Model-based Offline Reinforcement Learning with Local Misspecification

Kefan Dong[1] **Ramtin Keramati**[1] **Emma Brunskill**[1]

Abstract

In this paper we propose a model-based offline reinforcement learning algorithm that explicitly handles model misspecification and distribution mismatch. Theoretically, we prove a safe policy improvement theorem by establishing pessimism approximations to the value function. Our algorithm can output the best policy in the given policy class with interpretable error terms measuring misspecification level, distribution mismatch, and statistical deviation. In addition, as long as the model family can approximate the transitions of state-action pairs *visits* by a policy, we can approximate the value of that particular policy. We visualize the effect of error terms in the LQR setting, and show that the experiment results match our theory.

1. Introduction

Offline reinforcement learning (RL) could unlock an enormous amount of observational data and make it useful for data-driven decision making, such as marketing (Thomas et al., 2017), robotics (Quillen et al., 2018; Yu et al., 2020; 2021; Swazinna et al., 2020; Singh et al., 2020), recommendation systems (Swaminathan, Joachims, 2015), etc. Unlike online RL, which may involve unsafe exploration policies, an offline RL algorithm utilizes existing historical data, and outputs a policy with promising behavior. As a result, it is easier to deploy offline RL algorithms in real world applications.

However, there is a fundamental challenge of distribution shift between the data collected before and the type of data we might collect under a new policy. While extensive researches for offline RL in the tabular setting (Yin, Wang, 2020; Kallus, Uehara, 2020; Yin et al., 2020; 2021; Ren et al., 2021; Kidambi et al., 2020) and linear setting (Duan et al., 2020; Jin et al., 2020; Chen et al., 2021) yield promising results, offline RL with general function approximation is still a challenging task. In fact, in a large or continuous state space, it is non-trivial to even evaluate a given policy using only pre-collected data. To make progress we must rely on certain assumptions.

Existing work largely falls into one of two extremes. One extreme is to require only no confounding and minimal assumptions. Some of the works don't even require a Markov structure or a function class (Dudík et al., 2011; Li et al., 2015; Thomas, Brunskill, 2016). However, these approaches usually suffer from exponentially high variance (Liu et al., 2018b), and are hence hard to scale to long-horizon problems unless with enormous amounts of data. Marginalized importance sampling (MIS) methods (Liu et al., 2018a; Xie et al., 2019; Yin, Wang, 2020; Liu et al., 2020a) help address this but rely on system being Markov in the underlying state space.

Another extreme is to make strong assumptions on the domain satisfying the Markov assumption and that the modeling assumptions enable realizability everywhere. Most existing work in this space assumed not only the above two assumptions, but also required strong data coverage on any possible policy (Xie, Jiang, 2020; Chen, Jiang, 2019), which is extremely strong and unlikely to hold in real world applications. More recent work directly considers the limited data available and tries to find good policies within this set, using model-based (Yu et al., 2020; 2021; Kidambi et al., 2020), model-free (Liu et al., 2020b) or policy search methods (Curi et al., 2020; Hasselt van et al., 2019).

Such work still relies on realizability. For model-free approaches, a common assumption is that the value function is realizable, normally not just for optimal policy but all policies. Liu et al. (2020b) assume that the value function class is close under (modified) Bellman backups. A recent exception is the work of Xie, Jiang (2020), who only requires the optimal Q-function to be representable by the value function class. As a compromise, their sample complexity scales non-optimally (Xie, Jiang, 2020, Theorem 2), and they also make strong assumptions on the data coverage — essentially the dataset must visit all states with sufficient probability.

On the other hand, model-based approaches such as Malik et al. (2019) assume realizability of the dynamics class. The recent effort of Voloshin et al. (2021), while requiring

[1] Stanford University. Correspondence to: Kefan Dong <kefandong@stanford.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

Markov structures, take a model-based approach and can tolerate certain violations of realizability. Their MML algorithm minimizes a value-aware model error that upper bounds the difference of policy value in learned and real models. The model error remains an upper bound when the model misspecification is small, but it's unclear in which cases the model error can be optimized to zero with a misspecified model class.

Our insight is that the algorithm may be able to leverage misspecified models and still leverage Markov assumption for increased data efficiency. In particular, we only need to find dynamics that work well over the space of state-action pairs that a policy would visit. In other words, if modeling the whole domain is hard with our available dynamics class, we can still model parts of the domain, and optimize a policy thereafter.

In this paper, we build a lower bound for the value of every policy based on the pessimism principle. We prove a finite sample bound that directly accounts for model misspecification (see Lemma 3.3). With both theoretical and empirical evidence, we show that the misspecification error of our method is much tighter than other approaches, because we only look at model's ability to represent visited state-action pairs for a particular policy (see Section A and Section B). In that sense, we say our algorithm relies on small *local* misspecification. Because of the tighter pessimistic estimation, we can prove a novel safe policy improvement theorem (see Theorem 3.4) for offline policy optimization (OPO).

The key ingredient of our algorithm is to jointly optimize policy and dynamics. Prior model-based offline RL algorithms typically estimate dynamics first, and then optimize a policy w.r.t. the learned dynamics (Yu et al., 2020; 2021; Voloshin et al., 2021). But without realizability, there may not exist a unique "good dynamics" that can approximate the value of every policy. As a result, the learned policy may have a huge virtual reward (under learned dynamics), but still performs poorly in the real environment. Indeed, in Theorem A.1 we show that without realizability assumptions, decoupling the learning of policy and dynamics is suboptimal. Empirically, we also show that our algorithm achieves a much smaller off-policy optimization error compared with Voloshin et al. (2021) in a one-dimensional environment.

2. Problem Setup

A Markov Decision Process (MDP) is defined by a tuple $\langle T, r, \mathcal{S}, \mathcal{A}, \gamma \rangle$. $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces. $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition and $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ is the reward. $\gamma \in [0, 1)$ is the discount factor. For a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, the value function is defined as

$$V_T^\pi(s) = \mathbb{E}_{s_0=s, a_t\sim\pi(s_t), s_{t+1}\sim T(s_t,a_t)}[\textstyle\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]. \tag{1}$$

Let $R_{\max} \triangleq \max_{s,a} r(s, a)$ be the maximal reward and $V_{\max} \triangleq R_{\max}/(1-\gamma)$. Without loss of generality, we assume that the initial state is fixed as s_0. We use $\eta(T, \pi) \triangleq V_T^\pi(s_0)$ to denote the expected value of policy π. Let ρ_T^π be the normalized state-action distribution when we execute policy π in dynamics T.

In this paper we assume the reward function is *known*. An offline reinforcement learning algorithm takes a dataset $\mathcal{D} = \{(s_i, a_i, s_i')\}_{i=1}^n$ as input, where n is the size of the dataset. Each (s_i, a_i, s_i') tuple is drawn independently from a (known) behavior distribution μ. We assume that μ is consistent with the MDP in the sense that $\mu(\cdot \mid s, a) = T(s, a)$ for all (s, a). For simplicity, we use $\hat{\mathbb{E}}$ to denote the empirical distribution over the dataset $\mathcal{D}$.

The algorithm can access three (finite) function classes $\mathcal{G}, \mathcal{T}, \Pi$. $\mathcal{G}$ is a class of value functions, $\mathcal{T}$ a class of dynamics and Π a class of policies. We assume that $g(s, a) \in [0, V_{\max}]$ for all $g \in \mathcal{G}$. In the following, we use $T^\star$ to denote the ground-truth dynamics. Our goal is to compute a policy $\pi \in \Pi$ such that $\eta(T^\star, \pi)$ is maximized.

3. Main Results

We sketch our main results in this section. As a starting point, we invoke the simulation lemma.

Lemma 3.1 (Simulation Lemma (Yu et al., 2020; Kakade, Langford, 2002)). *Consider two MDPs with dynamics $T, T^\star$ and the same reward function. Then,*

$$\eta(T, \pi) - \eta(T^\star, \pi) = \frac{\gamma}{1-\gamma} \mathbb{E}_{(s,a)\sim\rho_T^\pi}\left[\mathbb{E}_{s'\sim T(s,a)}[V_{T^\star}^\pi(s')] - \mathbb{E}_{s'\sim T^\star(s,a)}[V_{T^\star}^\pi(s')]\right].$$

For a fixed ground-truth dynamics $T^\star$, define $G_T^\pi(s, a) = \mathbb{E}_{s'\sim T(s,a)}[V_{T^\star}^\pi(s')] - \mathbb{E}_{s'\sim T^\star(s,a)}[V_{T^\star}^\pi(s')]$. The simulation lemma states that the dynamics is good enough to estimate the value of a policy if $\mathbb{E}_{s'\sim T(s,a)}[V_{T^\star}^\pi(s')]$ matches $\mathbb{E}_{s'\sim T^\star(s,a)}[V_{T^\star}^\pi(s')]$. In other words, we want to minimize the model error $G_T^\pi(s, a)$ to have an accurate estimation,

For a value function $g \in \mathcal{G}$ and a dynamics T, let $f_T^g(s, a) \triangleq \mathbb{E}_{s'\sim T(s,a)}[g(s')]$. Recall that $\hat{\mathbb{E}}$ denotes the empirical expectation over dataset $\mathcal{D}$. For a density ratio $w : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$, the model loss is defined as

$$\ell_w(T, g) = |\hat{\mathbb{E}}[w(s, a)(f_T^g(s, a) - g(s'))]|. \tag{2}$$

We aim to upper bound policy evaluation error by the loss function even if there are state action pairs with small probability mass under μ (i.e., the offline dataset doesn't have a perfect coverage). Following Liu et al. (2020b), we treat the unknown state-action pairs pessimistically. Let ζ be a fixed parameter. For a policy π and dynamics T, we use

