On the Sample Complexity of Vanilla Model-Based Offline Reinforcement Learning with Dependent Samples

Mustafa O. Karabag, Ufuk Topcu

The University of Texas at Austin, Austin, TX, USA
{karabag, utopcu}@utexas.edu

Abstract

Offline reinforcement learning (offline RL) considers problems where learning is performed using only previously collected samples and is helpful for the settings in which collecting new data is costly or risky. In model-based offline RL, the learner performs estimation (or optimization) using a model constructed according to the empirical transition frequencies. We analyze the sample complexity of vanilla model-based offline RL with dependent samples in the infinite-horizon discounted-reward setting. In our setting, the samples obey the dynamics of the Markov decision process and, consequently, may have interdependencies. Under no assumption of independent samples, we provide a high-probability, polynomial sample complexity bound for vanilla model-based off-policy evaluation that requires partial or uniform coverage. We extend this result to the off-policy optimization under uniform coverage. As a comparison to the model-based approach, we analyze the sample complexity of off-policy evaluation with vanilla importance sampling in the infinite-horizon setting. Finally, we provide an estimator that outperforms the sample-mean estimator for almost deterministic dynamics that are prevalent in reinforcement learning.

Introduction

Offline reinforcement learning (RL) considers problems where a learner has access to only a dataset that is collected under a *behavior policy* in an environment and tries to evaluate (or optimize) a *target policy*. The learner typically has no control over the behavior policy, and the transition dynamics of the environment are unknown to the learner. Offline RL is helpful for settings where online learning may not be safe or previously collected data are abundant. The applications of offline RL include, but are not limited to healthcare (Shortreed et al. 2011; Tseng et al. 2017), robotics (Levine et al. 2018; Ebert et al. 2018; Zeng et al. 2018), natural language processing (Zhou et al. 2017; Henderson, Lemon, and Georgila 2008), and recommendation systems (Swaminathan et al. 2017; Gilotte et al. 2018).

We develop theoretical guarantees for offline RL. In detail, we use an infinite horizon Markov decision process (MDP) to model the environment and analyze the number of sample paths sufficient to achieve the desired accuracy for off-policy evaluation. We mainly focus on vanilla model-based off-policy evaluation, where a target policy is evaluated through a model based on the sample mean estimator of the transition dynamics.

Analyzing the theoretical properties of model-based off-policy evaluation is challenging due to the sequential nature of the MDP model and potentially dependent samples. These factors make model-based off-policy evaluation lack the unbiasedness property that importance-sampling-based off-policy methods have (Levine et al. 2018). The first source of bias is because the expected value is a non-linear function of the transition probabilities. Under the assumption that the transition probability estimates are unbiased, the bias in the value function estimate vanishes asymptotically with the increasing number of sample transitions. However, this bias is present with any finite number of samples (Mannor et al. 2004). The second source of bias is because of potentially biased transition probability estimates. In reality, the sample transitions come from time series data and are not necessarily independent. The sample mean estimator, consequently, is not guaranteed to be unbiased. Quantifying this bias requires knowing the model, which contradicts the motivations of RL.

We consider that the dataset is constructed using sample paths that are executions of an MDP under the behavior policy and derive a sample complexity upper bound for model-based off-policy evaluation. We overcome the first source of bias by using a robust MDP (Nilim and El Ghaoui 2005) that includes the true MDP with high probability. To overcome the second source of bias, we use a concentration bound that can handle random stopping times potentially dependent on the previous samples. We combine these methods and derive a sufficient condition on the number of sample paths to be collected to accurately estimate the value of the target policy with high probability. The bound shows that the vanilla model-based off-policy evaluation has performance guarantees under both partial and uniform coverage. We extend this sample complexity result to off-policy optimization under uniform coverage. In addition, as a comparison, we derive a sufficient condition on the number of samples for the vanilla importance sampling method. Finally, we give an estimator that outperforms the sample mean estimator in settings where transition dynamics of the MDP is almost deterministic, i.e., there is a probable next state for every state and action.

The main contributions of this paper are threefold:

1. We derive sufficient conditions on the number of sample paths for vanilla model-based off-policy evaluation and optimization. These bounds do not assume independence between the sample transitions.
2. We derive a sufficient condition on the number of sample paths for the importance-sampling-based off-policy evaluation in the infinite-horizon discounted reward setting.
3. We provide a new estimator for the transition probabilities that outperforms the sample mean estimator for the environments with a limited amount of stochasticity.

We remark that for the first two contributions, we aim to analyze the performance of vanilla off-policy methods in the discounted infinite horizon setting rather than building new algorithms with optimal sample complexities.

Preliminaries

A Markov decision process (MDP) is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, s_0)$ where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $\mathcal{P}(s, a, q)$ is the transition probability form state s to q under action a, $r(s, a)$ is the (random) reward of action a at state s, and s_0 is the initial state. We assume that the reward is normalized, i.e., $0 \leq \mathbb{E}[r(s, a)] \leq 1$ for all $s \in \mathcal{S}$, and $a \in \mathcal{A}$. S denotes the cardinality of $\mathcal{S}$ and A denotes the cardinality of $\mathcal{A}$. An absorbing state s transitions to itself under every action and has 0 reward, i.e., $\mathcal{P}(s, a, s) = 1$ and $r(s, a) = 0$ for all $a \in \mathcal{A}$. A (stationary) policy π assigns the same probabilities to actions given the state at every time step; $\pi(s, a)$ denotes the probability of taking action a at state s. An (infinite) path $\xi = s_0 a_0 r_0 s_1 a_1 r_1 \dots$ is a sequence of states, actions, and rewards. The value function $V^{\pi}_{\mathcal{M}}(s)$ denotes the expected total reward under policy π starting from s, i.e., $V^{\pi}_{\mathcal{M}}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} r(s_t, a_t)\right]$ where the expectation is over the randomness of the policy, transition dynamics, and rewards.

