AMORE: A Model-based Framework for Improving Arbitrary Baseline Policies with Offline Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

We propose a new model-based offline RL framework, called Adversarial Models for Offline Reinforcement Learning (AMORE), which can robustly learn policies to improve upon an arbitrary baseline policy regardless of data coverage. Based on the concept of relative pessimism, AMORE is designed to optimize for the worst-case relative performance when facing uncertainty. In theory, we prove that the learned policy of AMORE never degrades the performance of the baseline policy with *any* admissible hyperparameter, and can learn to compete with the best policy within data coverage when the hyperparameter is well tuned and the baseline policy is supported by the data. Such a robust policy improvement property makes AMORE especially suitable for building real-world learning systems, because in practice ensuring no performance degradation is imperative before considering any benefit learning can bring.

1 Introduction

Offline reinforcement learning (RL) is a technique for learning decision making policies from logged data (Jin et al., 2021; Xie et al., 2021a). In comparison with alternate learning techniques, such as off-policy RL and imitation learning, offline RL reduces the data assumption needed to learn good policies and does not require collecting new data. Theoretically, offline RL can learn the best policy that the given data can explain: as long as the offline data includes all scenarios that executing a near-optimal policy would encounter, an offline RL algorithm can learn a near-optimal policy, even when the data is collected by highly sub-optimal policies or is not diverse. Such robustness to data coverage quality makes offline RL a promising technique for solving real-world problems, because collecting diverse or expert-quality data in practice is expensive or simply infeasible.

The fundamental principle behind offline RL is the concept of pessimism in face of uncertainty, which considers worst-case outcomes for scenarios without data. In implementation, this is realized by (explicitly or implicitly) constructing performance lower bounds in policy learning, which penalizes the agent to take uncertain actions. Various designs have been proposed to construct such lower bounds, including behavior regularization (Fujimoto et al., 2019; Kumar et al., 2019; Wu et al., 2019; Laroche et al., 2019; Fujimoto and Gu, 2021), point-wise pessimism based on negative bonuses or truncation (Kidambi et al., 2020; Jin et al., 2021), value penalty (Kumar et al., 2020; Yu et al., 2020), or two-player games (Cheng et al., 2022; Xie et al., 2021a; Uehara and Sun, 2021). Conceptually, the tighter the lower bound is, the better the learned policy would perform, as the performance estimate is more accurate.

Despite these advances, offline RL still has not been widely adopted to build learning-based decision systems in practice. One reason we posit is that *achieving high performance in the worst case is not the full picture of designing real-world learning agents.*

Usually we apply machine learning to applications that are not completely unknown, but have some running policies. These policies are the decision rules that are currently used in the system (e.g., an engineered autonomous driving rule, or a heuristic-based system for diagnosis), and the goal of applying a learning algorithm is often to further improve upon these *baseline policies*. As a result, it is imperative that the policy learned by the agent does not lead to *performance degradation*. This criterion is especially critical for applications where the poor decision outcomes cannot be tolerated (such as health care, autonomous driving, and commercial resource allocation).

Although optimizing for absolute or relative performance is the same when full information is available, they can lead to different policies when we only have partial data coverage. In this case, the policy that has the best worst-case performance (which most existing offline RL aims to recover) would not necessarily perform better than the baseline policies when deployed in the real environment. Such performance degradation happens when the data do not cover *all* behaviors of the baseline policies, which can be due to finite samples or a coverage mismatch between the baselines and the data collection policies. As a result, running policies learned by existing offline RL algorithms could risk degrading performance.

In this work, we propose a new model-based offline RL framework, called **A**dversarial **M**odels for **O**ffline **Re**inforcement Learning (AMORE), which can robustly learn policies improving upon an arbitrary baseline policy. AMORE is designed based on the concept of relative pessimism (Cheng et al., 2022), which aims to optimize for the worst-case relative performance when facing uncertainty. In theory, we prove that the the learned policy from AMORE never degrades the performance of the baseline policy of comparison for a wide range of hyperparameters which are given beforehand, a property known as Robust Policy Improvement (RPI) (Cheng et al., 2022). In addition, we prove that, when the right hyperparameter is chosen and the baseline police is covered by the data, the learned policy of AMORE can also compete with any policy within data coverage in an absolute sense.

To our knowledge, RPI property of offline RL has so far be limited to comparing against the data collection policy (Cheng et al., 2022; Fujimoto et al., 2019; Kumar et al., 2019; Wu et al., 2019; Laroche et al., 2019; Fujimoto and Gu, 2021). However, it is quite common that the baseline policy of interest is different from the data collection policy. For example, in robotics manipulation, we often have a dataset of activities different from the target task that we wish to solve. In AMORE, by using models, we extend the technique of relative pessimism to achieve RPI with *arbitrary* baseline policies, regardless whether the baseline policies collected the data or not.