$w_{\pi,T}(s,a) \triangleq \mathbb{I}\left[\frac{\rho_T^\pi(s,a)}{\mu(s,a)} \le \zeta\right] \frac{\rho_T^\pi(s,a)}{\mu(s,a)}$ to denote the truncated density ratio. For a fixed policy π, when $w = w_{\pi,T}$,

$$\begin{aligned}
&\left|\mathbb{E}_{(s,a)\sim\rho_T^\pi}[G_T^\pi(s,a)]\right| \\
\le &\left|\mathbb{E}_{(s,a)\sim\rho_T^\pi}\left[\mathbb{I}\left[\frac{\rho_T^\pi(s,a)}{\mu(s,a)} \le \zeta\right] G_T^\pi(s,a)\right]\right| \\
&\quad + \left|\mathbb{E}_{(s,a)\sim\rho_T^\pi}\left[\mathbb{I}\left[\frac{\rho_T^\pi(s,a)}{\mu(s,a)} > \zeta\right] G_T^\pi(s,a)\right]\right| \\
\le &\left|\mathbb{E}_{(s,a)\sim\mu}[w(s,a)G_T^\pi(s,a)]\right| \\
&\quad + V_{\max}\left|\mathbb{E}_{(s,a)\sim\rho_T^\pi}\left[\mathbb{I}\left[\frac{\rho_T^\pi(s,a)}{\mu(s,a)} > \zeta\right]\right]\right|.
\end{aligned}$$

As a result, ignoring statistical error due to finite dataset, we can upper bound the estimation error $|\eta(T^\star,\pi) - \eta(T,\pi)|$ by

$$\frac{\gamma}{1-\gamma}\Bigg(\underbrace{\sup_{g\in\mathcal{G}}\left|\ell_{w_{\pi,T}}(g,T)\right|}_{(a)} \tag{3}$$

$$+ \underbrace{V_{\max}\mathbb{E}_{(s,a)\sim\rho_T^\pi}\left[\mathbb{I}\left[\frac{\rho_T^\pi(s,a)}{\mu(s,a)} > \zeta\right]\right]}_{(b)}\Bigg). \tag{4}$$

Intuitively, term (a) measures the error caused by imperfect dynamics T, and term (b) comes from distribution mismatch.

3.1. Pessimistic Policy Optimization with Model Misspecification

Now we explicitly consider misspecifications of the function class $\mathcal{G}, \mathcal{T}$. As argued in Sec 1, most of the prior works make strong assumptions regarding realizability.

Our key observation is that for a given dynamics T and policy π, computing the density ratio $w_{\pi,T}$ is *statistically* efficient. That is, to compute $w_{\pi,T}$ we don't need any sample from the true dynamics. Therefore, we can explicitly utilize the density ratio to get a relaxed realizability assumption.

Definition 3.2. The model misspecification error model misspecification error is defined by the following quantity.

$$\begin{aligned}
&\epsilon_V(T,\pi) \\
\triangleq &\inf_{g\in\mathcal{G}} |\mathbb{E}_{(s,a)\sim\mu}[w_{\pi,T}(s,a)(\mathbb{E}_{s'\sim T(s,a)}[(g - V_{T^\star}^\pi)(s')] \\
&\quad + \mathbb{E}_{s'\sim T^\star(s,a)}[(g - V_{T^\star}^\pi)(s')])]|.
\end{aligned}$$

The term ϵ_V measures the misspecification of the value function class. We call $\epsilon_V(T,\pi)$ local misspecification because the error is *not* evaluated on the state-action pairs that policy π doesn't visit.

With the local misspecification error, we can establish a pessimistic estimation of the true reward. Let $\mathcal{E}$ be a high probability event under which the loss function $\ell_{w_{\pi,T}}(T,g)$ is close to its expectation (randomness comes from the dataset $\mathcal{D}$). In Appendix D.1 we define the event formally and prove that $\Pr(\mathcal{E}) \ge 1-\delta$. The following lemma gives a lower bound on the true reward.

Lemma 3.3. *Let $\iota = \log(2|\mathcal{G}||\mathcal{T}||\Pi|/\delta)$. For any fixed dynamics T and policy π, define*

$$\mathrm{lb}(T,\pi) = \eta(T,\pi) - \frac{1}{1-\gamma}\Bigg(\sup_{g\in\mathcal{G}}\ell_{w_{\pi,T}}(g,T) \tag{5}$$

$$+ V_{\max}\mathbb{E}_{(s,a)\sim\rho_T^\pi}\left[\mathbb{I}\left[\frac{\rho_T^\pi(s,a)}{\mu(s,a)} > \zeta\right]\right]\Bigg). \tag{6}$$

Then under the event $\mathcal{E}$ we have

$$\eta(T^\star,\pi) \ge \mathrm{lb}(T,\pi) - \frac{\epsilon_V(T,\pi)}{1-\gamma} - \frac{2V_{\max}}{1-\gamma}\sqrt{\frac{\zeta\iota}{n}}. \tag{7}$$

Proof of Lemma 3.3 is deferred to Appendix D.2. Accordingly, we design an offline policy optimization algorithm, stated in Alg. 1.

Algorithm 1 Model-based Offline RL with Local Misspecification Error

```
Require: parameter $\zeta$.
for $\pi \in \Pi, T \in \mathcal{T}$ do
    Compute $w_{\pi,T}(s,a) = \mathbb{I}\left[\frac{\rho_T^\pi(s,a)}{\mu(s,a)} \le \zeta\right]\frac{\rho_T^\pi(s,a)}{\mu(s,a)}$.
    Compute the lower bound $\mathrm{lb}(T,\pi)$ by Eq. (5).
end for
$\pi \leftarrow \mathrm{argmax}_{\pi\in\Pi}\max_{T\in\mathcal{T}}\mathrm{lb}(T,\pi)$.
```

The algorithm enumerates every policy and dynamics and computes the truncated density ratio $w_{\pi,T}$. Note that computing $w_{\pi,T}$ doesn't require collecting new samples from the true dynamics $T^\star$. Then, the algorithm computes a lower bound $\mathrm{lb}(T,\pi)$. Finally, it outputs a policy that maximizes the lower bound. Compared with existing model-based algorithms (Yu et al., 2020; Voloshin et al., 2021), our algorithm optimizes dynamics and policy jointly, which potentially leads to a better performance.

3.2. Safe Policy Improvement

In the sequel we show that the Alg. 1 guarantees a safe policy improvement up to error terms given below. For a fixed policy π, define

$$\epsilon_\rho(\pi) \triangleq \inf_{T\in\mathcal{T}}\mathbb{E}_{(s,a)\sim\rho_{T^\star}^\pi}[\mathrm{TV}\left(T(s,a), T^\star(s,a)\right)], \tag{8}$$

$$\epsilon_\mu(\pi) \triangleq \mathbb{E}_{(s,a)\sim\rho_{T^\star}^\pi}\left[\mathbb{I}\left[\frac{\rho_{T^\star}^\pi(s,a)}{\mu(s,a)} > \zeta/2\right]\right]. \tag{9}$$

The term ϵ_ρ measures the quality of the dynamics class, and ϵ_μ measure the quality of the dataset. In the following theorem, we prove that the true value of the policy computed by Alg. 1 is lower bounded by that of the optimal policy in the function class with some error terms.

Theorem 3.4. *Consider a fixed parameter ζ. Let $\hat{\pi}$ be the policy computed by Algorithm 1 and $\hat{T} = \operatorname{argmax}_T \operatorname{lb}(T, \hat{\pi})$. Let $\iota = \log(2|\mathcal{G}||\mathcal{T}||\Pi|/\delta)$. Then with probability at least $1-\delta$ we have*

$$\eta(T^\star, \hat{\pi}) \geq \sup_{\pi} \left\{ \eta(T^\star, \pi) - \frac{6V_{\max}\epsilon_\rho(\pi)}{(1-\gamma)^2} - \frac{V_{\max}\epsilon_\mu(\pi)}{1-\gamma} \right\} - \frac{\epsilon_V(\hat{T}, \hat{\pi})}{1-\gamma} - \frac{4V_{\max}}{1-\gamma}\sqrt{\frac{\zeta\iota}{n}}. \qquad (10)$$

Compared with Voloshin et al. (2021) and Yu et al. (2020), our safe policy improvement theorem is novel. In fact, the estimation in Voloshin et al. (2021) could be over-pessimistic without realizability (see Section A). Thanks to the density ratio $w_{\pi,T}$ in our loss function, the terms $\epsilon_\rho(\pi)$ and $\epsilon_\mu(\pi)$ are also only evaluated on the state-action pairs that policy π visits.

To prove Theorem 3.4, we show that the lower bound computed by Alg. 1 is tight. In other words, the lower bound $\max_T \operatorname{lb}(T, \pi)$ is at least as high as the true value of the policy with some errors. Consequently, maximizing the lower bound also maximizes the true value of the policy. Formally speaking, we have the following Lemma.

Lemma 3.5. *For any policy $\pi \in \Pi$, under the event $\mathcal{E}$ we have*

$$\max_{T\in\mathcal{T}} \operatorname{lb}(T, \pi) \geq \eta(T^\star, \pi) - \frac{6V_{\max}\epsilon_\rho(\pi)}{(1-\gamma)^2} - \frac{V_{\max}\epsilon_\mu(\pi)}{1-\gamma} - \frac{2V_{\max}}{1-\gamma}\sqrt{\frac{\zeta\iota}{n}}.$$

In the sequel we present a proof sketch for Lemma 3.5. In this proof sketch we hide $1/(1-\gamma)$ factors in the big-O notation. For a fixed policy π, let $\hat{T}$ be the minimizer of Eq. (8). We prove Lemma 3.5 by analyzing the three terms in the definition of $\operatorname{lb}(\hat{T}, \pi)$ (Eq. (5)) separately.

i. Following the definition of Eq. (8), we can show that $\|\rho^\pi_{\hat{T}} - \rho^\pi_{T^\star}\|_1 \leq \mathcal{O}(\epsilon_\rho(\pi))$. Consequently we get $\eta(\hat{T}, \pi) \geq \eta(T^\star, \pi) - \mathcal{O}(\epsilon_\rho(\pi))$.

ii. Recall that we assume $g(s,a) \in [0, V_{\max}], \forall g \in \mathcal{G}$. As a result, for any (s,a) we have $\sup_{g\in\mathcal{G}} \left|\mathbb{E}_{s'\sim\hat{T}(s,a)} g(s') - \mathbb{E}_{s'\sim T^\star(s,a)} g(s')\right| \leq V_{\max}\mathrm{TV}\left(\hat{T}(s,a), T^\star(s,a)\right)$. Combining the definition of $\ell_w(g, T)$, Eq. (8) and statistical error we get $\sup_{g\in\mathcal{G}} \ell_{w_{\pi,T}}(g, T) \leq \widetilde{\mathcal{O}}(\epsilon_\rho(\pi)) + \mathcal{O}(1/\sqrt{n})$ under event $\mathcal{E}$.

iii. For the last term regarding distribution mismatch, we combine Eq. (9) and Lemma E.1. We can upper bound this term by $\mathcal{O}(\epsilon_\rho(\pi) + \epsilon_\mu(\pi))$.

