Minimax Model Learning

Cameron Voloshin [1] **Nan Jiang** [2] **Yisong Yue** [1]

Abstract

We present a novel off-policy loss function for learning a transition model in model-based reinforcement learning. Notably, our loss is derived from the off-policy policy evaluation objective with an emphasis on correcting distribution shift. We also apply it to policy learning. Compared to previous model-based techniques, our approach allows for greater robustness under model mis-specification or distribution shift induced by learning/evaluating policies that are distinct from the data-generating policy. We provide a theoretical analysis, finite sample guarantees, and show empirical improvements over existing model-based off-policy evaluation methods.

1. Introduction

We study the problem of learning a transition model in a batch, off-policy reinforcement learning (RL) setting, which is learning a function $P(s'|s,a)$ from a pre-collected dataset D. Contemporary approaches to model learning focus primarily on improving the performance of models learned through maximum likelihood estimation (MLE) (Sutton, 1990; Deisenroth & Rasmussen, 2011; Kurutach et al., 2018; Clavera et al., 2018; Chua et al., 2018; Luo et al., 2019) . The goal of MLE is to pick the model within some model class $\mathcal{P}$ that is most consistent with the observed data and, equivalently, most likely to have generated the data. This is done by minimizing negative log-loss (minimizing the KL divergence) summarized as follows:

$$\widehat{P}_{\text{MLE}} = \arg\max_{P \in \mathcal{P}} \frac{1}{n} \sum_{i=1}^{n} \log(P(s_i'|s_i, a_i)).$$

The main limitation with MLE is that it places a primary focus on picking a good model under the data distribution while ignoring how the model is actually used, i.e., that a sufficiently good model can be all-purpose. In an RL context, a model can be used to either learn a policy (policy learning/optimization) or evaluate some given policy (policy evaluation) without having to collect more data from the true environment. We call this actual objective the "decision problem." Interacting with the environment to solve the decision problem can be difficult, expensive and dangerous, whereas a model learned from batch data circumvents these issues. Yet, when the true model P^* does not belong to the model class $\mathcal{P}$, giving an error guarantee for the decision problem is non-trivial when using MLE for model learning.

Notable previous works that incorporate the decision problem into the model learning objective are Value-Aware Model Learning (VAML) and its variants (Farahmand et al., 2017; Farahmand, 2018; Abachi et al., 2020) These methods, however, still define their losses w.r.t. the data distribution as in MLE, and ignores the *distribution shift* from the data to the policy-induced distribution.

In contrast, we require the model to perform well under unknown distributions instead of the data distribution. To mitigate this issue, these previous methods often engage in online data collection and deviate from the batch learning setting. As such, we ask: *"From only pre-collected data, is there a model learning approach that naturally controls the decision problem error?"*

In this paper, we present a new loss function for model learning that (1) only relies on batch data, (2) takes into account the distribution shift effects, and (3) directly relates to the performance metrics for off-policy evaluation and learning under certain realizability assumptions. The design of our loss is inspired by recent advances in model-free off-policy evaluation (e.g., Liu et al., 2018; Uehara et al., 2019), and we build upon to develop our approach.

2. Preliminaries

We adopt the infinite-horizon discounted MDP framework specified by a tuple $(\mathcal{S}, \mathcal{A}, P, \mathcal{R}, \gamma)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition function, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \Delta([-R_{\max}, R_{\max}])$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor. Let $\mathcal{X} \equiv \mathcal{S} \times \mathcal{A}$. Given an MDP, a (stochastic)

[1]Caltech [2]UIUC. Correspondence to: Cameron Voloshin <cvoloshi@caltech.edu>.

Preliminary Work. Do not distribute.

policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ and a starting state distribution $d_0 \in \Delta(\mathcal{S})$ together determine a distribution over trajectories of the form $s_0, a_0, r_0, s_1, a_1, r_1, \ldots$, where $s_0 \sim d_0, a_t \sim \pi(s_t), r_t \sim \mathcal{R}(s_t, a_t)$, and $s_{t+1} \sim P(s_t, a_t)$ for $t \geq 0$. The performance of policy π is given by:

$$J(\pi, P) \equiv E_{s\sim d_0}[V_\pi^P(s)], \tag{1}$$

where, by the Bellman Equation,

$$V_\pi^P(s) \equiv E_{a\sim\pi(\cdot|s)}[E_{r\sim\mathcal{R}(\cdot|s,a)}[r] + \gamma E_{\tilde{s}\sim P(\cdot|s,a)}[V_\pi^P(\tilde{s})]]. \tag{2}$$

A useful equivalent measure of performance is:

$$J(\pi, P) = E_{(s,a,r)\sim d_{\pi,\gamma}^P R(s,a)}[r], \tag{3}$$

where $d_{\pi,\gamma}^P(s,a) \equiv \sum_{t=0}^\infty \gamma^t d_{\pi,t}^P(s,a)$ is the (discounted) distribution of state-action pairs induced by running π in P and $d_{\pi,t}^P \in \Delta(\mathcal{X})$ is the distribution of (s_t, a_t) induced by running π under P. The first term in $d_{\pi,\gamma}^P$ is $d_{\pi,0}^P = d_0$. $d_{\pi,t}^P$ has a recursive definition that we use in Section 3:

$$d_{\pi,t}^P(s,a) = \int d_{\pi,t-1}^P(\tilde{s},\tilde{a})P(s|\tilde{s},\tilde{a})\pi(a|s)d\nu(\tilde{s},\tilde{a}), \tag{4}$$

where ν is the Lebesgue measure. In the batch learning setting, we are given a dataset $D = \{(s_i, a_i, s_i')\}_{i=1}^n$, where $s_i \sim d_{\pi_b}(s)$, $a_i \sim \pi_b$, and $s_i' \sim P(\cdot|s_i, a_i)$, where π_b is some behavior policy that collects the data. For convenience, we write $(s, a, s') \sim D_{\pi_b}P$, where $D_{\pi_b}(s,a) = d_{\pi_b}(s)\pi_b(a|s)$. Let $E[\cdot]$ denote exact expectation and $E_n[\cdot]$ the empirical approximation using the n data points of D.

Finally, we also need three classes $\mathcal{W}, \mathcal{V}, \mathcal{P}$ of functions. $\mathcal{W} \subset (\mathcal{X} \to \mathbb{R})$ represents ratios between state-action occupancy, $\mathcal{V} \subset (\mathcal{S} \to \mathbb{R})$ represents value functions and $\mathcal{P} \subset (\mathcal{X} \to \Delta(\mathcal{S}))$ represents the class of models (or simulators) of the true environment.

3. Minimax Model Learning (MML) & Off-Policy Policy Evaluation (OPE)

We start with the off-policy evaluation (OPE) learning objective and derive the MML loss. In Section 5, we show that this loss also bounds policy optimization error through its connection with OPE. The OPE objective is to estimate $J(\pi, P^*)$, the performance of policy π in the true environment P^*, using only logging data $D \subset D_{\pi_b}P^*$. This is difficult because the actions in our dataset were chosen with π_b rather than π. Thus, any $\pi \neq \pi_b$ potentially induces a "shifted" state-action distribution $D_\pi \neq D_{\pi_b}$, and ignoring this distribution shift can lead to poor approximation.