2 Preliminaries

Markov Decision Process We consider an agent acting in an infinite-horizon discounted Markov Decision Process (MDP) M defined by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma \rangle$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition dynamics, $R : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a scalar reward function and $\gamma \in [0, 1)$ is the discount factor. The learner selects actions using a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. We denote by Π the space of all Markovian policies. Let, $d_M^\pi(s, a)$ denote the discounted state-action distribution obtained by running policy π on M, i.e $d_M^\pi(s,a) = (1-\gamma)\,\mathbb{E}\left[\sum_{t=0}^\infty \gamma^t \mathbb{1}\left(s_t = s, a_t = a | a_t \sim \pi\left(\cdot|s_t\right)\right)\right]$. Let $J_M(\pi) = \mathbb{E}_{\pi,M}\left[\sum_{t=0}^\infty \gamma^t r_t | a_t \sim \pi\right]$ be the expected discounted return of policy π on M. The goal of reinforcement learning is to find the policy that maximizes J. We define the value function as $V_M^\pi(s) = \mathbb{E}_{\pi,M}\left[\sum_{t=0}^\infty \gamma^t r_t | s_0 = s\right]$, and the related state-action value function (i.e., Q-function) as $Q_M^\pi(s,a) = \mathbb{E}_{\pi,M}\left[\sum_{t=0}^\infty \gamma^t r_t | s_0 = s, s_0 = a\right]$. We use $[0, V_{\max}]$ as the range of value functions.

Offline RL The aim of offline RL is to output strong policies from a fixed dataset collected using a behavior policy without further environmental interactions. We assume the dataset $\mathcal{D}$ consists of $\{(s_i, a_i, r_i, s_{i+1})\}_{i=1}^N$, where (s_i, a_i) is sampled i.i.d. from some distribution μ. We also abuse μ

as discounted state-action occupancy of behavior policy, i.e., $\mu \equiv d^{\mu}_{M}$, and we use $a \sim \mu(\cdot|s)$ to denote sampling from that behavior policy.

This paper is concerned with the model-based offline RL problem, and we use $\mathcal{M}$ to denote the model class. For each $M \in \mathcal{M}$, we use $P_M : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ and $R_M : \mathcal{S} \times \mathcal{A} \to [0,1]$ to denote the corresponding transition and reward function of M.

Assumption 1 (Realizability)**.** *We assume the ground truth model $M^\star$ is in the model class $\mathcal{M}$.*

3 Adversarial Models for Offline Reinforcement Learning (AMORE)

In this section, we introduce our proposed approach, Adversarially Trained Models (AMORE), in Algorithm 1, and the main theoretical results.

Algorithm 1 Adversarially Trained Models (AMORE)

Input: Batch data $\mathcal{D}$. Model class $\mathcal{M}$. Coefficient α. Policy class Π. Reference policy π_{ref}.

1: Construct version space for the model,

$$\mathcal{M}_\alpha = \left\{ M \in \mathcal{M} : \max_{M' \in \mathcal{M}} \mathcal{L}_{\mathcal{D}}(M') - \mathcal{L}_{\mathcal{D}}(M) \le \alpha \right\}, \tag{1}$$

$$\text{where} \quad \mathcal{L}_{\mathcal{D}}(M) \coloneqq \sum_{(s,a,r,s') \in \mathcal{D}} \left[\log \mathbb{P}_M(s'|s,a) - (R_M(s,a) - r)^2 \right], \ \forall M \in \mathcal{M}. \tag{2}$$

2: Conduct learning via relative pessimism,

$$\widehat{\pi} = \operatorname*{argmax}_{\pi \in \Pi} \min_{M \in \mathcal{M}_\alpha} J_M(\pi) - J_M(\pi_{\mathsf{ref}}). \tag{3}$$

AMORE can be viewed as a model-based extension of the ATAC algorithm by Cheng et al. (2022). In the next sections, we illustrate that AMORE is not only able to compete with the best data-covered policy as prior works (e.g., Xie et al., 2021a; Uehara and Sun, 2021; Cheng et al., 2022), but also enjoys a stronger robust policy improvement guarantee than (Cheng et al., 2022).

3.1 Theoretical Analysis

This section analyzes AMORE theoretically and presents guarantees on its absolute performance and the policy improvement over the reference policy π_{ref}. Before presenting the detailed guarantees, we introduce generalized single-policy concentrability, which measures the distribution shift over some arbitrary policy π and data distribution μ.