Proof of Lemma 3.5 is deferred to Appendix D.3. Theorem 3.4 follows directly from combining Lemma 3.3 and Lemma 3.5, and is shown in Appendix D.4.

3.3. Concrete Examples and Experiments

In Section A, we construct a concrete example where the dynamics class can only model a part of the state-action space. In this case, we prove that

i. it is suboptimal to first learn a dynamics, and then optimize a policy w.r.t. the learned dynamics (see Theorem A.1), and

ii. for off-policy estimation problem, the estimation error in Voloshin et al. (2021) is as large as a constant (see Proposition A.2).

Our safe policy improvement theorem together with these negative results show that Alg. 1 is more robust to model misspecification.

In Section B, we illustrate how we can use our approach to obtain the optimal policy in offline model-based reinforcement learning with model and distribution mismatch. We implement Alg. 1 in a misspecified one-dimensional LQR problem, and visualizes the result (see Figure 1). Table 1 shows that Alg. 1 indeed outperforms Voloshin et al. (2021) with model-misspecification and distribution mismatch.

4. Conclusion

This paper studies model-based offline reinforcement learning with local model misspecification errors, and proves a novel safe policy improvement theorem. By theory and experiment, we show that our algorithm outperforms existing ones in settings without realizability. For future work, we raise the following open questions:

1. We assume that the data distribution μ is known in order to compute the density ratio $w_{\pi,T}(s,a)$. Can we still guarantee safe policy improvement when μ is estimated, or even unknown?
2. The objective $\operatorname{lb}(T, \pi)$ involves the density ratio $w_{\pi,T}(s,a)$. In our implementation this term is not differentiable. As a result, the optimization problem in Line 5 of Alg. 1 is hard to solve for complex function classes such as neural networks. Can we design differentiable methods to compute the objective $\operatorname{lb}(T, \pi)$ and implement our algorithm for neural networks?

Acknowledgment

Research reported in this workshop paper was sponsored in part by the DEVCOM Army Research Laboratory under

Cooperative Agreement W911NF-17-2-0196 (ARL IoBT CRA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S.Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Argenson Arthur, Dulac-Arnold Gabriel. Model-based offline planning // arXiv preprint arXiv:2008.05556. 2020.

Bertsekas Dimitri P, others . Dynamic programming and optimal control: Vol. 1. 2000.

Buckman Jacob, Gelada Carles, Bellemare Marc G. The importance of pessimism in fixed-dataset policy optimization // arXiv preprint arXiv:2009.06799. 2020.

Chen Jinglin, Jiang Nan. Information-Theoretic Considerations in Batch Reinforcement Learning // International Conference on Machine Learning. 2019. 1042–1051.

Chen Lin, Scherrer Bruno, Bartlett Peter L. Infinite-Horizon Offline Reinforcement Learning with Linear Function Approximation: Curse of Dimensionality and Algorithm // arXiv preprint arXiv:2103.09847. 2021.

Curi Sebastian, Berkenkamp Felix, Krause Andreas. Efficient Model-Based Reinforcement Learning through Optimistic Policy Search and Planning // Advances in Neural Information Processing Systems. 2020. 33.

Duan Yaqi, Jia Zeyu, Wang Mengdi. Minimax-optimal off-policy evaluation with linear function approximation // International Conference on Machine Learning. 2020. 2701–2709.

Dudík Miroslav, Langford John, Li Lihong. Doubly robust policy evaluation and learning // Proceedings of the 28th International Conference on International Conference on Machine Learning. 2011. 1097–1104.

Farahmand Amir-massoud, Barreto Andre, Nikovski Daniel. Value-aware loss function for model-based reinforcement learning // Artificial Intelligence and Statistics. 2017. 1486–1494.

Fu Justin, Levine Sergey. Offline Model-Based Optimization via Normalized Maximum Likelihood Estimation // arXiv preprint arXiv:2102.07970. 2021.

Hasselt Hado P van, Hessel Matteo, Aslanides John. When to use parametric models in reinforcement learning? // Advances in Neural Information Processing Systems. 32. 2019.

Jiang Nan, Huang Jiawei. Minimax confidence interval for off-policy evaluation and policy optimization // arXiv preprint arXiv:2002.02081. 2020.

Jin Ying, Yang Zhuoran, Wang Zhaoran. Is Pessimism Provably Efficient for Offline RL? // arXiv preprint arXiv:2012.15085. 2020.

Kakade Sham, Langford John. Approximately Optimal Approximate Reinforcement Learning // Proceedings of the Nineteenth International Conference on Machine Learning. 2002. 267–274.

Kallus Nathan, Uehara Masatoshi. Double reinforcement learning for efficient off-policy evaluation in markov decision processes // Journal of Machine Learning Research. 2020. 21, 167. 1–63.

Kidambi Rahul, Rajeswaran Aravind, Netrapalli Praneeth, Joachims Thorsten. Morel: Model-based offline reinforcement learning // arXiv preprint arXiv:2005.05951. 2020.

Li Lihong, Munos Rémi, Szepesvári Csaba. Toward minimax off-policy value estimation // Artificial Intelligence and Statistics. 2015. 608–616.

Liu Qiang, Li Lihong, Tang Ziyang, Zhou Dengyong. Breaking the curse of horizon: infinite-horizon off-policy estimation // Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018a. 5361–5371.

Liu Y, Gottesman O, Raghu A, Komorowski M, Faisal A, Doshi-Velez F, Brunskill E. Representation Balancing MDPs for Off-Policy Policy Evaluation // Advances in neural information processing systems. 2018b.

Liu Yao, Bacon Pierre-Luc, Brunskill Emma. Understanding the curse of horizon in off-policy evaluation via conditional importance sampling // International Conference on Machine Learning. 2020a. 6184–6193.

Liu Yao, Swaminathan Adith, Agarwal Alekh, Brunskill Emma. Provably Good Batch Off-Policy Reinforcement Learning Without Great Exploration // Advances in Neural Information Processing Systems. 2020b. 33.

Malik Ali, Kuleshov Volodymyr, Song Jiaming, Nemer Danny, Seymour Harlan, Ermon Stefano. Calibrated model-based deep reinforcement learning // International Conference on Machine Learning. 2019. 4314–4323.

Matsushima Tatsuya, Furuta Hiroki, Matsuo Yutaka, Nachum Ofir, Gu Shixiang. Deployment-efficient reinforcement learning via model-based offline optimization // arXiv preprint arXiv:2006.03647. 2020.

DualDICE: Behavior-Agnostic Estimation of Discounted Stationary Distribution Corrections. // . 2019.

Quillen Deirdre, Jang Eric, Nachum Ofir, Finn Chelsea, Ibarz Julian, Levine Sergey. Deep reinforcement learning for vision-based robotic grasping: A simulated comparative evaluation of off-policy methods // 2018 IEEE International Conference on Robotics and Automation (ICRA). 2018. 6284–6291.

Rashidinejad Paria, Zhu Banghua, Ma Cong, Jiao Jiantao, Russell Stuart. Bridging Offline Reinforcement Learning and Imitation Learning: A Tale of Pessimism // arXiv preprint arXiv:2103.12021. 2021.

Ren Tongzheng, Li Jialian, Dai Bo, Du Simon S, Sanghavi Sujay. Nearly Horizon-Free Offline Reinforcement Learning // arXiv preprint arXiv:2103.14077. 2021.

Singh Avi, Yu Albert, Yang Jonathan, Zhang Jesse, Kumar Aviral, Levine Sergey. COG: Connecting New Skills to Past Experience with Offline Reinforcement Learning // arXiv preprint arXiv:2010.14500. 2020.

Swaminathan Adith, Joachims Thorsten. Batch learning from logged bandit feedback through counterfactual risk minimization // The Journal of Machine Learning Research. 2015. 16, 1. 1731–1755.

Swazinna Phillip, Udluft Steffen, Runkler Thomas. Overcoming Model Bias for Robust Offline Deep Reinforcement Learning // arXiv preprint arXiv:2008.05533. 2020.

Thomas Philip, Brunskill Emma. Data-efficient off-policy policy evaluation for reinforcement learning // International Conference on Machine Learning. 2016. 2139–2148.

Thomas Philip S, Theocharous Georgios, Ghavamzadeh Mohammad, Durugkar Ishan, Brunskill Emma. Predictive Off-Policy Policy Evaluation for Nonstationary Decision Problems, with Applications to Digital Marketing. // AAAI. 2017. 4740–4745.

Uehara Masatoshi, Huang Jiawei, Jiang Nan. Minimax weight and q-function learning for off-policy evaluation // International Conference on Machine Learning. 2020. 9659–9668.

Voloshin Cameron, Jiang Nan, Yue Yisong. Minimax Model Learning // International Conference on Artificial Intelligence and Statistics. 2021. 1612–1620.

Wang Ruosong, Foster Dean P, Kakade Sham M. What are the Statistical Limits of Offline RL with Linear Function Approximation? // arXiv preprint arXiv:2010.11895. 2020.

Wang Ruosong, Wu Yifan, Salakhutdinov Ruslan, Kakade Sham M. Instabilities of Offline RL with Pre-Trained Neural Representation // arXiv preprint arXiv:2103.04947. 2021.

Xiao Chenjun, Wu Yifan, Lattimore Tor, Dai Bo, Mei Jincheng, Li Lihong, Szepesvari Csaba, Schuurmans Dale. On the Optimality of Batch Policy Optimization Algorithms // arXiv preprint arXiv:2104.02293. 2021.

Xie T, Ma Y, Wang YX. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. // Advances in neural information processing systems. 2019.

Xie Tengyang, Jiang Nan. Batch value-function approximation with only realizability // arXiv preprint arXiv:2008.04990. 2020.

