Minimax Model Learning

Anonymous Author
Anonymous Institution

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. Compared to previous model-based techniques, our approach allows for greater robustness under model misspecification or distribution shift induced by learning/evaluating policies that are distinct from the data-generating policy. We provide a theoretical analysis and show empirical improvements over existing model-based off-policy evaluation methods. We provide further analysis showing our loss can be used for off-policy optimization (OPO) and demonstrate its integration with more recent improvements in OPO.

1 Introduction

We study the problem of learning a transition model in a batch, off-policy reinforcement learning (RL) setting, i.e., of learning a function $P(s'|s,a)$ from a pre-collected dataset $D = \{(s_i, a_i, s'_i)\}_{i=1}^n$ without further access to the environment. 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:

$$\widehat{P}_{\text{MLE}} = \arg\min_{P \in \mathcal{P}} \frac{1}{n} \sum_{(s_i,a_i,s'_i) \in D} -\log(P(s'_i|s_i,a_i)). \quad (1)$$

A key limitation of MLE is that it focuses on picking a good model under the data distribution while ignoring how the model is actually used.

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. Since MLE (1) does not optimize over the distribution of states induced by the policy from the decision problem, it thus does not prioritize solving the decision problem. 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 ignore the *distribution shift* from the data to the policy-induced distribution.

In contrast, we directly focus on requiring the model to perform well under unknown distributions instead of the data distribution. In other words, we are particularly interested in developing approaches that directly model the batch (offline) 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., 2020), which we build upon to develop our approach.

Preliminary work. Under review by AISTATS 2021. Do not distribute.

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) 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:

$$J(\pi, P) \equiv E_{s\sim d_0}[V_\pi^P(s)], \tag{2}$$

where, by the Bellman Equation,

$$V_\pi^P(s) \equiv E_{a\sim\pi(\cdot|s)}[E_{r\sim\mathcal{R}(\cdot|s,a)}[r] + \gamma E_{\tilde{s}\sim P(\cdot|s,a)}[V_\pi^P(\tilde{s})]]. \tag{3}$$

A useful equivalent measure of performance is:

$$J(\pi, P) = E_{(s,a,r)\sim d^P_{\pi,\gamma} R(s,a)}[r], \tag{4}$$

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

$$d^P_{\pi,t}(s, a) = \int d^P_{\pi,t-1}(\tilde{s}, \tilde{a}) P(s|\tilde{s}, \tilde{a}) \pi(a|s) d\nu(\tilde{s}, \tilde{a}), \tag{5}$$

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.

Note. Any Lemmas or Theorems presented without proof have full proofs in the Appendix.

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

3.1 Natural Derivation

We start with the off-policy evaluation (OPE) learning objective and derive the MML loss (Def 3.1). In Section 4, we show the loss also bounds off-policy optimization (OPO) error through its connection with OPE.

OPE Decision Problem. The OPE objective is to estimate:

$$J(\pi, P^*) \equiv E\left[\sum_{i=0}^\infty \gamma^i r_i \middle| \begin{smallmatrix} s_0\sim d_0 \\ a_i\sim\pi(\cdot|s_i) \\ s'_{i+1}\sim P^*(\cdot|s_i,a_i) \\ r_i\sim\mathcal{R}(\cdot|s,a) \end{smallmatrix}\right], \tag{6}$$

the performance of an evaluation policy π in the true environment P^*, using only logging data D with samples from $D_{\pi_b} P^*$. Solving this objective 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 shift can lead to poor estimation.

Model-Based OPE. Given a model class $\mathcal{P}$ and a desired evaluation policy π, we want to find a simulator $\widehat{P} \in \mathcal{P}$ using only logging data D such that:

$$\widehat{P} = \arg\min_{P\in\mathcal{P}} |J(\pi, P) - J(\pi, P^*)|. \tag{7}$$

Interpreting Eq. (7), 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^*.

Derivation. Using (2) and (4), we have:

$$\begin{aligned}\delta_\pi^{P,P^*} &= J(\pi, P) - J(\pi, P^*) \\ &= E_{s\sim d_0}[V_\pi^P(s)] - E_{(s,a,r)\sim d^{P^*}_{\pi,\gamma}(\cdot,\cdot)\mathcal{R}(\cdot|s,a)}[r].\end{aligned}$$

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

$$\begin{aligned}\delta_\pi^{P,P^*} &= E_{s\sim d_0}[V_\pi^P(s)] - E_{(s,a)\sim d^{P^*}_{\pi,\gamma}}[V_\pi^P(s)] &(8)\\ &\quad + E_{(s,a)\sim d^{P^*}_{\pi,\gamma}}[V_\pi^P(s) - E_{r\sim\mathcal{R}(\cdot|s,a)}[r]]. &(9)\end{aligned}$$

To simplify the above expression, we make the following observations. First, Eq. (9) can be simplified through the Bellman equation from Eq. (3). To see this, notice that $d^{P^*}_{\pi,\gamma}$ is equivalent to some $d(s)\pi(a|s)$ for an appropriate choice of $d(s)$. Thus,

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

Second, we can manipulate Eq. (8) using the definition

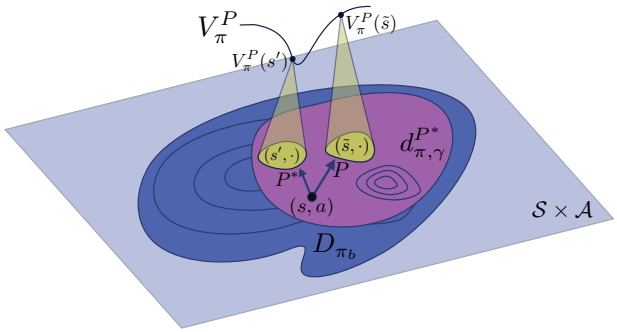

Figure 1: *A visual representation of the model-based OPE objective in Eq.* (10)*. The error at every point* (s,a) *in* D_{π_b} *is the difference between* $V_\pi^P(\tilde{s})$ *(induced by following* P*) and* $V_\pi^P(s')$ *(induced by following* P^**). We weight the over/under-counting of some points* (s,a) *in* D_{π_b} *to make it look like the data was actually generated by* $d^{P^*}_{\pi,\gamma}$*. Summing up all the errors exactly yields the OPE error of using* P *as a simulator.*

of $d^P_{\pi,\gamma}$ and recursive property of $d^P_{\pi,t}$ from Eq. (5):

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

Combining the above allows us to succinctly express:

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

Since D contains samples from D_{π_b} and not $d^{P^*}_{\pi,\gamma}$, we use importance sampling to simplify the right-hand side of δ_π^{P,P^*} to:

$$
\gamma \mathop{E}_{(s,a,s')\sim D_{\pi_b}P^*}\left[\frac{d^{P^*}_{\pi,\gamma}}{D_{\pi_b}}\left(\mathop{E}_{\tilde{s}\sim P(\cdot|s,a)}[V_\pi^P(\tilde{s})] - V_\pi^P(s')\right)\right]. \tag{10}
$$

Figure 1 gives a visual illustration of Eq. (10).

Define $w_\pi^P(s,a) \equiv \frac{d^P_{\pi,\gamma}(s,a)}{D_{\pi_b}(s,a)}$. 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 as:

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

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

When unambiguous, we will drop the MML subscript.

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.

With this intuition, we can formally guarantee that $J(\pi,P)\approx J(\pi,P^*)$ under the following *realizability conditions*:

Assumption 1 (Adequate Support)**.** $D_{\pi_b}(s,a)>0$ *whenever* $d^P_{\pi,\gamma}(s,a)>0$*. Define* $w_\pi^P(s,a)\equiv\frac{d^P_{\pi,\gamma}(s,a)}{D_{\pi_b}(s,a)}$.

Assumption 2 (OPE Realizability)**.** *For a given* π*,* $\mathcal{W}\times\mathcal{V}$ *contains at least one of* $(w_\pi^P, V_\pi^{P^*})$ *or* $(w_\pi^{P^*}, V_\pi^P)$ *for every* $P\in\mathcal{P}$*.*

Theorem 3.1 (MML & OPE)**.** *Under Assumption 2,*

$$
|J(\pi,\widehat{P}) - J(\pi,P^*)| \le \gamma \min_{P\in\mathcal{P}}\max_{w\in\mathcal{W},V\in\mathcal{V}}|\mathcal{L}(w,V,P)|, \tag{11}
$$

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. *We want to choose* $\mathcal{V},\mathcal{W},\mathcal{P}$ *carefully so that many* $P\in\mathcal{P}$ *satisfy* $\mathcal{L}(w,V,P)=0$ *and Assumption 2. By inspection,* $\mathcal{L}(w,V,P^*)=0$ *for any* $V\in\mathcal{V}, w\in\mathcal{W}$*.*

Remark 3.3. *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^**. In principle, we may redesign* $\mathcal{V}$ *to guarantee this condition.*

Remark 3.4. *When* $\gamma=0$*,* J *does not depend on a transition function, so* $J(\pi,P)=J(\pi,P^*)\ \forall P\in\mathcal{P}$*.*

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

Definition 3.2 (Minimax Model Learning (MML))**.**

$$
\widehat{P} = \arg\min_{P\in\mathcal{P}}\max_{w\in\mathcal{W},V\in\mathcal{V}}|\mathcal{L}_{MML}(w,V,P)|. \tag{12}
$$

3.2 Interpretation and Verifiability

Theorem 3.1 quantifies the error incurred by evaluating π in $\widehat{P}$ instead of P^*, assuming Assumption 2 holds. For OPE, $\widehat{P}$ is a reasonable proxy for P. In this sense, MML is a principled method approach for model-based OPE. See Appendix B.1 for a complete proof of Thm 3.1 and Appendix B.2 for the sample complexity analysis.

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 3.3, 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.

3.3 Comparison to Model-Free OPE

Recent model-free off-policy policy evaluation (OPE) literature (e.g., Liu et al., 2018; Uehara et al., 2020) has similar realizability assumptions to Assumption 2.

As an example, the method MWL (Uehara et al., 2020) takes the form of

$$J(\pi, P^*) \approx E_{(s,a,r)\sim D_{\pi_b}}[\widehat{w}(s,a)r]$$
$$\text{where } \widehat{w} = \arg\min_{w\in\mathcal{W}} \max_{Q\in\mathcal{Q}} |\mathcal{L}_{MWL}(w, Q)|,$$

requiring at least one of $w_\pi^{P^*}$ or $Q_\pi^{P^*}$ to be realized to control the OPE error. Here $\mathcal{Q}$ is analogous to our function class $\mathcal{V}$ where $E_{a\sim\pi(a|s)}[Q_\pi^{P^*}(s,a)] = V_\pi^{P^*}(s)$. The loss $\mathcal{L}_{MWL}$ has no dependence on P and is therefore model-free. MQL (Uehara et al., 2020) has analogous realizability conditions to MWL.

Our loss, $\mathcal{L}_{MML}$, has the same realizability assumptions in addition to one related to $\mathcal{P}$ (and not $\mathcal{P}^*$). As discussed in Remark 3.3, these $\mathcal{P}$-related assumptions can be verified a priori and in principle, satisfied by redesigning the function classes. Therefore, they do not pose a substantial theoretical challenge. See Section 6 for a practical discussion.