In model-based OPE, we run π in P to compute $J(\pi, P)$ as a proxy to $J(\pi, P^*)$. If we find some $P \in \mathcal{P}$ such that $|\delta_\pi^{P,P^*}| = |J(\pi, P) - J(\pi, P^*)|$ is small, then P is a good simulator for P^*. Using (1) and (3), we have:

$$\delta_\pi^{P,P^*} = E_{s\sim d_0}[V_\pi^P(s)] - E_{(s,a,r)\sim d_{\pi,\gamma}^{P^*}(\cdot,\cdot)\mathcal{R}(\cdot|s,a)}[r].$$

Adding and subtracting $E_{(s,a)\sim d_{\pi,\gamma}^{P^*}}[V_\pi^P(s)]$, we have:

$$\delta_\pi^{P,P^*} = E_{s\sim d_0}[V_\pi^P(s)] - E_{(s,a)\sim d_{\pi,\gamma}^{P^*}}[V_\pi^P(s)] \tag{5}$$
$$+ E_{(s,a)\sim d_{\pi,\gamma}^{P^*}}[V_\pi^P(s) - E_{r\sim\mathcal{R}(\cdot|s,a)}[r]]. \tag{6}$$

First, Eq (6) can be simplified through the Bellman equation from Eq (2). To see this, notice that $d_{\pi,\gamma}^{P^*}$ is the same as some $d(s)\pi(a|s)$ for an appropriate choice of $d(s)$. Thus,

$$\begin{aligned}
&E_{(s,a)\sim d_{\pi,\gamma}^{P^*}}[V_\pi^P(s) - E_{r\sim\mathcal{R}(\cdot|s,a)}[r]] \\
&\quad = E_{s\sim d(\cdot)}[E_{a\sim\pi(\cdot|s)}[V_\pi^P(s) - E_{r\sim\mathcal{R}(\cdot|s,a)}[r]]] \\
&\quad = E_{s\sim d(\cdot)}[E_{a\sim\pi(\cdot|s)}[E_{s'\sim P(\cdot|s,a)}[\gamma V_\pi^P(s)]]] \\
&\quad = \gamma E_{(s,a)\sim d_{\pi,\gamma}^{P^*}}[E_{s'\sim P(\cdot|s,a)}[V_\pi^P(s')]].
\end{aligned}$$

Second, for Eq (5), we use the definition of $d_{\pi,\gamma}^P$ and the recursive property of $d_{\pi,t}^P$ from (4):

$$\begin{aligned}
&E_{d_0}[V_\pi^P] - E_{(s,a)\sim d_{\pi,\gamma}^{P^*}}[V_\pi^P(s)] \\
&= -\sum_{t=1}^\infty \gamma^t \int d_{\pi,t}^{P^*}(s,a)V_\pi^P(s)d\nu(s,a) \\
&= -\gamma\sum_{t=0}^\infty \gamma^t \int d_{\pi,t+1}^{P^*}(s,a)V_\pi^P(s)d\nu(s,a) \\
&= -\gamma\sum_{t=0}^\infty \gamma^t \int d_{\pi,t}^{P^*}(\tilde{s},\tilde{a})P^*(s|\tilde{s},\tilde{a})\pi(a|s)V_\pi^P(s)d\nu(\tilde{s},\tilde{a},s,a) \\
&= -\gamma\sum_{t=0}^\infty \gamma^t \int d_{\pi,t}^{P^*}(s,a)P^*(s'|s,a)V_\pi^P(s')d\nu(s,a,s') \\
&= -\gamma E_{(s,a)\sim d_{\pi,\gamma}^{P^*}}[E_{s'\sim P^*(\cdot|s,a)}[V_\pi^P(s')]].
\end{aligned}$$

Combining the above allows us to succinctly express:

$$\begin{aligned}
\delta_\pi^{P,P^*} &= \gamma E_{(s,a)\sim d_{\pi,\gamma}^{P^*}}[E_{s'\sim P(\cdot|s,a)}[V_\pi^P(s')]] \\
&\quad - \gamma E_{(s,a)\sim d_{\pi,\gamma}^{P^*}}[E_{s'\sim P^*(\cdot|s,a)}[V_\pi^P(s')]]
\end{aligned}$$

Since D is sampled from D_{π_b} and not $d_{\pi,\gamma}^{P^*}$, we use importance sampling:

$$\delta_\pi^{P,P^*} = \gamma E_{(s,a,s')\sim D_{\pi_b}P^*}\left[\frac{d_{\pi,\gamma}^{P^*}}{D_{\pi_b}}\left(E_{x\sim P(\cdot|s,a)}[V_\pi^P(x)] - V_\pi^P(s')\right)\right].$$

Define $w_\pi^P(s,a) \equiv \frac{d_{\pi,\gamma}^P(s,a)}{D_{\pi_b}(s,a)}$. If we knew $w_\pi^{P^*}(s,a)$ and V_π^P (for every $P \in \mathcal{P}$), then we can select a $P \in \mathcal{P}$ to directly control δ_π^{P,P^*}. We encode this intuition in Def. 3.1.

Definition 3.1. [MML Loss] For $w \in \mathcal{W}, V \in \mathcal{V}, P \in \mathcal{P}$,

$$\begin{aligned}
\mathcal{L}(w, V, P) = &E_{(s,a,s')\sim D_{\pi_b}(\cdot,\cdot)P^*(\cdot|s,a)}[w(s,a) \cdot \\
&\left(E_{x\sim P(\cdot|s,a)}[V(x)] - V(s')\right)].
\end{aligned}$$

Here we have replaced $w_\pi^{P^*}(s,a)$ with w coming from function class $\mathcal{W}$ and V_π^P with V from class $\mathcal{V}$. The function class $\mathcal{W}$ represents the possible distribution shifts, while $\mathcal{V}$ represents the possible value functions.

It can be shown that $\mathcal{L}(w, V, P^*) = 0$ for any $V \in \mathcal{V}, w \in \mathcal{W}$. This solution is unique when $\mathcal{V}, \mathcal{W}$ are sufficiently rich (see Appendix Lemma A.1). However, any P satisfying $\mathcal{L}(w, V, P) = 0$ for any $V \in \mathcal{V}, w \in \mathcal{W}$ (assuming realizability) is desirable. Therefore, we want to choose $\mathcal{V}, \mathcal{W}$ carefully so that many $P \in \mathcal{P}$ satisfy the condition. With this intuition, we can see that $J(\pi, P) \approx J(\pi, P^*)$ under the *realizability conditions* that $\mathcal{W} \times \mathcal{V}$ contains either $(\frac{d^P_{\pi,\gamma}(s,a)}{D_{\pi_b}(s,a)}, V_\pi^{P^*})$ or $(\frac{d^{P^*}_{\pi,\gamma}(s,a)}{D_{\pi_b}(s,a)}, V_\pi^P)$ for every $P \in \mathcal{P}$:

Theorem 3.1 (MML & OPE)**.** *For a given π, if either $(w_\pi^P, V_\pi^{P^*})$ or $(w_\pi^{P^*}, V_\pi^P)$ is in $\mathcal{W} \times \mathcal{V}$, $\forall P \in \mathcal{P}$, then:*

$$|J(\pi, \widehat{P}) - J(\pi, P^*)| \leq \gamma \min_{P \in \mathcal{P}} \max_{w \in \mathcal{W}, V \in \mathcal{V}} |\mathcal{L}(w, V, P)|, \quad (7)$$

where $\widehat{P} = \arg\min_{P \in \mathcal{P}} \max_{w \in \mathcal{W}, V \in \mathcal{V}} |\mathcal{L}(w, V, P)|$.

Remark 3.2. *While $V_\pi^P \in \mathcal{V}\ \forall P \in \mathcal{P}$ appears strong, it can be verified for every $P \in \mathcal{P}$ before accessing the data, as the condition does not depend on P^*.*

Remark 3.3. *When $\gamma = 0$, J does not depend on a transition function, so $J(\pi, P) = J(\pi, P^*)$ for any $P \in \mathcal{P}$.*

$\mathcal{L}(w, V, P^*) = 0$ and Thm. 3.1 imply that the following learning procedure will be robust to any distribution shift in $\mathcal{W}$ and any value function in $\mathcal{V}$:

$$\widehat{P} = \arg\min_{P \in \mathcal{P}} \max_{w \in \mathcal{W}, V \in \mathcal{V}} |\mathcal{L}(w, V, P)|. \quad (8)$$

We call this approach *minimax model learning (MML)*.

Theorem 3.1 quantifies the error we will incur by evaluating π in $\widehat{P}$ instead of P^*, assuming the realizability conditions hold. For OPE, $\widehat{P}$ is a reasonable proxy for P. In this sense, MML is a principled method approach for model selection for OPE. For those interested in the sample complexity of the estimate see Appendix Thm A.4.