Yin Ming, Bai Yu, Wang Yu-Xiang. Near optimal provable uniform convergence in off-policy evaluation for reinforcement learning // arXiv preprint arXiv:2007.03760. 2020.

Yin Ming, Bai Yu, Wang Yu-Xiang. Near-Optimal Offline Reinforcement Learning via Double Variance Reduction // arXiv preprint arXiv:2102.01748. 2021.

Yin Ming, Wang Yu-Xiang. Asymptotically efficient off-policy evaluation for tabular reinforcement learning // International Conference on Artificial Intelligence and Statistics. 2020. 3948–3958.

Yu Tianhe, Kumar Aviral, Rafailov Rafael, Rajeswaran Aravind, Levine Sergey, Finn Chelsea. Combo: Conservative offline model-based policy optimization // arXiv preprint arXiv:2102.08363. 2021.

Yu Tianhe, Thomas Garrett, Yu Lantao, Ermon Stefano, Zou James, Levine Sergey, Finn Chelsea, Ma Tengyu. MOPO: Model-based Offline Policy Optimization // arXiv preprint arXiv:2005.13239. 2020.

Zanette Andrea. Exponential Lower Bounds for Batch Reinforcement Learning: Batch RL can be Exponentially Harder than Online RL // arXiv preprint arXiv:2012.08005. 2020.

Zhan Xianyuan, Zhu Xiangyu, Xu Haoran. Model-Based Offline Planning with Trajectory Pruning // arXiv preprint arXiv:2105.07351. 2021.

Zhang Ruiyi, Dai Bo, Li Lihong, Schuurmans Dale. GenDICE: Generalized Offline Estimation of Stationary Values // International Conference on Learning Representations. 2019.

Zhang Shangtong, Liu Bo, Whiteson Shimon. Gradientdice: Rethinking generalized offline estimation of stationary values // International Conference on Machine Learning. 2020. 11194–11203.

A. Concrete Examples

In this section we show a concrete example where Alg. 1 has better performance than existing approaches. The intuition is quite simple. We construct an MDP whose state space is partitioned into several parts. The function class is restricted in a sense that every function can only model one part of the state space. Transitions are designed so that every deterministic policy only visits one part of the state space. As a result, the local misspecification error is small. In contrast, if the dynamics is learned to fit the whole state space, the estimation error will be huge.

For a fixed parameter d, consider a tabular MDP where $\mathcal{S} = \{s_0, \cdots, s_d\} \cup \{s_g, s_b\}$. There are d actions for each state in $\{s_0, \cdots, s_d\}$, denoted by $a_1, \cdots, a_d$. We assume that there is an known feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, which will be specified later. The transition function is given as follows.

$$T^\star(s_0, a_i) = s_i, \tag{11}$$

$$T^\star(s_i, a_j) = \begin{cases} s_g, & \text{w.p. } \frac{1}{2}\big(1 + \phi(s_i, a_j)^\top \theta_i^\star\big), \\ s_b, & \text{w.p. } \frac{1}{2}\big(1 - \phi(s_i, a_j)^\top \theta_i^\star\big), \end{cases} \tag{12}$$

$$T^\star(s_g, a_i) = s_g, \forall i \in [d], \tag{13}$$

$$T^\star(s_b, a_i) = s_b, \forall i \in [d]. \tag{14}$$

And the reward function is non-zero only for state s_g. That is, $r(s_g, a_i) = 1, \forall i \in [d]$. Note that the state spaces are layered. The first layer contains state s_0. The second layer is $\{s_1, \cdots, s_d\}$ and the third layer is $\{s_g, s_d\}$. The feature map (for the second layer) is defined as $\phi(s_i, a_j) = e_j$.

The transition function class is parameterized by $\theta \in \mathbb{R}^d$ with $\|\theta\|_2 = 1$. For a fixed θ, the transition at second layer is given by

$$T_\theta(s_i, a_j) = \begin{cases} s_g, & \text{w.p. } \frac{1}{2}\big(1 + \phi(s_i, a_j)^\top \theta\big), \\ s_b, & \text{w.p. } \frac{1}{2}\big(1 - \phi(s_i, a_j)^\top \theta\big), \end{cases} \tag{15}$$

And the transition at other layers is exactly the same as $T^\star$. Note that for the ground-truth transition $T^\star$, the parameter $\theta_i^\star$ varies across different states. But transition T_θ in the function class must use the same parameter θ to approximate the dynamics in every state. As a result, the standard realizability assumptions do not hold in this example.

Optimal policy with respect to a learned dynamics is suboptimal. We consider a family of transitions denoted by $\{T_\kappa\}_{\kappa=1}^d$, where in T_κ we set $\theta_i^\star = \mathbb{I}[i = \kappa]\, e_i$. Consequently, the optimal action for state s_0 is a_κ. Note that for any fixed θ, the transitions at $s_1, \cdots, s_d$ are identical in T_θ. Therefore the optimal action induced by T_θ is suboptimal. In other words, algorithms that decouples the learning of dynamics and policy are suboptimal.

Theorem A.1. *Consider any (possibly stochastic) algorithm $\mathcal{A}$ that outputs an estimated policy T_θ. Let π_θ be the greedy policy w.r.t. T_θ (with ties breaking data-independently). Then for $d \geq 2$ we have*

$$\sup_{\mathcal{A}} \max_{\kappa} \Big(\max_{\pi} \eta(T_\kappa, \pi) - \eta(T_\kappa, \pi_\theta) \Big) \geq \frac{\gamma^2}{2(1-\gamma)}. \tag{16}$$

Proof of Theorem A.1 is deferred to Appendix D.5.

Error bounds in Voloshin et al. (2021) is not tight. Theorem A.1 proves the suboptimality of any algorithm (including MOPO (Yu et al., 2020) and MML (Voloshin et al., 2021)) that learns a dynamics first and output the optimal policy. In the sequel, we prove that for off-policy estimation problem, the estimation error in Voloshin et al. (2021) can be large without realizability.

In this case, we consider a fixed true dynamics $T^\star$ where $\theta_i^\star = e_i$. Voloshin et al. (2021) require an density ratio class $\mathcal{W} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$. Theorem 3.1 of (Voloshin et al., 2021) states that when $w_{\pi,T} \in \mathcal{W}$ and $V_{T^\star}^\pi \in \mathcal{G}$,

$$|\eta(T, \pi) - \eta(T^\star, \pi)| \leq \gamma \min_{T \in \mathcal{T}} \max_{w \in \mathcal{W}, g \in \mathcal{G}} |\ell_w(g, T)|. \tag{17}$$

But we have the following proposition.

Proposition A.2. *Consider the set the dynamics class $\mathcal{T} = \{T_\theta : \theta \in S^{d-1}\}$. Let $\Pi = \{\pi_x : x \in [d]\}$ where $\pi_x(s_i) = a_x$ for $0 \le i \le d$ and $\pi_x(s_g) = \pi_x(s_b) = a_1$. Let $\mathcal{W}$ be the density ratio induced by π. Even with $\mathcal{G} = \{V_{T^\star}^{\pi_x} : x \in [d]\}$ and infinite number of data, we have*

$$\min_{T \in \mathcal{T}} \max_{w \in \mathcal{W}, g \in \mathcal{G}} |\ell_w(g, T)| \ge \frac{\gamma}{4(1-\gamma)}. \tag{18}$$

Proposition A.2 is proved by explicitly computing the right hand side of Eq. (17). We defer the proof to Appendix D.6. In contrast, error terms in Theorem 3.4 converges to zero when $\zeta > \mathrm{poly}(d, 1/(1-\gamma))$ and $n \to \infty$ in the same setting.

B. Experiments

We illustrate how we can use our approach to obtain the optimal policy in offline model-based reinforcement learning with model and distribution mismatch. We empirically evaluate our method on Linear-Quadratic Regulator (LQR), a commonly used environment in optimal control theory (Bertsekas, others, 2000). We seek to answer the following question: *Does algorithm 1 return the optimal policy when we have both model and distribution mismatch?* We first describe the environment, the baseline algorithm we compare against, and then provide the results.

Linear-Quadratic Regulator (LQR) The Linear-Quadratic Regulator (LQR) is defined by a linear transition dynamics in the form of $s_{t+1} = As_t + Ba_t + \eta$, where $s_t \in \mathbb{R}^n$ and $a_t \in \mathbb{R}^m$ are state and action and time step t, respectively. $\eta \sim \mathcal{N}(0, \sigma^2 I)$ is random noise in the system. LQR has quadratic reward function $\mathcal{R}(s, a) = -(s^T Qs + a^T Ra)$ with $Q \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{m\times m}$ being positive semi-definite matrices, $Q, R \succeq 0$.

The optimal controller in LQR to maximize the sum of future rewards $\sum_{t=1}^{H} -(s_t^T Qs_t + a_t^T Ra_t)$ until the end of horizon H has the form $a_t = -Ks_t$ for some $K \in \mathbb{R}^{m\times n}$ (Bertsekas, others, 2000). Additionally, the value function is also a quadratic function, $V(s) = s^T Us + q$ for some constant q and positive semi-definite matrix $U \succeq 0$ (Voloshin et al., 2021).

Data Generation We use $1D$ version of LQR in our experiments with $A(x) = (1 + x/10), B(x) = (-0.5 - x/10)$, $Q = 1, R = 1$ and noise $\eta \sim \mathcal{N}(0, 0.01)$. The true model has $x = 6$ with the optimal policy $K = 1.0$. We use $K = 0.8$ to generate 100 trajectories each with 100 steps, starting from the initial state distribution $s_0 \sim \mathcal{N}(1.0, 0.1)$. This process generates the dataset $\mathcal{D}$ that we use for both our algorithm and the baseline to compute the optimal policy.

Baseline We compare our algorithm to minimizing MML loss as described in the OPO algorithm of Voloshin et al. (2021, Algorithm 2). MML strictly outperformed VAML (Farahmand et al., 2017) as shows in the experiments of Voloshin et al. (2021); hence, we only compare to MML in our experiments.