The occupancy measure $x^{\pi}(s, a)$ denotes the expected number of times that action a is taken at state s under policy π starting from s_0. Due to the linearity of expectation, we have $V^{\pi}_{\mathcal{M}}(s_0) = \sum_{s \in \mathcal{S}, a \in \mathcal{A}} x^{\pi}(s, a) \mathbb{E}[r(s, a)]$. Define $\rho^{\pi}(s, a)$ as the probability of taking action a at state s at least once under stationary policy π starting from s_0. Also, define $\lambda^{\pi}(s, a)$ as the probability of taking action a at state s again under stationary policy π given that the current state is s and current action is a. Due to the Markovianity of the transition dynamics and the stationarity of policy π, we have

$$x^{\pi}(s, a) = \rho^{\pi}(s, a) \sum_{i=1}^{\infty} (1 - \lambda^{\pi}(s, a)) \lambda^{\pi}(s, a)^{i-1} i = \frac{\rho^{\pi}(s, a)}{1 - \lambda^{\pi}(s, a)}$$

where i represents the number of times (s, a) is used.

Offline Reinforcement Learning Problem

We consider two offline reinforcement learning problems. The first problem is off-policy evaluation where the goal is to estimate the value $V^{\pi^t}_{\mathcal{M}}(s_0)$ of a known stationary target policy π^t given N sample paths that are collected under a known stationary behavior policy π^b. The second problem is off-policy optimization where the goal is to synthesize an optimal policy π^* that maximizes the value function $V^{\pi^*}_{\mathcal{M}}(s_0)$ given N sample paths that are collected under a known stationary behavior policy π^b.

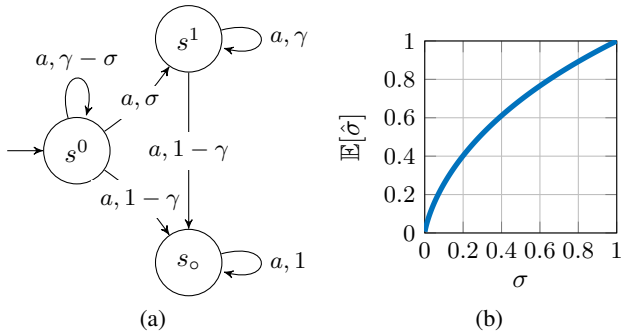

Figure 1: (a) An MDP with a single action. A label a, p of a directed edge from s to q means $\mathcal{P}(s, a, q) = p$. (b) The expected sample mean estimation $\mathbb{E}[\hat{\sigma}]$ of σ given a single path when $\gamma = 1$.

For both of these problems, we assume that there exists an absorbing final state $s_\circ$ such that every state in $\mathcal{S} \setminus \{s_\circ\}$ transitions to $s_\circ$ with probability $1 - \gamma$ under every action. State $s_\circ$ represents the effective end of the path. We note that transitioning to $s_\circ$ with a fixed probability is equivalent to having a discount factor γ and ensures the boundedness of the value function. On the other hand, the setting we consider is more disadvantaged compared to having infinite-length paths with discounted rewards since sample paths eventually end up at $s_\circ$ and the learner cannot access to further sample transitions from the other states.

Vanilla Model-Based Offline Learning

We analyze the vanilla model-based approach for the aforementioned offline reinforcement learning problems. In this section, we describe the model construction.

The model construction is fairly simple; we utilize the sample mean estimator to estimate the transition probabilities and rewards. Formally, let $n(s, a, q)$ be the total number of sample transitions in the sample paths from state s to state q under action a. Let $\tilde{\mathcal{P}}(s, a, q) = n(s, a, q) / \left(\sum_{q \in \mathcal{S}} n(s, a, q)\right)$ denote the empirical frequency of transitioning from s to q under a. The estimated transition probabilities $\hat{\mathcal{P}}(s, a, q)$ minimize the L_2 distance to the empirical frequencies subject to the constraint $\hat{\mathcal{P}}(s, a, s_\circ) = 1 - \gamma$. If state s has no sample transition, we set $\hat{\mathcal{P}}(s, a, s) = \gamma$ for all $a \in \mathcal{A}$. For simplicity, we assume that the mean reward $\mathbb{E}[r(s, a)]$ is known for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. The results we present in this paper can be extended to the unknown reward case by considering a sample mean estimator for the rewards as well.

Given the estimated model $\hat{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \hat{\mathcal{P}}, r, s_0)$ and the target policy π^t, the value of the target policy can be estimated by solving a set of equations or by value iteration.

We assume that this computation can be performed exactly. $V^{\pi^t}_{\hat{\mathcal{M}}}(s_0)$ denotes the model-based value estimate for π^t.

Bias Issues in Model-Based Offline Learning

The main challenges associated with model-based off-policy evaluation are due to the biases in the estimation of transition probabilities and the estimation of the value function.

Bias in the estimation of transition probabilities We note that the sample-mean estimator is the maximum likelihood estimator of the transition probabilities. However, this estimator is biased since the outcomes of the future samples are dependent on the previous samples. For example, consider the MDP given in Figure 1a. A transition from s^0 to s^1 implies that all previous transitions are from s^0 to s^0. The (potential) dependencies between the samples, i.e., the dependency between the number of samples and the outcomes of the sample transitions, make the sample mean estimate biased. As shown in Figure 1b, the bias of the sample mean estimator can be as large as 0.22 given a single sample path. In general, this bias occurs if the successor states of an origin state have different return probabilities to the origin state.