If the exploratory state distribution d_{π_b} and π_b are known then D_{π_b} is known. In this case, we can also verify that $w_\pi^P \in \mathcal{W}$ for every $P \in \mathcal{P}$ a priori. Together with Remark 3.2, we may assume that both $w_\pi^P \in \mathcal{W}$ and $V_\pi^P \in \mathcal{V}$ for all $P \in \mathcal{P}$. Consequently, only one of $V_\pi^{P^*} \in \mathcal{V}$ or $w_\pi^{P^*} \in \mathcal{W}$ has to be realizable for Theorem 3.1 to hold.

Instead of checking for realizability apriori, we can perform post-verification that $w_\pi^{\widehat{P}} \in \mathcal{W}$ and $V_\pi^{\widehat{P}} \in \mathcal{V}$. Together with the terms depending on P^*, realizability of these are also sufficient for Theorem 3.1 to hold. This relaxes the strong "for all $P \in \mathcal{P}$" condition. See Appendix A.1 for complete proof.

4. Comparison to Model-Free OPE

Recent model-free OPE literature (e.g., Liu et al., 2018; Uehara et al., 2019) has similar realizability assumptions to our method. For example, two model-free OPE methods MQL/MWL require either $w_\pi^{P^*}$ or $V_\pi^{P^*}$ (respectively) to be realized to control the OPE error (Uehara et al., 2019). MML requires the same, in addition to terms that depend on $\mathcal{P}$. Computation aside, these terms are verifiable, so they do not pose a substantial theoretical challenge.

An advantage of model-free approaches is that when both $w_\pi^{P^*}, V_\pi^{P^*}$ are realized, they return an exact OPE point estimate. In contrast, MML additionally requires some $P \in \mathcal{P}$ that makes the loss zero for any $w \in \mathcal{W}, V \in \mathcal{V}$. The advantage of MML is the increased flexibility of a model, enabling OPO (Section 5) and visualization of results through simulation (leading to more transparency).

While recent model-free OPE and our method both take a minimax approach, the classes $\mathcal{W}, \mathcal{V}, \mathcal{P}$ play different roles. In the model-free case, minimization is w.r.t either $\mathcal{W}$ or $\mathcal{V}$ and maximization is w.r.t the other. In our case, $\mathcal{W}, \mathcal{V}$ are on the same (maximization) team, while minimization is over $\mathcal{P}$. This allows us to treat $\mathcal{W} \times \mathcal{V}$ as a single unit, and represents distribution-shifted value functions. A member of this class, $E_{\text{data}}[wV]$ $(= E_{(s,a)\sim D_{\pi_b}}[\frac{d^{P^*}_{\pi,\gamma}}{D_{\pi_b}} V_\pi^P(s)])$, ties together the OPE estimate.

5. Off-Policy Optimization (OPO)

In this section we examine MML's interaction with the policy learning/optimization objective. In this setting, the goal is to find a good policy with respect to the true environment P^* without interacting with P^*. The traditional model-based approach is to build a simulator $\widehat{P}$ and subsequently learn $\pi^*_{\widehat{P}}$ in the simulator through any policy optimization algorithm of choice. The hope is that the ideal in-simulator policy $\pi^*_{\widehat{P}}$ and the actual best (true environment) policy $\pi^*_{P^*}$ perform competitively: $J(\pi^*_{\widehat{P}}, P^*) \approx J(\pi^*_{P^*}, P^*)$.

Beginning with the objective, we add zero twice:

$$J(\pi^*_{P^*}, P^*) - J(\pi^*_P, P^*) = \underbrace{J(\pi^*_{P^*}, P^*) - J(\pi^*_{P^*}, P)}_{(a)} + \underbrace{J(\pi^*_{P^*}, P) - J(\pi^*_P, P)}_{(b)} + \underbrace{J(\pi^*_P, P) - J(\pi^*_P, P^*)}_{(c)}.$$

Term (b) is non-positive since π^*_P is optimal in P ($\pi^*_{P^*}$ is suboptimal), so we remove it with appropriate inequality. Term (a) is the OPE estimate of $\pi^*_{P^*}$ and term (c) the OPE estimate of π^*_P, implying that we should use Theorem 3.1. With this intuition, we have

Theorem 5.1. *[MML & OPO] If $w^{P^*}_{\pi^*_{P^*}}, w^{P^*}_{\pi^*_P} \in \mathcal{W}$ and*

$V^P_{\pi^*_{P^*}}, V^P_{\pi^*_P} \in \mathcal{V}$ *for every* $P \in \mathcal{P}$ *then:*

$$|J(\pi^*_{P^*}, P^*) - J(\pi^*_{\widehat{P}}, P^*)| \leq 2\gamma \min_P \max_{w,V} |\mathcal{L}(w, V, P)|.$$

The statement also holds if, instead, $w^P_{\pi^*_{P^*}}, w^P_{\pi^*_P} \in \mathcal{W}$ *and* $V^{P^*}_{\pi^*_{P^*}}, V^{P^*}_{\pi^*_P} \in \mathcal{V}$ *for every* $P \in \mathcal{P}$.

Theorem 5.1 compares two different policies in the same (true) environment, since $\pi^*_{\widehat{P}}$ will be run in P^* rather than $\widehat{P}$. In contrast, Theorem 3.1 compared the same policy in two different environments. The derivation of Thm. 5.1 (see Appendix A.1) shows that having a good bound on the OPE objective is sufficient for OPO. MML shows how to learn a model that exploits this relationship.

Furthermore, the realizability assumptions of Theorem 5.1 relax the requirements of an OPE oracle. Rather than requiring the OPE estimate for every π, it is sufficient to have the OPE estimate of $\pi^*_{P^*}$ and π^*_P (for every $P \in \mathcal{P}$) when there is a $P \in \mathcal{P}$ such that $\mathcal{L}(w, V, P)$ is small for any $w \in \mathcal{W}, V \in \mathcal{V}$.

For OPO, apriori verification of realizability is not possible. Whereas the target policy π was known for OPE, now $\pi^*_{\widehat{P}}$ is not known. Fortunately, like in OPE, we can perform post-verification of $w^P_{\pi^*_{\widehat{P}}} \in \mathcal{W}$ and $V^P_{\pi^*_{\widehat{P}}} \in \mathcal{V}$. If they do not hold, then we can modify the function class $\mathcal{P}$ until they do. This greatly relaxes the "for every $P \in \mathcal{P}$" condition and leaves only two unverifiable quantities relating to P^*.

For those interested in the sample complexity of the estimate see Appendix Thm A.2.3.

6. Comparison to Model-Free OPO

For minimax model-free OPO, (Chen & Jiang, 2019) have developed a minimax variant of Fitted Q Iteration (FQI) (Ernst et al., 2005). FQI is a commonly used model-free OPO method. In addition to realizability assumptions, these methods also maintain a completeness assumption: the function class of interest is closed under bellman update. Increasing the function class size can only help realizability but may break completeness. It is unknown if the completeness assumption of FQI is removable (Chen & Jiang, 2019). MML only has realizability requirements.

7. Computational Considerations

A discussion about computational issues can be found in Appendix A.3. We show there that, like (Uehara et al., 2019), we can simplify the minimax problem to minimization in a closed form when $\mathcal{WV}$ correspond to a reproducing kernel Hilbert space (RKHS). We also show that the entire minimax procedure has a closed form when the classes consist of only linear functions (e.g. tabular setting).

8. MML and Residual Dynamics

Suppose we already had some baseline model P_0 of P^*. Alternatively, we may view this as the real world starting with (approximately) known dynamics P_0 and drifting to P^*. We can modify MML to incorporate knowledge of P_0 to find the residual dynamics that are distributionally and decision-aware as follows:

Definition 8.1. [Residual MML Loss] For $w \in \mathcal{W}, V \in \mathcal{V}, P \in \mathcal{P}$,

$$\mathcal{L}(w, V, P) = E_{(s,a,s') \sim D_{\pi_b}(\cdot,\cdot) P^*(\cdot|s,a)} [w(s,a) \cdot \left(E_{x \sim P_0(\cdot|s,a)} [\frac{P_0(x|s,a) - P(x|s,a)}{P_0(x|s,a)} V(x)] - V(s') \right)]$$

9. Experiments & Results

In this section we perform brief experiments to verify our theoretical results for OPE. We use the standard Cartpole benchmark (OpenAI, (Brockman et al., 2016)). Details on the environment and policy specifics can be found in Appendix A.4.