B.1. Results and Discussion

We use the following model class $T(x) \in \mathcal{T}$ parametrized by x,

$$T(x) = \begin{cases} s_t < 0.05: & s_{t+1} = (1 + x/10)s_t - (0.5 + x/10)a_t \\ s_t \ge 0.05: & s_{t+1} = -(1 + x/10)s_t - (0.5 + x/10)a_t \end{cases}$$

These models are mis-specified in two way, first they are deterministic, whereas the true model is stochastic and second, for $s \ge 0.05$ they show a completely different behaviour. We consider $x \in \{0, 2, 4, 6, 8\}$ in our experiments. Test function class $\mathcal{G}$ consist of quadratic function corresponding to value functions of different model, policy pairs (we used the same class for value function V in MML). Similarly for the MML loss, we compute the weight function $w(s, a)$ corresponding to the same model, policy pair. For more information reader may refer to the supplementary materials.

Table 1 shows the result of our experiments. We compare the OPO loss, $|\eta(T^\star, \pi^\star_{\hat{T}}) - \eta(T^\star, \pi^\star_{T^\star})|$, that is the difference of the return in the true model of the optimal policy versus the best policy based on the method. These results shows that our method, will pick the optimal policy for the true model. This is in contrast to MML that picks the wrong model and therefore the wrong policy. The 0.148 difference is about 5.7% difference in the return. We note that, optimization in our experiment is over a discrete set so there is no randomness caused by initialization. In 10 runs it obtained the same optimal value. That is both ours and MML picked the same policy over 10 different runs. We reported average difference between the policy values in Table 1.

Hence we did not put any confidence interval, as mentioned in the main text.

Table 1. LQR OPO error

Algorithm	MML(Voloshin et al., 2021)	Ours
OPO Error	0.148	0

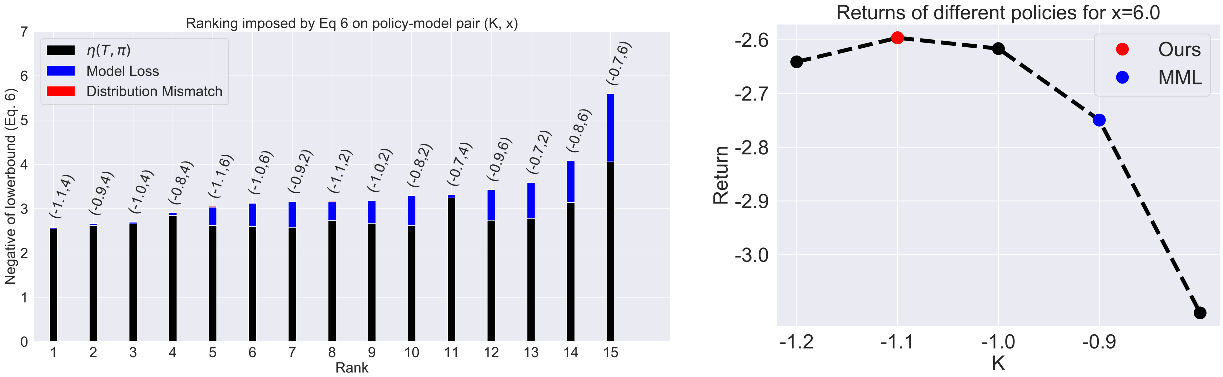

Figure 1. **Left:** Visualization of negative lower bounds $\text{lb}(T,\pi)$ for different policies and models. Distribution mismatch error in this case is mild. Note that in our algorithm, the model loss also depends on the policy. **Right:** Visualization of true policy value $\eta(T^\star,\pi)$. Our algorithm picks the optimal policy, whereas MML picks a suboptimal policy.

This result highlights the main benefit of our method over the baseline. Since all models in the model class are misspecified, the maximization over weight function w in the MML loss results in an unrealistically large loss value for some models. However, if the chosen policy does not explore the part of the state space with a large model mismatch, there is no need to incur a high penalty. This is exactly why our method accounts for weights separately and can extract the optimal policy even with the misspecified model.

B.2. Visualization of Lower Bounds

We visualize the effect of three terms in the definition of $\text{lb}(T,\pi)$ in Figure 1. We can see that the model loss for different policy is different (e.g., model loss for $(K,x)=(-0.7,6)$ is significantly larger than $(-1.1,6)$, even if the dynamics is the same). This is because the model loss is evaluated with a different density ratio. The MML algorithm evaluates model loss by taking a maximum over all possible density ratios, and consequently, over penalize imperfect models.

C. Related Works

The most related result to our paper is Voloshin et al. (2021). The loss function for model learning in (Voloshin et al., 2021) is almost the same as ours, except that we explicitly compute the density ratio where Voloshin et al. (2021) maximize over all possible ratios. We also consider distribution mismatch explicitly. The differences enable us to show the tightness of our estimation (see Lemma 3.5 and Proposition A.2). Algorithmically, we optimize dynamics and policy at the same time. The MML algorithm (Voloshin et al., 2021) learns the dynamics first, and then optimizes a policy. We also show in Theorem A.1 that this decoupling in the learning of dynamics and policy is suboptimal.

There are extensive researches on model-free methods for offline reinforcement learning. Nachum et al. (2019) and their follow-ups (Zhang et al., 2019; 2020) learn a distribution correction term, on top of which they perform evaluation or policy optimization tasks. Uehara et al. (2020); Jiang, Huang (2020) study the duality between learning Q-functions and learning importance weights. Liu et al. (2020b) explicitly consider the distribution shift in offline reinforcement learning and propose conservative Bellman equations. There are also several papers that study the pessimism principle in different settings (Jin et al., 2020; Rashidinejad et al., 2021; Xiao et al., 2021; Buckman et al., 2020).

Another line of research uses model-based methods (Kidambi et al., 2020; Yu et al., 2021; Matsushima et al., 2020; Swazinna et al., 2020; Fu, Levine, 2021). Instead of using maximum likelihood loss, Farahmand et al. (2017) propose a value-aware loss function for model learning using linear function approximator. Yu et al. (2020) build an uncertainty quantification on top of the learned dynamics, and select a policy that optimizes the lower confidence bound. Other papers focus on policy optimization instead of model learning (Argenson, Dulac-Arnold, 2020; Zhan et al., 2021).

We also consider the effect of finite dataset. In Table 2, we compare the finite sample error bounds. Our bound is better or at least comparable with existing results.

Table 2. Comparison of statistical errors

Algorithm	VAML(Farahmand et al., 2017)	MBS-PI(Liu et al., 2020b)	MML(Voloshin et al., 2021)	Ours
Statistical Error	$\widetilde{\mathcal{O}}\left(\frac{p}{\sqrt{n}}\right)$[1]	$\widetilde{\mathcal{O}}\left(\frac{V_{\max}\zeta}{(1-\gamma)^2\sqrt{n}}\right)$	$\mathfrak{R}_n$[2]	$\widetilde{\mathcal{O}}\left(\frac{V_{\max}}{1-\gamma}\sqrt{\frac{\zeta}{n}}\right)$

In addition to the flourish researches in offline reinforcement learning algorithms, there are also several lower bounds with linear function approximation (Zanette, 2020; Wang et al., 2020; 2021). Our results do not contradict these lower bounds because we assume that the offline dataset covers the state-action space, where the lower bounds focus on covering the feature space.

D. Missing Proofs

D.1. High Probability Events

In this section we introduce concentration inequalities and define the high probability events.

Define the following quantities

$$L(\pi, g, T) = \mathbb{E}_{(s,a,s')\sim\mu}\left[w_{\pi,T}(s,a)(\mathbb{E}_{x\sim T(s,a)}[g(x)] - \mathbb{E}_{x\sim T^\star(s,a)}[g(x)])\right], \tag{19}$$

$$l(\pi, g, T) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}}[w_{\pi,T}(s,a)(f_T^g(s,a) - g(s'))]. \tag{20}$$

Recall that $\iota = \log(2|\mathcal{G}||\mathcal{T}||\Pi|/\delta)$. Consider the event

$$\mathcal{E} = \left\{|L(\pi, g, T) - l(\pi, g, T)| \le 2V_{\max}\sqrt{\frac{\zeta\iota}{n}}, \quad \forall \pi \in \Pi, g \in \mathcal{G}, T \in \mathcal{T}\right\}. \tag{21}$$

In the following we show that

$$\Pr(\mathcal{E}) \ge 1 - \delta. \tag{22}$$

Recall that $\mathcal{D} = \{(s_i, a_i, s_i')\}_{i=1}^n$ where $(s_i, a_i, s_i') \sim \mu$ are i.i.d. samples from distribution μ. For fixed $\pi \in \Pi, g \in \mathcal{G}, T \in \mathcal{T}$, we have $\mathbb{E}[\hat{l}(\pi, g, T)] = l(\pi, g, T)$. Meanwhile, note that

$$|w_{\pi,T}(s,a)(f_T^g(s,a) - g(s'))| \le V_{\max}\zeta, \tag{23}$$

$$\mathbb{E}_{(s,a,s')\sim\mu}[w_{\pi,T}(s,a)^2(f_T^g(s,a) - g(s'))^2] \tag{24}$$

$$\le \mathbb{E}_{(s,a,s')\sim\rho_T^\pi}[w_{\pi,T}(s,a)(f_T^g(s,a) - g(s'))^2] \le V_{\max}^2\zeta. \tag{25}$$

By Bernstein inequality, with probability at least $1 - \delta/(|\mathcal{G}||\mathcal{T}||\Pi|)$,

$$|L(\pi, g, T) - l(\pi, g, T)| \le \sqrt{\frac{2V_{\max}^2\zeta\log(2|\mathcal{G}||\mathcal{T}||\Pi|/\delta)}{n}} + \frac{V_{\max}\zeta}{3n}\log(2|\mathcal{G}||\mathcal{T}||\Pi|/\delta) \tag{26}$$

Recall that $\iota = \log(2|\mathcal{G}||\mathcal{T}||\Pi|/\delta)$. When $n \ge \zeta$ we have

$$|L(\pi, g, T) - l(\pi, g, T)| \le 2V_{\max}\sqrt{\frac{\zeta\iota}{n}}. \tag{27}$$

Note that when $n < \zeta$, $\mathcal{E}$ trivially holds. As a result, applying union bound we prove Eq. (22).

[2] p is the dimension of the feature vector.

[2] Rademacher complexity of an induced function class. For finite hypothesis, the best known upper bound is in the same order of ours.