While the model-based off-policy estimation is provably good when the estimates for the transition probabilities are unbiased (Mannor et al. 2004), it is challenging to obtain unbiased estimates with low variance since the learning data usually consists of samples that have dependencies. A provably unbiased estimator utilizes only the first sample from the origin state. However, this estimator suffers from high variance due to the low number of samples. Another option to overcome the bias issue is fixing a number of samples per state-action pair as in the PAC-MDP literature (Strehl and Littman 2008). However, this approach may result in a large number of "unknown" state-action pairs and is wasteful in that not all samples are utilized.

We overcome the bias issue by considering a concentration bound that can work with a random number of samples and handle the dependencies between the outcomes.

Bias in the value function estimation Even when the transition probability can be estimated without a bias, the estimate for the value function is, in general, biased. Given the estimates for the transition probabilities are unbiased, the bias and variance of the value function estimate vanish as the number of samples per state approach to ∞ (Mannor et al. 2004). A way to overcome the bias in the value function estimation is to use a robust MDP model that uses a possible set of transition probabilities (Yu et al. 2020, 2021). The robust model is then used to compute upper and lower bounds on the value function. We also follow this approach and use a robust MDP to show that the value function estimate is accurate with high probability.

Theoretical Guarantees for Vanilla Model-Based Off-Policy Evaluation and Optimization

In this section, we analyze the performance of vanilla model-based off-policy evaluation and optimization.

We derive a bound on the number N of required sample paths that relates the estimation accuracy to the distance between the behavior and target policies. The bound is polynomial in the statistics of the behavior and target policies, and the size of the MDP.

Theorem 1. *Let N be the number of sample paths that are independently collected under π^b. Define*

$$D = \left\{ (s,a) \middle| x^{\pi^t}(s,a) \geq \frac{(\varepsilon/2)^{1/\beta}(1-\gamma)^{(2-\beta)/\beta}}{SA} \right\}.$$

If

$$N \geq \tilde{\mathcal{O}}\left(\min_{\beta\in[0,1]} \max_{(s,a)\in D} \left(\frac{S^{1+2\beta}A^{2\beta}}{(1-\gamma)^{4-2\beta}} \cdot \frac{x^{\pi^t}(s,a)^{2\beta}}{x^{\pi^b}(s,a)} \cdot \frac{1}{\varepsilon^2}, \frac{1}{\gamma\rho^{\pi^b}(s,a)} \right) \right)$$

then with probability at least $1-\delta$, we have

$$|V^{\pi^t}_{\mathcal{M}}(s_0) - V^{\pi^t}_{\hat{\mathcal{M}}}(s_0)| \leq \varepsilon$$

where the dependency on $1/\delta$ is logarithmic.

The proof is given in (Karabag and Topcu 2023).

The bound in Theorem 1 holds for every $\beta \in [0,1]$. This implies that vanilla model-based off-policy evaluation works under both uniform coverage and partial coverage. For $\beta = 0$, the bound depends on how uniformly π^b covers the state-action space, i.e., $\max 1/x^{\pi^b}(s,a)$. If $\beta > 0$, the bound depends on the distributional shift between the policies; it is sufficient that π^b covers the state-action pairs that are frequently visited by π^t. We note that we do not need to decide on the value of β a priori. We also note that the maximum is over the set D of state-action pairs for which the target policy has a sufficiently large occupancy measure. This implies that vanilla model-based estimation remains to be accurate for pathological cases where some parts of the MDP are unreachable or reached with a very low probability under both policies.

The first fraction in the bound given in Theorem 1 shows that as the occupancy measures of the behavior and target policies get close to each other in terms of ratio, then the off-policy estimation gets more accurate. In detail, the sufficient number of sample paths increase when the size of the MDP S, the maximum occupancy measure $1/1-\gamma$, the desired accuracy $1/\varepsilon$ or the distributional shift $x^{\pi^t}(s,a)/x^{\pi^b}(s,a)$ between the behavior and target policies increase. The last fraction in the bound is a natural consequence of rejection sampling: $\tilde{\mathcal{O}}(1/\gamma\rho^{\pi^b}(s,a))$ sample paths are required to ensure that there is at least one path that has a sample from (s,a) to $\mathcal{S}\setminus\{s_\circ\}$.

In the extreme case where the behavior and target policies are the same, the bound has $1/\varepsilon^2$ dependence on the desired optimality gap. We note that the expected reward of a random path is subgaussian, and the $1/\varepsilon^2$ dependency matches the Chernoff bound.

Proof sketch for Theorem 1 We follow the steps shown in Figure 2 to prove Theorem 1. We first decide on the sufficient accuracy level for the transition probability estimates

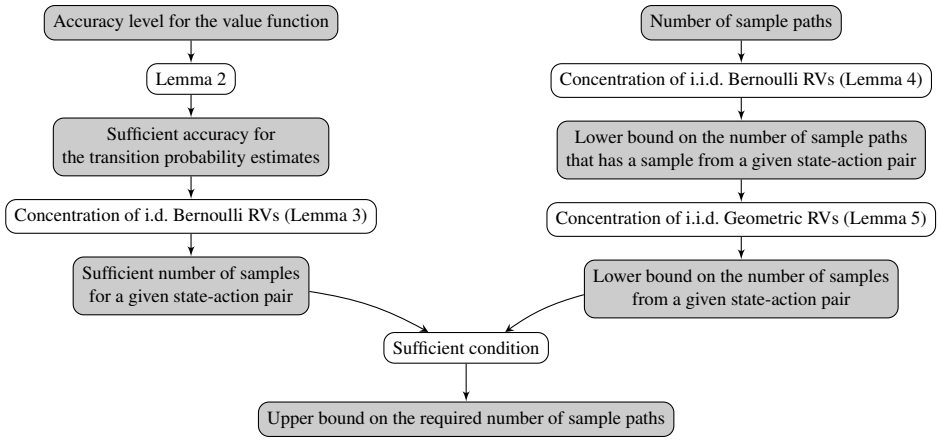

Figure 2: The flowchart for the proof of Theorem 1. Gray boxes are the relevant quantities, and white boxes are the relations between these quantities.

for a given level of accuracy for the value function. Next, we decide on a sufficient number of samples from a given state-action pair to achieve the accuracy level for the transition probability estimates. Finally, we decide on a sufficient number of sample paths to collect the sufficient number of samples from given state-action pairs.