Definition 1 (Generalized Single-policy Concentrability)**.** *We define the generalized single-policy concentrability for policy π for model class $\mathcal{M}$ and offline data distribution μ as*

$$\mathfrak{C}_{\mathcal{M}}(\pi) \coloneqq \sup_{M \in \mathcal{M}} \frac{\mathbb{E}_{d^\pi}\left[D_{\mathrm{TV}}\left(P_M(\cdot|s,a), P^\star(\cdot|s,a)\right)^2 + \left(R_M(s,a) - R^\star(s,a)\right)^2 \right]}{d_\mu \left[D_{\mathrm{TV}}\left(P_M(\cdot|s,a), P^\star(\cdot|s,a)\right)^2 + \left(R_M(s,a) - R^\star(s,a)\right)^2 \right]}.$$

Note that $\mathfrak{C}_{\mathcal{M}}(\pi)$ is always upper bounded by the standard single-policy concentrability coefficient $\|d^\pi / \mu\|_\infty$ (e.g., Jin et al., 2021; Rashidinejad et al., 2021; Xie et al., 2021b), but it can be smaller in general with model class $\mathcal{M}$. It can also be viewed as a model-based analog of the one in Xie et al. (2021a), and the detailed discussion around $\mathfrak{C}_{\mathcal{M}}(\pi)$ refers to Uehara and Sun (2021).

We are now ready to present the absolute performance guarantee of AMORE.

Theorem 1 (Absolute performance guarantee)**.** *Under Assumption 1, there exists an absolute constant c such that for any $\delta \in (0,1]$, if we choose $\alpha = c \cdot (\log(|\mathcal{M}|/\delta))$ in Algorithm 1, then for arbitrary reference policy π_{ref} and comparator policy $\pi^\dagger \in \Pi$, with probability $1-\delta$, the policy $\widehat{\pi}$*

learned by Algorithm 1 satisfies

$$J(\pi^\dagger) - J(\widehat{\pi}) \leq \mathcal{O}\left(\left[\sqrt{\mathfrak{C}_{\mathcal{M}}(\pi^\dagger)} + \sqrt{\mathfrak{C}_{\mathcal{M}}(\pi_{\mathsf{ref}})}\right] \cdot \frac{V_{\max}}{1-\gamma}\sqrt{\frac{\log(|\mathcal{M}|/,\delta)}{n}}\right).$$

Roughly speaking, Theorem 1 shows that $\widehat{\pi}$ learned by Algorithm 1 could compete with any policy π with a large enough dataset, as long as the offline data μ has good coverage on comparator policy $\pi^\dagger$ (since the reference policy π_{ref} is the input of Theorem 1, one can set $\pi_{\mathsf{ref}} = \mu$ (data collection policy) as $\mathfrak{C}_{\mathcal{M}}(\mu) \leq \mathfrak{C}_{\mathcal{M}}(\pi^\dagger)$). Compared to the closest model-based offline RL work (Uehara and Sun, 2021), if we set $\pi_{\mathsf{ref}} = \mu$ (data collection policy), Theorem 1 leads to the almost same guarantee as Uehara and Sun (2021, Theorem 1) (up to constant factors).

In addition to the guarantee on the absolute performance above, below we show that, if Assumption 1 is satisfied and $\pi_{\mathsf{ref}} \in \Pi$, AMORE is always guaranteed to improve over $J(\widehat{\pi})$ for *a wide range choice of pessimistic parameter* α.

Theorem 2 (Robust strong policy improvement)**.** *Under Assumption 1, there exists an absolute constant c such that for any $\delta \in (0,1]$, if: i) $\alpha \geq c \cdot (\log(|\mathcal{M}|/\delta))$ in Algorithm 1; ii) $\pi_{\mathsf{ref}} \in \Pi$, then with probability $1-\delta$, the policy $\widehat{\pi}$ learned by Algorithm 1 satisfies $J(\pi_{\mathsf{ref}}) \geq J(\widehat{\pi})$.*

3.2 Discussion

Improving over some reference policy has been long studied in the literature. To highlight the advantage of AMORE, we formally give the definition of different policy improvement properties.

Definition 2 (Robust policy improvement)**.** *Suppose $\widehat{\pi}$ is the learned policy from an algorithm. We say the algorithm has the policy improvement (PI) guarantee if $J(\pi_{\mathsf{ref}}) - J(\widehat{\pi}) \leq o(n)/n$ is guaranteed for some reference policy π_{ref} with offline data $\mathcal{D} \sim \mu$, where $n = |\mathcal{D}|$. We use the following two criteria w.r.t. π_{ref} and μ to define different kinds PI:*

(i) The PI is strong if π_{ref} can be selected arbitrarily from policy class Π regardless of the choice data-collection policy μ; otherwise, PI is weak (e.g., $\pi_{\mathsf{ref}} \equiv \mu$ is required).

(ii) The PI is robust if it can be achieved by a range of hyperparameters with a known subset.

Weak policy improvement is also known as *safe policy improvement* in the literature (Fujimoto et al., 2019; Laroche et al., 2019). It requires the reference policy to be also the behavior policy that collects the offline data. In comparison, strong policy improvement imposes a stricter requirement on the algorithm, which requires policy improvement *regardless* of how the data were collected. This condition is motivated by the common situation where the reference policy is not the data collection policy. For example, in a multi-task problem with shared dynamics, the data are collected by policies for different tasks, and the reference policy we wish to improve on is task specific. In this case, weak policy improvement is meaningless because the behavior policy, which is the average of policies from all tasks, does not have meaningful performance in the target task.

