 && (51) \\
= & \left(\mathbb{E}_{\widehat{T}} V^{\pi}_{\widehat{M},H-t-1}(s') - \mathbb{E}_T V^{\pi}_{\widehat{M},H-t-1}(s')\right)^2 + \left(\mathbb{E}_T V^{\pi}_{\widehat{M},H-t-1}(s') - V^{\pi}_{\widehat{M},H-t-1}(s'_t)\right)^2 \\
& +2\left(\mathbb{E}_{\widehat{T}} V^{\pi}_{\widehat{M},H-t-1}(s') - \mathbb{E}_T V^{\pi}_{\widehat{M},H-t-1}(s')\right)\left(\mathbb{E}_T V^{\pi}_{\widehat{M},H-t-1}(s') - V^{\pi}_{\widehat{M},H-t-1}(s'_t)\right) && (52)
\end{aligned}$$

Note that

$$\mathbb{E}_{s'_t\sim T}\left[\mathbb{E}_{s'\sim T} V^{\pi}_{\widehat{M},H-t-1}(s') - V^{\pi}_{\widehat{M},H-t-1}(s'_t)\right] = \mathbb{E}_{s'\sim T} V^{\pi}_{\widehat{M},H-t-1}(s') - \mathbb{E}_{s'_t\sim T} V^{\pi}_{\widehat{M},H-t-1}(s'_t) = 0$$

So the last term here will also be zero when we apply integral over s'_t to $\ell_T(s_t, a_t, s'_t, \widehat{M})$ and we have that:

$$\int_{\mathcal{S}} \sum_{a_t} \int_{\mathcal{S}} \ell_T(s_t, a_t, s'_t, \widehat{M}) p(s'_t|s_t, a_t, s'_t) q(s_t, a_t) ds'_t ds_t$$

$$= \int_{\mathcal{S}} \sum_{a_t} \bar{\ell}_T(s_t, a_t, \widehat{M}) q(s_t, a_t) ds_t + \sigma_{q,t}(T) \quad (53)$$

Thus we finished the proof by combining the ℓ_r part with ℓ_T part. □

Now we are going to bound the expected value of ℓ_r and ℓ_T terms by the empirical mean of it. We restate our definition about empirical risk and add the definition about corresponding generalization risk:

Definition 5. *(Restate)*

$$\begin{aligned} R_\mu(\widehat{M}) &= \sum_{t=0}^{H-1} \int_{\mathcal{S}} \sum_{a_t} \left(\int_{\mathcal{R}} \ell_r(s_t, a_t, r, \widehat{M}) p(r|s_t, a_t) dr \right. \\ &\quad \left. + \int_{\mathcal{S}} \ell_T(s_t, a_t, \widehat{M}) p(s'_t|s_t, a_t) ds'_t \right) p_{\mu,M}(s_t, a_t) ds_t \\ \widehat{R}_\mu(\widehat{M}) &= \sum_{t=0}^{H-1} \int_{\mathcal{S}} \sum_{a_t} \left(\int_{\mathcal{R}} \ell_r(s_t, a_t, r, \widehat{M}) p(r|s_t, a_t) dr \right. \\ &\quad \left. + \int_{\mathcal{S}} \ell_T(s_t, a_t, \widehat{M}) p(s'_t|s_t, a_t) ds'_t \right) \widehat{p}_{\mu,M}(s_t, a_t) ds_t \\ &= \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{H-1} \ell_r(s_t^{(i)}, a_t^{(i)}, r^{(i)}, \widehat{M}) + \ell_T(s_t^{(i)}, a_t^{(i)}, s_t^{\prime(i)}, \widehat{M}), \end{aligned}$$

where n is the number of trajectories and $s_t^{(i)}$ is the state of the tth step in the ith trajectory. Similarly we define $R_{\pi,u}$ and $\widehat{R}_{\pi,u}$:

$$\begin{aligned} R_{\pi,u}(\widehat{M}) &= \sum_{t=0}^{H-1} \int_{\mathcal{S}} \sum_{a_t} \frac{\mathbb{1}(a_{0:t} = \pi)}{u_{0:t}} \left(\int_{\mathcal{R}} \ell_r(s_t, a_t, r, \widehat{M}) p(r|s_t, a_t) dr \right. \\ &\quad \left. + \int_{\mathcal{S}} \ell_T(s_t, a_t, \widehat{M}) p(s'_t|s_t, a_t) ds'_t \right) p_{M,\mu}(s_t, a_t) ds_t \\ \widehat{R}_{\pi,u}(\widehat{M}) &= \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{H-1} \frac{\mathbb{1}(a_{0:t}^{(i)} = \pi)}{\widehat{u}_{0:t}} \left[\ell_r(s_t^{(i)}, a_t^{(i)}, r^{(i)}, \widehat{M}) + \ell_T(s_t^{(i)}, a_t^{(i)}, s_t^{\prime(i)}, \widehat{M}) \right], \end{aligned}$$

where $\widehat{u}_{0:t} = \sum_{i=1}^{n} \frac{\mathbb{1}(a_{0:t}^{(i)} = \pi)}{n}$

We could bound R_μ and $R_{\pi,u}$ and pseudo-dimension which is a complexity term of the model class. We will use the learning bound about importance sampling and weighted importance sampling in Cortes et al. [4] to bound R_μ and $R_{\pi,u}$. The following lemma is an immediate consequence of Corollary 2 and section 6 in Cortes et al. [4].