We first decide on the required accuracy for the transition probability estimates from each state to make an accurate estimation for the value function. Lemma 2 shows that if the transition probabilities of state-action pairs are accurate proportionally to their occupancy measures, then the estimated value function is accurate. We note that this lemma is similar to simulation lemma (Strehl and Littman 2008); however, unlike the simulation lemma, we do not assume a fixed accuracy level for every state-action pair for the estimation of transition probabilities.

Lemma 2. *For any $\alpha \geq 0$ and $0 \leq \beta \leq 1$, if*

$$\sum_{q \in \mathcal{S}} |\hat{\mathcal{P}}(s,a,q) - \mathcal{P}(s,a,q)| \leq \frac{\alpha}{x^{\pi^t}(s,a)^\beta}$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, then

$$|V^{\pi^t}_{\mathcal{M}}(s_0) - V^{\pi^t}_{\hat{\mathcal{M}}}(s_0)| \leq \frac{\alpha (SA)^\beta}{(1-\gamma)^{2-\beta}}.$$

We set $\alpha = {\varepsilon(1-\gamma)^{(2-\beta)}}/{(SA)^\beta}$ in Lemma 2 to achieve ε accuracy. Given Lemma 2, our goal is to determine the number of sample paths that guarantee a desired estimation accuracy. In order to do so, we first determine the number of sample transitions that is sufficient to estimate the transition probabilities accurately. Lemma 3 provides an upper bound on the number of samples from a state-action pair to estimate transition probabilities within a desired accuracy.

Lemma 3. *For any $0 < \delta' < 1$, if*

$$\sum_{q \in \mathcal{S} \setminus \{s_\circ\}} n(s,a,q) \geq \frac{40S}{(\varepsilon')^2} \log\left(\frac{1}{\varepsilon'}\right) \log\left(\frac{5}{3\delta}\right),$$

then $\sum_{q \in \mathcal{S}} |\hat{\mathcal{P}}(s,a,q) - \mathcal{P}(s,a,q)| \leq \gamma\varepsilon'$ with probability at least $1-\delta'$.

To prove Lemma 3, we use a concentration bound that can handle a random number of samples and possible dependencies between the samples. The bound is a random stopping time generalization of the i.i.d. concentration inequality given in (Weissman et al. 2003) for the categorical distributions. Thanks to this bound, we overcome the aforementioned bias problem in estimating transition probabilities during our analysis. By Lemma 3, to achieve the accuracy given in Lemma 2, $\tilde{\mathcal{O}}\left({S^{1+2\beta}A^{2\beta}\gamma^2 x^{\pi^t}(s,a)^{2\beta}}/{\varepsilon^2(1-\gamma)^{4-2\beta}}\right)$ samples from (s,a) to $\mathcal{S} \setminus \{s_\circ\}$ are sufficient.

In the second part of the proof, we decide on the required number of paths to have enough samples from every state. We first decide on the number of sample paths that has a sample from (s,a) to $\mathcal{S} \setminus \{s_\circ\}$. Lemma 4 provides a lower bound on the number of sample paths that has a sample from a given state-action pair.

Lemma 4. *Let N be the number of sample paths that are independently collected under π^b. For any $0 < \delta' < 1$, $N' \geq 0$, and $(s,a) \in \mathcal{S} \times \mathcal{A}$, if*

$$N \geq \frac{6}{\gamma \rho^{\pi^b}(s,a)} \max\left(N', \log({1}/{\delta'})\right),$$

then the number of sample paths that has a sample from state s to $\mathcal{S} \setminus \{s_\circ\}$ under action a, is at least N' with probability at least $1-\delta'$.

By Lemma 4, to have N' sample paths that has a sample from (s,a) to $\mathcal{S} \setminus \{s_\circ\}$, $\tilde{\mathcal{O}}\left({N'}/{\gamma\rho^{\pi^b}(s,a)}\right)$ sample paths are sufficient.

We note that if a path has a sample from (s,a) to $\mathcal{S} \setminus \{s_\circ\}$, then the number of samples from (s,a) to $\mathcal{S} \setminus \{s_\circ\}$ follows a geometric distribution due to the stationarity of π^b. By a tail bound for the sum of geometric random variables (Janson 2018), Lemma 5 provides an upper bound on the number of sample paths that has a sample from a given state-action pair, in order to ensure a desired number of sample transitions from the state-action pair.