Since we are learning policies offline, without online interactions, it is not straightforward to tune the hyperparameter directly. Therefore, it is desirable that we can design algorithms with these properties in a robust manner in terms of hyperparameter selection. Formally, Definition 2 requires the policy improvement to be achievable by a set of hyperparameters that is known before learning.

Theorem 2 indicates the robust strong policy improvement of AMORE. On the other hand, algorithms with robust weak policy improvement are available in the literature (Cheng et al., 2022; Fujimoto et al., 2019; Kumar et al., 2019; Wu et al., 2019; Laroche et al., 2019; Fujimoto and Gu, 2021); this is usually achieved by designing the algorithm to behave like imitation learning (IL) for a known set of hyperparameter (e.g., behavior regularization algorithms have a weight that can turn off the RL behavior and regress to IL). However, the absolute performance guarantee of achieving the best data-covered policy of the IL-like algorithm is challenging due to its imitating nature. To our best knowledge, ATAC (Cheng et al., 2022) is the only algorithm that achieves robust (weak) policy improvement as well as guarantees absolute performance.

References

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained actor critic for offline reinforcement learning. *arXiv preprint arXiv:2202.02446*, 2022.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pages 2052–2062, 2019.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. In *NeurIPS*, 2020.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 32:11784–11794, 2019.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

Romain Laroche, Paul Trichelair, and Remi Tachet Des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, pages 3652–3661. PMLR, 2019.

Qinghua Liu, Alan Chung, Csaba Szepesvári, and Chi Jin. When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, volume 178, pages 5175–5220. PMLR, 2022.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2021.

Sara A van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021a.

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021b.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.

Tong Zhang. From ε-entropy to kl-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006.

Appendix

A Proofs from Section 3

A.1 Technical Tools

Lemma 3 (Simulation lemma)**.** *Consider any two MDP model M and M', and any $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, we have*

$$|J_M(\pi) - J_{M'}(\pi)| \le \frac{V_{\max}}{1-\gamma}\mathbb{E}_{d^\pi}\left[D_{\mathrm{TV}}\left(P_M(\cdot|s,a), P_{M'}(\cdot|s,a)\right)\right] + \frac{1}{1-\gamma}\mathbb{E}_{d^\pi}\left[|R_M(s,a) - R_{M'}(s,a)|\right].$$

Lemma 3 is the standard simulation lemma in model-based reinforcement learning literature, and its proof can be found in, e.g., Uehara and Sun (2021, Lemma 7).

A.2 Guarantees about Version Space

Lemma 4. *Let $M^\star$ be the ground truth model. Then, with probability at least $1-\delta$, we have*

$$\max_{M \in \mathcal{M}} \mathcal{L}_{\mathcal{D}}(M) - \mathcal{L}_{\mathcal{D}}(M^\star) \le \mathcal{O}\left(\log(|\mathcal{M}|/\delta)\right),$$

where $\mathcal{L}_{\mathcal{D}}$ is defined in Eq. (2).

Proof of Lemma 4. By Lemma 6, we know

$$\max_{M \in \mathcal{M}} \log \ell_{\mathcal{D}}(M) - \log \ell_{\mathcal{D}}(M^\star) \le \log(|\mathcal{M}|/\delta). \tag{4}$$

In addition, by Xie et al. (2021a, Theorem A.1) (with setting $\gamma = 0$), we know w.p. $1-\delta$,

$$\sum_{(s,a,r,s') \in \mathcal{D}} \left(R^\star(s,a) - r\right)^2 - \min_{M \in \mathcal{M}} \sum_{(s,a,r,s') \in \mathcal{D}} \left(R_M(s,a) - r\right)^2 \lesssim \log(|\mathcal{M}|/\delta). \tag{1}$$

Combining the Eqs. (1) and (4), we have w.p. $1-\delta$,

$$\begin{aligned}
&\max_{M \in \mathcal{M}} \mathcal{L}_{\mathcal{D}}(M) - \mathcal{L}_{\mathcal{D}}(M^\star) \\
\le\ &\max_{M \in \mathcal{M}} \log \ell_{\mathcal{D}}(M) - \min_{M \in \mathcal{M}} \sum_{(s,a,r,s') \in \mathcal{D}} \left(R_M(s,a) - r\right)^2 - \mathcal{L}_{\mathcal{D}}(M^\star) \\
\lesssim\ &\log(|\mathcal{M}|/\delta).
\end{aligned}$$

This completes the proof. $\square$

Lemma 5. *For any $M \in \mathcal{M}$, we have with probability at least $1-\delta$,*

$$\begin{aligned}
&\mathbb{E}_\mu\left[D_{\mathrm{TV}}\left(P_M(\cdot|s,a), P^\star(\cdot|s,a)\right)^2 + \left(R_M(s,a) - R^\star(s,a)\right)^2\right] \\
&\quad \le \mathcal{O}\left(\frac{\max_{M' \in \mathcal{M}} \mathcal{L}_{\mathcal{D}}(M') - \mathcal{L}_{\mathcal{D}}(M) + \log(|\mathcal{M}|/\delta)}{n}\right),
\end{aligned}$$

where $\mathcal{L}_{\mathcal{D}}$ is defined in Eq. (2).