An advantage of model-free approaches is that when both $w_\pi^{P^*}, Q_\pi^{P^*}$ 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 4) 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, $E_{\text{data}}[wV]$ $(= E_{(s,a)\sim D_{\pi_b}}[\frac{d_{\pi,\gamma}^{P^*}}{D_{\pi_b}} V_\pi^P(s)])$, ties together the OPE estimate.

3.4 Misspecification of $\mathcal{P}, \mathcal{V}, \mathcal{W}$

Suppose Assumption 2 does not hold and $P^* \notin \mathcal{P}$. Define a new function $h(s,a,s') \in \mathcal{H} = \{w(s,a)V(s') | (w,V) \in \mathcal{W} \times \mathcal{V}\}$ then we redefine $\mathcal{L}$:

$$\mathcal{L}(h,P) = E_{(s,a,s')\sim D_{\pi_b}(\cdot,\cdot)P^*(\cdot|s,a)}[$$
$$E_{x\sim P(\cdot|s,a)}[h(s,a,x)] - h(s,a,s')].$$

Proposition 3.5 (Misspecification discrepancy for OPE)**.** *Let $\mathcal{H} \subset (\mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R})$ be a set of functions on (s,a,s'). Denote $(WV)^* = w_\pi^{P^*}(s,a)V_\pi^P(s')$ (or, equivalently, $(WV)^* = w_\pi^P(s,a)V_\pi^{P^*}(s')$).*

$$|J(\pi,\widehat{P}) - J(\pi,P^*)| \le \gamma \min_P \max_{h\in\mathcal{H}} |\mathcal{L}(h,P)| + \gamma\epsilon_{\mathcal{H}}, \tag{13}$$

where $\epsilon_{\mathcal{H}} = \max_{P\in\mathcal{P}} \min_{h\in\mathcal{H}} |\mathcal{L}((WV)^* - h, P)|$.

$\mathcal{L}(WV^* - h, P)$ measures the difference between h and $(WV)^*$. Another interpretation of Prop 3.5 is if $\arg\max_{\mathcal{H}\cup\{(WV)^*\}} \mathcal{L}(h,P) = (WV)^*$ for some $P \in \mathcal{P}$ then MML returns a value $\gamma\epsilon_{\mathcal{H}}$ below the true upper bound, otherwise the output of MML remains the upperbound. This result illustrates that realizability is sufficient but not necessary for MML to be an upperbound on the loss.

3.5 Application to the Online Setting

While the main focus of MML is batch OPE and OPO, we will make a few remarks relating to the online setting. In particular, if we assume we can engage in online data collection then $\mathcal{W} = \{1\}$, representing no distribution shift since $\pi_b = \pi$. When VAML and MML share the same function class V, we can show that $\min_{\mathcal{P}} \max_{\mathcal{W},\mathcal{V}} \mathcal{L}_{MML}(w,V,P)^2 \le \min_P \mathcal{L}_{VAML}(\mathcal{V},P)$ for any $\mathcal{V}, \mathcal{P}$. In other words, MML is a tighter decision-aware loss even in online data collection. In addition, MML enables greater flexibility in the choice of $\mathcal{V}$. See Appendix B.4 for further details.

4 Off-Policy Optimization (OPO)

4.1 Natural Derivation

In this section we examine how our MML approach can be integrated into 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^*.

OPO Decision Problem. Given a policy class Π and access to only a logging dataset D with samples from $D_{\pi_b}P^*$, find a policy $\pi \in \Pi$ that is competitive with the unknown optimal policy $\pi_{P^*}^*$:

$$\widehat{\pi}^* = \arg\min_{\pi\in\Pi} |J(\pi,P^*) - J(\pi_{P^*}^*, P^*)|. \tag{14}$$

Note: No additional exploration is allowed.

Model-Based OPO. Given a model class $\mathcal{P}$, we want to find a simulator $\widehat{P} \in \mathcal{P}$ using only logging data D and subsequently learn $\pi^*_{\widehat{P}} \in \Pi$ in $\widehat{P}$ through any policy optimization algorithm which we call Planner$(\cdot)$.

Algorithm 1 Standard Model-Based OPO

Input: $D = D_{\pi_b} P^*$, Modeler, Planner
1: Learn $\widehat{P} \leftarrow$ Modeler(D)
2: Learn $\widehat{\pi}^*_P \leftarrow$ Planner$(\widehat{P})$
3: **return** $\widehat{\pi}^*_P$

In Algorithm 1, Modeler$(\cdot)$ refers to any (batch) model learning procedure. The hope for model-based OPO is that the ideal in-simulator policy $\pi^*_{\widehat{P}}$ and the actual best (true environment) policy $\pi^*_{P^*}$ perform competitively: $J(\pi^*_{\widehat{P}}, P^*) \approx J(\pi^*_{P^*}, P^*)$. Hence, instead of minimizing Eq (14) over all $\pi \in \Pi$, we can focus $\Pi = \{\pi^*_P\}_{P \in \mathcal{P}}$.

Derivation. 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 can drop it in an upper bound. 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 4.1 (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 $P \in \mathcal{P}$.

4.2 Interpretation and Verifiability

Theorem 4.1 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 Theorem 4.1 (see Appendix C.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 4.1 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 $w \in \mathcal{W}, V \in \mathcal{V}$.

We could have instead examined the quantity $\min_\pi |J(\pi^*_{P^*}, P^*) - J(\pi, P^*)|$ directly from Eq (14). What we would find is that the upper bound is $2 \min_P \max_{w,V} |E_{d_0}[V] - \mathcal{L}(w, V, P)|$ and the realizability requirements would be that $V^P_\pi \in \mathcal{V}, w^{P^*}_\pi \in \mathcal{W}$ for every π in some policy class. This is a much stronger requirement than in Theorem 4.1.

For OPO, apriori verification of realizability is possible by enumerating over $P \in \mathcal{P}$. Whereas the target policy π was fixed in OPE, now π^*_P varies for each $P \in \mathcal{P}$. It may be more practical to, as in OPE, perform post-verification that $w^P_{\pi^*_{\widehat{P}}} \in \mathcal{W}$ and $V^P_{\pi^*_{\widehat{P}}} \in \mathcal{V}$. If they do not hold, then we can modify the function classes until they do. This greatly relaxes the "for every $P \in \mathcal{P}$" condition and leaves only a few unverifiable quantities relating to P^*.

Sample complexity and function class misspecification results for OPO can be found in Appendix C.2, C.3.

4.3 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.

5 Scenarios & Considerations

In this section we investigate a few scenarios where we can calculate the class $\mathcal{V}$ and $\mathcal{W}$ or modify the loss based on structured knowledge of $\mathcal{P}, \mathcal{W}$, and $\mathcal{V}$.

In examining the scenarios, we aim to verify that MML gives *sensible* results. For example, in scenarios where we know MLE to be optimal, MML should coincide. Indeed, we show this to be the case for the tabular function class and Linear-Quadratic Regulators. Other scenarios include showing that MML is compatible with incorporating prior knowledge using either a nominal dynamics model or a kernel.

The proofs for any Lemmas in this section can be found in Appendix E.

5.1 Linear & Tabular Function Classes

When $\mathcal{W}, \mathcal{V}, \mathcal{P}$ are linear function classes then the entire minimax optimization has a closed form solution. In particular, $\mathcal{P}$ takes the form $P = \phi(s,a,s')^T\alpha$ where $\phi \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}\times\mathcal{S}|}$ is some basis of features with $\alpha \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}\times\mathcal{S}|}$ its parameters and $(w(s,a), V(s')) \in \mathcal{WV} = \{\psi(s,a,s')^T\beta : \|\beta\|_\infty < +\infty\}$ where $\psi \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}\times\mathcal{S}|}$.

Proposition 5.1 (Linear Function classes)**.** *Let $P = \phi(s,a,s')^T\alpha$ where $\phi \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}\times\mathcal{S}|}$ is some basis of features with α its parameters. Let $(w(s,a), V(s')) \in \mathcal{WV} = \{\psi(s,a,s')^T\beta : \|\beta\|_\infty < +\infty\}$. Then,*

$$\widehat{\alpha} = E_n^{-T}\left[\int \phi(s,a,s')\psi(s,a,s')^T d\nu(s')\right] E_n[\psi(s,a,s')], \tag{15}$$

if $E_n\left[\int \phi(s,a,s')\psi(s,a,s')^T d\nu(s')\right]$ has full rank.

The tabular setting, when the state-action space is finite, is a common special case. We can choose

$$\psi(s,a,s') = \phi(s,a,s') = e_i \tag{16}$$

as the ith standard basis vector where $i = s|\mathcal{A}||\mathcal{S}| + a|\mathcal{S}| + s'$. There is no model misspecification in the tabular setting (i.e., $P^* \in \mathcal{P}$), therefore $\widehat{P} = P^*$ in the case of infinite data.

Proposition 5.2 (Tabular representation)**.** *Let $P = \phi(s,a,s')^T\alpha$ with $\phi \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}\times\mathcal{S}|}$ as in Eq (16) and α its parameters. Let $(w(s,a), V(s')) \in \mathcal{WV} = \{\phi(s,a,s')^T\beta : \|\beta\|_\infty < +\infty\}$. Assume we have at least one data point from every (s,a) pair. Then*

$$\widehat{P}_n(s'|s,a) = \frac{\#\{(s,a,s') \in D\}}{\#\{(s,a,\cdot) \in D\}}. \tag{17}$$

Prop. 5.2 shows that MML and MLE coincide, even in the finite-data regime. Both models are simply the observed propensity of entering state s' from tuple (s,a).

5.2 Linear Quadratic Regulator (LQR)

The Linear Quadratic Regulator (LQR) is defined as linear transition dynamics $P^*(s'|s,a) = A^*s + B^*a + w^*$ where w^* is random noise and a quadratic reward function $\mathcal{R}(s,a) = s^TQs + a^TRa$ for $Q, R \geq 0$ symmetric positive semi-definite. For ease of exposition we assume that $w^* \sim N(0, \sigma^{*2}I)$. We assume that (A^*, B^*) is controllable. Exploiting the structure of this problem, we can check that every $V \in \mathcal{V}$ takes the form $V(s) = s^TUs + q$ for some symmetric semi-positive definite U and constant q (Appendix Lemma E.1).

Furthermore, we know controllers of the form $\pi(a|s) = -Ks$ where $K \in \mathbb{R}^{k\times n}$ are optimal in LQR (Bertsekas et al., 2005). We consider determistic and therefore *misspecified* models of the form $P(s'|s,a) = As + Ba$. $\mathcal{W}$ is a Gaussian mixture and we can write $\mathcal{L}_{MML}$ as a function of U, K and (A,B) (Appendix E.2).

Proposition 5.3 (MML + MLE Coincide for LQR)**.** *Let $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times k}, K \in \mathbb{R}^{k\times n}$. Let $U \in \mathcal{S}^n$ be positive semi-definite. Set $k = 1$, a single input system. Then,*

$$\begin{aligned}\arg\min_{(A,B)}\max_{K,U}|\mathcal{L}_{MML}(K,U,(A,B))| &= (A^*,B^*)\\ &= \arg\min_{(A,B)}\mathcal{L}_{MLE}(A,B).\end{aligned}$$

Despite model misspecification and representing two different loss functions, both MLE and MML give the correct parameters $(\widehat{A}, \widehat{B}) = (A^*, B^*)$. We leave showing that MML and MLE coincide in multi-input $(k > 1)$ LQR systems for future work.