Lemma 5. *Let N' be the number of sample paths that has a sample from state s to $\mathcal{S} \setminus \{s_\circ\}$ under action a. For any*

$0 < \delta' < 1$ *and* $k \geq 0$*, if*

$$N' \geq \max\left(8k(1-\lambda^{\pi^b}(s,a)), \log(1/\delta')\right)$$

then the number of sample transitions from state s *to* $\mathcal{S}\setminus\{s_\circ\}$ *under action* a*, is at least* k *with probability at least* $1-\delta'$*.*

By Lemma 5, if N' paths has a sample from (s,a) to $\mathcal{S}\setminus\{s_\circ\}$, then we have $\tilde{\mathcal{O}}(N'/(1-\lambda^{\pi^b}(s,a)))$ sample transitions from (s,a) to $\mathcal{S}\setminus\{s_\circ\}$ with high probability. Combining these bounds and the sufficient number of samples per state-action pair, we conclude that if the number N of paths is greater than or equal to

$$N \geq \tilde{\mathcal{O}}\left(\min_{\beta\in[0,1]} \max_{\substack{s\in\mathcal{S}\setminus\{s_\circ\}\\ a\in\mathcal{A}}} \left(\frac{S^{1+2\beta}A^{2\beta}}{(1-\gamma)^{4-2\beta}} \cdot \frac{x^{\pi^t}(s,a)^{2\beta}}{x^{\pi^b}(s,a)} \cdot \frac{1}{\varepsilon^2}, \frac{1}{\gamma\rho^{\pi^b}(s,a)}\right)\right),$$

then the model-based off-policy estimate is ε-accurate with high probability. Finally, we note that if $x^{\pi^t}(s,a)^\beta \leq \varepsilon(1-\gamma)^{1-\beta}/2(SA)^\beta$, then the transition probability estimates for (s,a) trivially satisfies the condition given in Lemma 2, and β can be arbitrarily chosen between 0 and 1. Hence, if

$$N \geq \tilde{\mathcal{O}}\left(\min_{\beta\in[0,1]} \max_{(s,a)\in D} \left(\frac{S^{1+2\beta}A^{2\beta}}{(1-\gamma)^{4-2\beta}} \cdot \frac{x^{\pi^t}(s,a)^{2\beta}}{x^{\pi^b}(s,a)} \cdot \frac{1}{\varepsilon^2}, \frac{1}{\gamma\rho^{\pi^b}(s,a)}\right)\right),$$

then the model-based off-policy estimate is ε-accurate with high probability where

$$D = \left\{(s,a)\middle| x^{\pi^t}(s,a) \leq \frac{(\varepsilon/2)^{1/\beta}(1-\gamma)^{(1-\beta)/\beta}}{SA}\right\}.$$

Comparison with importance sampling As a comparison for the bound given in Theorem 1, we derive a bound on the sufficient number of sample paths that need to be collected for the vanilla off-policy estimation with importance sampling in the infinite-horizon discounted-reward setting.

The importance sampling estimate is

$$\hat{V}^{\pi^t}_{\mathcal{M}}(s_0) = \frac{1}{N}\sum_{i=1}^{N} \frac{\Pr(\xi^i = s_0^i a_0^i \ldots|\pi^t)}{\Pr(\xi^i = s_0^i a_0^i \ldots|\pi^b)} \sum_{t=0}^{\infty} \mathbb{E}\left[r(s_t^i, a_t^i)\right]$$

where $\xi^1, \ldots, \xi^N$ are sample paths collected under π^b. Let Γ^t and Γ^b denote the distribution of paths under the target and behavior policies, respectively. To accurately estimate the expected value of a given function with respect to the probability measure Γ^t with high probability using the sample paths that are collected under policy Γ^b, approximately

$$\exp\left(\mathbb{E}_{\xi\sim\Gamma^{\pi^t}}\left[l(\xi)\right] + \mathcal{O}\left(\mathrm{Std}_{\xi\sim\Gamma^{\pi^t}}\left(l(\xi)\right)\right)\right)$$

samples are sufficient for importance sampling where $l(\xi) = \log\left(\Pr(\xi|\pi^t)/\Pr(\xi|\pi^b)\right)$ (Chatterjee and Diaconis 2018). Considering that π^t and π^b are stationary, we have

$$\mathbb{E}_{\xi\sim\Gamma^{\pi^t}}\left[l(\xi)\right] \leq \frac{1}{1-\gamma}\log\left(\max_{\substack{s\in\mathcal{S}\setminus s_\circ\\ a\in\mathcal{A}}} \frac{\pi^t(s,a)}{\pi^b(s,a)}\right)$$

and

$$\mathrm{Std}_{\xi\sim\Gamma^{\pi^t}}(l(\xi)) \leq \frac{\sqrt{2}}{1-\gamma}\log\left(\max_{\substack{s\in\mathcal{S}\setminus s_\circ\\ a\in\mathcal{A}}} \frac{\pi^t(s,a)}{\pi^b(s,a)}\right).$$

The details of this derivation are given in (Karabag and Topcu 2023). As a result, approximately

$$\exp\left(\mathbb{E}_{\xi\sim\Gamma^{\pi^t}}\left[l(\xi)\right] + \mathcal{O}\left(\mathrm{Std}_{\xi\sim\Gamma^{\pi^t}}\left(l(\xi)\right)\right)\right) \leq \left(\max_{\substack{s\in\mathcal{S}\setminus s_\circ\\ a\in\mathcal{A}}} \frac{\pi^t(s,a)}{\pi^b(s,a)}\right)^{\mathcal{O}\left(\frac{1}{1-\gamma}\right)}.$$

sample paths are sufficient for the vanilla off-policy estimation with importance sampling in the infinite-horizon discounted-reward case.

A form of the maximum distributional shift between the policies, $\max_{(s,a)\in D}\left(x^{\pi^t}(s,a)^{2\beta}/x^{\pi^b}(s,a)\right)$ for the model-based method and $\max_{\substack{s\in\mathcal{S}\setminus s_\circ\\ a\in\mathcal{A}}}\left(\pi^t(s,a)/\pi^b(s,a)\right)$ for the importance sampling method, appears in the upper bounds for both model-based off-policy estimation and off-policy estimation via importance sampling: As the inherit distance between the target and behavior policies increase, the problem of off-policy estimation becomes more challenging.