We compare the methods using the log-relative MSE metric: $\log(\frac{(J(\pi,\widehat{P}) - J(\pi,P^*))^2}{(J(\pi_b,P^*) - J(\pi,P^*))^2})$, which is negative when the OPE estimate $J(\pi, \widehat{P})$ is superior to the on-policy estimate $J(\pi_b, \widehat{P})$. The more negative, the better the estimate.

As seen in Figure 1, our method outperforms the other model-learning approaches. This experiment validates the theoretical observation that previous value-aware work (VAML) is not distributionally robust and that MLE suffers from lacking decision-awareness.

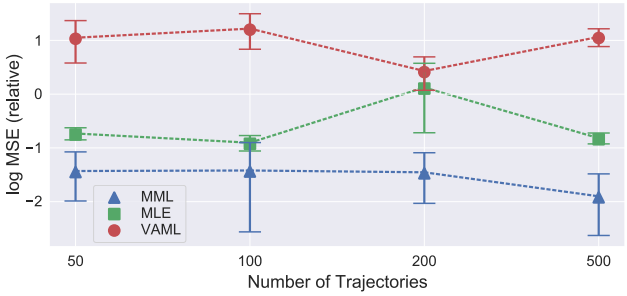

Figure 1. Comparison of model-based approaches for OPE with function-approx. MML outperforms other model-based methods.

10. Future Work

We wish to better understand the role of the class $\mathcal{W} \times \mathcal{V}$ and to examine if there are more implications regarding verifiability: "Are there environments where we know the proper $\mathcal{W}, \mathcal{V}$ classes?" We also wish to study when MML and MLE/VAML coincide. Finally, we want to find a theoretical connection between MML and MLE and run additional OPO experiments.

References

Abachi, R., Ghavamzadeh, M., and massoud Farahmand, A. Policy-aware model learning for policy gradient methods, 2020.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. In *Proceedings of the 14th Annual Conference on Computational Learning Theory and and 5th European Conference on Computational Learning Theory*, COLT '01/EuroCOLT '01, pp. 224–240, Berlin, Heidelberg, 2001. Springer-Verlag. ISBN 3540423435.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *CoRR*, abs/1606.01540, 2016.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1042–1051, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Chua, K., Calandra, R., McAllister, R., and Levine, S. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 4754–4765. Curran Associates, Inc., 2018.

Clavera, I., Rothfuss, J., Schulman, J., Fujita, Y., Asfour, T., and Abbeel, P. Model-based reinforcement learning via meta-policy optimization. In *2nd Annual Conference on Robot Learning, CoRL 2018, Zürich, Switzerland, 29-31 October 2018, Proceedings*, volume 87 of *Proceedings of Machine Learning Research*, pp. 617–629. PMLR, 2018.

Deisenroth, M. P. and Rasmussen, C. E. Pilco: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 465–472, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

Ernst, D., Geurts, P., and Wehenkel, L. Tree-based batch mode reinforcement learning. *J. Mach. Learn. Res.*, 6: 503–556, December 2005. ISSN 1532-4435.

Farahmand, A.-m. Iterative value-aware model learning. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 9072–9083. Curran Associates, Inc., 2018.

Farahmand, A.-M., Barreto, A., and Nikovski, D. Value-Aware Loss Function for Model-based Reinforcement Learning. In Singh, A. and Zhu, J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1486–1494, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc., 2014.

Jiang, N. and Huang, J. Minimax confidence interval for off-policy evaluation and policy optimization, 2020.

Kurutach, T., Clavera, I., Duan, Y., Tamar, A., and Abbeel, P. Model-ensemble trust-region policy optimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pp. 5361–5371, 2018.

Luo, Y., Xu, H., Li, Y., Tian, Y., Darrell, T., and Ma, T. Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

MacKay, D. J. C. *Information Theory, Inference Learning Algorithms*. Cambridge University Press, USA, 2002. ISBN 0521642981.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2012.

Sutton, R. S. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *In Proceedings of the Seventh International Conference on Machine Learning*, pp. 216–224. Morgan Kaufmann, 1990.

Uehara, M., Huang, J., and Jiang, N. Minimax weight and q-function learning for off-policy evaluation, 2019.

A. Appendix

A.1. Proofs

Proof for Theorem 3.1. Assume $(w_\pi^{P^*}, V_\pi^P) \in \mathcal{W} \times \mathcal{V}$. Fix some $P \in \mathcal{P}$. We use both definitions of J as follows

$$\begin{aligned}
J(\pi, P) - J(\pi, P^*) &= E_{d_0}[V_\pi^P] - E_{(s,a)\sim d_{\pi,\gamma}^{P^*}, r\sim\mathcal{R}(\cdot|s,a)}[r] \\
&= E_{(s,a)\sim d_{\pi,\gamma}^{P^*}}[V_\pi^P(s) - E_{r\sim\mathcal{R}(\cdot|s,a)}[r]] + E_{d_0}[V_\pi^P] - E_{(s,a)\sim d_{\pi,\gamma}^{P^*}}[V_\pi^P(s)] \\
&= E_{(s,a)\sim d_{\pi,\gamma}^{P^*}}[V_\pi^P(s) - E_{r\sim\mathcal{R}(\cdot|s,a)}[r]] - \sum_{t=1}^{\infty}\gamma^t \int d_{\pi,t}^{P^*}(s,a) V_\pi^P(s) d\nu(s,a) \\
&= E_{(s,a)\sim d_{\pi,\gamma}^{P^*}}[\gamma E_{s'\sim P(\cdot|s,a)}[V_\pi^P(s')]] - \gamma\sum_{t=0}^{\infty}\gamma^t \int d_{\pi,t+1}^{P^*}(s,a) V_\pi^P(s) d\nu(s,a) \\
&= \gamma E_{(s,a)\sim d_{\pi,\gamma}^{P^*}}[E_{s'\sim P(\cdot|s,a)}[V_\pi^P(s')]] - \gamma\sum_{t=0}^{\infty}\gamma^t \int d_{\pi,t}^{P^*}(\tilde{s},\tilde{a}) P^*(s|\tilde{s},\tilde{a})\pi(a|s) V_\pi^P(s) d\nu(\tilde{s},\tilde{a},s,a) \\
&= \gamma E_{(s,a)\sim d_{\pi,\gamma}^{P^*}}[E_{s'\sim P(\cdot|s,a)}[V_\pi^P(s')]] - \gamma\sum_{t=0}^{\infty}\gamma^t \int d_{\pi,t}^{P^*}(s,a) P^*(s'|s,a) V_\pi^P(s') d\nu(s,a,s') \\
&= \gamma E_{(s,a)\sim d_{\pi,\gamma}^{P^*}}[E_{s'\sim P(\cdot|s,a)}[V_\pi^P(s')]] - \gamma E_{(s,a)\sim d_{\pi,\gamma}^{P^*}}[E_{s'\sim P^*(\cdot|s,a)}[V_\pi^P(s')]] \\
&= \gamma E_{(s,a)\sim d_{\pi,\gamma}^{P^*}}[E_{s'\sim P(\cdot|s,a)}[V_\pi^P(s')] - E_{s'\sim P^*(\cdot|s,a)}[V_\pi^P(s')]] \\
&= \gamma E_{(s,a,s')\sim D_{\pi_b}P^*(\cdot|s,a)}[\frac{d_{\pi,\gamma}^{P^*}(s,a)}{D_{\pi_b}(s,a)}\left(E_{x\sim P(\cdot|s,a)}[V_\pi^P(x)] - V_\pi^P(s')\right)]] \\
&= \gamma E_{(s,a,s')\sim D_{\pi_b}P^*(\cdot|s,a)}[w_\pi^{P^*}(s,a)\left(E_{x\sim P(\cdot|s,a)}[V_\pi^P(x)] - V_\pi^P(s')\right)]] \\
&= \gamma L(w_\pi^{P^*}, V_\pi^P, P),
\end{aligned}$$

where the first equality is definition. The second equality is addition of 0. The third equality is simplification. The fourth equality is change of bounds. The fifth is definition. The sixth is relabeling of the integration variables. The seventh and eighth are simplification. The ninth is importance sampling. The tenth and last is definition. Since $(w_\pi^{P^*}, V_\pi^P) \in \mathcal{W} \times \mathcal{V}$ then

$$|J(\pi, P) - J(\pi, P^*)| = \gamma|L(w_\pi^{P^*}, V_\pi^P, P)| \leq \gamma \max_{w\in\mathcal{W}, V\in\mathcal{V}} |L(w, V, P)| \leq \gamma \min_{P\in\mathcal{P}} \max_{w\in\mathcal{W}, V\in\mathcal{V}} |L(w, V, P)|,$$

where the last inequality holds because P was selected in $\mathcal{P}$ arbitrarily.