5.3 Residual Dynamics & Environment Shift

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:

Definition 5.1. [Residual MML Loss] For $w \in \mathcal{W}, V \in \mathcal{V}, P \in \mathcal{P}$,

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

This solution form matches the intuition that having prior knowledge in the form of P_0 focuses the learning objective on the difference between P^* and P_0.

5.4 Incorporating Kernels

Our approach is also compatible with incorporating kernels (which is a way of encoding domain knowledge such as smoothness) to learn in a Reproducing Kernel Hilbert Space (RKHS). For example, we may derive a closed form for $\max_{(w,V)\in\mathcal{WV}}\mathcal{L}(w,V,P)^2$ when $\mathcal{W}\times\mathcal{V}$ corresponds to an RKHS and use standard gradient descent to find $\widehat{P} \in \mathcal{P}$, making the minimax problem much more tractable. See Appendix E.3 for a detailed discussion on RKHS, computational issues relating to sampling from P and alternative approaches to solving the minimax problem.

6 Experiments

In our experiments, we seek to answer the following questions: (1) Does MML prefer models that minimize the OPE objective? (2) What can we expect when we

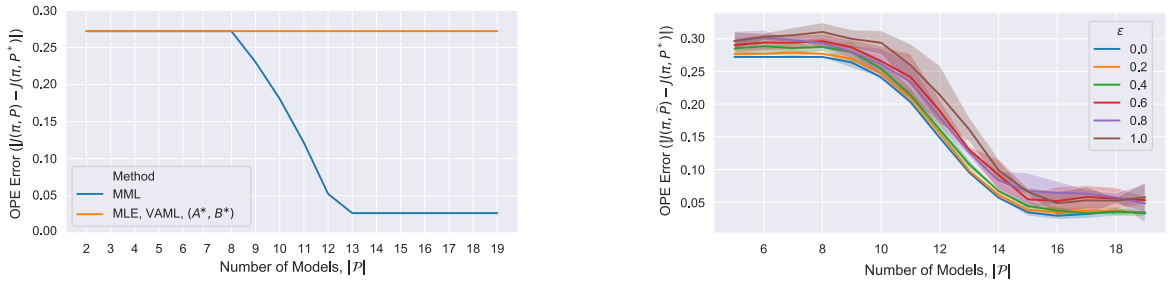

Figure 2: *LQR. (Left, OPE Error)* MML finds the $P \in \mathcal{P}$ with the lowest OPE error as $\mathcal{P}$ gets richer. Since calculations are done in expectation, no error bars are included. *(Right, Verifiability)* The OPE error (smoothed) increases with misspecification in $\mathcal{V}$ parametrized by ϵ, the expected MSE between the true $V_\pi^{P^*} \notin \mathcal{V}$ and the approximated $\widehat{V}_\pi^{P^*} \in \mathcal{V}$. Nevertheless, directionally they all follow the same trajectory as $\mathcal{P}$ gets richer.

have misspecification in $\mathcal{V}$? (3) How does MML perform against MLE and VAML in OPE? (4) Does our approach complement modern offline RL approaches? For this last question, we consider integrating MML with the recently proposed MOREL (Kidambi et al., 2020) approach for offline RL. See Appendix F.3 for details on MOREL.

6.1 Brief Environment Description/Setup

We perform our experiments in three different domains. A thorough description of the environments and setup can be found in Appendix F.

Linear-Quadratic Regulator (LQR). The LQR domain is a 1D environment with stochastic dynamics $P^*(s'|s,a)$. We use a finite class $\mathcal{P}$ consisting of deterministic policies.

Cartpole (Brockman et al., 2016). The reward function is modified to be a function of angle and location rather than 0/1 to make the OPE problem more challenging. Each $P \in \mathcal{P}$ is a parametrized NN that outputs a mean, and logvariance representing a normal distribution around the next state.

Inverted Pendulum (IP) (Dorobantu & Taylor, 2020). This IP environment has a Runge-Kutta(4) integrator rather than Forward Euler (Runge-Kutta(1)) as in OpenAI (Brockman et al., 2016), producing significantly more realistic data. Each $P \in \mathcal{P}$ is a deterministic model parametrized with a neural network.

6.2 Results

Does MML prefer OPE? We vary the size of the model class Figure 2 (left) testing to see if MML will pick up on the models which have better OPE performance. When the sizes of $|\mathcal{P}|$ are small, each method selects (A^*, B^*) (e.g. $P(s'|s,a) = A^*s + B^*a$), the deterministic version of the optimal model. However, as we increase the richness of $\mathcal{P}$, MML begins to pick up on models that are able to better evaluate π.

Two remarks are in order. In LQR, policy optimization in (A^*, B^*) coincides with policy optimization in P^*. Therefore, if we tried to do policy optimization in our selected model then our policy would be suboptimal in P^*. Secondly, MML deliberately selects a model other than (A^*, B^*) because a good OPE estimate relies on appoximating the contribution from the stochastic part of P^*.

A Tradeoff? There is a trade-off between the OPE objective and the OPO objective. MML's preference is dependent on the capacities of $\mathcal{P}, \mathcal{W}, \mathcal{V}$. Figure 2 (left) illustrates OPE is preferred for $\mathcal{W}$ fixed. Appendix Figure 5 explores the OPO objective and shows that if we increase $\mathcal{W}$ then OPO becomes favored. One interpretation of this is we are asking MML to be robust to many more OPE problems as $|\mathcal{W}| \uparrow$ and therefore the performance on any single one decreases but overall we are more likely to be able to do OPO.

Misspecification and Verifiability? To check verifiability in practice, we would run π in a few $P \in \mathcal{P}$ and calculate V_π^P. Then we would check if $V_\pi^P \in \mathcal{V}$ by fitting $\widehat{V}_\pi^P$ and measuring the empirical gap $E[(\widehat{V}_\pi^P - V_\pi^P)^2] = \epsilon^2$. *Do we have to close this gap?*

Figure 2 (right) shows how MML performs when $V_\pi^P \notin \mathcal{V}$ but we do have $\widehat{V}_\pi^P(s) = V_\pi^P(s) + \mathcal{N}(0, \epsilon) \in \mathcal{V}$. Since $E[(\widehat{V}_\pi^P - V_\pi^P)^2] = \epsilon^2$ then ϵ is the root-mean squared error between the two functions. Directionally all of the errors go down as $|\mathcal{P}| \uparrow$, however it is clear that ϵ has a noticeable effect. We speculate that if this error not distributed around zero and instead is dependent on the state then the effects can be worse.

MML for OPE? In addition to Figure 2 (left), Figure 3 also illustrates that our method outperforms the other model-learning approaches in OPE. The envi-

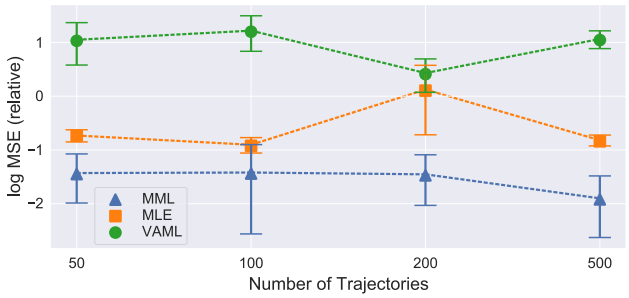

Figure 3: *(Cartpole, OPE Error) Comparison of model-based approaches for OPE with function-approx. Lower is better. MML outperforms others.*

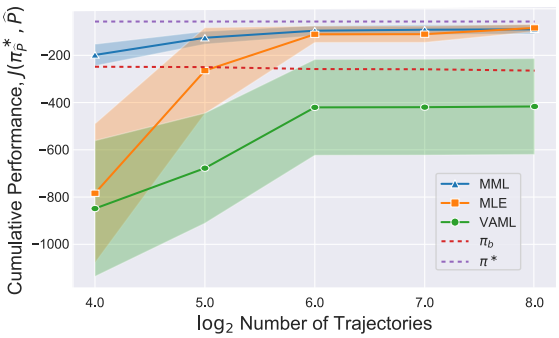

Figure 4: *(Invert. Pend., OPO Performance) Comparison of model-based approaches for OPO with function-approx using Algorithm 2. Higher is better. MML performs competitively even in low data regimes.*

ronment and reward function is challenging, requiring function approximation. This further validates that MML is a good choice for model-learning with an OPE objective. Despite the added complexity of solving a minimax problem, doing so gives nearly an order of magnitude improvement over MLE and many orders over VAML.

Algorithm 2 OPO Algorithm (based on MOREL (Kidambi et al., 2020))

Input: D, $\mathcal{L}$ among {MML, MLE, VAML}

1: Learn an ensemble of dynamics $P_1, \ldots, P_4 \in \mathcal{P}$ using $P_i = \arg\min_{P \in \mathcal{P}} \mathcal{L}(D)$

2: Construct a pessimistic MDP $\mathcal{M}$ (see Appendix F.3) with $P(s,a) = \frac{1}{4}\sum_{i=1}^{4} P_i(s,a)$.

3: $\widehat{\pi} \leftarrow$ PPO($\mathcal{M}$) (Best of 3) (Schulman et al., 2017)

MML for OPO? We integrate MML, VAML, and MLE with MOREL as in Algorithm 2. Consequently, Figure 4 shows that MML performs competitively with the other methods, achieving near-optimal performance as the number of trajectories increases. MML has good performance even in the low-data regime, whereas other methods perform worse than π_b. Performance in the low-data regime is of particular interest since sample efficiency is highly desirable.

Algorithm 2 forms a pessimistic MDP where a policy is penalized if it enters a state where there is disagreement between $P_1, \ldots, P_4$. Given that MML performs well in low-data, we can reason that MML produces models with support that stays within the dataset D or generalize well slightly outside this set. The other models poor performance is suggestive of incorrect over-confidence outside of D and PPO produces a policy which takes advantage of this.

7 Other Related Work

Minimax and Model-Based RL. Rajeswaran et al. (2020) introduce an iterative minimax approach to simultaneously find the optimal-policy and a model of the environment. Despite distribution-shift correction, online data collection is required and is not comparable to MML, where we focus on the batch setting.

Batch (Offline) Model-Based RL Recent improvements in batch model-based RL focus primarily on the issue of policies taking advantage of errors in the model (Kidambi et al., 2020; Deisenroth & Rasmussen, 2011; Chua et al., 2018; Janner et al., 2019). These improvements typically involve uncertainty quantification to keep the agent in highly certain states to avoid model exploitation. These improvements are independent of the loss function involved.

8 Discussion and Future Work

We have presented a novel approach to learning a model for batch, off-policy model-based reinforcement learning. Our approach follows naturally from the definitions of the OPE and OPO objectives and enjoys distributional robustness and decision-awareness. We examined different scenarios under which our method coincided with other methods as well as when closed form solutions were available. We provided sample complexity analysis and misspecification analysis. Finally, we empirically validated that our method was competitive with current model learning approaches.