Lemma 6. *For a hypothesis class $\mathcal{H}$ over input space $\mathcal{X}$, let d be the pseudo-dimension of a real valued loss function class $\{\ell_h(x), h \in \mathcal{H}, x \in \mathcal{X}\}$. $w(x)$ is a weighting function such that $\mathbb{E}_p[w(x)] = 1$. Let $\widehat{p}$ be the empirical distribution over n samples, and $\widehat{w}(x_i) = nw(x_i)/\sum_{i=1}^{n} w(x_i)$ is the normalized weights. For any ℓ in the loss function class. with probability $1 - \delta$*

$$|\mathbb{E}_p[w(x)\ell(x)] - \mathbb{E}_{\widehat{p}}[\widehat{w}(x)\ell(x)]| \leq \mathbb{V}_{p,\widehat{p}}[w, \ell] \frac{\mathcal{C}_{n,\delta}^{\mathcal{M}}}{n^{3/8}} + \ell_{\max} \mathbb{V}_{p,\widehat{p}}[w, 1] \frac{\mathcal{C}_{n,\delta}^{\mathcal{M}}}{n^{3/8}} \quad (54)$$

where $\mathcal{C}_{n,\delta}^{\mathcal{H}} = 2^{5/4} \left(d \log(2ne/d) + \log(4/\delta) \right)^{3/8}$, $\mathbb{V}_{p,\widehat{p}}[w, \ell] = \max\{\sqrt{\mathbb{E}[w(x)^2 \ell(x)^2]}, \sqrt{\widehat{\mathbb{E}}[w(x)^2 \ell(x)^2]}\}$*, and* $\ell_{\max} = \max_x |\ell(x)|$.

Proof. We can decompose the gap into two parts by adding $\mathbb{E}_{\widehat{p}}[w(x)\ell(x)]$:

$$\begin{aligned}&|\mathbb{E}_p[w(x)\ell(x)] - \mathbb{E}_{\widehat{p}}[\widehat{w}(x)\ell(x)]| \\ &\qquad \leq |\mathbb{E}_p[w(x)\ell(x)] - \mathbb{E}_{\widehat{p}}[w(x)\ell(x)]| + |\mathbb{E}_{\widehat{p}}[w(x)\ell(x)] - \mathbb{E}_{\widehat{p}}[\widehat{w}(x)\ell(x)]| \end{aligned} \tag{55}$$

For the first part, we can bound it by Corollary 2 from Cortes et al. [4]:

$$|\mathbb{E}_p[w(x)\ell(x)] - \mathbb{E}_{\widehat{p}}[w(x)\ell(x)]| \quad \leq \quad \mathbb{V}_{p,\widehat{p}}[w, \ell] \frac{\mathcal{C}^{\mathcal{M}}_{n,\delta}}{n^{3/8}} \tag{56}$$

For the second part, according to section 6 from Cortes et al. [4], we have that [2]:

$$\left| \frac{1}{n} \left(\widehat{w}(x_i) - w(x_i) \right) \right| = \frac{w(x_i)}{W} \left| 1 - \frac{W}{n} \right| \leq \frac{w(x_i)}{W} \mathbb{V}_{p,\widehat{p}}[w, 1] \frac{\mathcal{C}^{\mathcal{M}}_{n,\delta}}{n^{3/8}} \tag{57}$$

where $W = \sum_i w(x_i)$. Then:

$$\begin{aligned} |\mathbb{E}_{\widehat{p}}[(w(x) - \widehat{w}(x))\ell(x)]| \quad &\leq \quad \ell_{\max} \mathbb{E}_p[|w(x) - \widehat{w}(x)|] && (58) \\ &= \quad \ell_{\max} \left[\frac{1}{n} \sum_i |w(x_i) - \widehat{w}(x_i)| \right] && (59) \\ &\leq \quad \ell_{\max} \mathbb{V}_{p,\widehat{p}}[w, 1] \frac{\mathcal{C}^{\mathcal{M}}_{n,\delta}}{n^{3/8}} && (60) \end{aligned}$$

Thus we finished the proof. □

We will apply this lemma to the risk at each time step t separately. Since $\mathbb{E}_{M,\mu}[\frac{\mathbb{1}(a^{(i)}_{0:t}=\pi)}{u_{0:t}}] = 1$ for each t, we can let $w(x) = \frac{\mathbb{1}(a^{(i)}_{0:t}=\pi)}{u_{0:t}}$. In that case $\frac{\mathbb{1}(a^{(i)}_{0:t}=\pi)}{\widehat{u}_{0:t}}$ is the normalized weights $\widehat{w}(x)$. We can also bound of R_μ from this as well, by setting the weight function to be one. In that case $w = \widehat{w} = 1$ and $\mathbb{V}_{p,\widehat{p}}[1, \ell] \leq \ell_{\max} \leq R^2_{\max} + V^2_{\max,t}$ for the tth step loss function ℓ.