We also note that the upper bound for the importance sampling has an exponential dependency on the expected time horizon, i.e., $1/1-\gamma$. Similar to the finite-horizon case, the variance of the estimates can grow exponentially with the expected time horizon in the infinite-horizon discounted-reward case.

Off-policy optimization The estimated model can be used to maximize a known reward function. By letting $\beta = 0$ in Theorem 1, we have the following result, which provides a bound on the number of paths that need to be collected for off-policy optimization.

Corollary 6. *Let* N *be the number of sample paths that are independently collected under* π^b*. Also, let* π' *denote optimal policy for the estimated model* $\hat{\mathcal{M}}$*, and* π^* *denote optimal policy for the true model* $\mathcal{M}$*. If the number* N *of paths is greater than or equal to*

$$\tilde{\mathcal{O}}\left(\max_{\substack{s\in\mathcal{S}\setminus\{s_\circ\}\\ a\in\mathcal{A}}}\left(\frac{S\gamma}{(1-\gamma)^4}\cdot\frac{1}{x^{\pi^b}(s,a)}\cdot\frac{1}{\varepsilon^2}, \frac{1}{\gamma\rho^{\pi^b}(s,a)}\right)\right),$$

then with probability at least $1-\delta$*, we have*

$$|V^{\pi'}_{\mathcal{M}}(s_0) - V^{\pi^*}_{\mathcal{M}}(s_0)| \leq \varepsilon,$$

where the dependency on $1/\delta$ *is logarithmic.*

The proof is given in (Karabag and Topcu 2023).

Corollary 6 shows the accuracy of vanilla model-based optimization under uniform coverage, i.e., dependency on $1/x^{\pi^b}(s,a)$. Related works such as (Yan et al. 2022; Rashidinejad et al. 2021) provide sample complexity bounds that require only partial coverage, i.e., $x^{\pi^*}(s,a)/x^{\pi^b}(s,a)$. The difference is because the vanilla model-based estimation is performed using the sample-mean estimates whereas the aforementioned works use a pessimistic MDP (as in the proof of

Theorem 1). We can recover the same sample complexity bound given in Theorem 1 for off-policy optimization with a pessimistic MDP built using the confidence bounds given in Lemma 3.

Estimation of the Accuracy Level The learner is agnostic to the value of the bound given in Theorem 1. Computing the lower bound requires knowing the model fully, which inherently contradicts the motivation of off-policy learning. Consequently, the learner cannot directly know the accuracy level for a given confidence level.

To estimate the level of accuracy, we can use the occupancy measure values that are computed using the estimated model. This approach gives an asymptotically consistent estimator of the accuracy level. Since the estimated model becomes more accurate with the increasing number of samples, the estimated occupancy measures become more accurate and, consequently, the estimate for the level of accuracy becomes more accurate.

In order to estimate the accuracy level with non-asymptotical guarantees, we need to compute the occupancy measures of the behavior and target policies. Computing the occupancy measures of the behavior policy is relatively easy and does not require the model construction: The sample paths are direct samples for the occupancy measures. Since the number of samples from a state-action pair in a path is a subexponential random variable, we can use Bernstein's inequality to compute a lower bound on the occupancy measures of the behavior policy. On the other hand, computing the occupancy measures of the target policy is challenging since the distribution of interest and the source of samples are not the same. For this computation, we can use robust MDPs. Lemma 3 gives a possible set of transition probabilities for a given confidence level. Using this lemma, we can build a robust MDP that contains the true transition probabilities with high probability. Then, we compute the largest possible occupancy measures for the target policy. Finally, given the robust estimates for the occupancy measures and the number of sample paths, we can compute a provably correct accuracy level using the bound given in Theorem 1.

Beyond Sample Mean Estimators

Sample mean estimators are preferred due to their asymptotic consistency and low variance. But, can we do better? The answer is positive given prior knowledge of the transition dynamics of the system. Reinforcement learning is often concerned with almost deterministic systems that have noisy dynamics, but the amount of noise is limited. For these systems, there exists an asymptotically consistent estimator that has a lower error than the sample mean estimator in the regime of low number of samples.

In our context, an MDP is almost deterministic if for all $s \in \mathcal{S} \setminus \{s_\circ\}$ and $a \in \mathcal{A}$ there exists a $q \in \mathcal{S} \setminus \{s_\circ\}$ such that $\mathcal{P}(s,a,q) \geq \gamma(1-\epsilon)$ with $\epsilon \approx 0$.

Let X be a categorical random variable with support $\{1,\dots,K\}$ and probability distribution $[p_1,\dots,p_K]$. The sample mean estimator of p_i is $\hat{p}_i^{SM} = {}^{N_i}/_{N}$ where N is the number of i.i.d. samples and N_i is the number of samples with value i. Let $d = \arg\max_i N_i$. We define the deterministic-favored estimator as

$$\hat{p}_i^{DF} = \frac{N_i + \sqrt{N}\mathbb{1}_d(i)}{N + \sqrt{N}}$$

where $\mathbb{1}_d(i)$ is 1 if $i = d$ and 0 otherwise. We note that the deterministic-favored estimator does the opposite of the minimax estimator: while the minimax estimator favors a uniform posterior (Wasserman 2006), the deterministic-favored estimator boosts the estimated probability of the most frequent observation.

The deterministic-favored estimator is asymptotically consistent. Furthermore, despite being biased, it has a lower mean squared error (MSE) than the sample mean estimator when X is almost deterministic.