D.2. Proof of Lemma 3.3

Proof. In the following we consider a fixed policy π and dynamics $T \in \mathcal{T}$. We use w to denote $w_{\pi,T}$ when the context is clear.

By basic algebra we get

$$\begin{aligned}
&\left|\mathbb{E}_{(s,a)\sim\rho_T^\pi}[G_T^\pi(s,a)]\right| && (28)\\
\le& \left|\mathbb{E}_{(s,a)\sim\rho_T^\pi}\left[\mathbb{I}\left[\frac{\rho_T^\pi(s,a)}{\mu(s,a)} \le \zeta\right] G_T^\pi(s,a)\right]\right| + \mathbb{E}_{(s,a)\sim\rho_T^\pi}\left[\mathbb{I}\left[\frac{\rho_T^\pi(s,a)}{\mu(s,a)} > \zeta\right] \left|G_T^\pi(s,a)\right|\right] && (29)\\
\le& \left|\mathbb{E}_{(s,a)\sim\mu}[w(s,a)G_T^\pi(s,a)]\right| + V_{\max}\mathbb{E}_{(s,a)\sim\rho_T^\pi}\left[\mathbb{I}\left[\frac{\rho_T^\pi(s,a)}{\mu(s,a)} > \zeta\right]\right]. && (30)
\end{aligned}$$

Consequently, in the following we prove

$$\left|\mathbb{E}_{(s,a)\sim\mu}[w(s,a)G_T^\pi(s,a)]\right| \le \sup_{g\in\mathcal{G}} \ell_w(g,T) + \epsilon_V(T,\pi) + 2V_{\max}\sqrt{\frac{\zeta\iota}{n}}.$$

Let $L_w(g,T) = \left|\mathbb{E}_{(s,a,s')\sim\mu}\left[w(s,a)(\mathbb{E}_{x\sim T(s,a)}[g(x)] - \mathbb{E}_{x\sim T^\star(s,a)}[g(x)])\right]\right|$ be the population error. Recall that under the high probability event $\mathcal{E}$ in Eq. (21), for any $g \in \mathcal{G}$ and $T \in \mathcal{T}$

$$|L_w(g,T) - \ell_w(g,T)| \le 2V_{\max}\sqrt{\frac{\zeta\iota}{n}}. \tag{31}$$

Now by the definition of $G_T^\pi(s,a)$, for any $g \in \mathcal{G}$ we have

$$\begin{aligned}
&\left|\mathbb{E}_{(s,a)\sim\mu}[w(s,a)G_T^\pi(s,a)]\right| && (32)\\
=& \left|\mathbb{E}_{(s,a)\sim\mu}\left[w(s,a)\big(\mathbb{E}_{s'\sim T(s,a)}[V_{T^\star}^\pi(s')] - \mathbb{E}_{s'\sim T^\star(s,a)}[V_{T^\star}^\pi(s')]\big)\right]\right| && (33)\\
\le& \left|\mathbb{E}_{(s,a)\sim\mu}\left[w(s,a)\big(\mathbb{E}_{s'\sim T(s,a)}[g(s')] - \mathbb{E}_{s'\sim T^\star(s,a)}[g(s')]\big)\right]\right| && (34)\\
&+ \left|\mathbb{E}_{(s,a)\sim\mu}\left[w(s,a)\big(\mathbb{E}_{s'\sim T(s,a)}[g(s') - V_{T^\star}^\pi(s')] + \mathbb{E}_{s'\sim T^\star(s,a)}[g(s') - V_{T^\star}^\pi(s')]\big)\right]\right|. && (35)
\end{aligned}$$

Define

$$\hat{g} = \operatorname*{argmin}_{g\in\mathcal{G}} \left|\mathbb{E}_{(s,a)\sim\mu}\left[w(s,a)\left(\mathbb{E}_{s'\sim T(s,a)}[g(s') - V_{T^\star}^\pi(s')] + \mathbb{E}_{s'\sim T^\star(s,a)}[g(s') - V_{T^\star}^\pi(s')]\right)\right]\right|.$$

Since g is arbitrarily, continuing Eq. (35) and recalling Definition 3.2 we get

$$\begin{aligned}
&\left|\mathbb{E}_{(s,a)\sim\mu}[w(s,a)G_T^\pi(s,a)]\right| && (36)\\
\le& \left|\mathbb{E}_{(s,a)\sim\mu}\left[w(s,a)\big(\mathbb{E}_{s'\sim T(s,a)}[\hat{g}(s')] - \mathbb{E}_{s'\sim T^\star(s,a)}[\hat{g}(s')]\big)\right]\right| + \epsilon_V(T,\pi) && (37)\\
\le& \sup_{g\in\mathcal{G}}\left|\mathbb{E}_{(s,a)\sim\mu}\left[w(s,a)\big(\mathbb{E}_{s'\sim T(s,a)}[g(s')] - \mathbb{E}_{s'\sim T^\star(s,a)}[g(s')]\big)\right]\right| + \epsilon_V(T,\pi). && (38)
\end{aligned}$$

Combining Eq. (38) and Eq. (31) we get,

$$\begin{aligned}
\left|\mathbb{E}_{(s,a)\sim\mu}[w(s,a)G_T^\pi(s,a)]\right| &\le \sup_{g\in\mathcal{G}} L_w(g,T) + \epsilon_V(T,\pi) && (39)\\
&\le \sup_{g\in\mathcal{G}} \ell_w(g,T) + \epsilon_V(T,\pi) + 2V_{\max}\sqrt{\frac{\zeta\iota}{n}}. && (40)
\end{aligned}$$

Now plugging in Eq. (30) we get,

$$\begin{aligned}
&\left|\mathbb{E}_{(s,a)\sim\rho_T^\pi}[G_T^\pi(s,a)]\right|\\
\le& \sup_{g\in\mathcal{G}} \ell_w(g,T) + \epsilon_V(T,\pi) + 2V_{\max}\sqrt{\frac{\zeta\iota}{n}} + V_{\max}\mathbb{E}_{(s,a)\sim\rho_T^\pi}\left[\mathbb{I}\left[\frac{\rho_T^\pi(s,a)}{\mu(s,a)} > \zeta\right]\right].
\end{aligned}$$

Finally, combining with simulation lemma (Lemma 3.1) we finish the proof. □

D.3. Proof of Lemma 3.5

Proof of Lemma 3.5. Consider a fixed $\pi \in \Pi$. When the context is clear, we use ϵ_ρ and ϵ_μ to denote $\epsilon_\rho(\pi)$ and $\epsilon_\mu(\pi)$ respectively.

Consider the dynamics

$$\hat{T} = \operatorname*{argmin}_{T\in\mathcal{T}} \mathbb{E}_{(s,a)\sim\rho^\pi_{T^\star}}[\mathrm{TV}\left(T(s,a), T^\star(s,a)\right)]. \tag{41}$$

By the definition of ϵ_ρ we get

$$\mathbb{E}_{(s,a)\sim\rho^\pi_{T^\star}}\left[\mathrm{TV}\left(\hat{T}(s,a), T^\star(s,a)\right)\right] \le \epsilon_\rho.$$

Applying Lemma E.2 we get

$$\left\|\rho^\pi_{\hat{T}} - \rho^\pi_{T^\star}\right\|_1 \le \frac{\epsilon_\rho}{(1-\gamma)}. \tag{42}$$

The rest of the proof is organized in the following way. We bound the three terms in RHS of Eq. (5) respectively as follows

$$\eta(\hat{T},\pi) \ge \eta(T^\star,\pi) - \frac{V_{\max}}{1-\gamma}\epsilon_\rho, \tag{43}$$

$$\sup_{g\in\mathcal{G}} \ell_w(g,\hat{T}) \le \frac{2V_{\max}\epsilon_\rho}{1-\gamma} + 2V_{\max}\sqrt{\frac{\zeta\iota}{n}}, \tag{44}$$

$$\mathbb{E}_{(s,a)\sim\rho^\pi_{\hat{T}}}\left[\mathbb{I}\left[\frac{\rho^\pi_{\hat{T}}(s,a)}{\mu(s,a)} > \zeta\right]\right] \le \epsilon_\mu + \frac{3\epsilon_\rho}{(1-\gamma)}. \tag{45}$$

Then we combine these inequalities together to prove Lemma 3.5.

Step 1: Proving Eq. (43). Note that for every T and π, $\eta(T,\pi) = \frac{1}{1-\gamma}\left\langle \rho^\pi_T, r\right\rangle$ where r is the reward function. Then we have

$$\eta(T^\star,\pi) - \eta(\hat{T},\pi) = \frac{1}{1-\gamma}\left\langle \rho^\pi_{T^\star} - \rho^\pi_{\hat{T}}, r\right\rangle \le \frac{1}{1-\gamma}\left\|\rho^\pi_{T^\star} - \rho^\pi_{\hat{T}}\right\|_1 \|r\|_\infty. \tag{46}$$

Combining with Eq. (42) we get Eq. (43).

Step 2: Proving Eq. (44). For any fixed function $g \in \mathcal{G}$. Let $w = w_{\pi,\hat{T}}$ be a shorthand. Define

$$L_w(g,T) = \left|\mathbb{E}_{(s,a,s')\sim\mu}[w(s,a)(f^g_T(s,a) - g(s'))]\right|$$

to be the population error. Then we have

$$\begin{aligned}
& L_w(g,\hat{T}) \\
&= \left|\mathbb{E}_{(s,a)\sim\mu}\left[w(s,a)\Big(\mathbb{E}_{s'\sim\hat{T}(s,a)}[g(s')] - \mathbb{E}_{s'\sim T^\star(s,a)}[g(s')]\Big)\right]\right| \\
&= \left|\mathbb{E}_{(s,a)\sim\rho^\pi_{\hat{T}}}\left[\mathbb{I}\left[\frac{\rho^\pi_{\hat{T}}(s,a)}{\mu(s,a)} \le \zeta\right]\left(\mathbb{E}_{s'\sim\hat{T}(s,a)}[g(s')] - \mathbb{E}_{s'\sim T^\star(s,a)}[g(s')]\right)\right]\right| && \text{(By the definition of } w.\text{)} \\
&\le V_{\max}\mathbb{E}_{(s,a)\sim\rho^\pi_{\hat{T}}}\left[\mathbb{I}\left[\frac{\rho^\pi_{\hat{T}}(s,a)}{\mu(s,a)} \le \zeta\right]\mathrm{TV}\left(\hat{T}(s,a), T^\star(s,a)\right)\right] \\
&\le V_{\max}\mathbb{E}_{(s,a)\sim\rho^\pi_{T^\star}}\left[\mathrm{TV}\left(\hat{T}(s,a), T^\star(s,a)\right)\right] + \frac{V_{\max}\epsilon_\rho}{1-\gamma} && \text{(By Eq. (42))} \\
&\le V_{\max}\left(\epsilon_\rho + \frac{\epsilon_\rho}{1-\gamma}\right) \le \frac{2V_{\max}\epsilon_\rho}{1-\gamma}.
\end{aligned}$$

Under event $\mathcal{E}$ we have

$$\ell_w(g,\hat{T}) \le L_w(g,\hat{T}) + 2V_{\max}\sqrt{\frac{\zeta\iota}{n}}. \tag{47}$$

Because g is arbitrary, we get Eq. (44).