For the IPM term, using norm-1 reproducing kernel Hilbert space (RKHS) function class for G leads to IPM being the maximum mean discrepancy (MMD) distance. We can bound the gap between MMD distance and its empirical estimation using the following lemma in Sriperumbudur et al.'s work [20]. There are many other choice such as of 1-Lipschitz functions, leading to Wasserstein distance, and l_∞ norm unit ball, leading to total variation distance. There are similar results with those function class and distance measure, with worse bounds. We also use norm-1 RKHS functions and MMD metric in our experiment section.

Lemma 7. *(Theorem 11 from Sriperumbudur et al. [20]) Let $\mathcal{X}$ be a measurable space. Suppose k is measurable kernel such that $\sup_{x\in\mathcal{X}} k(x, x) \leq C \leq \infty$ and $\mathcal{H}$ the reproducing kernel Hilbert space induced by k. Let $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} = 1\}$, and $\nu = \sup_{x\in\mathcal{X}, f\in\mathcal{F}} |f(x)| < \infty$. Then, with $\widehat{p}$, $\widehat{q}$ the empirical distributions of p, q from m_1 and m_2 samples respectively, and with probability at least 1-δ,*

$$|IPM_{\mathcal{F}}(p, q) - IPM_{\mathcal{F}}(\widehat{p}, \widehat{q})| \leq \sqrt{18\nu^2 \ln(4/\delta) C} \left(\frac{1}{\sqrt{m_1}} + \frac{1}{\sqrt{m_2}} \right) \tag{61}$$

Theorem 2. *(Restated) Suppose $\mathcal{M}_\Phi$ is a model class of model MDP models based on twice-differentiable, invertible state representation ϕ's: $\widehat{M_\phi} = \langle \widehat{r}(s, a), \widehat{T}(s', s, a) \rangle = \langle h_r(\phi(s), a), h_T(\phi(s'), \phi(s), a) \rangle$. Given n H-step trajectories sampled from policy μ, let the loss function for (s_t, a_t) pair at t^th^ step be $\ell_t(s_t, a_t, \widehat{M_\phi}) = \ell_r(s_t, a_t, r_t, \widehat{M}) + \ell_T(s_t, a_t, s'_t, \widehat{M})$. Let 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 $\mathcal{H}$ the reproducing kernel Hilbert space induced by k such that $\sup_{z\in\mathcal{Z}} k(z, z) \leq C \leq \infty$, and $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} = 1\}$, and $\nu = \sup_{z\in\mathcal{X}, f\in\mathcal{F}} |f(z)| < \infty$. Assume there exist a constant $B_{\phi,t}$ such*

[2] Note that the definition of $\widehat{w}(x_i)$ in Cortes et al. [4] is different with ours by a constant n. Here $\widehat{w}(x)$ follows our definition.

that $\frac{1}{B_{\phi,t}}\ell_t(\psi(z),\pi(\psi(z)),\widehat{M}_\phi)\in\mathcal{F}$. *Then with probability* $1-3\delta$, *for any* $\widehat{M}\in\mathcal{M}_\Phi$*:*

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

$m_{t,1}$ *and* $m_{t,2}$ *are the number of samples that used to estimate* $\widehat{p}^{\phi,F}_{M,\mu}(z_t)$ *and* $\widehat{p}^{\phi,CF}_{M,\mu}(z_t)$ *respectively.* $\mathcal{D}^{\mathcal{F}}_\delta=\sqrt{18\nu^2\ln(4/\delta)C}$. $\mathcal{C}^{\mathcal{M}}_{n,\delta,t}=2^{5/4}\left(d_t\log(2ne/d_t)+\log(4/\delta)\right)^{3/8}$. $\mathbb{V}_{p.\hat{p}}[w,\ell_t]=\max\{\sqrt{\mathbb{E}_{p_{M,\mu}}[w(s_t,a_t)^2\ell_t(s_t,a_t)^2]},\sqrt{\mathbb{E}_{\hat{p}_{M,\mu}}[w(s_t,a_t)^2\ell_t(s_t,a_t)^2]}\}$. $\ell_{t,\max}=\max_{s_t,a_t}|\ell_t(s_t,a_t)|\le R^2_{\max}+V^2_{\max,t}$.