A key component throughout this paper has been the function class $\mathcal{W} \times \mathcal{V}$. Finding other interpretations for this term may prove to be useful outside of MML and is of interest in future work. Furthermore, MML remains part of a two-step OPO pipeline: first learn the model, then return the optimal policy in that model. Another direction of future research is to have a single-shot batch OPO objective that returns both a model and the optimal policy simultaneously, in effect combining MML with the minimax algorithm in Rajeswaran et al. (2020).

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*, Berlin, Heidelberg, 2001. Springer-Verlag. ISBN 3540423435.

Bertsekas, D. P., Bertsekas, D. P., Bertsekas, D. P., and Bertsekas, D. P. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 2005.

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*, 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*. 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*. 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*, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

Dorobantu, V. and Taylor, A. Lyapy. `https://github.com/vdorobantu/lyapy`, 2020.

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*, 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*, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

Feng, Y., Li, L., and Liu, Q. A kernel loss for solving the bellman equation. In *Advances in Neural Information Processing Systems*, 2019.

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*, 2672–2680. Curran Associates, Inc., 2014.

Janner, M., Fu, J., Zhang, M., and Levine, S. When to trust your model: Model-based policy optimization. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, 12519–12530. Curran Associates, Inc., 2019.

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. Morel : Model-based offline reinforcement learning, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

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*, 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.

Raffin, A., Hill, A., Ernestus, M., Gleave, A., Kanervisto, A., and Dormann, N. Stable baselines3. `https://github.com/DLR-RM/stable-baselines3`, 2019.

Rajeswaran, A., Mordatch, I., and Kumar, V. A game theoretic framework for model based reinforcement learning, 2020.

Schaefer, F. and Anandkumar, A. Competitive gradient descent. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, 7625–7635. Curran Associates, Inc., 2019.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

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*. Morgan Kaufmann, 1990.

Uehara, M., Huang, J., and Jiang, N. Minimax Weight and Q-Function Learning for Off-Policy Evaluation. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Contents

A Glossary of Terms

Table 1: Glossary of terms

Acronym	Term
OPE	Off Policy (Policy) Evaluation
OPO	Off Policy (Policy) Optimization. Also goes by batch off-policy reinforcement learning.
$\mathcal{S}$	State Space
$\mathcal{A}$	Action Space
P	Transition Function
P^*	True Transition Function
$\mathcal{R}$	Reward Function
$\mathcal{X}$	State-Action Space $\mathcal{S} \times \mathcal{A}$
γ	Discount Factor
π	Policy
$J(\pi, P)$	Performance of π in P
V_π^P	Value Function of π with respect to P
d_0	Initial State Distribution
$d_\pi^{P,\gamma}$	(Discounted) Distribution of State-Action Pairs Induced by Running π in P
w_π^P	Distribution Shift ($w_\pi^P(s,a) = \frac{d_\pi^{P,\gamma}(s,a)}{D_{\pi_b}(s,a)}$)
ν	Lebesgue measure
d_{π_b}	Behavior state distribution
π_b	Behavior policy
D_{π_b}	Behavior data ($d_{\pi_b}\pi_b$)
D	Dataset containing samples from $D_{\pi_b}P^*$
$E_n[\cdot]$	Empirical approximation using D
$E[\cdot]$	Exact expectation
$\mathcal{W}$	Distribution Shifts Function Class (e.g. $\frac{d_\pi^P(s,a)}{D_\pi(s,a)}$)
$\mathcal{V}$	Value Function Class (e.g. $V_\pi^P \in \mathcal{V}$)
$\mathcal{P}$	Model Function Class (e.g. $P \in \mathcal{P}$)
$\mathcal{L}$	Model Learning Loss Function
$\widehat{P}$	Best Model w.r.t $\mathcal{L}$
$\epsilon_{\mathcal{H}}$	Misspecification Error
π_P^*	Optimal Policy in P
RKHS	Reproducing Kernel Hilbert Space
LQR	Linear Quadratic Regulator
IP	Inverted Pendulum
MML	Minimax Model Learning (Ours)
MLE	Maximum Likelihood Estimation
VAML	Value-Aware Model Learning

B OPE

In this section we explore the OPE results in the order in which they were presented in the main paper.

B.1 Main Result

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

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

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

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

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,

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

$$
\begin{aligned}
&= \gamma \sum_{t=0}^{\infty} \gamma^t \int d^P_{\pi,t}(s,a) P(s'|s,a) V_\pi^{P^*}(s') d\nu(s,a,s') - \gamma E_{(s,a)\sim d^P_{\pi,\gamma}}[E_{s'\sim P^*(\cdot|s,a)}[V_\pi^{P^*}(s')]] \\
&= \gamma E_{(s,a)\sim d^P_{\pi,\gamma}}[E_{s'\sim P(\cdot|s,a)}[V_\pi^{P^*}(s')]] - \gamma E_{(s,a)\sim d^P_{\pi,\gamma}}[E_{s'\sim P^*(\cdot|s,a)}[V_\pi^{P^*}(s')]] \\
&= \gamma E_{(s,a)\sim d^P_{\pi,\gamma}}[E_{s'\sim P(\cdot|s,a)}[V_\pi^{P^*}(s')] - E_{s'\sim P^*(\cdot|s,a)}[V_\pi^{P^*}(s')]] \\
&= \gamma E_{(s,a,s')\sim D_{\pi_b} P^*(\cdot|s,a)}[\frac{d^P_{\pi,\gamma}(s,a)}{D_{\pi_b}(s,a)} \left(E_{x\sim P(\cdot|s,a)}[V_\pi^{P^*}(x)] - V_\pi^{P^*}(s')\right)]] \\
&= \gamma E_{(s,a,s')\sim D_{\pi_b} P^*(\cdot|s,a)}[w^P_\pi(s,a) \left(E_{x\sim P(\cdot|s,a)}[V_\pi^{P^*}(x)] - V_\pi^{P^*}(s')\right)]] \\
&= \gamma \mathcal{L}(w^P_\pi, V_\pi^{P^*}, P),
\end{aligned}
$$

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

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

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

B.2 Sample Complexity for OPE

We do not have access to exact expectations, so we must work with $\widehat{P}_n = \arg\min_P \max_{w,V} E_n[\ldots]$ instead of $\widehat{P} = \arg\min_P \max_{w,V} E[\ldots]$. 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 B.1. *[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 Theorem 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 B.1. By definition and triangle inequality,

$$
\begin{aligned}
|J_{T,m}(\pi, \widehat{P}_n) - J(\pi, P^*)| &= |J_{T,m}(\pi, \widehat{P}_n) - J(\pi, \widehat{P}_n) + J(\pi, \widehat{P}_n) - J(\pi, P^*)| \\
&\leq \underbrace{|J_{T,m}(\pi, \widehat{P}_n) - J(\pi, \widehat{P}_n)|}_{(a)} + \underbrace{|J(\pi, \widehat{P}_n) - J(\pi, P^*)|}_{(b)}
\end{aligned} \tag{18}
$$

Define $\widehat{V}^P_{\pi,T}(s_0^i) \equiv \sum_{t=0}^{T} \gamma^t r_t^i$ 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),

$$
\begin{aligned}
|J_{T,m}(\pi, \widehat{P}_n) - J(\pi, \widehat{P}_n)| &= \left| \frac{1}{m}\sum_{i=1}^{m} \widehat{V}^{\widehat{P}_n}_{\pi,T}(s_0^i) - \frac{1}{m}\sum_{i=1}^{m} \widehat{V}^{\widehat{P}_n}_{\pi,\infty}(s_0^i) + \frac{1}{m}\sum_{i=1}^{m} \widehat{V}^{\widehat{P}_n}_{\pi,\infty}(s_0^i) - E_{d_0}[V_\pi^{\widehat{P}_n}] \right| \\
&\leq \left| \frac{1}{m}\sum_{i=1}^{m} \widehat{V}^{\widehat{P}_n}_{\pi,T}(s_0^i) - \frac{1}{m}\sum_{i=1}^{m} \widehat{V}^{\widehat{P}_n}_{\pi,\infty}(s_0^i) \right| + \left| \frac{1}{m}\sum_{i=1}^{m} \widehat{V}^{\widehat{P}_n}_{\pi,\infty}(s_0^i) - E_{d_0}[V_\pi^{\widehat{P}_n}] \right| \\
&\leq \frac{2R_{\max}}{1-\gamma}\gamma^{T+1} + \frac{2R_{\max}}{1-\gamma}\sqrt{\log(2/\delta)/(2m)},
\end{aligned} \tag{19}
$$

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,

$$
\begin{aligned}
&|J(\pi,\widehat{P}_n) - J(\pi,P^*)| \\
&\quad = \gamma |L(w_\pi^{P^*}, V^{\widehat{P}_n}, \widehat{P}_n)| \\
&\quad \leq \gamma \max_{w,V} |L(w,V,\widehat{P}_n)| \\
&\quad = \gamma(\max_{w,V} |L(w,V,\widehat{P}_n)| - \max_{w,V} |L_n(w,V,\widehat{P}_n)| + \max_{w,V} |L_n(w,V,\widehat{P}_n)| - \max_{w,V} |L(w,V,\widehat{P})| + \max_{w,V} |L(w,V,\widehat{P})|) \\
&\quad \leq \gamma(2 \max_{w,V,P} ||L(w,V,P)| - |L_n(w,V,P)|| + \min_P \max_{w,V} |L(w,V,P)|) \\
&\quad \leq \gamma(2\mathfrak{R}'_n(\mathcal{W},\mathcal{V},\mathcal{P}) + 2K\sqrt{\log(2/\delta)/n} + \min_P \max_{w,V} |L(w,V,P)|) \\
&\quad \leq \gamma(4\mathfrak{R}_n(\mathcal{W},\mathcal{V},\mathcal{P}) + 2K\sqrt{\log(2/\delta)/n} + \min_P \max_{w,V} |L(w,V,P)|) \qquad (20)
\end{aligned}
$$

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 $K = 2C_wC_V$ 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. □

The first term can be thought of as the estimate under infinite data, the second term as the penalty for using function classes that are too rich, and the remaining terms as the price we pay for finite data/ finite calculations.

B.3 Misspecification for OPE

When the assumptions behind MML do not hold, our method underbounds the true error. The following is the proof for this Proposition.

Proof for Prop. 3.5. We have shown already that $J(\pi,\widehat{P}) - J(\pi,P^*) = \gamma\mathcal{L}(w_\pi^{P^*}, V_\pi^P, P)$ $(= \gamma\mathcal{L}((WV)^*, P))$. Therefore, by linearity of $\mathcal{L}$ in $\mathcal{H}$, we have

$$
\begin{aligned}
|\mathcal{L}((WV)^*,P)| &= |\mathcal{L}(h,P) + \mathcal{L}((WV)^* - h, P)| \quad \forall h \in \mathcal{H}, P \in \mathcal{P} \\
&\leq |\mathcal{L}(h,P)| + |\mathcal{L}((WV)^* - h, P)| \\
&\leq \min_P \max_h |\mathcal{L}(h,P)| + |\mathcal{L}(h - (WV)^*, P)| \\
&\leq \min_P \max_h |\mathcal{L}(h,P)| + \max_P \min_h |\mathcal{L}((WV)^* - h, P)|
\end{aligned}
$$

where $\epsilon_{\mathcal{H}} = \max_P \min_h |\mathcal{L}((WV)^* - h, P)|$. Therefore $|J(\pi,\widehat{P}) - J(\pi,P^*)| \leq \gamma(\min_P \max_h |\mathcal{L}(h,P)| + \epsilon_{\mathcal{H}})$, as desired. □

B.4 Application to the Online Setting and Brief VAML Comparison

Algorithm 3 is the prototypical online model-based RL algorithm. In contrast to the batch setting, we allow for online data collection. We require a function called PLANNER, which can take a model P_k and find the optimal solution π_k in P_k.

