Near-Optimal Provable Uniform Convergence in Offline Policy Evaluation for Reinforcement Learning

Ming Yin
Department of Statistics
UC Santa Barbara
Santa Barbara, CA
ming_yin@ucsb.edu

Yu Bai
Salesforce Research
Palo Alto, CA
yu.bai@salesforce.com

Yu-Xiang Wang
Department of Computer Science
UC Santa Barbara
Santa Barbara, CA
yuxiangw@cs.ucsb.edu

Abstract

The problem of *Offline Policy Evaluation* (OPE) in Reinforcement Learning (RL) is a critical step towards applying RL in real life applications. Existing work on OPE mostly focus on evaluating a *fixed* target policy π, which does not provide useful bounds for offline policy learning as π will then be data-dependent. We address this problem by *simultaneously* evaluating all policies in a policy class Π — uniform convergence in OPE — and obtain nearly optimal error bounds for a number of global / local policy classes. Our results imply that the model-based planning achieves an optimal episode complexity of $\widetilde{O}(H^3/d_m\epsilon^2)$ in identifying an ϵ-optimal policy under the *time-inhomogeneous episodic* MDP model (H is the planning horizon, d_m is a quantity that reflects the exploration of the logging policy μ). To the best of our knowledge, this is the first time the optimal rate is shown to be possible for the offline RL setting and the paper is the first that systematically investigates the uniform convergence in OPE.

1 Introduction

In offline reinforcement learning (offline RL), there are mainly two fundamental problems: *offline policy evaluation* (OPE) and *offline learning* (also known as *batch RL*) (Sutton & Barto, 2018). OPE addresses to the statistical estimation problem of predicting the performance of a fixed target policy π with only data collected by a logging/behavioral policy μ. On the other hand, offline learning is a *statistical learning* problem that aims at learning a near-optimal policy using an offline dataset alone (Lange et al., 2012).

As offline RL methods do not require interacting with the task environments or having access to a simulator, they are more suitable for real-world applications of RL such as those in marketing (Thomas et al., 2017), targeted advertising (Bottou et al., 2013; Tang et al., 2013), finance (Bertoluzzo & Corazza, 2012), robotics (Quillen et al., 2018; Dasari et al., 2020), language (Jaques et al., 2019) and health care (Ernst et al., 2006; Raghu et al., 2017, 2018; Gottesman et al., 2019). In these tasks, it is usually not feasible to deploy an online RL algorithm to trials-and-error with the environment. Instead, we are given a large offline dataset of historical interaction to come up with a new policy π and to demonstrate that this new policy π will perform better using the same dataset without actually testing it online.

In this paper, we present our solution via a statistical learning perspective by studying the *uniform convergence* in OPE under the *non-stationary transition, finite horizon, episodic Markov decision process (MDP)* model with finite states and actions. Informally, given a policy class Π and a logging policy μ, uniform convergence problem in OPE (Uniform OPE for short) focuses on coming up with OPE estimator $\widehat{v}^\pi$ and characterizing the number of episodes n we need (from μ) in order for $\widehat{v}^\pi$ to

Offline Reinforcement Learning Workshop at Neural Information Processing Systems, 2020.

satisfies that with high probability

$$\sup_{\pi\in\Pi} |\widehat{v}^{\pi} - v^{\pi}| \le \epsilon.$$

The focus of research would be to characterizing the *episode complexity*: the number of episodes n needed as a function of ϵ, failure probability δ, the parameters of the MDP as well as the logging policy μ.

We highlight that even though uniform convergence is the main workhorse in statistical learning theory (see, e.g., Vapnik, 2013), few analogous results have been established for the offline reinforcement learning problem. The overarching theme of this work is to understand what a natural complexity measure is for policy classes in reinforcement learning and its dependence in the size of the state-space and planning horizon.

In addition, uniform OPE has two major consequences (which we elaborate in detail in the following motivation section): (1) allowing any accurate planning algorithm to work as sample efficient offline learning algorithm with our model-based method; (2) providing finite sample guarantee for offline evaluation uniformly for all policies in the policy class.