Proof. Applying Lemma 5 to the result in Corollary 1 and plugging the definition of R_μ and $R_{\pi,u}$ in, we have that:

$$\begin{aligned}&\mathbb{E}_{s_0}\left[V^\pi_{\widehat{M}}(s_0)-V^\pi_M(s_0)\right]^2\\ =\;&2H\left(\sum_{t=0}^{H-1}B_{\phi,t}\mathrm{IPM}_{\mathcal{F}}\left(p^{\phi,F}_{M,\mu}(z_t),p^{\phi,CF}_{M,\mu}(z_t)\right)+R_\mu(\widehat{M})+R_{\pi,u}(\widehat{M})-\sigma\right) &(62)\\ \le\;&2H\left(\sum_{t=0}^{H-1}B_{\phi,t}\mathrm{IPM}_{\mathcal{F}}\left(p^{\phi,F}_{M,\mu}(z_t),p^{\phi,CF}_{M,\mu}(z_t)\right)+R_\mu(\widehat{M})+R_{\pi,u}(\widehat{M})\right) &(63)\end{aligned}$$

where σ is $\sum_{t=0}^{H-1}\sigma_{p_{M,\mu},t}+\sigma_{p_{M,\mu}(\cdot|a_{0:t}=\pi),t}\ge 0$. We will work term by term. First, we can use Lemma 7 for the IPM term:

$$\mathrm{IPM}_{\mathcal{F}}\left(p^{\phi,F}_{M,\mu}(z_t),p^{\phi,CF}_{M,\mu}(z_t)\right)\le\mathrm{IPM}_{\mathcal{F}}\left(\widehat{p}^{\phi,F}_{M,\mu}(z_t),\widehat{p}^{\phi,CF}_{M,\mu}(z_t)\right)+\mathcal{D}^{\mathcal{F}}_\delta\left(\frac{1}{\sqrt{m_{t,1}}}+\frac{1}{\sqrt{m_{t,2}}}\right)\quad(64)$$

At the same time, we know that for any two distribution p,q, $0\le\mathrm{IPM}_{\mathcal{F}}(p,q)\le 2\nu$. So:

$$\begin{aligned}&\mathrm{IPM}_{\mathcal{F}}\left(p^{\phi,F}_{M,\mu}(z_t),p^{\phi,CF}_{M,\mu}(z_t)\right)\\ &\quad\le\mathrm{IPM}_{\mathcal{F}}\left(\widehat{p}^{\phi,F}_{M,\mu}(z_t),\widehat{p}^{\phi,CF}_{M,\mu}(z_t)\right)+\min\left\{\mathcal{D}^{\mathcal{F}}_\delta\left(\frac{1}{\sqrt{m_{t,1}}}+\frac{1}{\sqrt{m_{t,2}}}\right),2\nu\right\}\quad(65)\end{aligned}$$

For R_μ, if we plug $w(s,a)=\widehat{w}(s,a)=1$, $\ell=\ell_t(s_t,a_t,\widehat{M})$, and $p=p_{M,\mu}(s_t,a_t)$ into Lemma 6, we have that:

$$\begin{aligned}R_\mu=\sum_{t=0}^{H-1}\mathbb{E}_p[\ell_t(s_t,a_t)]\;&\le\;\sum_{t=0}^{H-1}\left(\mathbb{E}_{\widehat{p}}[\ell_t(s_t,a_t)]+\frac{\mathcal{C}^{\mathcal{M}}_{n,\delta,t}}{n^{3/8}}\mathbb{V}_{p,\widehat{p}}[1,\ell_t]\right) &(66)\\ &=\;\widehat{R}_\mu+\sum_{t=0}^{H-1}\frac{\mathcal{C}^{\mathcal{M}}_{n,\delta,t}}{n^{3/8}}\mathbb{V}_{p,\widehat{p}}[1,\ell_t] &(67)\end{aligned}$$

An analogous argument can be made for $R_{\pi,u}$ by letting $w(s_t,a_t)=\frac{\mathbb{1}(a_{0:t}=\pi)}{u_{0:t}}$ which leads to that $\widehat{w}(s_t,a_t)=\frac{\mathbb{1}(a_{0:t}=\pi)}{\widehat{u}_{0:t}}$:

$$\begin{aligned}R_{\pi,u}\;&=\;\sum_{t=0}^{H-1}\mathbb{E}_p[w(s_t,a_t)\ell_t(s_t,a_t)] &(68)\\ &\le\;\sum_{t=0}^{H-1}\left(\mathbb{E}_{\widehat{p}}[\widehat{w}(s_t,a_t)\ell_t(s_t,a_t)]+\mathbb{V}_{p,\widehat{p}}[\widehat{w},\ell_t]\frac{\mathcal{C}^{\mathcal{M}}_{n,\delta}}{n^{3/8}}+\ell_{t,\max}\mathbb{V}_{p,\widehat{p}}[w,1]\frac{\mathcal{C}^{\mathcal{M}}_{n,\delta}}{n^{3/8}}\right) &(69)\\ &=\;\widehat{R}_{\pi,u}+\sum_{t=0}^{H-1}\left(\mathbb{V}_{p,\widehat{p}}[\widehat{w},\ell_t]\frac{\mathcal{C}^{\mathcal{M}}_{n,\delta}}{n^{3/8}}+\ell_{t,\max}\mathbb{V}_{p,\widehat{p}}[w,1]\frac{\mathcal{C}^{\mathcal{M}}_{n,\delta}}{n^{3/8}}\right) &(70)\end{aligned}$$