Step 3: Proving Eq. (45). Note that

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim\rho^{\pi}_{\hat{T}}}\left[\mathbb{I}\left[\frac{\rho^{\hat{\pi}}_{T}(s,a)}{\mu(s,a)} > \zeta\right]\right] && (48)\\
=\ &\mathbb{E}_{(s,a)\sim\rho^{\pi}_{\hat{T}}}\left[\mathbb{I}\left[\frac{\rho^{\pi}_{\hat{T}}(s,a)}{\rho^{\pi}_{T^\star}(s,a)}\frac{\rho^{\pi}_{T^\star}(s,a)}{\mu(s,a)} > \zeta\right]\right] && (49)\\
\le\ &\mathbb{E}_{(s,a)\sim\rho^{\pi}_{\hat{T}}}\left[\mathbb{I}\left[\frac{\rho^{\pi}_{\hat{T}}(s,a)}{\rho^{\pi}_{T^\star}(s,a)} > 2\right]\right] + \mathbb{E}_{(s,a)\sim\rho^{\pi}_{\hat{T}}}\left[\mathbb{I}\left[\frac{\rho^{\pi}_{T^\star}(s,a)}{\mu(s,a)} > \zeta/2\right]\right]. && (50)
\end{aligned}$$

With the help of Lemma E.1, we can upper bound the first term of Eq. (50) by the total variation between $\rho^{\pi}_{\hat{T}}$ and $\rho^{\pi}_{T^\star}$. Combining Lemma E.1 and Eq. (42) we get

$$\mathbb{E}_{(s,a)\sim\rho^{\pi}_{\hat{T}}}\left[\mathbb{I}\left[\frac{\rho^{\hat{\pi}}_{T}(s,a)}{\rho^{\pi}_{T^\star}(s,a)} > 2\right]\right] \le \frac{2\epsilon_\rho}{1-\gamma}. \tag{51}$$

On the other hand, combining Eq. (42) and definition of ϵ_μ we get

$$\mathbb{E}_{(s,a)\sim\rho^{\pi}_{\hat{T}}}\left[\mathbb{I}\left[\frac{\rho^{\pi}_{T^\star}(s,a)}{\mu(s,a)} > \zeta/2\right]\right] \le \mathbb{E}_{(s,a)\sim\rho^{\pi}_{T^\star}}\left[\mathbb{I}\left[\frac{\rho^{\pi}_{T^\star}(s,a)}{\mu(s,a)} > \zeta/2\right]\right] + \frac{\epsilon_\rho}{1-\gamma} \le \epsilon_\mu + \frac{\epsilon_\rho}{1-\gamma}.$$

Consequently, we get Eq. (45).

Now we stitch Eq. (42), Eq. (43) and Eq. (44) together. Combining with the definition of $\mathrm{lb}(\hat{T},\pi)$ in Eq. (5), we have

$$\begin{aligned}
\mathrm{lb}(\hat{T},\pi) &= \eta(\hat{T},\pi) - \frac{1}{1-\gamma}\left(\sup_{g\in\mathcal{G}}\left|\ell_{w_{\pi,T}}(g,\hat{T})\right| + V_{\max}\mathbb{E}_{(s,a)\sim\rho^{\pi}_{T}}\left[\mathbb{I}\left[\frac{\rho^{\pi}_{\hat{T}}(s,a)}{\mu(s,a)} > \zeta\right]\right]\right)\\
&\ge \eta(T^\star,\pi) - \frac{V_{\max}\epsilon_\rho}{1-\gamma} - \frac{2V_{\max}\epsilon_\rho}{(1-\gamma)^2} + \frac{2V_{\max}}{1-\gamma}\sqrt{\frac{\zeta\iota}{n}} - \frac{V_{\max}}{1-\gamma}\left(\frac{3\epsilon_\rho}{1-\gamma} + \epsilon_\mu\right)\\
&\ge \eta(T^\star,\pi) - \frac{6V_{\max}\epsilon_\rho}{(1-\gamma)^2} - \frac{V_{\max}\epsilon_\mu}{1-\gamma} - \frac{2V_{\max}}{1-\gamma}\sqrt{\frac{\zeta\iota}{n}}.
\end{aligned}$$

Note that $\hat{T}\in\mathcal{T}$, we have

$$\max_{T\in\mathcal{T}} \mathrm{lb}(T,\pi) \ge \mathrm{lb}(\hat{T},\pi), \tag{52}$$

which finishes the proof. □

D.4. Proof of Theorem 3.4

Proof of Theorem 3.4. Let $\hat{T},\hat{\pi} \leftarrow \mathrm{argmax}_{T\in\mathcal{T},\pi\in\Pi}\,\mathrm{lb}(T,\pi)$ be the dynamics and policy that maximizes the lower bound. Note hat $\hat{\pi}$ is the output of Algorithm 1.

Now under the event $\mathcal{E}$, by Lemma 3.5, for any policy π we have

$$\max_{T\in\mathcal{T}} \mathrm{lb}(T,\pi) \ge \eta(T^\star,\pi) - \frac{6V_{\max}\epsilon_\rho(\pi)}{(1-\gamma)^2} - \frac{V_{\max}\epsilon_\mu(\pi)}{1-\gamma} - \frac{2V_{\max}}{1-\gamma}\sqrt{\frac{\zeta\iota}{n}}. \tag{53}$$

On the other hand, under the event $\mathcal{E}$, by Lemma 3.3 we get

$$\eta(T^\star,\pi) \ge \mathrm{lb}(\hat{T},\hat{\pi}) - \frac{\epsilon_V(\hat{T},\hat{\pi})}{1-\gamma} - \frac{2V_{\max}}{1-\gamma}\sqrt{\frac{\zeta\iota}{n}}. \tag{54}$$

By the optimality of $\hat{T}, \hat{\pi}$, we have $\mathrm{lb}(\hat{T}, \hat{\pi}) \ge \sup_{T \in \mathcal{T}} \mathrm{lb}(T, \pi)$ for any π. As a result, combining with Eq. (53) and Eq. (54) we get

$$\begin{aligned} \eta(T^\star, \hat{\pi}) &\ge \mathrm{lb}(\hat{T}, \hat{\pi}) - \frac{\epsilon_V(\hat{T}, \hat{\pi})}{1-\gamma} - \frac{2V_{\max}}{1-\gamma}\sqrt{\frac{\zeta\iota}{n}} && (55)\\ &\ge \sup_{\pi \in \Pi} \sup_{T \in \mathcal{T}} \mathrm{lb}(T, \pi) - \frac{\epsilon_V(\hat{T}, \hat{\pi})}{1-\gamma} - \frac{2V_{\max}}{1-\gamma}\sqrt{\frac{\zeta\iota}{n}} && (56)\\ &\ge \sup_{\pi} \left\{ \eta(T^\star, \pi) - \frac{6V_{\max}\epsilon_\rho(\pi)}{(1-\gamma)^2} - \frac{V_{\max}\epsilon_\mu(\pi)}{1-\gamma} \right\} - \frac{\epsilon_V(\hat{T}, \hat{\pi})}{1-\gamma} - \frac{4V_{\max}}{1-\gamma}\sqrt{\frac{\zeta\iota}{n}}. && (57) \end{aligned}$$

□

D.5. Proof of Theorem A.1

Proof of Theorem A.1. Note that for any $\theta \in \mathbb{R}^d$, the transition function for state $s_1, \cdots, s_d$ are identical. As a result, $V^\pi_{T_\theta}(s_1) = V^\pi_{T_\theta}(s_2) = \cdots = V^\pi_{T_\theta}(s_d)$ for any policy π. Since the tie-breaking strategy is data-independent, there exists a distribution $p \in \Delta(\mathcal{A})$ such that $\pi_\theta(s_0) = p, \forall \theta$.

Now consider $\kappa = \mathrm{argmin}_{i \in [d]}\, p(i)$. We claim that $\eta(T_\kappa, \pi_\theta)$ is suboptimal. Indeed, notice that $V^\star_{T_\kappa}(s_i) = \frac{\gamma}{1-\gamma}\mathbb{I}\left[i = \kappa\right]$ for all $i \in [d]$. As a result, $\eta(T_\kappa, \pi_\theta) \le \frac{\gamma^2}{1-\gamma} p(i) \le \frac{\gamma^2}{2(1-\gamma)}$. On the other hand, we have $\max_\pi \eta(T_\kappa, \pi) = \frac{\gamma^2}{1-\gamma}$. As a result,

$$\sup_{\mathcal{A}} \max_{\kappa} \Big(\max_{\pi} \eta(T_\kappa, \pi) - \eta(T_\kappa, \pi_\theta) \Big) \ge \frac{\gamma^2}{2(1-\gamma)}. \tag{58}$$

□

D.6. Proof of Proposition A.2

In the following we show that, even with $\mathcal{G} = \{V^\pi_{T^\star}\}$ and infinite number of data, the upper bound given by Eq. (17) is loose. Recall that we set the dynamics class $\mathcal{T} = \{T_\theta : \theta \in S^{d-1}\}$. Let $\Pi = \{\pi_x : x \in [d]\}$ where $\pi_x(s_i) = a_x$ for $0 \le i \le d$ and $\pi_x(s_g) = \pi_x(s_b) = a_1$. Let $\mathcal{W}$ be the density ratio induced by π. For any $x \in [d]$, we can compute

$$\begin{aligned} &\rho^{\pi_x}_{T^\star}(s_0, a_i) = (1-\gamma)\mathbb{I}\left[i = x\right], \quad \rho^{\pi_x}_{T^\star}(s_i, a_j) = \gamma(1-\gamma)\mathbb{I}\left[i = x, j = x\right], && (59)\\ &\rho^{\pi_x}_{T^\star}(s_g, a_j) = \rho^{\pi}_{T^\star}(s_b, a_j) = \gamma^2(1-\gamma)\mathbb{I}\left[j = 1\right]. && (60) \end{aligned}$$

Let μ be uniform distribution over $3d + d^2$ state action pairs. Then we can define $\mathcal{W} = \{w_x : x \in [d]\}$ where $w_x(s, a) \triangleq \frac{1}{1-\gamma}\frac{\rho^{\pi_x}_{T^\star}(s,a)}{\mu(s,a)}$.