Now, instead, assume $(w_\pi^P, V_\pi^{P^*}) \in \mathcal{W} \times \mathcal{V}$. Fix some $P \in \mathcal{P}$. Then, similarly,

$$\begin{aligned}
J(\pi, P) - J(\pi, P^*) &= E_{(s,a)\sim d_{\pi,\gamma}^{P}, r\sim\mathcal{R}(\cdot|s,a)}[r] - E_{d_0}[V_\pi^{P^*}] \\
&= E_{(s,a)\sim d_{\pi,\gamma}^{P}}[V_\pi^{P^*}(s)] - E_{d_0}[V_\pi^{P^*}] - E_{(s,a)\sim d_{\pi,\gamma}^{P}}[V_\pi^{P^*}(s) - E_{r\sim\mathcal{R}(\cdot|s,a)}[r]] \\
&= \sum_{t=1}^{\infty}\gamma^t \int d_{\pi,t}^{P}(s,a) V_\pi^{P^*}(s) d\nu(s,a) - E_{(s,a)\sim d_{\pi,\gamma}^{P}}[V_\pi^{P^*}(s) - E_{r\sim\mathcal{R}(\cdot|s,a)}[r]] \\
&= \gamma\sum_{t=0}^{\infty}\gamma^t \int d_{\pi,t+1}^{P}(s,a) V_\pi^{P^*}(s) d\nu(s,a) - E_{(s,a)\sim d_{\pi,\gamma}^{P}}[\gamma E_{s'\sim P^*(\cdot|s,a)}[V_\pi^{P^*}(s')]] \\
&= \gamma\sum_{t=0}^{\infty}\gamma^t \int d_{\pi,t}^{P}(\tilde{s},\tilde{a}) P(s|\tilde{s},\tilde{a})\pi(a|s) V_\pi^{P^*}(s) d\nu(\tilde{s},\tilde{a},s,a) - \gamma E_{(s,a)\sim d_{\pi,\gamma}^{P}}[E_{s'\sim P^*(\cdot|s,a)}[V_\pi^{P^*}(s')]] \\
&= \gamma\sum_{t=0}^{\infty}\gamma^t \int d_{\pi,t}^{P}(s,a) P(s'|s,a) V_\pi^{P^*}(s') d\nu(s,a,s') - \gamma E_{(s,a)\sim d_{\pi,\gamma}^{P}}[E_{s'\sim P^*(\cdot|s,a)}[V_\pi^{P^*}(s')]] \\
&= \gamma E_{(s,a)\sim d_{\pi,\gamma}^{P}}[E_{s'\sim P(\cdot|s,a)}[V_\pi^{P^*}(s')]] - \gamma E_{(s,a)\sim d_{\pi,\gamma}^{P}}[E_{s'\sim P^*(\cdot|s,a)}[V_\pi^{P^*}(s')]] \\
&= \gamma E_{(s,a)\sim d_{\pi,\gamma}^{P}}[E_{s'\sim P(\cdot|s,a)}[V_\pi^{P^*}(s')] - E_{s'\sim P^*(\cdot|s,a)}[V_\pi^{P^*}(s')]]
\end{aligned}$$

$$
\begin{aligned}
&= \gamma E_{(s,a,s')\sim D_{\pi_b} P^*(\cdot|s,a)}[\frac{d^P_{\pi,\gamma}(s,a)}{D_{\pi_b}(s,a)} \left(E_{x\sim P(\cdot|s,a)}[V^{P^*}_\pi(x)] - V^{P^*}_\pi(s')\right)]] \\
&= \gamma E_{(s,a,s')\sim D_{\pi_b} P^*(\cdot|s,a)}[w^P_\pi(s,a) \left(E_{x\sim P(\cdot|s,a)}[V^{P^*}_\pi(x)] - V^{P^*}_\pi(s')\right)]] \\
&= \gamma L(w^P_\pi, V^{P^*}_\pi, P),
\end{aligned}
$$

where we follow the same steps as in the previous derivation. Since $(w^P_\pi, V^{P^*}_\pi) \in \mathcal{W} \times \mathcal{V}$ then

$$
|J(\pi, P) - J(\pi, P^*)| = \gamma |L(w^P_\pi, V^{P^*}_\pi, P)| \leq \gamma \max_{w\in\mathcal{W}, V\in\mathcal{V}} |L(w, V, P)| \leq \gamma \min_{P\in\mathcal{P}} \max_{w\in\mathcal{W}, V\in\mathcal{V}} |L(w, V, P)|,
$$

where the last inequality holds because P was selected in $\mathcal{P}$ arbitrarily. □

Proof for Theorem 5.1. Fix some $P \in \mathcal{P}$. Through addition of 0, we get

$$
\begin{aligned}
J(\pi^*_{P^*}, P^*) - J(\pi^*_P, P^*) = &J(\pi^*_{P^*}, P^*) - J(\pi^*_{P^*}, P) \\
&+ J(\pi^*_{P^*}, P) - J(\pi^*_P, P) \\
&+ J(\pi^*_P, P) - J(\pi^*_P, P^*)
\end{aligned}
$$

Since π^*_P is optimal in P then $J(\pi^*_{P^*}, P) - J(\pi^*_P, P) \leq 0$ which implies

$$
J(\pi^*_{P^*}, P^*) - J(\pi^*_P, P^*) \leq J(\pi^*_{P^*}, P^*) - J(\pi^*_{P^*}, P) + J(\pi^*_P, P) - J(\pi^*_P, P^*)
$$

Taking the absolute value of both sides, triangle inequality and invoking Lemma 3.1 yields:

$$
|J(\pi^*_{P^*}, P^*) - J(\pi^*_{\widehat{P}}, P^*)| \leq 2\gamma \max_{w,V} |L(w, V, \widehat{P})| = 2\gamma \min_P \max_{w,V} |L(w, V, P)|
$$

when $w^{P^*}_{\pi^*_{P^*}}, w^{P^*}_{\pi^*_P} \in \mathcal{W}$ and $V^P_{\pi^*_{P^*}}, V^P_{\pi^*_P} \in V$ for every $P \in \mathcal{P}$, or alternatively $w^P_{\pi^*_{P^*}}, w^P_{\pi^*_P} \in \mathcal{W}$ and $V^{P^*}_{\pi^*_{P^*}}, V^{P^*}_{\pi^*_P} \in V$ for every $P \in \mathcal{P}$. □

A.2. Additional theory

A.2.1. NECESSARY AND SUFFICIENT CONDITIONS FOR UNIQUENESS OF $|\mathcal{L}(w, V, P)| = 0$

When $\mathcal{W}, \mathcal{V}$ are in L^2 then $|\mathcal{L}| = 0$ is uniquely determined:

Lemma A.1. *[Necessary and Sufficient]* $\mathcal{L}(w, V, P) = 0$ *for all* $w \in L^2(\mathcal{X}, \nu) = \{g : \int g^2(x, a)d\nu(x, a) < \infty\}, V \in L^2(\mathcal{S}, \nu) = \{f : \int f^2(x)d\nu(x) < \infty\}$ *if and only if* $P = P^*$ *wherever* $D_{\pi_b}(s, a) \neq 0$.