Thus we finish the proof by combining IPM terms, R_μ and $R_{\pi,u}$ together. □

D Proof of Section 5

Corollary 2. *Let $\widehat{M}^*_{\phi^*} = \arg\min_{\widehat{M}_\phi} \mathcal{L}(\widehat{M}_\phi; \alpha_t)$ for a large enough α such that $\alpha_t > B_{\phi^*,t}$. Under the same definition and assumption in Theorem 2, we have that:*

$$\mathbb{E}_{s_0}\left[V^\pi_{\widehat{M}^*_{\phi^*}}(s_0) - V^\pi_M(s_0)\right]^2 \leq O\left(\frac{1}{n^{3/8}}\right) + \sum_{t=1}^{H-1} O\left(\frac{1}{\sqrt{m_{t,1}}} + \frac{1}{\sqrt{m_{t,2}}}\right)$$
$$+ 2H \min_{\widehat{M}_\phi \in \mathcal{M}} \left(R_\mu(\widehat{M}_\phi) + R_{\pi,u}(\widehat{M}_\phi) + \sum_{t=0}^{H-1} \alpha_t \mathit{IPM}_{\mathcal{F}}\left(p^{\phi,F}_{M,\mu}(z_t), p^{\phi,CF}_{M,\mu}(z_t)\right)\right)$$

Proof. Let $M^*_{\phi^*_0}$ be the model that minimizes the expected risk and IPM term:

$$M^*_{\phi^*_0} = \arg\min_{\widehat{M}_\phi} \left(R_\mu(\widehat{M}_\phi) + R_{\pi,u}(\widehat{M}_\phi) + \sum_{t=0}^{H-1} \alpha_t \mathrm{IPM}_{\mathcal{F}}\left(p^{\phi,F}_{M,\mu}(z_t), p^{\phi,CF}_{M,\mu}(z_t)\right)\right)$$

From theorem 2 we have that:

$$\begin{aligned}
&\mathbb{E}_{s_0}\left[V^\pi_{\widehat{M}^*_{\phi^*}}(s_0) - V^\pi_M(s_0)\right]^2 \\
\leq\ & O\left(\frac{1}{n^{3/8}}\right) + \sum_{t=1}^{H-1} O\left(\frac{1}{\sqrt{m_{t,1}}} + \frac{1}{\sqrt{m_{t,2}}}\right) + 2H\mathcal{L}(\widehat{M}^*_{\phi^*}; B_{\phi^*,t}) \qquad (71)\\
\leq\ & O\left(\frac{1}{n^{3/8}}\right) + \sum_{t=1}^{H-1} O\left(\frac{1}{\sqrt{m_{t,1}}} + \frac{1}{\sqrt{m_{t,2}}}\right) + 2H\mathcal{L}(\widehat{M}^*_{\phi^*}; \alpha_t) \qquad (72)\\
\leq\ & O\left(\frac{1}{n^{3/8}}\right) + \sum_{t=1}^{H-1} O\left(\frac{1}{\sqrt{m_{t,1}}} + \frac{1}{\sqrt{m_{t,2}}}\right) + 2H\mathcal{L}(M^*_{\phi^*_0}; \alpha_t) \qquad (73)
\end{aligned}$$

The first step follows from Theorem 2. The second step is from the fact that $\alpha_t > B_{\phi^*,t}$. The third step is from that $\widehat{M}^*_{\phi^*} = \arg\min_{\widehat{M}_\phi} \mathcal{L}(\widehat{M}_\phi; \alpha_t)$. Then we can bound the empirical loss term $\mathcal{L}(M^*_{\phi^*_0}; \alpha_t)$ by the expected value of it:

$$\begin{aligned}
\mathcal{L}(M^*_{\phi^*_0}; \alpha_t) \ =\ & \widehat{R}_\mu(M^*_{\phi^*_0}) + \widehat{R}_{\pi,u}(M^*_{\phi^*_0}) + \sum_{t=0}^{H-1} \alpha_t \mathrm{IPM}_{\mathcal{F}}\left(\widehat{p}^{\phi^*_0,F}_{M,\mu}(z_t), \widehat{p}^{\phi^*_0,CF}_{M,\mu}(z_t)\right)\\
& + \frac{\mathfrak{R}(\widehat{M}_\phi)}{n^{3/8}} \qquad (74)\\
\leq\ & R_\mu(M^*_{\phi^*_0}) + R_{\pi,u}(M^*_{\phi^*_0}) + \sum_{t=0}^{H-1} \alpha_t \mathrm{IPM}_{\mathcal{F}}\left(p^{\phi^*_0,F}_{M,\mu}(z_t), p^{\phi^*_0,CF}_{M,\mu}(z_t)\right)\\
& + O\left(\frac{1}{n^{3/8}}\right) + O\left(\frac{1}{\sqrt{m_{t,1}}} + \frac{1}{\sqrt{m_{t,2}}}\right) \qquad (75)
\end{aligned}$$

This follows from using Lemma 6 and Lemma 7 similarly with Equation 65, 67, 70 but in different direction, together with the fact that $\frac{\mathfrak{R}(\widehat{M}_\phi)}{n^{3/8}} = O\left(\frac{1}{n^{3/8}}\right)$.