Algorithm 3 Online Model-Based RL

Input: $\pi_0 = \pi_b$. PLANNER$(\cdot)$

1: **for** $k = 0, 1, \ldots, K$ **do**
2: Collect data D_k by interacting with the true environment using π_k.
3: Fit $P_k \leftarrow \arg\min_{P \in \mathcal{P}} \max_{w, V \in \mathcal{W}, \mathcal{V}} \mathcal{L}_{MML}(w, V, P)$ where $D_{\pi_b} = D_k$
4: Fit $\pi_k \leftarrow$ PLANNER(P_k)
5: **return** (P_K, π_K)

Here we show that MML lower bounds the VAML error in online model-based RL, where VAML is designed.

Proposition B.2. *Let* $\mathcal{W} = \{1\}$. *Then*

$$\min_{P \in \mathcal{P}} \max_{w \in \mathcal{W}, V \in \mathcal{V}} \mathcal{L}_{MML}(w, V, P)^2 \leq \min_{P \in \mathcal{P}} \mathcal{L}_{VAML}(\mathcal{V}, P),$$

for every $\mathcal{V}, \mathcal{P}$.

Proof. Pick an arbitrary $P \in \mathcal{P}$. Then, by definition, $\mathcal{L}_{MML}(w, V, P) = E_{(s,a,s') \sim D_{\pi_b} P^*}[w(s,a)(E_{x \sim P(\cdot|s,a)}[V(x)] - V(s'))]$. Since $\mathcal{W} = \{1\}$, then we can eliminate this dependence and get $\mathcal{L}_{MML}(1, V, P) = E_{(s,a,s') \sim D_{\pi_b} P^*}[E_{x \sim P(\cdot|s,a)}[V(x)] - V(s')]$. Explicitly,

$$\begin{aligned}
\mathcal{L}_{MML}(1, V, P)^2 &= \Big(\int \left(\int P(x|s,a) V(x) d\nu(x) - \int P^*(s'|s,a) V(s') d\nu(s') \right) d\nu(s,a) \Big)^2 \\
&= \Big(\int \left(\int (P(x|s,a) - P^*(x|s,a)) V(s') d\nu(x) \right) d\nu(s,a) \Big)^2 \\
&\leq \int \left(\int (P(x|s,a) - P^*(x|s,a)) V(x) d\nu(x) \right)^2 d\nu(s,a), \qquad \text{Cauchy Schwarz}
\end{aligned}$$

Taking the $\max_{V \in \mathcal{V}}$ on both sides and noting $\max_V \int f(V) \leq \int \max_V f(V)$ for any f, V then

$$\begin{aligned}
\max_{V \in \mathcal{V}} \mathcal{L}_{MML}(1, V, P)^2 &\leq \int \max_{V \in \mathcal{V}} \left(\int (P(x|s,a) - P^*(x|s,a)) V(x) d\nu(x) \right)^2 d\nu(s,a) \qquad (21) \\
&= \mathcal{L}_{VAML}(\mathcal{V}, P). \qquad (22)
\end{aligned}$$

Since we chose P arbitrarily, then eq 21 holds for any $P \in \mathcal{P}$. In particular, if $\widehat{P}_{VAML} = \arg\min_{P \in \mathcal{P}} \mathcal{L}_{VAML}(\mathcal{V}, P)$ then

$$\min_{P \in \mathcal{P}} \max_{V \in \mathcal{V}} \mathcal{L}_{MML}(1, V, P)^2 \leq \max_{V \in \mathcal{V}} \mathcal{L}_{MML}(1, V, \widehat{P}_{VAML})^2 \leq \min_{P \in \mathcal{P}} \mathcal{L}_{VAML}(\mathcal{V}, P)$$

□

Prop B.2 reflects that the MML loss function is a tighter loss in the online model-based RL case than VAML. In a sense, this reflects that MML should be the preferred decision-aware loss function even in online model-based RL. An argument in favor of VAML is that it is more computationally tractable given an assumption that $\mathcal{V}$ is the set of linear function approximators. However, if we desire to use more powerful function approximation VAML suffers the same computational issues as MML. In general the pointwise supremum within VAML presents a substantial computational challenge while the uniform supremum from MML is much more mild, can be formulated as a two player game and solved via higher-order gradient descent (see Section E.3).

Lastly, VAML defines the pointwise loss with respect to the L^2 norm of the difference between P and P^*. The choice is justified in that it is computationally friendlier but it is noted that L^1 may also be reasonable (Farahmand et al., 2017). We show in the following example that, actually, VAML may not work with a pointwise L^1 error.

Example B.1. Let $\mathcal{S} = A \cup B$, a disjoint partition of the state space. For simplicity, assume no dependence on actions. Suppose our models $\mathcal{P} = \{P_\alpha\}_{\alpha \in [0,1]}$ take the form

$$P_\alpha(s'|s) = \begin{cases} \alpha & s' \in A \\ 1-\alpha & s' \in B \end{cases}$$

Suppose also that $P^*_{\alpha^*} \in \mathcal{P}$ for some $\alpha^* \in [0,1]$. Let $\mathcal{V} = \{x\mathbf{1}_{s\in A}(s) + y\mathbf{1}_{s\in B}(s) | x, y < M \in \mathbb{R}^+\}$ be all bounded piecewise constant value functions with $\|V\|_\infty = M \in \mathbb{R}^+$. Then the empirical VAML loss with L^1 pointwise distance does not choose P^* when $\alpha \neq \frac{1}{2}$ and cannot differentiate between P^* and any other $P \in \mathcal{P}$ when $\alpha^* = \frac{1}{2}$. MML does not have this issue.

Proof. To show this, first fix $P \in \mathcal{P}$. Then the empirical VAML loss (in expectation) is given by

$$\begin{aligned} E_{s\sim P^*}[\max_V |E_{x\sim P}[V(x)] - V(s)|] &= \alpha^* \max_V |E_{x\sim P}[V(x)] - V(A)| + (1-\alpha^*) \max_V |E_{x\sim P}[V(x)] - V(B)| \\ &= \alpha^* \max_{x,y\in[0,M]} |\alpha x + (1-\alpha)y - x| + (1-\alpha^*) \max_{x,y\in[0,M]} |\alpha x + (1-\alpha)y - y| \\ &= \alpha^* \max_{x,y\in[0,M]} |(\alpha-1)(x-y)| + (1-\alpha^*) \max_{x,y\in[0,M]} |\alpha(x-y)| \\ &= (\alpha^*|\alpha - 1| + (1-\alpha^*)|\alpha|)M \end{aligned}$$

If $\alpha^* < .5$ then the minimizer of the above quantity is $\alpha = 0$, if $\alpha^* > .5$ then the minimizer is $\alpha = 1$. Therefore, if $\alpha^* \notin (0, .5) \cup (.5, 1)$ then VAML picks the wrong model $\alpha \neq \alpha^*$. Additionally, in the case that $\alpha^* = .5$ then the loss is $\frac{M}{2}$ for every $P \in \mathcal{P}$. In this case, VAML with L^1 cannot differentiate between any model; all models are perfectly identical.

On the other hand, we repeat this process with MML:

$$\begin{aligned} |E_{s\sim P^*}[E_{x\sim P}[V(x)] - V(s)]| &= |\alpha^*(E_{x\sim P}[V(x)] - V(A)) + (1-\alpha^*)(E_{x\sim P}[V(x)] - V(B))| \\ &= |\alpha^*(\alpha x + (1-\alpha)y - x) + (1-\alpha^*)(\alpha x + (1-\alpha)y - y)| \\ &= |\alpha^*(\alpha-1)(x-y) + (1-\alpha^*)\alpha(x-y)| \\ &= |\alpha - \alpha^*||x-y| \end{aligned}$$

Clearly $\min_{\alpha\in[0,1]} \max_{x,y\in[0,M]} |\alpha - \alpha^*||x-y| = 0$ where $\alpha = \alpha^*$. □

We do not have to worry about the choice of norm for MML because we know that the OPE error is precisely $\mathcal{L}_{MML}$. On the other hand, as shown in the example, this is not the case for VAML.

C OPO

In this section we explore the OPO results in the order in which they were presented in the main paper.

C.1 Main Result

Proof for Theorem 4.1. 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 $P \in \mathcal{P}$. □

C.2 Sample Complexity for OPO

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

Theorem C.1 (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 Theorem 4.1 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 C.1. By Theorem 4.1,

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

We have shown in the proof of Theorem 3.1 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. □

This bound has the same interpretation as in the OPO case, see Section B.2.

C.3 Misspecification

Similarly as in Section B.3, we show the misspecification gap for OPO in the following result.

Lemma C.2 (OPO Misspecification)**.** *Let $\mathcal{H} \subset (\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R})$ be functions on (s, a, s'). Denote $(WV)^*_{P^*} = w^{P^*}_{\pi^*_{P^*}}(s, a)V^P_{\pi^*_{P^*}}(s')$ and $(WV)^*_P = w^{P^*}_{\pi^*_P}(s, a)V^P_{\pi^*_P}(s')$.*

$$|J(\pi, \widehat{P}) - J(\pi, P^*)| \leq 2\gamma \left(\min_{P \in \mathcal{P}} \max_{h \in \mathcal{H}} |\mathcal{L}(h, P)| + \epsilon_{\mathcal{H}} \right) \tag{23}$$

where $\epsilon_{\mathcal{H}} = \max(\max_{P \in \mathcal{P}} \min_{h \in \mathcal{H}} |\mathcal{L}((WV)^*_{P^*} - h, P)|, \max_{P \in \mathcal{P}} \min_{g \in \mathcal{H}} |\mathcal{L}((WV)^*_P - g, P)|)$.