Corollary A.2. *The same result holds if* $w \cdot V \in L^2(\mathcal{X} \times \mathcal{S}, \nu) = \{h : \int h^2(x, a, x')d\nu(x, a, x') < \infty\}$.

Proof for Lemma A.1 and Corollary A.2. We begin with definition 8.1 and expand the expectation.

$$
\begin{aligned}
L(w, V, P) =&E_{(s,a,s')\sim D_{\pi_b}(\cdot,\cdot)P^*(\cdot|s,a)}[w(s, a) \left(E_{x\sim P(\cdot|s,a)}[V(x)] - V(s')\right)] \\
&= E_{(s,a)\sim D_{\pi_b}(\cdot,\cdot)P^*(\cdot|s,a)}[w(s, a) \left(E_{s'\sim P(\cdot|s,a)}[V(s')] - E_{s'\sim P(\cdot|s,a)}[V(s')]\right)] \\
&= \int D_{\pi_b}(s, a)w(s, a)(V(s')(P(s'|s, a) - P^*(s'|s, a))\, d\nu(s, a, s').
\end{aligned}
$$

($\Rightarrow$) Clearly if $P = P^*$ then $L(w, V, P) = 0$. ($\Leftarrow$) For the other direction, suppose $L(w, V, P) = 0$. By assumption, $w(s, a)$ can take on any function in $L^2(\mathcal{X}, \nu)$ and therefore if $L(w, V, P) = 0$ then

$$
\int V(s')(P(s'|s, a) - P^*(s'|s, a))\, d\nu(s') = 0, \tag{9}
$$

wherever $D_{\pi_b}(s, a) \neq 0$. Similarly, $V(s')$ can take on any function in $L^2(\mathcal{S}, \nu)$ and therefore if equation (9) holds then $P = P^*$. For the corollary, let $(w, V) \in \mathcal{W}\mathcal{V}$ take on any function in $L^2(\mathcal{X} \times \mathcal{S}, \nu)$. If $L(w, V, P) = 0$ then $P(s'|s, a) - P^*(s'|s, a) = 0$, as desired. □

In an RKHS, when the kernel corresponds to an integrally strict positive definite kernel (ISPD), $P = P^*$ remains the unique minimizer of the MML Loss:

Lemma A.3 (Realizability means zero loss even in RKHS)**.** *$\mathcal{L}(w, f, P) = 0$ if and only if $P = P^*$ for all $(w, V) \in \{(w(s,a), V(s')) : \langle wV, wV\rangle_{\mathcal{H}_k} \leq 1, w : X \times A \to \mathbb{R}, V : X \to \mathbb{R}\}$ in an RKHS with an integrally strict positive definite (ISPD) kernel.*

Proof for Lemma A.3. (Uehara et al., 2019) prove an analogous result and proof here is included for reader convenience. From Mercer's theorem (Mohri et al., 2012), there exists an orthonormal basis $(\phi_j)_{j=1}^{\infty}$ of $L^2(\mathcal{X} \times \mathcal{S}, \nu)$ such that RKHS is represented as

$$\mathcal{WV} = \left\{ w \cdot V = \sum_{j=1}^{\infty} b_j \phi_j \;\middle|\; (b_j)_{j=1}^{\infty} \in l^2(\mathbb{N}) \text{ with } \sum_{j=1}^{\infty} \frac{b_j^2}{\mu_j} < \infty \right\}$$

where each μ_j is a positive value since kernel is ISPD. Suppose there exists some $P \in \mathcal{P}$ such that $L(w, V, P) = 0$ for all $(w, V) \in \mathcal{WV}$ and $P \neq P^*$. Then, by taking $b_j = 1$ when $(j = j')$ and $b_j = 0$ when $(j \neq j')$ for any $j' \in \mathbb{N}$, we have $L(\phi_j, P) = 0$ where we treat $w \cdot V$ as a single input to L. This implies $L(w, V, P) = 0$ for all $w \cdot V \in L^2(\mathcal{X} \times \mathcal{S}, \nu) = 0$. This contradicts corollary A.2, concluding the proof. □

A.2.2. Sample Complexity for OPE

We do not have access to exact expectations, so we must work with $\widehat{P}_n = \min\max E_n[\ldots]$ instead of $\widehat{P} = \min\max E[\ldots]$. Furthermore, $J(\pi, \widehat{P})$ requires exact expectation of an infinite sum: $E_{d_0}[\sum_{t=0}^{\infty} \gamma^t r_t]$ where we collect r_t by running π in simulation $\widehat{P}$. Instead, we can only estimate an empirical average over a finite sum in $\widehat{P}_n$: $J_{T,m}(\pi, \widehat{P}_n) = \frac{1}{m}\sum_{j=1}^{m}\sum_{t=0}^{T} \gamma^t r_t^j$, where each j indexes rollouts starting from $s_0 \sim d_0$ and the simulation is over $\widehat{P}_n$. Our OPE estimate is therefore bounded as follows:

Theorem A.4. *[OPE Error] Let the functions in $\mathcal{V}$ and $\mathcal{W}$ be uniformly bounded by $C_{\mathcal{V}}$ and $C_{\mathcal{W}}$ respectively. Assume the conditions of Lemma 3.1 hold and $|\mathcal{R}| \leq R_{max}, \gamma \in [0, 1)$. Then, with probability $1 - \delta$,*

$$\begin{aligned} |J_{T,m}(\pi, \widehat{P}_n) - J(\pi, P^*)| &\leq \gamma \min_{P} \max_{w,V} |\mathcal{L}(w, V, P)| \\ &+ 4\gamma \mathfrak{R}_n(\mathcal{W}, \mathcal{V}, \mathcal{P}) + \frac{2R_{max}}{1-\gamma}\gamma^{T+1} \\ &+ \frac{2R_{max}}{1-\gamma}\sqrt{\log(2/\delta)/(2m)} + 4\gamma C_{\mathcal{W}} C_{\mathcal{V}} \sqrt{\log(2/\delta)/n} \end{aligned}$$

where $\mathfrak{R}_n(\mathcal{W}, \mathcal{V}, \mathcal{P})$ is the Rademacher complexity of the function class

$$\begin{aligned} \{(s, a, s') \mapsto & w(s,a)(E_{x\sim P}[V(x)] - V(s')) : \\ & w \in \mathcal{W}, V \in \mathcal{V}, P \in \mathcal{P}\}. \end{aligned}$$

Proof for Theorem A.4. By definition and triangle inequality,

$$\begin{aligned} |J_{T,m}(\pi, \widehat{P}_n) - J(\pi, P^*)| &= |J_{T,m}(\pi, \widehat{P}_n) - J(\pi, \widehat{P}_n) + J(\pi, \widehat{P}_n) - J(\pi, P^*)| \\ &\leq \underbrace{|J_{T,m}(\pi, \widehat{P}_n) - J(\pi, \widehat{P}_n)|}_{(a)} + \underbrace{|J(\pi, \widehat{P}_n) - J(\pi, P^*)|}_{(b)} \end{aligned} \tag{10}$$

Define $\widehat{V}_{\pi,T}^{P}(s_0^i) \equiv \sum_{t=0}^{T} \gamma^t r_t^i$ for some trajectory indexed by $i \in \mathbb{N}$ where r_t^i is the reward obtained by running π in P at time $t \leq T$ starting at s_0^i. For (a),

$$\begin{aligned} |J_{T,m}(\pi, \widehat{P}_n) - J(\pi, \widehat{P}_n)| &= \left| \frac{1}{m}\sum_{i=1}^{m} \widehat{V}_{\pi,T}^{\widehat{P}_n}(s_0^i) - \frac{1}{m}\sum_{i=1}^{m} \widehat{V}_{\pi,\infty}^{\widehat{P}_n}(s_0^i) + \frac{1}{m}\sum_{i=1}^{m} \widehat{V}_{\pi,\infty}^{\widehat{P}_n}(s_0^i) - E_{d_0}[V_{\pi}^{\widehat{P}_n}] \right| \\ &\leq \left| \frac{1}{m}\sum_{i=1}^{m} \widehat{V}_{\pi,T}^{\widehat{P}_n}(s_0^i) - \frac{1}{m}\sum_{i=1}^{m} \widehat{V}_{\pi,\infty}^{\widehat{P}_n}(s_0^i) \right| + \left| \frac{1}{m}\sum_{i=1}^{m} \widehat{V}_{\pi,\infty}^{\widehat{P}_n}(s_0^i) - E_{d_0}[V_{\pi}^{\widehat{P}_n}] \right| \end{aligned}$$

$$\leq \frac{2R_{\max}}{1-\gamma}\gamma^{T+1} + \frac{2R_{\max}}{1-\gamma}\sqrt{\log(2/\delta)/(2m)}, \tag{11}$$

with probability $1-\delta$, where the last inequality is definition of $\widehat{V}_{\pi,T}$ and Hoeffding's inequality ().