Put this into equation 73, we have that

$$\mathbb{E}_{s_0}\left[V^\pi_{\widehat{M}^*_{\phi^*}}(s_0) - V^\pi_M(s_0)\right]^2 \leq 2H\left(R_\mu(M^*_{\phi^*_0}) + R_{\pi,u}(M^*_{\phi^*_0})\right.$$
$$\left. + \sum_{t=0}^{H-1} \alpha_t \mathrm{IPM}_{\mathcal{F}}\left(p^{\phi^*_0,F}_{M,\mu}(z_t), p^{\phi^*_0,CF}_{M,\mu}(z_t)\right)\right) + O\left(\frac{1}{n^{3/8}}\right) + \sum_{t=1}^{H-1} O\left(\frac{1}{\sqrt{m_{t,1}}} + \frac{1}{\sqrt{m_{t,2}}}\right) \qquad (76)$$

Thus we finished the proof. □

Under assumption about support set of μ, $m_{t,1}, m_{t,2} \to \infty$ when $n \to \infty$. Then an immediate consequence from this corollary is that, if there exists an MDP and representation model in our model class that could achieve no generalization error,

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

then $\lim_{n\to\infty} \mathbb{E}_{s_0} \left[V^\pi_{\widehat{M}^*_{\phi^*}}(s_0) - V^\pi_M(s_0) \right]^2 \to 0$ and estimator $V^\pi_{\widehat{M}^*_{\phi^*}}(s_0)$ is a consistent estimator for any s_0.

E Details of Experiment

E.1 Synthetic domain: Cart Pole

We will clarify the details of the Cart Pole experiment and provide results from additional OPPE methods.

Details of the domain We follow the same settings as in the OpenAI Gym [3] CartPole-v0 environment. The state consists of 4 features: position, speed, angle and angular speed. The agent can take two actions: move to the left or to the right. The trajectory will end either when the time step is larger than 200 or when the absolute value of position or angle is larger than the threshold. The goal in this domain is to control a cart as long as possible. We will receive the reward after each time step if the cart is under control, and the trajectory ends when the cart falls.

We include two different variants in this domain: long horizon and short horizon. For long horizon, we learn a near-optimal Q function, and use the greedy policy as evaluation policy and $\epsilon-$greedy policy with $\epsilon = 0.2$ as behavior policy. The average value, which is also the average length of trajectories, of the evaluation policy is 195 and the average value of the behavior policy is 190. For shorter horizon, we learn a weaker Q function and generate the policies in the same way, with the average value of 23.8 and 24 respectively. The reason that we learn a near optimal but not optimal policy for the long horizon is that the optimal policy can always hold the cart for 200 steps (max length), which makes it easy to estimate since there is no possibility of overestimating it.

Details of our model Our model has three parts: a representation module, a reward module, and a transition module. The representation module is a one layer feed-forward network that takes the state as input and outputs a 32-dimension representation. The reward module takes the representation as input and outputs $A = 2$ predictions, corresponding to 2 different actions. The transition module is similar to the reward module, but it predicts the difference between state and next state which is a widely-used trick for transition dynamics modeling. Both the reward module and transition module are feed-forward networks with no hidden layer. We optimize the model using Adam. Since this domain has variable length of trajectories, we also learn the condition of terminal state. The only domain prior we assume is that we know the maximum length of a trajectory is 200.

We also need to explain the details of transition loss ℓ_T. Since this domain is a deterministic domain, the loss function turns to be:

$$\ell_T(s_t, a_t, s'_t, \widehat{M}) = \left(V^\pi_{\widehat{M},H-t-1}(s'_t) - V^\pi_{\widehat{M},H-t-1}(s') \right)^2, \tag{77}$$

where s' is the predicted next state prediction and s'_t is the logged next state in dataset. Since repeatedly performing planning at training time is very computationally-intensive, it is difficult to get the function $V^\pi_{\widehat{M},H-t-1}(s)$. It is also challenging to compute the the derivative of this with respect to s. If we assume the resulting value is L-Lipschitz, then this loss can be bounded by $L(s' - s'_t)^2$. This is slightly different to the algorithmic part in the main body but it will still be an upper bound of ℓ_T in the main body. In this experiment we set $L = 1$.

If we are in a discrete state space, the transition loss ℓ_T turns to be:

$$\ell_T(s_t, a_t, s'_t, \widehat{M}) = \sum_{s' \in \mathcal{S}} V^\pi_{\widehat{M},H-t-1}(s') \left(\widehat{T}(s'|s,a) - \mathbb{1}(s'_t - s') \right)^2 \tag{78}$$

We can use similar trick with double Q learning for DQN: doing value iteration to generate a target value vector $V^{\pi}_{\widehat{M},H-t-1}(s')$, and view this as constant vector when we compute derivative with s'. Then this loss becomes a weighted MSE loss. We can update the target value vector $V^{\pi}_{\widehat{M},H-t-1}(s')$ every several episodes.