Proof for Lemma C.2. From the proof of Theorem 4.1, $J(\pi^*_{P^*}, P^*) - J(\pi^*_P, P^*) \leq J(\pi^*_{P^*}, P^*) - J(\pi^*_{P^*}, P) + J(\pi^*_P, P) - J(\pi^*_P, P^*) = \mathcal{L}(w^{P^*}_{\pi^*_{P^*}}, V^P_{\pi^*_{P^*}}, P) + \mathcal{L}(w^{P^*}_{\pi^*_P}, V^P_{\pi^*_P}, P)$. Using the result from proof of Lemma 3.5,

$$\begin{aligned}
|\mathcal{L}(w^{P^*}_{\pi^*_{P^*}}, V^P_{\pi^*_{P^*}}, P) + \mathcal{L}(w^{P^*}_{\pi^*_P}, V^P_{\pi^*_P}, P)| &\leq |\mathcal{L}(h, P) + \mathcal{L}((WV)^*_{P^*} - h, P)| + |\mathcal{L}(g, P) + \mathcal{L}((WV)^*_P - g, P)| \\
&\leq 2 \min_P \max_{h \in \mathcal{H}} |\mathcal{L}(h, P)| + \max_P \min_{h \in \mathcal{H}} |\mathcal{L}((WV)^*_{P^*} - h, P)| \\
&\qquad + \max_P \min_{g \in \mathcal{H}} |\mathcal{L}((WV)^*_P - g, P)| \\
&\leq 2(\min_P \max_{h \in \mathcal{H}} |\mathcal{L}(h, P)| + \epsilon_{\mathcal{H}})
\end{aligned}$$

where $\epsilon_{\mathcal{H}} = \max(\max_P \min_h |\mathcal{L}((WV)^*_{P^*} - h, P)|, \max_P \min_g |\mathcal{L}((WV)^*_P - g, P)|)$. Therefore $|J(\pi, \widehat{P}) - J(\pi, P^*)| \leq 2\gamma(\min_P \max_h |\mathcal{L}(h, P)| + \epsilon_{\mathcal{H}})$, as desired. □

D Additional theory

In this section, we provide additional results that were not covered in the paper. Specifically, we show that as we make $\mathcal{W}, \mathcal{V}$ too rich then the only model with zero loss is P^* itself, which may not be in $\mathcal{P}$.

D.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 D.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 D.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 D.1 and Corollary D.2. We begin with definition 5.1 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

$$\int V(s')(P(s'|s, a) - P^*(s'|s, a))\, d\nu(s') = 0, \tag{24}$$

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 (24) 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 D.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 D.3. Uehara et al. (2020) 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

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

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 $(w, V) \in \mathcal{W}\mathcal{V}$ 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 D.2, concluding the proof. □

E Scenarios & Considerations

In this section we give proof for the various propositions for the corresponding section in the main paper.

E.1 Linear Function Classes

Proof for Prop. 5.1. Given $w(s,a)V(s') = \psi(s,a,s')^T\beta$ and $P(s'|s,a) = \phi(s,a,s')^T\alpha$ then

$$\begin{aligned}
L_n(w,V,P) &= E_n[E_{x\sim P}[\psi(s,a,x)^T\beta] - \psi(s,a,s')^T\beta], \\
&= E_n\left[\int \alpha\phi(s,a,x)^T\psi(s,a,x)^T\beta d\nu(x) - \psi(s,a,s')^T\beta\right], \\
&= E_n[\alpha^T\left(\int \phi(s,a,s')\psi(s,a,s')^T d\nu(s')\right)\beta - \psi(s,a,s')^T\beta],
\end{aligned}$$

which is linear in β. $L_n^2(w,V,P) = 0$ is achieved through $E_n[\alpha^T\left(\int \phi(s,a,s')\psi(s,a,s')^T d\nu(s')\right) - \psi(s,a,s')^T] = 0$. Thus,

$$\widehat{\alpha}^T = E_n[\psi(s,a,s')^T]E_n\left[\int \phi(s,a,s')\psi(s,a,s')^T d\nu(s')\right]^{-1},$$

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

Proof for Prop. 5.2. We begin with $\phi(s,a,s') = e_{(s,a,s')}$, the (s,a,s')-th standard basis vector and $\psi = \phi$. Then

$$X(s,a) = (\sum_{x\in\mathcal{S}} \phi(s,a,x)\phi(s,a,x)^T)_{i,j} = \begin{cases} 1 & i = s|\mathcal{A}||\mathcal{S}| + a|\mathcal{S}|, i = j \\ 0 & \text{otherwise} \end{cases}.$$

Notice that $X(s,a)$ is a diagonal matrix and is the discrete counter-part to $\int \phi(s,a,s')\psi(s,a,s')^T d\nu(x)$. Therefore, $E_n[X(s,a)] = \frac{1}{N}\sum_{(s,a,s')\in D} X(s,a)$, which is a diagonal matrix of the average number of times (s,a) appears in the dataset D. Similarly, $E_n[\phi(s,a,s')]$ is the average number of times that (s,a,s') appears in the dataset D. Hence, by Prop 5.1,

$$\widehat{\alpha}_{s,a,s'} = \frac{\#\{(s,a,s') \in D\}}{\#\{(s,a,x) \in D : \forall x \in \mathcal{S}\}}.$$

Therefore $P(s'|s,a) = \phi(s,a,s')^T\widehat{\alpha} = \widehat{\alpha}_{s,a,s'}$, as desired.

□

E.2 LQR

In order to provide proof that MML gives the LQR-optimal solution, we begin with a few Lemmas. First, we show that the value function is quadratic.

Lemma E.1 (Value Function is Quadratic)**.** *Let $s_{t+1} = As_t + Ba_t + w$ with $w \sim N(0,\sigma^{*2}I)$ be the dynamics, $\pi_K(a|s) = -Ks + w_K$ where $w_K \sim N(0,\sigma_K^2 I)$ be the policy. Let $\gamma \in (0,1]$ be the discount factor. Then $V(s) = s^TUs + q$ where*

$$\begin{aligned}
U &= Q + K^TRK + \gamma(A-BK)^TU(A-BK) \\
q &= \frac{1}{1-\gamma}(\sigma_K^2 tr(R) + \gamma\sigma_K^2 tr(B^TUB) + \gamma\sigma^{*2}tr(U)).
\end{aligned}$$

Proof for Lemma E.1. The value function is given by:

$$\begin{aligned}
x^TUx + q &= x^TQx + E_{N(-Kx,\sigma_K^2 I)}[u^TRu + \gamma E_{N(Ax+Bu,\sigma^{*2}I)}[V(s')]] \\
&= x^TQx + E_{N(-Kx,\sigma_K^2 I)}[u^TRu + \gamma(Ax+Bu)^TU(Ax+Bu) + \gamma q + \gamma\sigma^{*2}tr(U)] \\
&= x^TQx + x^TK^TRKx + \sigma_K^2 tr(R) + \gamma x^T(A-BK)^TJ(A-BK)x \\
&\quad + \gamma\sigma_K^2 tr(B^TUB) + \gamma q + \gamma\sigma^{*2}tr(U)
\end{aligned}$$

Thus, the quadratic terms satisfy

$$U = Q + K^T RK + \gamma(A - BK)^T U(A - BK)$$

and the linear term satisfies

$$q = \frac{1}{1-\gamma}(\sigma_K^2 tr(R) + \gamma\sigma_K^2 tr(B^T UB) + \gamma\sigma^{*2} tr(U))$$

The final value is given by:

$$J(\pi, P^*) = E_{N(s_0,\sigma_0^2 I)}[U] = s_0^T U s_0 + q + \sigma_0^2 tr(U)$$

Existence and uniqueness of U, q is heavily studied (Bertsekas et al., 2005). □

Under the same assumptions as Lemma E.1, we can simplify $\mathcal{L}$ into a reduced form:

Lemma E.2 (LQR Loss Simplified)**.** *In addition to the assumptions of Lemma E.1, let $d_0 = s_0 + w_{d_0}$ where $w_{d_0} \sim N(0, \sigma_{d_0}^2 I)$ be the initial state distribution. Let $P = As + Ba \in \mathcal{P}$ where $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times k}$ and (A, B) is controllable. Let $K \in \mathbb{R}^{k\times n}$ represent all linear policies and $U \in \mathbb{S}_+^n$ be all symmetric positive semi-definite matrices.*

$$\begin{aligned}
&\min_P \max_{w,V} |\mathcal{L}(w, V, P)| \\
&= \min_{A,B} \max_{K,U} \sum_i \gamma^i [s_0^T (A^* - B^*K)^{iT} \Delta (A^* - B^*K)^i s_0 \\
&\quad + tr(\Delta\Sigma_i)] + \sigma_K^2 tr(B^T UB - B^{*T} UB^*) - \sigma^{*2} tr(U),
\end{aligned}$$

where $\Delta = (A-BK)^T U(A-BK) - (A^*-B^*K)^T U(A^*-B^*K)$ and $\Sigma_i = \sigma^*(I + \ldots + F^{i-1}F^{(i-1)T}) + \sigma_K(B^*B^{*T} + \ldots F^{i-1}B^*B^{*T}F^{(i-1)T}) + \sigma_0 F^i F^{iT}$ for $i > 0$ and $\Sigma_0 = \sigma_0 I$, $F = A^* - B^*K$.

Proof for Lemma E.2. We first show that the evolution of dynamics P^* under gaussian noise, with a linear gaussian controller is a gaussian mixture $\sum_i N((A^* - B^*K)^i s_0, \Sigma_i)$, where $\Sigma_i = \sigma^*(I + \ldots + F^{i-1}F^{(i-1)T}) + \sigma_K(B^*B^{*T} + \ldots F^{i-1}B^*B^{*T}F^{(i-1)T}) + \sigma_0 F^i F^{iT}$ for $i > 0$ and $\Sigma_0 = \sigma_0 I$, $F = A^* - B^*K$.

It's clear $s_0 \sim N(s_0, \sigma_0^2 I)$, the base case. Suppose for induction $s_n \sim N((A^* - B^*K)^n s_0, \Sigma_n)$ holds for some $n \geq 0$. Then

$$\begin{aligned}
s_{n+1} &= A^* s_n + B^*(-Ks_n + w_K) + w^* \\
&= (A^* - B^*K)s_n + B^* w_k + w^* \\
&\sim N((A^* - B^*K)^{n+1} s_0, (A^* - B^*K)\Sigma_n (A^* - B^*K)^T + B^* B^{*T} + \sigma^* I) \\
&= N((A^* - B^*K)^{n+1} s_0, \Sigma_{n+1}),
\end{aligned}$$

completing the inductive step. Notice every step s_t is sampled from a gaussian distribution, therefore

$$d_{\pi,\gamma}^{P^*}(s, a) = \sum_{i=0}^{\infty} \gamma^i N(s; F^i s_0, \Sigma_i) N(a; -Ks, \sigma_K^2 I), \tag{25}$$

is a gaussian mixture. Let $w = \frac{d_{\pi,\gamma}^{P^*}}{D}$. We know V is quadratic, given by $U \in \mathcal{S}_+^n$. Therefore,

$$\begin{aligned}
&\min_P \max_{w,V} \mathcal{L}(w, V, P) = \min_{A,B} \max_{w,V} E_{(s,a)\sim D}[w[E_P[V] - E_{P^*}[V]]] \\
&= \min_{A,B} \max_{w,U} E_{(s,a)\sim D}[w[(As + Bu)^T U(As + Bu) - (A^*s + B^*u)^T U(A^*s + B^*u) - \sigma^{*2} tr(U)]] \\
&= \min_{A,B} \max_{K,U} E_{\sum_i \gamma^i N((A^*-B^*K)^i s_0, \Sigma_i)}[E_{u\sim N(-Ks, \sigma_k^2 I)}[\ldots]] \\
&= \min_{A,B} \max_{K,U} E_{\sum_i \gamma^i N((A^*-B^*K)^i s_0, \Sigma_i)}[s^T[(A - BK)^T U(A - BK) - (A^* - B^*K)^T U(A^* - B^*K)]s \\
&\qquad + \sigma_K^2 tr(B^T UB) - \sigma_K^2 tr(B^{*T} UB^*) - \sigma^{*2} tr(U)] \\
&= \min_{A,B} \max_{K,U} E_{\sum_i \gamma^i N((A^*-B^*K)^i s_0, \Sigma_i)}[s^T[\Delta(A, B, A^*, B^*, U, K)]s + \sigma_K^2 tr(B^T UB - B^{*T} UB^*) - \sigma^{*2} tr(U)] \\
&= \min_{A,B} \max_{K,U} \sum_i \gamma^i [s_0^T (A^* - B^*K)^{iT} \Delta (A^* - B^*K)^i s_0 + tr(\Delta\Sigma_i)] + \sigma_K^2 tr(B^T UB - B^{*T} UB^*) - \sigma^{*2} tr(U)
\end{aligned}$$

where $\Delta = (A - BK)^T U(A - BK) - (A^* - B^*K)^T U(A^* - B^*K)$. □

First, Lemma E.2 supposes that there is model mismatch $P^* \notin \mathcal{P}$ since $\mathcal{P}$ are deterministic simulators and P^* is stochastic. Second, we notice that K takes the position of w, which is to say that the policy K directly specifies w, as expected. We will need the previous two results in the experiments. We may now prove Prop 5.3 that says MML yields the true parameters of LQR in expectation:

Proof for Prop 5.3. Consider two linear, controllable systems with parameters $P_1 = (A_1, B_1)$ and $P_2 = (A_2, B_2)$. Then there exists a controller K that stabilizes P_1 (i.e, $J(P_1), K) < \infty$) but destabilizes P_2 (i.e, $J(P_2, K) = \infty$). We show this by analyzing the characteristic polynomial of both $A_1 - B_1K$ and $A_2 - B_2K$. There exists an invertible matrix T_1, T_2 that put $(A_1, B_1), (A_2, B_2)$ into controllable canonical forms (CCF), respectively Bertsekas et al. (2005). Thus, we will assume, wlog, that $(\tilde{A}_1, \tilde{B}_1), (\tilde{A}_2, \tilde{B}_2)$ are already in CCF. Hence,

$$\tilde{A}_1 = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & & 1 \\ -a_0 & -a_1 & -a_2 & \dots & -a_{n-1} \end{bmatrix}, \quad \tilde{B}_1 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

and

$$\tilde{A}_2 = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & & 1 \\ -b_0 & -b_1 & -b_2 & \dots & -b_{n-1} \end{bmatrix}, \quad \tilde{B}_2 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

We will find a controller in the form $K = K_1T_1 = K_2T_2$ for some K_1, K_2 for T_1, T_2 that put the systems into CCF. Consider a desired characteristic polynomial of $f(s) = (s+\epsilon)^{n-1}(s+\lambda)$ for $\epsilon, \lambda \in \mathbb{R}^+(> 0)$. This polynomial has eigenvalues equal to $-\epsilon, -\lambda$ and therefore a system with this polynomial is asymptotically stable (converges to 0 exponentially fast). Take $K_1 = [k_{1,0}, k_{1,1}, \dots, k_{1,n-1}]$. Then $det(sI - (\tilde{A}_1 - \tilde{B}_1K_1)) = s^n + (a_{n-1} + k_{1,n-1})s^{n-1} + \cdots + (a_0 + k_{1,0})$. By selecting $k_{1,i} = \left(\binom{n-1}{i}\lambda + \binom{n-1}{i-1}\epsilon\right)\epsilon^{n-1-i} - a_i$ then $det(sI - (\tilde{A}_1 - \tilde{B}_1K_1)) = f(s)$. Hence, $(\tilde{A}_1, \tilde{B}_1)$ is asymototically stable with eigenvalues $-\lambda, -\epsilon$ for any λ, ϵ strictly positive. Therefore $K = K_1T_1$ makes the system (A_1, B_1) asymptotically stable.

Now we consider $K_2 = K_1T_1T_2^{-1}$. Let us denote $T_1T_2^{-1} = T$ which is also invertible since T_1, T_2 are invertible. Then by taking the last term of $det(sI - (\tilde{A}_2 - \tilde{B}_2K_2))$, we can examine the product $\prod_{i=0}^{n-1}\lambda_i$ of the eigenvalues of the closed loop system $\tilde{A}_2 - \tilde{B}_2K_2$. Namely, $b_0 + \sum_{i=0}^{n-1} k_{1,i}T_{i,n}$ is the product of eigenvalues. We may simplify this via some algebra as follows:

$$\begin{aligned} \prod_{i=0}^{n-1}\lambda_i &= b_0 + \sum_{i=0}^{n-1} k_{1,i}T_{i,n} \\ &= b_0 + \sum_{i=0}^{n-1} T_{i,n}\left(\left(\binom{n-1}{i}\lambda + \binom{n-1}{i-1}\epsilon\right)\epsilon^{n-1-i} - a_i\right) \\ &= \underbrace{b_0 - \sum_{i=0}^{n-1} a_i + \sum_{i=0}^{n-1} T_{i,n}\binom{n-1}{i-1}\epsilon^{n-i}}_{\bar{b}} + \lambda \underbrace{\sum_{i=0}^{n-1} T_{i,n}\binom{n-1}{i}\epsilon^{n-1-i}}_{c} \\ &= \bar{b} + \lambda c \end{aligned}$$

We may select $\epsilon > 0$ so that $c \neq 0$ otherwise $T_{i,n} = 0$ for all i which would contradict invertibility of T. Therefore $\prod_{i=0}^{n-1}\lambda_i$ is linear in λ. By driving $\lambda \to \infty$, then $|\prod_{i=0}^{n-1}\lambda_i| \to \infty$ is unbounded. Select λ so that $|\bar{b} + \lambda c| > 1$. By the pigeonhole principle, at least one of the eigenvalues of $\tilde{A}_2 - \tilde{B}_2K_2$ must have a magnitude greater than

1 and therefore the system is unstable. Therefore the controller $K_2T_2 = K_1T_1T_2^{-1}T_2 = K_1T_1 = K$ makes the system (A_2, B_2) unstable. Hence, K simultaneously stabilizes (A_1, B_1) but destabilizes (A_2, B_2).

According to Lemma E.2, when $(A, B) = (A^*, B^*)$ then for any K, $\max_U \mathcal{L}((A, B), K, U) = \max_U |\sigma^{*2} tr(U)| < \infty$ since U are bounded by assumption. Furthermore, we have just shown that there always exists a K that destabilizes any controller $(A, B) \neq (A^*, B^*)$ while stabilizing (A^*, B^*). Therefore $\max_{K,U} \mathcal{L}((A, B), K, U) = \infty$ for any system $(A, B) \neq (A^*, B^*)$. Therefore $\min_{(A,B)} \max_{K,U} \mathcal{L}((A, B), K, U) = (A^*, B^*)$.

It is well known that ordinary least squares is a consistent estimator when the noise is exogenous, as it is here. Therefore the maximum likelihood solution also yields (A^*, B^*) in expectation. □

We will use the following lemma in our experiments:

Lemma E.3 (VAML Loss in 1-d LQR)**.** *Consider a 1-dimensional LQR problem. Let $P^*(s'|s, a) = A^*s + B^*a + w^*$ where $w^* \sim N(0, \sigma)$. Suppose $P(s'|s, a) = As + Ba \in \mathcal{P}$ deterministic and $P(s'|s, a) = A^*s + B^*a \in \mathcal{P}$. Let $V(s) = s^T Us + q \in \mathcal{V}$ for some $U > 0$, $U \in \mathcal{U} \subset \mathbb{R}$ and $q \in \mathbb{R}$. Let $\mathcal{V}$ be a set containing $V(s)$. Then*

$$\arg\min_{P \in \mathcal{P}} \mathcal{L}_{VAML}(\mathcal{V}, P) = (A^*, B^*)$$

Proof.

$$\begin{aligned}
\mathcal{L}_{VAML}(\mathcal{V}, P) &= \int d\nu(s, a) \max_{V \in \mathcal{V}} [(E_P[V] - E_{P^*}[V])^2] \\
&= \int d\nu(s, a) \max_{U \in \mathcal{U}} [((As + Ba)^T U(As + Ba) - (A^*s + B^*a)^T U(A^*s + B^*a) - \sigma^2 tr(U))^2] \\
&= \int d\nu(s, a) \max_{u \in \mathcal{U}} [(f(s, a)^2 u - g(s, a)^2 u - \sigma^2 u)^2] \quad (f(s, a) = As + Ba, g(s, a) = (A^*s + B^*a)) \\
&= C \int d\nu(s, a)(f(s, a)^2 - g(s, a)^2 - \sigma^2)^2 \quad (C = \max_{u \in \mathcal{U}} u^2)
\end{aligned}$$

Since the integrand is positive for any (s, a) pair then the $\arg\min_{P \in \mathcal{P}} \mathcal{L}_{VAML}(\mathcal{V}, P)$ occurs when $f(s, a) = g(s, a)$, and thus $\widehat{P}_{VAML} = (A^*, B^*)$. □

E.3 RKHS & Practical Implementation

Since $P \in \mathcal{P}$ is a stochastic model in general, then the inner expectation of the loss in def (5.1) 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 (12) requires solving a minimax problem which is often computationally challenging. One approach might be to set-up neural networks in a GAN-like fashion and use a higher order gradient descent (Goodfellow et al., 2014; Schaefer & Anandkumar, 2019).

If we have access to a kernel, say radial basis function (RBF), 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 (Liu et al., 2018; Feng et al., 2019; Uehara et al., 2020) we have

Proposition E.4 (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:*

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

Proof for Prop E.4. 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 5.1,

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

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. □

F Experiments

F.1 Environment Descriptions

F.1.1 LQR

The LQR domain is a 1D stochastic environment with true dynamics: $P^*(s'|s,a) = s - .5a + w^*$ where $w^* \sim N(0, .01)$. We let $x_0 \sim N(1, .01^2)$. The reward function is $R(s,a) = -(s+a)$ and $\gamma = .9$. We use a finite class $\mathcal{P}$ consisting of all deterministic policies $\mathcal{P} = \{P_x(s'|s,a) = (1 + x/10)s - (.5 + x/10)a | x \in [0, M]\}$ where we vary $M \in \{2, 3, \ldots, 19\}$. We write $(A^*, B^*) = P_0(s'|s,a) = A^*s + B^*a$, the deterministic version of P^*.

F.1.2 Cartpole

We use the standard Cartpole benchmark (OpenAI, Brockman et al. (2016)). The state space is a tuple $(x, \dot{x}, \theta, \dot{\theta})$ representing the position of the cart, velocity of the cart, angle of the pole and angular velocity of the pole, respectively. The action space is discrete given by pushing the car to the left or pushing the car to the right. We add $N(0, .001^2)$ Gaussian noise to each component of the state to make the dynamics stochastic. We consider the infinite horizon setting with $\gamma = .98$. The reward function is modified to be a function of angle and location $R(s, \theta) = (2 - \theta/\theta_{\max}) * (2 - s/s_{\max}) - 1)$ rather than 0/1 to make the OPE problem more challenging.