Proof of Lemma 5. By Lemma 7, we have w.p. $1-\delta$,

$$n \cdot \mathbb{E}_\mu\left[D_{\mathrm{TV}}\left(P_M(\cdot|s,a), P^\star(\cdot|s,a)\right)^2\right] \lesssim \log \ell_{\mathcal{D}}(M^\star) - \log \ell_{\mathcal{D}}(M) + \log(|\mathcal{M}|/\delta). \tag{5}$$

Also, we have

$$\begin{aligned}
&n \cdot \mathbb{E}_\mu\left[\left(R_M(s,a) - R^\star(s,a)\right)^2\right] \qquad (6)\\
=\ &n \cdot \mathbb{E}_\mu\left[\left(R_M(s,a) - r\right)^2\right] - n \cdot \mathbb{E}_\mu\left[\left(R^\star(s,a) - r\right)^2\right] \\
&\text{(see, e.g., Xie et al., 2021a, Eq. (A.10) with } \gamma = 0\text{)}
\end{aligned}$$

$$\lesssim \sum_{(s,a,r,s')\in\mathcal{D}} (R_M(s,a)-r)^2 - \sum_{(s,a,r,s')\in\mathcal{D}} (R^\star(s,a)-r)^2 + \log(|\mathcal{M}|/\delta),$$

where the last inequality is a direct implication of Xie et al. (2021a, Lemma A.4) and $1 = 1$. Combining Eqs. (5) and (6), we obtain

$$\begin{aligned}
& n \cdot \mathbb{E}_\mu \left[D_{\mathrm{TV}} \left(P_M(\cdot|s,a), P^\star(\cdot|s,a)\right)^2 + \left(R_M(s,a) - R^\star(s,a)\right)^2 \right] \\
\lesssim{} & \log \ell_{\mathcal{D}}(M^\star) - \sum_{(s,a,r,s')\in\mathcal{D}} (R^\star(s,a)-r)^2 - \log \ell_{\mathcal{D}}(M) + \sum_{(s,a,r,s')\in\mathcal{D}} (R_M(s,a)-r)^2 + \log(|\mathcal{M}|/\delta) \\
={} & \mathcal{L}_{\mathcal{D}}(M^\star) - \mathcal{L}_{\mathcal{D}}(M) + \log(|\mathcal{M}|/\delta) \\
\leq{} & \max_{M'\in\mathcal{M}} \mathcal{L}_{\mathcal{D}}(M') - \mathcal{L}_{\mathcal{D}}(M) + \log(|\mathcal{M}|/\delta).
\end{aligned}$$

This completes the proof.

$\square$

A.3 MLE Guarantees

We use $\ell_{\mathcal{D}}(M)$ to denote the likelihood of model $M = (P, R)$ with offline data $\mathcal{D}$, where

$$\ell_{\mathcal{D}}(M) = \prod_{(s,a,r,s')\in\mathcal{D}} P_M(s'|s,a). \tag{7}$$

For the analysis around maximum likelihood estimation, we largely follow the proving idea of Agarwal et al. (2020); Liu et al. (2022), which is inspired by Zhang (2006).

The next lemma shows that the ground truth model $M^\star$ has a comparable log-likelihood compared with MLE solution.

Lemma 6. *Let $M^\star$ be the ground truth model. Then, with probability at least $1-\delta$, we have*

$$\max_{M\in\mathcal{M}} \log \ell_{\mathcal{D}}(M) - \log \ell_{\mathcal{D}}(M^\star) \leq \log(|\mathcal{M}|/\delta). \tag{8}$$