Theorem 7. *The MSE of the deterministic-favored estimator $\hat{p}_i^{DF}$ satisfies*

$$\mathbb{E}\left[\left(\hat{p}_i^{DF} - p_i\right)^2\right] \leq \frac{N(1-p_i)}{(N+\sqrt{N})^2} + \exp\left(-\frac{(2p_i-1)^2 N}{12(1-p_i)}\right)$$

if $p_i > {}^1/_2$.

The proof is given in (Karabag and Topcu 2023).

We note that the second term decays exponentially fast as N or p_i increases. On the other hand, the MSE of the sample mean estimator is ${}^{Np_i(1-p_i)}/_{N^2}$. The deterministic favored estimator performs better than the sample mean estimator if $p_i \geq 1-\epsilon$ where $\epsilon \approx {}^{N^2}/_{(N+\sqrt{N})^2}$. For instance, $1 - {}^{N^2}/_{(N+\sqrt{N})^2} = 0.826$ for $N = 100$. Symmetrically, for small p_i, the deterministic-favored estimator performs better than the sample mean estimator if there exists j such that $p_j \geq 1-\epsilon$ where $\epsilon \approx {}^{N^2}/_{(N+\sqrt{N})^2}$. Overall, the deterministic-favored estimator outperforms the sample mean estimator for almost-deterministic random variables in the regime of low number of samples.

Related Work

Off-policy evaluation literature mainly focuses on finding good estimators with a low bias and variance. The previous methods to solve this problem include importance sampling (Liu et al. 2018; Xie, Ma, and Wang 2019), variants of dynamic programming (Hao et al. 2021; Munos and Szepesvári 2008), and model-based approaches (Rashidinejad et al. 2021; Yan et al. 2022; Mannor et al. 2004). Model-based approaches are also used for offline policy optimization (Yu et al. 2020, 2021).

Yin, Bai, and Wang (2021) studied the sample complexity of vanilla model-based off-policy evaluation and optimization in the finite-horizon setting under uniform coverage. Different from the infinite horizon setting, the sample transitions are i.i.d. in the finite-horizon setting. For off-policy evaluation, Yin, Bai, and Wang (2021) showed a path sample complexity of $\tilde{\mathcal{O}}\left({}^{H^2}/_{\min_{\substack{s\in\mathcal{S}\\a\in\mathcal{A}}} x^{\pi^b}(s,a)\varepsilon^2} + {}^{H^2\sqrt{SA}}/_{\varepsilon}\right)$ using martingale concentration inequalities where H is the time horizon. For the infinite-horizon setting, the bound that we give in Theorem 1, does not have any non-logarithmic dependencies in the number of actions and has a higher order dependency in the time horizon when $\beta = 0$, i.e., uniform coverage. Similarly, for off-policy optimization Yin,

Bai, and Wang (2021) showed a path sample complexity of $\tilde{\mathcal{O}}\left(H^4 S/\min_{\substack{s\in\mathcal{S}\\ a\in\mathcal{A}}} x^{\pi^b}(s,a)\varepsilon^2\right)$. The bound that we give in Corollary 6 matches the finite-horizon bound of (Yin, Bai, and Wang 2021). We also remark that our bound matches the sample complexity bound of vanilla asynchronous Q-learning (Li et al. 2021).

The sample complexity of model-based off-policy optimization is studied extensively using pessimistic models (Uehara and Sun 2022; Ross and Bagnell 2012; Li et al. 2022; Rashidinejad et al. 2021). Recently, Li et al. (2022) showed a lower bound of $\tilde{\mathcal{O}}\left(S\max_{\substack{s\in\mathcal{S}\setminus s_\circ\\ a\in\mathcal{A}}} \frac{x^{\pi^t}(s,a)}{x^{\pi^b}(s,a)}/(1-\gamma)^3\varepsilon^2\right)$ (independent) sample transitions and provided an algorithm with a matching sample complexity. The algorithm given in (Li et al. 2022) uses a pessimistic model, whereas we analyze the sample complexity of vanilla model-based off-policy optimization without pessimistic penalties.

Different from the majority of model-based off-policy evaluation and optimization works, we consider that the samples are coming from time series data, i.e., paths, that obey the dynamics of the MDP. In the model-free setting, Uehara, Huang, and Jiang (2020); Yan et al. (2022); Li et al. (2021) consider that the samples come from an underlying Markov chain and showed that the independence assumption can be removed by assuming a mixing property for the underlying Markov chain. Unlike these works, we consider that the samples are (unbounded length) episodes of the MDP and do not require the underlying Markov chain to be ergodic. Consequently, we do not have the burn-in sampling costs due to the mixing time.

Existing literature on finite-horizon off-policy importance sampling (Liu et al. 2018; Xie, Ma, and Wang 2019) shows that the variance of the importance sampling estimates grows exponentially with the horizon length. As a similar result, we show a sample complexity upper bound for the discounted infinite-horizon setting that has an exponential dependency in the expected time horizon.

To the best of our knowledge, previous model-based offline RL works use sample-mean estimators to estimate the transition probabilities. For the almost-deterministic environments, we propose an estimator that is motivated by the minimax estimator of the categorical random variables (Wasserman 2006).

Conclusion

We analyzed the sample complexity of vanilla model-based off-policy reinforcement learning with dependent samples. Despite its simple nature, the sample complexities of the vanilla model-based method are comparable to those of optimal algorithms. While the sample mean estimator becomes biased with dependent samples, our results show the order of the sample complexity remains the same compared to the case with independent samples. We also give an estimator that outperforms the sample mean estimator for almost-deterministic random variables.

Acknowledgments

This work was supported in part by ARO W911NF2110009 and NSF 1652113.