For (b), by Theorem 3.1,

$$\begin{aligned}
&|J(\pi, \widehat{P}_n) - J(\pi, P^*)| \\
&\quad = \gamma |L(w_\pi^{P^*}, V^{\widehat{P}_n}, \widehat{P}_n)| \\
&\quad \leq \gamma \max_{w,V} |L(w, V, \widehat{P}_n)| \\
&\quad = \gamma(\max_{w,V} |L(w, V, \widehat{P}_n)| - \max_{w,V} |L_n(w, V, \widehat{P}_n)| + \max_{w,V} |L_n(w, V, \widehat{P}_n)| - \max_{w,V} |L(w, V, \widehat{P})| + \max_{w,V} |L(w, V, \widehat{P})|) \\
&\quad \leq \gamma(2 \max_{w,V,P} ||L(w, V, P)| - |L_n(w, V, P)|| + \min_P \max_{w,V} |L(w, V, P)|) \\
&\quad \leq \gamma(2\mathfrak{R}'_n(\mathcal{W}, \mathcal{V}, \mathcal{P}) + 2K\sqrt{\log(2/\delta)/n} + \min_P \max_{w,V} |L(w, V, P)|) \\
&\quad \leq \gamma(4\mathfrak{R}_n(\mathcal{W}, \mathcal{V}, \mathcal{P}) + 2K\sqrt{\log(2/\delta)/n} + \min_f \max_{w,V} |L(w, f, V)|)
\end{aligned} \tag{12}$$

where $\mathfrak{R}'_n(\mathcal{W}, \mathcal{V}, \mathcal{P})$ is the Rademacher complexity of the function class

$$\{(s, a, s') \mapsto |w(s, a)(E_{x\sim P}[V(x)] - V(s'))| : w \in \mathcal{W}, V \in \mathcal{V}, P \in \mathcal{P}\}$$

noting that $K = 2C_w C_V$ uniformly bounds $|w(s, a)(E_{x\sim P(\cdot|s,a)}[V(x)] - V(s'))|$ (Theorem 8 (Bartlett & Mendelson, 2001)). Furthermore since absolute value is 1-Lipshitz (by reverse triangle ineq), then $\mathfrak{R}'_n < 2\mathfrak{R}_n$ (Theorem 12 (Bartlett & Mendelson, 2001)) where $\mathfrak{R}_n(\mathcal{W}, \mathcal{V}, \mathcal{P})$ is the Rademacher complexity of the function class

$$\{(s, a, s') \mapsto w(s, a)(E_{x\sim P(\cdot|s,a)}[V(x)] - V(s'))) : w \in \mathcal{W}, V \in \mathcal{V}, P \in \mathcal{P}\}.$$

Altogether, combining (1), (2), (3) we get our result. □

A.2.3. Sample Complexity for Learning

Since we will only have access to the empirical version $\widehat{P}_n$ rather than $\widehat{P}$, we provide the following bound

Theorem A.5 (Learning Error)**.** *Let the functions in $\mathcal{V}$ and $\mathcal{W}$ be uniformly bounded by C_V and C_W respectively. Assume the conditions of Lemma 5.1 hold and $|\mathcal{R}| \leq R_{max}, \gamma \in [0, 1)$. Then, with probability $1-\delta$,*

$$\begin{aligned}
|J(\pi^*_{\widehat{P}_n}, P^*) - J(\pi^*_{P^*}, P^*)| \leq 2\gamma \min_P \max_{w,V} |L(w, V, P)| \\
+ 8\gamma\mathfrak{R}_n(\mathcal{W}, \mathcal{V}, \mathcal{P}) + 8\gamma C_{\mathcal{W}} C_{\mathcal{V}} \sqrt{\log(2/\delta)/n}
\end{aligned}$$

where $\mathfrak{R}_n(\mathcal{W}, \mathcal{V}, \mathcal{P})$ is the Rademacher complexity of the function class

$$\begin{aligned}
\{(s, a, s') \mapsto & w(s, a)(E_{x\sim P}[V(x)] - V(s')) : \\
& w \in \mathcal{W}, P \in \mathcal{P}, V \in \mathcal{V}\}.
\end{aligned}$$

Proof for Theorem A.5. By Theorem 5.1,

$$|J(\pi^*_{\widehat{P}_n}, P^*) - J(\pi^*_{P^*}, P^*)| \leq 2\gamma \max_{w,V} |L(w, V, \widehat{P}_n)|.$$

We have shown in the proof of Theorem A.4 that

$$\max_{w,V} |L(w, V, \widehat{P}_n)| \leq \min_P \max_{w,V} |L(w, V, P)| + 4\mathfrak{R}_n(\mathcal{W}, \mathcal{V}, \mathcal{P}) + 4C_{\mathcal{W}} C_{\mathcal{V}} \sqrt{\log(2/\delta)/n}.$$

Combining the two completes the proof. □

A.3. Computational Considerations

Since $P \in \mathcal{P}$ is a model, the inner expectation of the loss in def (8.1) over P involves sampling x from $P(\cdot|s,a)$ and computing the empirical average of $V(x)$. In general this can be computationally demanding if $\mathcal{S}$ is high dimensional and P does not have a closed form, requiring MCMC estimates or variational estimates (MacKay, 2002; Goodfellow et al.). However, in practice, most parametrizations of models use nice distributions, such as gaussians, from which sampling is efficient. This issue is similarly present in other decision-aware literature (e.g (Farahmand et al., 2017)).

The estimator based on Eq (8) requires solving a minimax problem which is often computationally challenging. One approach might be to set up adversarial networks in a GAN-like fashion (Goodfellow et al., 2014). If Eq (8) instead were minimax of $\mathcal{L}^2$ then the inner maximization over w, V has a closed form when $\mathcal{W} \times \mathcal{V}$ correspond to a reproducing kernel Hilbert space (RKHS), H_K with kernel K. In particular, in similar spirit to (Uehara et al., 2019; Liu et al., 2018) we have

Lemma A.6. *[Closed form exists in RKHS] Assume* $\mathcal{WV} = \{(w(s,a), V(s')) : \langle wV, wV\rangle_{\mathcal{H}_K} \leq 1, w : \mathcal{X} \to \mathbb{R}, V : \mathcal{S} \to \mathbb{R}\}$. *Let* $\langle\cdot,\cdot\rangle_{\mathcal{H}_K}$ *be an inner product on* $\mathcal{H}_K$ *satisfying the reproducing kernel property* $w(s,a)V(s') = \langle wV, K((s,a,s'),\cdot)\rangle_{\mathcal{H}_K}$. *The term* $\max_{(w,V)\in\mathcal{WV}} \mathcal{L}(w,V,P)^2$ *has a closed form:*

$$\max_{(w,V)\in\mathcal{WV}} \mathcal{L}(w,f,V)^2 = E_{(s,a,s')\sim D_{\pi_b}P^*,(\tilde{s},\tilde{a},\tilde{s}')\sim D_{\pi_b}P^*}\Big[E_{x\sim P,\tilde{x}\sim P}[K((s,a,x),(\tilde{s},\tilde{a},\tilde{x}))] - 2E_{x\sim P}[K((s,a,x),(\tilde{s},\tilde{a},\tilde{s}'))] + K((s,a,s'),(\tilde{s},\tilde{a},\tilde{s}'))\Big]$$