Proof of Lemma 6. The proof of this lemma is obtained by a standard argument of MLE (see, e.g., van de Geer, 2000). For any $M \in \mathcal{M}$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\log \ell_{\mathcal{D}}(M) - \log \ell_{\mathcal{D}}(M^\star)\right)\right] &= \mathbb{E}\left[\frac{\ell_{\mathcal{D}}(M)}{\ell_{\mathcal{D}}(M^\star)}\right] \\
&= \mathbb{E}\left[\frac{\prod_{(s,a,r,s')\in\mathcal{D}} \mathbb{P}_M(s'|s,a)}{\prod_{(s,a,r,s')\in\mathcal{D}} \mathbb{P}_{M^\star}(s'|s,a)}\right] \\
&= \mathbb{E}\left[\prod_{(s,a,r,s')\in\mathcal{D}} \frac{\mathbb{P}_M(s'|s,a)}{\mathbb{P}_{M^\star}(s'|s,a)}\right] \\
&= \mathbb{E}\left[\prod_{(s,a)\in\mathcal{D}} \mathbb{E}\left[\frac{\mathbb{P}_M(s'|s,a)}{\mathbb{P}_{M^\star}(s'|s,a)} \,\middle|\, s,a\right]\right] \\
&= \mathbb{E}\left[\prod_{(s,a)\in\mathcal{D}} \sum_{s',r} \mathbb{P}_M(s'|s,a)\right] \\
&= 1.
\end{aligned} \tag{9}$$

Then by Markov's inequality, we obtain

$$\begin{aligned}
& \mathbb{P}\left[\left(\log \ell_{\mathcal{D}}(M) - \log \ell_{\mathcal{D}}(M^\star)\right) > \log(1/\delta)\right] \\
\leq{} & \underbrace{\mathbb{E}\left[\exp\left(\log \ell_{\mathcal{D}}(M) - \log \ell_{\mathcal{D}}(M^\star)\right)\right]}_{=1 \text{ by Eq. (9)}} \cdot \exp\left[-\log(1/\delta)\right] = \delta.
\end{aligned}$$

Therefore, taking a union bound over $\mathcal{M}$, we obtain

$$\mathbb{P}\left[(\log \ell_{\mathcal{D}}(M) - \log \ell_{\mathcal{D}}(M^\star)) > \log(|\mathcal{M}|/\delta)\right] \le \delta.$$

This completes the proof. $\square$

The following lemma shows that, the on-support error of any model $M \in \mathcal{M}$ can be captured via its log-likelihood (by comparing with the MLE solution).

Lemma 7. *For any $M = (P, R)$, we have with probability at least $1 - \delta$,*

$$\mathbb{E}_\mu\left[D_{\mathrm{TV}}\left(P(\cdot|s,a), P^\star(\cdot|s,a)\right)^2\right] \le \mathcal{O}\left(\frac{\log \ell_{\mathcal{D}}(M^\star) - \log \ell_{\mathcal{D}}(M) + \log(|\mathcal{M}|/\delta)}{n}\right),$$

where $\ell_{\mathcal{D}}(\cdot)$ is defined in Eq. (7).

Proof of Lemma 7. By Agarwal et al. (2020, Lemma 25), we have

$$\mathbb{E}_\mu\left[D_{\mathrm{TV}}\left(P(\cdot|s,a), P^\star(\cdot|s,a)\right)^2\right] \le -2\log \mathbb{E}_{\mu\times P^\star}\left[\exp\left(-\frac{1}{2}\log\left(\frac{P^\star(s'|s,a)}{P(s'|s,a)}\right)\right)\right] \tag{10}$$

$$\mathbb{E}_\mu\left[D_{\mathrm{TV}}\left(R(\cdot|s,a), R^\star(\cdot|s,a)\right)^2\right] \le -2\log \mathbb{E}_{\mu\times R^\star}\left[\exp\left(-\frac{1}{2}\log\left(\frac{R^\star(r|s,a)}{R(r|s,a)}\right)\right)\right],$$

where $\mu \times P^\star$ and $\mu \times R^\star$ denote the ground truth offline joint distribution of (s, a, s') and (s, a, r).

Let $\widetilde{\mathcal{D}} = \{(\widetilde{s}_i, \widetilde{a}_i, \widetilde{r}_i, \widetilde{s}'_i)\}_{i=1}^n \sim \mu$ be another offline dataset that is independent to $\mathcal{D}$. Then,

$$\begin{aligned}
& -n\cdot \log \mathbb{E}_{\mu\times P^\star}\left[\exp\left(-\frac{1}{2}\log\left(\frac{P^\star(s'|s,a)}{P(s'|s,a)}\right)\right)\right] \\
&= -\sum_{i=1}^n \log \mathbb{E}_{(\widetilde{s}_i,\widetilde{a}_i,\widetilde{s}'_i)\sim\mu}\left[\exp\left(-\frac{1}{2}\log\left(\frac{P^\star(\widetilde{s}'_i|\widetilde{s}_i,\widetilde{a}_i)}{P(\widetilde{s}'_i|\widetilde{s}_i,\widetilde{a}_i)}\right)\right)\right] \\
&= -\log \mathbb{E}_{\widetilde{\mathcal{D}}\sim\mu}\left[\exp\left(\sum_{i=1}^n -\frac{1}{2}\log\left(\frac{P^\star(\widetilde{s}'_i|\widetilde{s}_i,\widetilde{a}_i)}{P(\widetilde{s}'_i|\widetilde{s}_i,\widetilde{a}_i)}\right)\right)\,\middle|\, \mathcal{D}\right] \\
&= -\log \mathbb{E}_{\widetilde{\mathcal{D}}\sim\mu}\left[\exp\left(\sum_{(s,a,s')\in\widetilde{\mathcal{D}}} -\frac{1}{2}\log\left(\frac{P^\star(s'|s,a)}{P(s'|s,a)}\right)\right)\,\middle|\, \mathcal{D}\right].
\end{aligned} \tag{11}$$

