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

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

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,

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

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

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,

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

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

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

where $\epsilon_{\mathcal{H}} = \max_P \min_h |\mathcal{L}((WV)^* - h, P)|$. Therefore $|J(\pi,\widehat{P}) - J(\pi,P^*)| \le \gamma(\min_P \max_h |\mathcal{L}(h,P)| + \epsilon_{\mathcal{H}})$, 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. Fix $P \in \mathcal{P}$. 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 $\alpha^* \in (0, .5) \cup (.5, 1)$ 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.

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

($\Rightarrow$) 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 $w = \frac{d^{P^*}_{\pi,\gamma}}{D}$. We know V is quadratic, given by $U \in \mathcal{S}^n_+$. Therefore,

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

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 $\tilde{A}_2 - \tilde{B}_2 K_2$ must have a magnitude greater than 1 and therefore the system is unstable. Therefore the controller $K_2 T_2 = K_1 T_1 T_2^{-1} T_2 = K_1 T_1 = K$ 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. □

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.3 (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:*

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

Proof for Prop E.3. 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,

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

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

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

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

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

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 $w^* \sim N(0, .01^2)$. We let $x_0 \sim N(1, .1^2)$. The reward function is $R(s,a) = -(s+a)$ and $\gamma = .9$. We use a finite class $\mathcal{P}$ consisting of all deterministic models $\mathcal{P} = \{P_x(s'|s,a) = (1 + x/10)s - (.5 + x/10)a | x \in [0, M]\}$ 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 $R([\theta, \dot{\theta}], a) = \min(((\theta + \pi) \bmod 2\pi - \pi)^2 + .1\dot{\theta}^2 + .001u^2, 100)$. 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 $\pi(a|s) = N(1.3s, .1^2)$. 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). We derive 1-d equations for VAML analogous to Lemma E.2). Finally, we know MLE gives (A^*, B^*) in expectation (see Prop 5.3).

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 gives (A^*, B^*) in expectation (see Prop 5.3) and we evaluate VAML using equations analogous to those in Lemma E.2). 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.3 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.3.

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: $\log(\frac{(J(\pi,\widehat{P})-J(\pi,P^*))^2}{(J(\pi_b,P^*)-J(\pi,P^*))^2})$, 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.3, 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

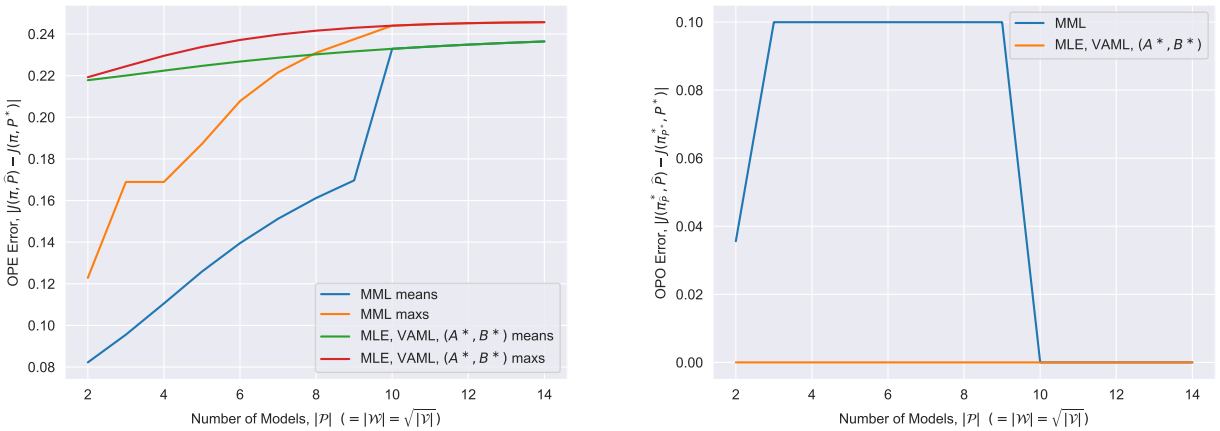

Figure 5: *(LQR) As we increase $|W|, |V|$ then MML is forced to be robust to too many OPE problems and settles for the system (A^*, B^*) since this is the only system robust to the most OPE problems.*

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^*).