References

Chatterjee, S.; and Diaconis, P. 2018. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2): 1099–1135.

Ebert, F.; Finn, C.; Dasari, S.; Xie, A.; Lee, A. X.; and Levine, S. 2018. Visual Foresight: Model-Based Deep Reinforcement Learning for Vision-Based Robotic Control. *CoRR*, abs/1812.00568.

Gilotte, A.; Calauzènes, C.; Nedelec, T.; Abraham, A.; and Dollé, S. 2018. Offline a/b testing for recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 198–206.

Hao, B.; Ji, X.; Duan, Y.; Lu, H.; Szepesvari, C.; and Wang, M. 2021. Bootstrapping Fitted Q-Evaluation for Off-Policy Inference. In *Proceedings of the 37th International Conference on Machine Learning*, 4074–4084.

Henderson, J.; Lemon, O.; and Georgila, K. 2008. Hybrid reinforcement/supervised learning of dialogue policies from fixed data sets. *Computational Linguistics*, 34(4): 487–511.

Janson, S. 2018. Tail bounds for sums of geometric and exponential variables. *Statistics & Probability Letters*, 135: 1–6.

Karabag, M. O.; and Topcu, U. 2023. On the Sample Complexity of Vanilla Model-Based Offline Reinforcement Learning with Dependent Samples. *arXiv preprint arXiv:2303.04268*.

Levine, S.; Pastor, P.; Krizhevsky, A.; Ibarz, J.; and Quillen, D. 2018. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International journal of robotics research*, 37(4-5): 421–436.

Li, G.; Cai, C.; Chen, Y.; Gu, Y.; Wei, Y.; and Chi, Y. 2021. Is Q-learning minimax optimal? a tight sample complexity analysis. *arXiv preprint arXiv:2102.06548*.

Li, G.; Shi, L.; Chen, Y.; Chi, Y.; and Wei, Y. 2022. Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*.

Liu, Q.; Li, L.; Tang, Z.; and Zhou, D. 2018. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in Neural Information Processing Systems*, 31.

Mannor, S.; Simester, D.; Sun, P.; and Tsitsiklis, J. N. 2004. Bias and variance in value function estimation. In *Proceedings of the 21st International Conference on Machine Learning*, 72.

Munos, R.; and Szepesvári, C. 2008. Finite-Time Bounds for Fitted Value Iteration. *Journal of Machine Learning Research*, 9(5).

Nilim, A.; and El Ghaoui, L. 2005. Robust Control of Markov Decision Processes with Uncertain Transition Matrices. *Operations Research*, 53(5): 780–798.

Rashidinejad, P.; Zhu, B.; Ma, C.; Jiao, J.; and Russell, S. 2021. Bridging offline reinforcement learning and imitation

learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34.

Ross, S.; and Bagnell, J. A. 2012. Agnostic system identification for model-based reinforcement learning. In *Proceedings of the 29th International Conference on Machine Learning*, 1905–1912.

Shortreed, S. M.; Laber, E.; Lizotte, D. J.; Stroup, T. S.; Pineau, J.; and Murphy, S. A. 2011. Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Machine learning*, 84(1): 109–136.

Strehl, A. L.; and Littman, M. L. 2008. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8): 1309–1331.

Swaminathan, A.; Krishnamurthy, A.; Agarwal, A.; Dudik, M.; Langford, J.; Jose, D.; and Zitouni, I. 2017. Off-policy evaluation for slate recommendation. *Advances in Neural Information Processing Systems*, 30.

Tseng, H.-H.; Luo, Y.; Cui, S.; Chien, J.-T.; Ten Haken, R. K.; and Naqa, I. E. 2017. Deep reinforcement learning for automated radiation adaptation in lung cancer. *Medical physics*, 44(12): 6690–6705.

Uehara, M.; Huang, J.; and Jiang, N. 2020. Minimax weight and q-function learning for off-policy evaluation. In *Proceedings of the 37th International Conference on Machine Learning*, 9659–9668. PMLR.

Uehara, M.; and Sun, W. 2022. Pessimistic Model-based Offline Reinforcement Learning under Partial Coverage. In *International Conference on Learning Representations*.

Wasserman, L. 2006. *All of nonparametric statistics*. Springer Science & Business Media.

Weissman, T.; Ordentlich, E.; Seroussi, G.; Verdu, S.; and Weinberger, M. J. 2003. Inequalities for the L1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*.

Xie, T.; Ma, Y.; and Wang, Y.-X. 2019. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *Advances in Neural Information Processing Systems*, 32.

Yan, Y.; Li, G.; Chen, Y.; and Fan, J. 2022. The efficacy of pessimism in asynchronous Q-learning. *arXiv preprint arXiv:2203.07368*.

Yin, M.; Bai, Y.; and Wang, Y.-X. 2021. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, 1567–1575. PMLR.

Yu, T.; Kumar, A.; Rafailov, R.; Rajeswaran, A.; Levine, S.; and Finn, C. 2021. Combo: Conservative offline model-based policy optimization. *Advances in Neural Information Processing Systems*, 34.

Yu, T.; Thomas, G.; Yu, L.; Ermon, S.; Zou, J. Y.; Levine, S.; Finn, C.; and Ma, T. 2020. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33: 14129–14142.

Zeng, A.; Song, S.; Welker, S.; Lee, J.; Rodriguez, A.; and Funkhouser, T. 2018. Learning synergies between pushing and grasping with self-supervised deep reinforcement learning. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4238–4245. IEEE.

Zhou, L.; Small, K.; Rokhlenko, O.; and Elkan, C. 2017. End-to-end offline goal-oriented dialog policy learning via policy gradient. *arXiv preprint arXiv:1712.02838*.