We use $\ell_P(s,a,s')$ as the shorthand of $-\frac{1}{2}\log\left(\frac{P^\star(s|s,a)}{P(s'|s,a)}\right)$, for any $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$. By Agarwal et al. (2020, Lemma 24) (see also Liu et al., 2022, Lemma 15), we know

$$\mathbb{E}_{\mathcal{D}\sim\mu}\left[\exp\left(\sum_{(s,a,s')\in\mathcal{D}} \ell_P(s,a,s') - \log \mathbb{E}_{\widetilde{\mathcal{D}}\sim\mu}\left[\exp\left(\sum_{(s,a,s')\in\widetilde{\mathcal{D}}} \ell_P(s,a,s')\right)\,\middle|\,\mathcal{D}\right] - \log|\mathcal{M}|\right)\right] \le 1.$$

Thus, we can use Chernoff method as well as a union bound on the equation above to obtain the following exponential tail bound: with probability at least $1-\delta$, we have for all $(P, R) = M \in \mathcal{M}$,

$$-\log \mathbb{E}_{\widetilde{\mathcal{D}}\sim\mu}\left[\exp\left(\sum_{(s,a,s')\in\widetilde{\mathcal{D}}} \ell_P(s,a,s')\right)\,\middle|\,\mathcal{D}\right] \le -\sum_{(s,a,s')\in\mathcal{D}} \ell_P(s,a,s') + 2\log(|\mathcal{M}|/\delta). \tag{12}$$

Plugging back the definition of ℓ_P and combining Eqs. (10) to (12), we obtain

$$n\cdot\mathbb{E}_\mu\left[D_{\mathrm{TV}}\left(P(\cdot|s,a), P^\star(\cdot|s,a)\right)^2\right] \le \frac{1}{2}\sum_{(s,a,s')\in\mathcal{D}} \log\left(\frac{P^\star(s|s,a)}{P(s'|s,a)}\right) + 2\log(|\mathcal{M}|/\delta). \tag{13}$$

By the same steps of obtaining to Eq. (13), we also have

$$n\cdot\mathbb{E}_\mu\left[D_{\mathrm{TV}}\left(R(\cdot|s,a), R^\star(\cdot|s,a)\right)^2\right] \le \frac{1}{2}\sum_{(s,a,r')\in\mathcal{D}} \log\left(\frac{R^\star(s|s,a)}{R(s'|s,a)}\right) + 2\log(|\mathcal{M}|/\delta). \tag{14}$$

Combining Eqs. (13) and (14), we obtain

$$
\begin{aligned}
& n \cdot \mathbb{E}_\mu \left[D_{\mathrm{TV}} \left(P(\cdot|s,a), P^\star(\cdot|s,a) \right)^2 + D_{\mathrm{TV}} \left(R(\cdot|s,a), R^\star(\cdot|s,a) \right)^2 \right] \\
\lesssim & \sum_{(s,a,s') \in \mathcal{D}} \log \left(\frac{P^\star(s|s,a)}{P(s'|s,a)} \right) + \sum_{(s,a,r') \in \mathcal{D}} \log \left(\frac{R^\star(s|s,a)}{R(s'|s,a)} \right) + \log(|\mathcal{M}|/\delta) \\
= & \log \ell_{\mathcal{D}}(M^\star) - \log \ell_{\mathcal{D}}(M) + \log(|\mathcal{M}|/\delta). \qquad (\ell_{\mathcal{D}}(\cdot) \text{ is defined in Eq. (7)})
\end{aligned}
$$

This completes the proof. □

A.4 Proof of Main Theorems

Proof of Theorem 1. By the optimality of $\widehat{\pi}$ (from Eq. (3)), we have

$$
\begin{aligned}
J(\pi^\dagger) - J(\widehat{\pi}) &= J(\pi^\dagger) - J(\pi_{\mathsf{ref}}) - [J(\widehat{\pi}) - J(\pi_{\mathsf{ref}})] \\
&\leq J(\pi^\dagger) - J(\pi_{\mathsf{ref}}) - \min_{M \in \mathcal{M}_\alpha} [J_M(\widehat{\pi}) - J_M(\pi_{\mathsf{ref}})] \\
&\qquad\qquad (\text{by Lemma 6, we have } M^\star \in \mathcal{M}_\alpha) \\
&\leq J(\pi^\dagger) - J(\pi_{\mathsf{ref}}) - \min_{M \in \mathcal{M}_\alpha} \left[J_M(\pi^\dagger) - J_M(\pi_{\mathsf{ref}}) \right], \qquad (15)
\end{aligned}
$$