Methods We compare several different methods: **1) RepBM** The proposed method. **2) AM** We compare our method $RepBM$ with a baseline approximate model, which uses the exactly same model class as our model, with the objective of minimizing the on-policy loss R_{μ}. This is a straight-forward way to fit a regression model without any off-policy adjustment. **3) MRDR** we also compare with the more robust doubly robust (MRDR) method, which proposed a new way to train a Q function and use it in doubly robust. MRDR trains the Q function to minimize:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{H-1}(w_{0:t}^{(i)})^2\frac{1-\mu(a_t^{(i)}|s_t^{(i)})}{\mu(a_t^{(i)}|s_t^{(i)})}\left(\bar{R}_{t:H-1}^{(i)}-\widehat{Q}^{\pi}(s_t^{(i)},a_t^{(i)})\right), \tag{79}$$

where $\bar{R}_{t:H-1}^{(i)}=\sum_{j=t+1}^{H-1}w_{t+1:j}^{(i)}r_j^{(i)}$ is the per-decision IS return from $t+1$ to $H-1$, and w's are IS weights. **4) MRDR-WIS** Since this objective function can be very noisy and hard to fit when IS weights are high-variance, we also test another variant of MRDR by changing $\bar{R}_{t:H-1}^{(i)}$ to a weighted per-decision IS return from $t+1$ to $H-1$.

Within each one of the methods above, we test five different kinds of estimator. We have a pure MDP/Q model estimator, doubly robust (DR) using that MDP/Q model and weighted doubly robust (WDR) using that MDP/Q model. We evaluate a deterministic evaluation policy, which will result in most of the IS weights being zero; once an IS weight at one timestep is zero, then the product of all IS weights after that step will be zero. This setting is challenging for importance sampling and DR. We also test a very simple idea to avoid this problem – we add a slight noise perturbation into the evaluation policy (ϵ = 0.01), and treat it as the true evaluation policy to generate IS weights for DR and WDR. The additional noise is small enough so that the error introduced by this is negligible compared with the MSEs of estimators. We call these variants of DR and WDR soft DR and soft WDR respectively.

We also compare with importance sampling (IS), weighted IS (WIS), soft IS, soft WIS, per-decision importance sampling (PDIS), weighted PDIS (WPDIS), soft PDIS, soft WPDIS. The soft methods are produced by changing the IS weights using the soft evaluation policy.

We report the results in Table 3. Note that in the long horizon case, IS weights are all zero so WIS estimator is not defined. Though it is clear that for a single individual in continuous state space, IS and DR would not produce meaningful results due to the fact they only estimate from one trajectory, here we still include the IS and DR estimates for MSE for individual policy values. Not surprisingly we observe that those results are enormous which verifies that plain IS and DR are not reasonable estimators for individual value.

Evaluation Thomas and Brunskill [22] discussed that it is not obvious how to use the trajectories to fairly compare DR, IS and AM estimators, in Appendix D.4 from [22]. There are three ways that are reasonable: the first way is that AM and DR estimators should be provided with additional trajectories that are not available to IS, which are used to learn the model. This can be viewed as the additional domain prior knowledge. This is the setting in MRDR's experiment [6]. The second way is that all methods should have the same amount of data. DR methods should split the data into two parts to learn the model and IS weights separately. That partition keeps the unbiasedness of DR, but reduces the size of available samples for model learning in DR. The third way is that all methods should have the same amount of data. The DR method reuses the data to learn the model and compute IS weights. This helps DR methods to achieve best empirical performance in Thomas and Brunskill [22]. There is not necessarily a "correct" answer to this question. We follow the third setting to make both DR and IS stronger baselines.

We sample 1024 trajectories to generate off-policy estimators. For our method and AM method, we split the data into a training set (90%) and a validation set (10%) and use the validation set to tune the model structure and optimization settings. To compute the MSE of an individual value, we record the initial state of the 1024 trajectories and roll-out from true environment to get the true policy over those initial states as ground truth. We use the average policy value over these initial states as the

Table 3: Root MSE for Cart Pole

	Long horizon		Short horizon	
	MSE (mean)	MSE (individual)	MSE (mean)	MSE (individual)
RepBM	**0.412**	**1.033**	0.078	**0.481**
DR(RepBM)	1.359	40.820	0.021	0.789
WDR(RepBM)	0.619	17.760	0.026	0.857
Soft DR(RepBM)	1.608	53.390	**0.020**	0.686
Soft WDR(RepBM)	0.730	24.95	20.59	634.7
AM	0.754	1.313	0.125	0.551
DR(AM)	1.786	58.66	0.024	0.863
WDR(AM)	0.706	19.73	0.025	0.929
Soft DR(AM)	1.613	52.56	**0.020**	0.744
Soft WDR(AM)	0.848	29.71	20.28	640.4
MRDR's Q	151.1	151.9	3.013	3.823
MRDR	202.0	7055	0.258	8.266
WMRDR	123.6	1049	2.343	59.640
Soft MRDR	813.8	2590	0.211	6.758
Soft WMRDR	92.00	2669	22.550	601.4
MRDR-WIS's Q	143.9	145.1	2.486	3.440
MRDR-WIS	190.9	6106	0.248	8.075
WMRDR-WIS	122.0	1054	2.599	68.60
Soft MRDR-WIS	746.9	23610	0.199	6.626
Soft WMRDR-WIS	108.3	2992	21.26	570.6
IS	194.500	194.7	2.860	93.87
WIS	-	-	0.505	93.86
Soft IS	187.9	1115	2.179	70.78
Soft WIS	8.144	4698	0.380	70.55
PSIS	477.5	1526	1.083	36.67
WPSIS	125.9	622.5	1.819	63.59
Soft PSIS	215.2	6853	0.903	30.08
Soft WPSIS	4.225	1983	24	678.2