Now for any fixed $\theta \in S^{d-1}$, consider

$$\max_{w \in \mathcal{W}, g \in \mathcal{G}} |\ell_w(g, T_\theta)|. \tag{61}$$

Let $x = \mathrm{argmin}_i [\theta]_i$.[3] We claim that

$$\ell_{w_x}(V^{\pi_x}_{T^\star}, T_\theta) \ge \frac{\gamma}{4(1-\gamma)}.$$

Indeed, with infinite data we have

$$\begin{aligned} \ell_{w_x}(V^{\pi_x}_{T^\star}, T_\theta) &= \left|\mathbb{E}_{(s,a)\sim\mu}\left[w_x(s,a)\left(\mathbb{E}_{s'\sim T(s,a)}[V^{\pi_x}_{T^\star}(s')] - \mathbb{E}_{s'\sim T^\star(s,a)}[V^{\pi_x}_{T^\star}(s')]\right)\right]\right|\\ &= \frac{1}{1-\gamma}\left|\mathbb{E}_{(s,a)\sim\rho^{\pi_x}_{T^\star}}\left[\left(\mathbb{E}_{s'\sim T(s,a)}[V^{\pi_x}_{T^\star}(s')] - \mathbb{E}_{s'\sim T^\star(s,a)}[V^{\pi_x}_{T^\star}(s')]\right)\right]\right|. \end{aligned}$$

Recall that $T_\theta = T^\star$ for states in the first and third layer. As a result, we continue the equation by

$$\frac{1}{1-\gamma}\left|\mathbb{E}_{(s,a)\sim\rho^{\pi_x}_{T^\star}}\left[\left(\mathbb{E}_{s'\sim T(s,a)}[V^{\pi_x}_{T^\star}(s')] - \mathbb{E}_{s'\sim T^\star(s,a)}[V^{\pi_x}_{T^\star}(s')]\right)\right]\right|$$

[3] We use $[\theta]_i$ to denote the i-th coordinate of θ.

$$
\begin{aligned}
&= \gamma\left|\mathbb{E}_{s'\sim T(s_x,a_x)}[V_{T^\star}^{\pi_x}(s')] - \mathbb{E}_{s'\sim T^\star(s_x,a_x)}[V_{T^\star}^{\pi_x}(s')]\right| && \text{(by the definition of } \rho\text{)}\\
&= \gamma\left|\frac{1}{2}(1+[\theta]_x)V_{T^\star}^{\pi_x}(s_g) + \frac{1}{2}(1-[\theta]_x)V_{T^\star}^{\pi_x}(s_b) - V_{T^\star}^{\pi_x}(s_g)\right| && \text{(by the definition of } T_\theta\text{)}\\
&= \frac{\gamma}{2}(1-[\theta]_x)(V_{T^\star}^{\pi_x}(s_g) - V_{T^\star}^{\pi_x}(s_b)).
\end{aligned}
$$

By basic algebra, $V_{T^\star}^{\pi_x}(s_g) = (1-\gamma)^{-1}$ and $V_{T^\star}^{\pi_x}(s_b) = 0$. As a result, we get

$$
\ell_{w_x}(V_{T^\star}^{\pi_x}, T_\theta) \ge \frac{\gamma}{2(1-\gamma)}(1-[\theta]_x) \ge \frac{\gamma}{4(1-\gamma)}, \tag{62}
$$

where the last inequality comes from the choice of x.

E. Helper Lemmas

In this section, we present several helper lemmas used in Appendix D.

Lemma E.1. *For two distribution p, q over $x \in \mathcal{X}$, if we have $\|p-q\|_1 \le \epsilon$, then for any $\zeta > 1$,*

$$
\mathbb{E}_{x\sim p}\left[\mathbb{I}\left[\frac{p(x)}{q(x)} > \zeta\right]\right] \le \frac{\zeta}{\zeta-1}\epsilon.
$$

Proof. Define $E(x) = \mathbb{I}\left[\frac{p(x)}{q(x)} > \zeta\right]$. Note that under event $E(x)$ we have

$$
p(x) > q(x)\zeta \implies p(x) - q(x) > q(x)(\zeta - 1). \tag{63}
$$

As a result,

$$
\begin{aligned}
\epsilon \ge \|p-q\|_1 &\ge \int |p(x)-q(x)|E(x)\,dx && (64)\\
&\ge \int (\zeta-1)q(x)E(x)\,dx = \mathbb{E}_{x\sim q}[E(x)](\zeta-1) && (65)\\
&\ge (\mathbb{E}_{x\sim p}[E(x)] - \epsilon)(\zeta-1). && (66)
\end{aligned}
$$

By algebraic manipulation we get $\mathbb{E}_{x\sim p}[E(x)] \le \frac{\zeta}{\zeta-1}\epsilon$. □

Lemma E.2. *Consider a fixed policy π and two dynamics model $T, \bar{T}$. Suppose*

$$
\mathbb{E}_{(s,a)\sim\rho_T^\pi}\left[\mathrm{TV}\left(T(s,a), \bar{T}(s,a)\right)\right] \le \epsilon,
$$

we get

$$
\left\|\rho_T^\pi - \rho_{\bar{T}}^\pi\right\|_1 \le \frac{1}{1-\gamma}\epsilon. \tag{67}
$$

Proof. First of all let $G, \bar{G}$ be the transition kernel from $\mathcal{S}\times\mathcal{A}$ to $\mathcal{S}\times\mathcal{A}$ induced by T, π and $\bar{T}, \pi$ respectively. Then for any distribution $\rho \in \Delta(\mathcal{S}\times\mathcal{A})$ we have

$$
\left\|G\rho - \bar{G}\rho\right\|_1 \le \mathbb{E}_{(s,a)\sim\rho}\left[\mathrm{TV}\left(\bar{T}(s,a), T(s,a)\right)\right]. \tag{68}
$$

Let ρ_h (or $\bar{\rho}_h$) be the state-action distribution on step h under dynamics T (or $\bar{T}$). Then we have

$$
\rho_h - \bar{\rho}_h = \left(G^h - \bar{G}^h\right)\rho_0 = \sum_{h'=0}^{h-1} \bar{G}^{h-h'-1}\left(G - \bar{G}\right)G^{h'}\rho_0. \tag{69}
$$

As a result,

$$
\|\rho_h - \bar{\rho}_h\|_1 \le \sum_{h'=0}^{h-1}\left\|\bar{G}^{h-h'-1}\left(G-\bar{G}\right)G^{h'}\rho_0\right\|_1 \tag{70}
$$

$$\leq \sum_{h'=0}^{h-1} \left\| (G - \bar{G}) G^{h'} \rho_0 \right\|_1 \leq \sum_{h'=0}^{h-1} \mathbb{E}_{(s,a)\sim\rho_{h'}} \left[\mathrm{TV}\left(\bar{T}(s,a), T(s,a) \right) \right]. \tag{71}$$

It follows that

$$\left\| \rho^{\pi}_{T} - \rho^{\pi}_{\bar{T}} \right\|_1 \leq (1-\gamma) \sum_{h=0}^{\infty} \gamma^h \left\| \rho_h - \bar{\rho}_h \right\|_1 \tag{72}$$

$$\leq (1-\gamma) \sum_{h=0}^{\infty} \gamma^h \sum_{h'=0}^{h-1} \mathbb{E}_{(s,a)\sim\rho_{h'}} \left[\mathrm{TV}\left(\bar{T}(s,a), T(s,a) \right) \right] \tag{73}$$

$$\leq (1-\gamma) \sum_{h=0}^{\infty} \frac{\gamma^h}{1-\gamma} \mathbb{E}_{(s,a)\sim\rho_{h}} \left[\mathrm{TV}\left(\bar{T}(s,a), T(s,a) \right) \right] \tag{74}$$

$$= \sum_{h=0}^{\infty} \gamma^h \mathbb{E}_{(s,a)\sim\rho_{h}} \left[\mathrm{TV}\left(\bar{T}(s,a), T(s,a) \right) \right] \tag{75}$$

$$= \frac{1}{1-\gamma} \mathbb{E}_{(s,a)\sim\rho^{\pi}_{T}} \left[\mathrm{TV}\left(\bar{T}(s,a), T(s,a) \right) \right]. \tag{76}$$

□

F. Experimental Details

In our experiments as described in the main text, we used the function class $T(x) \in \mathcal{T}$ parameterized by x.

$$T(x) = \begin{cases} s_t < 0.05: \; s_{t+1} = (1 + x/10)s_t - (0.5 + x/10)a_t \\ s_t \geq 0.05: \; s_{t+1} = -(1 + x/10)s_t - (0.5 + x/10)a_t \end{cases}$$

We used $X = \{0, 2, 4, 6, 8\}$. In order to compute the weight function and the value function, we considered policies with values, $K = \{0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2\}$. That is we run the policies $k \in K$ in the transition dynamics generated by the $x \in X$,

$$s_{t+1} = (1 + x/10)s_t - (0.5 + x/10)a_t$$

and the reward function generated by $Q = 1, R = 1$. For our model, we discretized the state action space with 10 bins, and the support of LQR is $s \in \{-0.3, 0.3\}$.