Proof for Lemma A.6. Recall that by the reproducing property of kernel K in the RKHS space H_K then $\langle f, K\rangle_{H_K}$ for any $f \in H_K$. Starting from definition 8.1,

$$\begin{aligned}
L(w,V,P)^2 =& E_{(s,a,s')\sim D_{\pi_b}(\cdot,\cdot)P^*(\cdot|s,a)}[w(s,a)\left(E_{x\sim P(\cdot|s,a)}[V(x)] - V(s')\right)]^2 \\
=& E_{(s,a,s',x)\sim D_{\pi_b}(\cdot,\cdot)P^*(\cdot|s,a)P(\cdot|s,a)}[w(s,a)V(x) - w(s,a)V(s')]^2 \\
=& E_{(s,a,s',x)\sim D_{\pi_b}(\cdot,\cdot)P^*(\cdot|s,a)P(\cdot|s,a)}[\langle wV, K((s,a,x),\cdot)\rangle_{\mathcal{H}_k} - \langle wV, K((s,a,s'),\cdot)\rangle_{\mathcal{H}_k}]^2 \\
=& \langle wV, (wV)^*\rangle^2_{\mathcal{H}_k}
\end{aligned}$$

where $(wV)^*(\cdot) = E_{(s,a,s',x)\sim D_{\pi_b}(\cdot,\cdot)P^*(\cdot|s,a)P(\cdot|s,a)}[K((s,a,x),\cdot) - K((s,a,s',\cdot)]$. By Cauchy-Schwarz and the fact that wV is within a unit ball, then

$$\max_{w,V\in\mathcal{WV}} L(w,f,V)^2 = \max_{w,V\in\mathcal{WV}} \langle wV, (wV)^*\rangle^2_{\mathcal{H}_k} = \|(wV)^*\|^2 = \langle (wV)^*, (wV)^*\rangle_{\mathcal{H}_k}.$$

Expanding,

$$\begin{aligned}
\max_{w,V\in\mathcal{WV}} L(w,f,V)^2 &= \langle (wV)^*, (wV)^*\rangle_{\mathcal{H}_k} \\
&= \langle E_{(s,a,s',x)\sim D_{\pi_b}(\cdot,\cdot)P^*(\cdot|s,a)P(\cdot|s,a)}[K((s,a,x),\cdot) - K((s,a,s',\cdot)], \\
&\quad E_{(\tilde{s},\tilde{a},\tilde{s}',\tilde{x})\sim D_{\pi_b}(\cdot,\cdot)P^*(\cdot|\tilde{s},\tilde{a})P(\cdot|\tilde{s},\tilde{a})}[K((\tilde{s},\tilde{a},\tilde{x}),\cdot) - K((\tilde{s},\tilde{a},\tilde{s}',\cdot)]\rangle_{\mathcal{H}_k} \\
&= \Big\langle \int D_{\pi_b}(s,a)P^*(s'|s,a)P(x|s,a)(K((s,a,x),\cdot) - K((s,a,s'),\cdot)), \\
&\quad \int D_{\pi_b}(\tilde{s},\tilde{a})P^*(\tilde{s}'|\tilde{s},\tilde{a})P(\tilde{x}|\tilde{s},\tilde{a})(K((\tilde{s},\tilde{a},\tilde{x}),\cdot) - K((\tilde{s},\tilde{a},\tilde{s}'),\cdot)), \Big\rangle_{\mathcal{H}_k} \\
&= \int D_{\pi_b}(s,a)P^*(s'|s,a)P(x|s,a)D_{\pi_b}(\tilde{s},\tilde{a})P^*(\tilde{s}'|\tilde{s},\tilde{a})P(\tilde{x}|\tilde{s},\tilde{a}) \\
&\quad \times \langle K((s,a,x),\cdot) - K((s,a,s'),\cdot), K((\tilde{s},\tilde{a},\tilde{x}),\cdot) - K((\tilde{s},\tilde{a},\tilde{s}',\cdot)\rangle_{\mathcal{H}_k}
\end{aligned}$$

By linearity of the inner product, the reproducing kernel property we get

$$\begin{aligned}\max_{(w,V)\in\mathcal{WV}} L(w,f,V)^2 &= E_{(s,a,s',x)\sim D_{\pi_b}P^*P,(\tilde{s},\tilde{a},\tilde{s}',\tilde{x})\sim D_{\pi_b}P^*P}[K((s,a,x),(\tilde{s},\tilde{a},\tilde{x})) - K((s,a,x),(\tilde{s},\tilde{a},\tilde{s}')) \\ &\quad - K((s,a,s'),(\tilde{s},\tilde{a},\tilde{x})) + K((s,a,s'),(\tilde{s},\tilde{a},\tilde{s}'))] \\ &= E_{(s,a,s',x)\sim D_{\pi_b}P^*P,(\tilde{s},\tilde{a},\tilde{s}',\tilde{x})\sim D_{\pi_b}P^*P}[K((s,a,x),(\tilde{s},\tilde{a},\tilde{x})) - 2K((s,a,x),(\tilde{s},\tilde{a},\tilde{s}')) \\ &\quad + K((s,a,s'),(\tilde{s},\tilde{a},\tilde{s}'))],\end{aligned}$$

where for the last equality we used the fact that K is symmetric. □

Finally, when $\mathcal{WV},\mathcal{P}$ are linear function classes with the same features then the entire minimax optimization has a closed form. In the tabular setting, $P^*\in\mathcal{P}$ and therefore $\widehat{P}=P^*$ in the case of infinite data. In practice we only have finite data and so the $\widehat{P}$ coincides with the MLE estimate. The following example shows how to compute $\widehat{P}$ in such a case.

Example A.1 (Linear Function classes and tabular representation)**.** Let $P=\phi(s,a,s')^T\alpha$ where $\phi\in\mathbb{R}^d$ is some basis of features with α its parameters. Let $(w(s,a),V(s'))\in\mathcal{WV}=\{\phi(s,a,s')^T\beta:\|\beta\|_2\leq 1\}$. Then

$$\widehat{\alpha}=E_n^{-1}[\phi(s,a,s')\phi(s,a,s')^T]E_n[\phi(s,a,s')], \tag{13}$$

assuming $\phi(s,a,s')\phi(s,a,s')^T$ has full rank.

Proof for Example A.1. Given $w(s,a)V(s')=\phi(s,a,s')^T\beta$ and $P(s'|s,a)=\phi(s,a,s')^T\alpha$ then

$$L(w,V,P)=E_n[\alpha^T\phi(s,a,s')\phi(s,a,s')^T\beta-\phi(s,a,s')^T\beta],$$

which is linear in β. $L^2(w,V,P)=0$ is achieved through $E_n[\alpha^T\phi(s,a,s')\phi(s,a,s')^T-\phi(s,a,s')^T]=0$. Thus,

$$\widehat{\alpha}^T=E_n[\phi(s,a,s')^T]E_n[\phi(s,a,s')\phi(s,a,s')^T]^{-1},$$

assuming $E_n[\phi(s,a,s')\phi(s,a,s')^T]$ is full rank. Taking the transpose completes the proof. □

A.4. Experiments

A.4.1. Environment Description

Each environment model takes the form $s'\sim\mathcal{N}(\mu(s,a),\sigma(s,a))$, where a NN outputs a mean, and logvariance representing a normal distribution around the next state.

For Cartpole we follow (Jiang & Huang, 2020) and consider the infinite horizon setting with $\gamma=.98$. The reward function is modified to be a function of angle and location rather than 0/1 to make the OPE problem a bit more challenging. We generate the behavior and target policy using a near-perfect DDQN-based policy Q with a final softmax layer and adjustable parameter τ: $\pi(a|s)\propto\exp(Q(s,a)/\tau)$. The behavior policy has $\tau=1$, while the target policy has $\tau=1.5$. We truncate all rollouts at 1000 time steps and we calculate the true expected value using the MC average of 10000 rollouts.