ground truth for MSE of mean value. We repeat the whole process for $N = 100$ runs and report the square root of averaged MSEs (for both individual and mean).

Effect of parameter α

We study the effect of the hyper-paramter α in the IPM terms on the estimation results. We show the MSE of RepBM trained using different α.

Table 4: Root MSEs of RepBM with different α for the cartpole domain

Long horizon	$\alpha = 0$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 1$	$\alpha = 10$
Mean	0.554	0.412	0.406	**0.389**	2.287
Individual	1.178	1.033	**1.008**	1.023	3.903
Short horizon	$\alpha = 0$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 1$	$\alpha = 10$
Mean	0.114	**0.078**	0.114	0.357	0.365
Individual	0.672	**0.481**	0.702	1.684	1.545

E.2 Real data: Sepsis management domain

Details of the domain The data for each patient consists of 47 features such as measurements of vitals, lab tests and demographics, recorded at 4 hour intervals. To account for the possible influence of a patient's history on their current response to treatment, we define the patient state as the 188 continuous features vector consisting of a concatenation of the last 4 observations. The actions we consider are the prescription of IV fluids and vasopressors. Each of the two treatments is binned into 5 discrete actions, with four bins representing which quartile the given dosage belongs to, and one bin representing not prescribing the relevant treatment. The combination of two treatments with 5 bins representing each one results in 25 possible actions. The rewards are defined as the change of an acuity measure known as the SOFA score following an action (a negative change in SOFA score corresponds to an improvement in the patient's condition). We treat the data as a task with a fixed time horizon by only observing the first 10 transitions of every patient.

The model details are similar to the model used for the cartpole example, except that for sepsis our hidden representation includes 64 dimensions.

The effect of α The results for the population RMSE of the estimation error for RepBM with different values of α are given in Table 5. We also include the approximate model method as comparison. We note that for the sepsis example, the smallest RMSE was obtained using $\alpha = 0$, indicating that the improvement in learning the model comes from reweighting the trajectories consistent with the evaluation policy, rather than from learning a representation of the data.

Table 5: Root MSE for Sepsis

	AM	$\alpha = 0$	$\alpha = 0.01$	$\alpha = 0.03$	$\alpha = 0.1$
RMSE	0.23	**0.09**	**0.09**	0.13	0.15

Approximators for policies and its effect on OPPE results. In section 6.2, we approximate the evaluation and behavior policies by using a nearest-neighbors classifier with 10 neighbors. The small number of neighbors allows us to accurately approximate the evaluation policy by only using patients which are very similar to each other in order to estimate the action used to treat them. However, for this number of neighbors, many state-action observations in our dataset have probability zero under the approximated behavior policy. RepBM is not affected by this issue, since it is independent of the behavior policy, but the IS based methods are not applicable in that case. In order to compare our method with IS estimators we increase the number of neighbors to 150 for IS estimators so that no state-action pairs with probability zero under the behavior policy are observed. Here we include the results in Section 6.2, as well as the results for RepBM and AM using 150 neighbors. This results in the inclusion of far away neighbors in the classifier which decreases the quality of our policy estimation (through a worse estimation of the evaluation policy, for which we increase the number of neighbors as well). However, in table 6 we show that even after we handicap RepBM in order to compare it with IS methods, it still outperforms both the IS methods and the approximate model method.

Table 6: Root MSE for Sepsis

	RepBM	RepBM (k=150)	DR (RepBM)	AM	AM (k=150)	DR (AM)	IS	WDR (RepBM)	WDR (AM)
Mean	**0.09**	0.14	$3 \cdot 10^5$	0.23	0.19	$4 \cdot 10^5$	$5 \cdot 10^5$	0.8	0.73

E.3 Further discussion

An interesting issue is about the effect of the horizon. Although the "marginal" IS weights have less variance than IS weights, there is still a concern when the horizon is very long and the overlap of the behavior policy and the evaluation policy is small. That also has an effect on the IPM term: we would not have enough factual/counterfactual samples to estimate the IPMs, for large time steps t. In that case, the IPMs only effectively adjust the representation for the earlier of the trajectories. Both of the experimental domains actually encounter this case, and the experimental results show that RepBM still outperforms other methods. In the Cart Pole domain it is clear that RepBM can still benefit from IPM, but for the sepsis domain it is not clear since we can observe that we achieve the best MSE when $\alpha = 0$.