F.1.3 Inverted Pendulum (IP)

We consider the infinite horizon setting with $\gamma = .98$. The state space is a tuple $(\theta, \dot{\theta})$ representing the angle of the pole and angular velocity of the pole, respectively. The action space $\mathcal{A} = \mathbb{R}$ is continuous representing a clockwise or counterclockwise force. The reward function is a clipped quadratic function $R([\theta, \dot{\theta}], a) = \min(((\theta + \pi) \mod 2\pi - \pi)^2 + .1\dot{\theta}^2 + .001u^2, 100)$. This IP environment has a Runge-Kutta(4) integrator (Dorobantu & Taylor, 2020) rather than Forwrd Euler and, thus, produces more realistic data. The mass of the rod is .25 and the length .5.

F.2 Experiment Descriptions

F.2.1 LQR OPE/OPO

OPE. We aim to evaluate $\pi(a|s) = N(1.3s, .01)$. We ensure $V_\pi^P \in \mathcal{V}$ for all $P \in \mathcal{P}$ by solving the equations in Lemma E.1. We ensure $W_\pi^{P^*} \in \mathcal{W}$ using Equation (25). MLE and VAML both give (A^*, B^*) in expectation (see Prop 5.3, Lemma E.3). **Metric:** We compute $|(J(\pi, \widehat{P}) - J(\pi, P^*))|$, the OPE error.

OPO. Similarly as in OPE, we ensure that all MML realizability assumptions hold. This means as we increase $\mathcal{P}$ then we have to increase the sizes of both $\mathcal{W}$ and $\mathcal{V}$ now instead of just $\mathcal{V}$ as in OPE. Once again MLE and VAML both give (A^*, B^*) in expectation (see Prop 5.3, Lemma E.3). With this, we produce Figure 5 (right). By increasing $\mathcal{P}$, we also have more policies $\{\pi_P^*\}_{P \in \mathcal{P}}$ we may consider. Instead of selecting one for OPE, for each $\pi \in \{\pi_P^*\}_{P \in \mathcal{P}}$ we calculate the OPE error. We aggregate across all $\{\pi_P^*\}_{P \in \mathcal{P}}$ by taking the average of the OPE errors and the worst-case, which can be seen in Figure 5 (left). **Metric:** We compute $|(J(\pi_{\widehat{P}}^*, \widehat{P}) - J(\pi_{P^*}^*, P^*))|$, the OPO error.

Note: All calculations in LQR OPE/OPO are in expectation so no error bars need be included.

Verifiability. With the same setup as in OPE, now randomly sample 100k points in the interval $[-3, 3] \times [-3, 3]$, which is the support of the LQR system. We rerun the same experiment as in OPE except now we add $w \sim \mathcal{N}(0, \epsilon)$ noise to $V \in \mathcal{V}$ where $\epsilon \in \{0, .2, \ldots, .8, 1\}$. We evaluate the error $|(J(\pi, \widehat{P}) - J(\pi, P^*))|$ over the 100k samples rather than in expectation as before. We run 5 seeds and present the mean over the seeds with standard error. We smooth the resulting mean with a moving average filter of size 3. The result can be seen in Figure 2 (right).

F.2.2 Cartpole OPE

Each $P \in \mathcal{P}$ 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. Each model has a two hidden layers and with 64 units each and ReLU activation with final linear layer. 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 monte-carlo average of 10000 rollouts.

We model the class $\mathcal{WV}$ as a RKHS as in Lemma E.4 with an RBF kernel. We do the same for VAML. The RKHS kernel we use for MML and VAML is given by $K(s,a,s') = K_1(s)K_2(a)K_3(s')$ and $K_3(s')$ respectively where K_i are Gaussian Radial Basis Function (RBF) kernels with a bandwidth equal to the median of the pair-wise distances for each coordinate (s, a, s' independently) over the batch.

For MML, we sample from P a total of 5 times and take the empirical mean to calculate the expectation over P for the RKHS formula given in E.4.

We run 20000 batches of size 128 and normalize the data over the batch. Our learning rate is 10^{-3} and we use Adam (Kingma & Ba, 2015) optimizer. The estimate we use is the mean over the last 10 batches. We run 5 random seeds per dataset size, and plot the log-relative MSE with standard error in Figure 3.

Note: These hyperparameters remain the same across the different loss functions.

Metric: We compare the methods using the log-relative MSE metric: $\log(\frac{(J(\pi,\widehat{P})-J(\pi,P^*))^2}{(J(\pi_b,\widehat{P^*})-J(\pi,P^*))^2})$, 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. To calculate $J(\pi,\widehat{P})$ we run 100 trajectories in $\widehat{P}$ and take the mean.

F.2.3 Inverted Pendulum OPO

We generate the behavior data using a noisy feedback-linearized controller: $\pi_b(a|s)$ is uniformly random with probability .3 and is a feedback-linearized LQR controller (FLC) with probability .7 where we use the FLC corresponding to LQR matrices $Q = 2I_{2\times 2}, R = I_{2\times 2}$. We truncate all rollouts at 200 time steps. We fit 4 feed-forward neural networks representing $P_1, \ldots, P_4$ where each is a deterministic model with two layers of 16 weights and a Tanh activation followed with Linear. We use Adam (Kingma & Ba, 2015) optimizer with 10^{-3} as the learning rate. Using different batches of size 64 on each P_i and perform 5000 iterations for each model.

The RKHS kernel we use for MML and VAML is given by $K(s,a,s') = K_1(s)K_2(a)K_3(s')$ and $K_3(s')$ respectively where K_i are Gaussian Radial Basis Function (RBF) kernels with a bandwidth equal to 1.

For MML, we only sample from P once to calculate the expectation over P for the RKHS formula given in E.4, since P is deterministic.

Now we have $P(s'|s,a) = \frac{1}{4}\sum_{i=1}^{4} P_i(s'|s,a)$. We calculate $\alpha = Median(\{\|P_j(s,a) - s'\|_2 : j \in [1,\ldots,4], (s,a,s') \in X \subset D\})$ where X is 10000 random samples from the dataset. We form an α-USAD (see MOREL Section F.3) and construct a pessimistic MDP $(\tilde{P}, \tilde{R})$ (see Section F.3). We use PPO as our policy optimizer with the default settings from (Raffin et al., 2019). We run PPO three times in the pessimistic MDP and take the policy that performs the best and report its performance. We keep track of the running maximum as we increase the dataset size. We plot the mean of the running maximums over the five seeds including standard error bars in Figure 4.

Note: These hyperparameters remain the same across the different loss functions.

Metric: We look at the performance $J(\pi^*_{\widehat{P}}, P^*)$ of a policy and compare it to π^*, learned from PPO. To calculate $J(\pi^*_{\widehat{P}}, P^*)$ we run 100 trajectories in P^* and take the mean.

F.3 MOREL

We give a brief explanation of MOREL (Kidambi et al., 2020) and its construction. The objective of MOREL is to make sure that the policy we learn does not take advantage of the errors in the simulator P. If there are errors in P then a policy may think the agent can perform a particular state transition (s,a,s') and $R(s',a')$ has high reward for some action a'. However, it's possible that such a transition (s,a,s') may not occur in the true environment. Therefore, we modify our model $P(s'|s,a)$ in the following way:

$$\tilde{P}(s'|s,a) = \begin{cases} \text{Terminate episode} & U^\alpha(s,a) = 1 \\ P(s'|s,a) & \text{otherwise} \end{cases}$$

where $U^\alpha(s,a) = 1$ if $\max_{i\in\{1,2,3,4\}} \|P_i(s'|s,a) - P(s'|s,a)\| \geq \alpha$, otherwise 0. In other words, we've modified the transition dynamics so that we do not trust our model P unless all the P_i are in agreement. We also modify our reward to be

$$\tilde{R}(s,a) = \begin{cases} \text{-100} & U^\alpha(s,a) = 1 \\ R(s,a) & \text{otherwise} \end{cases}$$

where -100 is chosen this value is well below any reward that the Inverted Pendulum environment generates. Similarly, we penalize our policy for entering a state where we are uncertain. Together, this creates a pessimistic MDP.

F.4 Additional Experiments

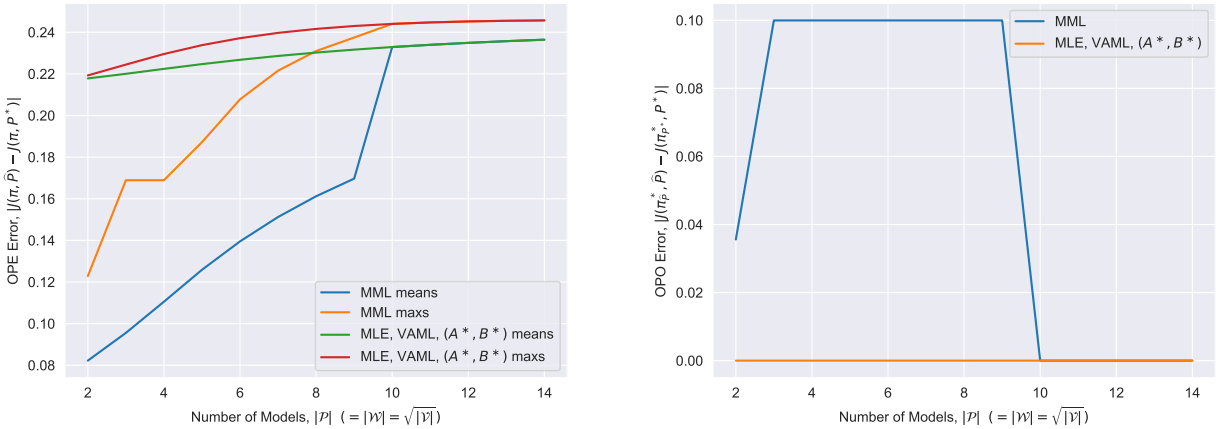

Figure 5: *(LQR) As we increase $|W|, |V|$ then MML is forced to be robust to too many OPE problems and settles for the system (A^*, B^*) since this is the only system robust to the most OPE problems.*

In the experiments for Figure 5, we consider what happens when we satisfy the realizability conditions for OPO. As we increase $|\mathcal{P}|$, we must also increase $|\mathcal{W}|, |\mathcal{V}|$ because each $P \in \mathcal{P}$ induces an optimal policy π_P^* to which we have to make sure $w_{\pi_P^*}^{P^*} \in \mathcal{W}$ and $V_{\pi_P^*}^{P_i} \in \mathcal{V}$ for $\forall P_i \in \mathcal{P}$. In a sense, we are adding more OPE problems for MML to be robust to. In particular, we now have more policies $\{\pi_P^*\}_{P\in\mathcal{P}}$ to consider. As described earlier, for each $\pi \in \{\pi_P^*\}_{P\in\mathcal{P}}$ we calculate the OPE error. We aggregate across all $\{\pi_P^*\}_{P\in\mathcal{P}}$ by taking the average of the OPE errors and the worst-case, which can be seen in Figure 5 (left). We plot the OPO error in Figure 5 (right). What we see is that while $|\mathcal{P}|$ is small, MML is able to be robust to a certain number of OPE problems. But as we increase the number of OPE problems the average and max error increases until all methods select the same model, which is the OPO-optimal model, (A^*, B^*).