where the last step is because of $\pi^\dagger \in \Pi$ By the simulation lemma (Lemma 3), we know for any policy π and any $M \in \mathcal{M}_\alpha$,

$$
\begin{aligned}
|J(\pi) - J_M(\pi)| &\leq \frac{V_{\max}}{1-\gamma} \mathbb{E}_{d^\pi} \left[D_{\mathrm{TV}} \left(P_M(\cdot|s,a), P^\star(\cdot|s,a) \right) \right] + \frac{1}{1-\gamma} \mathbb{E}_{d^\pi} \left[|R_M(s,a) - R^\star(s,a)| \right] \\
&\leq \frac{V_{\max}}{1-\gamma} \sqrt{\mathbb{E}_{d^\pi} \left[D_{\mathrm{TV}} \left(P_M(\cdot|s,a), P^\star(\cdot|s,a) \right)^2 \right]} + \frac{1}{1-\gamma} \sqrt{\mathbb{E}_{d^\pi} \left[\left(R_M(s,a) - R^\star(s,a) \right)^2 \right]} \\
&\lesssim \frac{V_{\max}}{1-\gamma} \sqrt{\mathbb{E}_{d^\pi} \left[D_{\mathrm{TV}} \left(P_M(\cdot|s,a), P^\star(\cdot|s,a) \right)^2 + \left(R_M(s,a) - R^\star(s,a) \right)^2 \right]} \\
&\qquad\qquad (a \lesssim b \text{ means } a \leq \mathcal{O}(b)) \\
&\leq \frac{V_{\max} \sqrt{\mathfrak{C}_{\mathcal{M}}(\pi)}}{1-\gamma} \sqrt{\mathbb{E}_\mu \left[D_{\mathrm{TV}} \left(P_M(\cdot|s,a), P^\star(\cdot|s,a) \right)^2 + \left(R_M(s,a) - R^\star(s,a) \right)^2 \right]} \\
&\lesssim \frac{V_{\max} \sqrt{\mathfrak{C}_{\mathcal{M}}(\pi)}}{1-\gamma} \sqrt{\frac{\max_{M' \in \mathcal{M}} \mathcal{L}_{\mathcal{D}}(M') - \mathcal{L}_{\mathcal{D}}(M) + \log(|\mathcal{M}|/\delta)}{n}} \\
&\qquad\qquad (\text{by Lemma 5}) \\
&\lesssim \frac{V_{\max} \sqrt{\mathfrak{C}_{\mathcal{M}}(\pi)}}{1-\gamma} \sqrt{\frac{\log(|\mathcal{M}|/,\delta)}{n}} \qquad (16)
\end{aligned}
$$

where the last step is because $\max_{M' \in \mathcal{M}} \mathcal{L}_{\mathcal{D}}(M') - \mathcal{L}_{\mathcal{D}}(M) \leq \alpha = \mathcal{O}(\log(|\mathcal{M}|/\delta)/n$ by Eq. (1).

Combining Eqs. (15) and (16), we obtain

$$
J(\pi^\dagger) - J(\widehat{\pi}) \lesssim \left[\sqrt{\mathfrak{C}_{\mathcal{M}}(\pi^\dagger)} + \sqrt{\mathfrak{C}_{\mathcal{M}}(\pi_{\mathsf{ref}})} \right] \cdot \frac{V_{\max}}{1-\gamma} \sqrt{\frac{\log(|\mathcal{M}|/,\delta)}{n}}.
$$

This completes the proof. □

Proof of Theorem 2.

$$
\begin{aligned}
J(\pi_{\mathsf{ref}}) - J(\widehat{\pi}) &= J(\pi_{\mathsf{ref}}) - J(\pi_{\mathsf{ref}}) - [J(\widehat{\pi}) - J(\pi_{\mathsf{ref}})] \\
&\leq -\min_{M \in \mathcal{M}_\alpha} [J_M(\widehat{\pi}) - J_M(\pi_{\mathsf{ref}})] \qquad (\text{by Lemma 6, we have } M^\star \in \mathcal{M}_\alpha) \\
&= -\max_{\pi \in \Pi} \min_{M \in \mathcal{M}_\alpha} [J_M(\pi) - J_M(\pi_{\mathsf{ref}})] \qquad (\text{by the optimality of } \widehat{\pi} \text{ from Eq. (3)}) \\
&\leq -\min_{M \in \mathcal{M}_\alpha} [J_M(\pi_{\mathsf{ref}}) - J_M(\pi_{\mathsf{ref}})] \qquad (\pi_{\mathsf{ref}} \in \Pi) \\
&= 0.
\end{aligned}
$$

□