The Motivation. Existing research in offline RL usually focuses on designing specific algorithms that learn the optimal policy $\pi^\star := \mathrm{argmax}_\pi v^\pi$ with given static offline data $\mathcal{D}$. In the rich literature of statistical learning theory, however, learning bounds are often obtained via a stronger uniform convergence argument which ensures an arbitrary learner to output a model that generalizes. Specifically, the *empirical risk minimizer* (ERM) that outputs the *empirical optimal policy* has been shown to be sufficient and necessary for efficiently learning almost all learnable problems (Vapnik, 2013; Shalev-Shwartz et al., 2010).

The natural analogy of ERM in the RL setting would be to find the *empirical optimal policy* $\widehat{\pi}^\star := \mathrm{argmax}_\pi \widehat{v}^\pi$ for some OPE estimator $\widehat{v}^\pi$. If we could establish a uniform convergence bound for $\widehat{v}^\pi$, then it implies that $\widehat{\pi}^\star$ is nearly optimal too via

$$0 \le v^{\pi^\star} - v^{\widehat{\pi}^\star} = v^{\pi^\star} - \widehat{v}^{\pi^\star} + \widehat{v}^{\widehat{\pi}^\star} - v^{\widehat{\pi}^\star} \le |v^{\pi^\star} - \widehat{v}^{\pi^\star}| + |\widehat{v}^{\widehat{\pi}^\star} - v^{\widehat{\pi}^\star}| \le 2\sup_{\pi} |v^\pi - \widehat{v}^\pi|.$$

Thus, uniform OPE is a stronger setting than offline learning with the additional benefit of accurately evaluating any other (possibly heuristic) policy optimization algorithms that are used in practice.

From the OPE perspective, there is often a need to evaluate the performance of a *data-dependent* policy, and uniform OPE becomes useful. For example, when combined with existing methods, it will allow us to evaluate policies selected by safe-policy improvements, proximal policy optimization, UCB-style exploration-bonus as well as any heuristic exploration criteria such as curiosity, diversity and reward-shaping techniques.

Model-based estimator for OPE. The OPE estimator we consider in this paper is the standard model-based estimator, i.e., estimating the transition dynamics and immediate rewards, then simply plug in the parameters of empirically estimated MDP $\widehat{M}$ to obtain $\hat{v}^\pi$ for any π. This model-based approach has several benefits. **1.** It enables flexible choice of policy search methods since it converts the problem to planning over the estimated MDP $\widehat{M}$. **2.** Uniform OPE with model-based estimator avoids the use of data-splitting that leads to inefficient data use. For example, Sidford et al. (2018) learns the ϵ-optimal policy with the optimal rate in the generative model setting, where in each subroutine new independent data $s^{(1)}_{s,a}, ..., s^{(m)}_{s,a}$ need to be sampled to estimate $P_{s,a}$ and samples from previous rounds cannot be reused. A uniform convergence result could completely avoid data splitting during the learning procedure.

Our contribution. Our main contributions are summarized as follows.

- For the global policy class (deterministic or stochastic), we use fully model-based OPEMA estimator to obtain an ϵ-uniform OPE with episode complexity $\widetilde{O}(H^4S/d_m\epsilon^2)$ (Theorem 3.1) and in some cases this can be reduced to $\widetilde{O}(H^4/d_m\epsilon^2)$, where d_m is minimal marginal state-action occupancy probability depending on logging policy μ.
- For the global deterministic policy class, we obtain an ϵ-uniform OPE with episode complexity $\widetilde{O}(H^3S/d_m\epsilon^2)$ with an optimal dependence on H (Theorem 3.3).

- For a (data-dependent) local policy class that cover all policies are in the $O(\sqrt{H}/S)$-neighborhood of the *empirical* optimal policy (see the definition in Section 2.1), we obtain ϵ-uniform OPE with $\widetilde{O}(H^3/d_m\epsilon^2)$ episodes (Theorem 3.4).
- We prove a information-theoretical lower bound of $\Omega(H^3/d_m\epsilon^2)$ for OPE (Theorem 3.5) which certifies that results for local policy class is optimal.
- Our uniform OPE over the local policy class implies that ERM (VI or PI with empirically estimated MDP), as well as any sufficiently accurate model-based planning algorithm, has an optimal episode complexity of $\widetilde{O}(H^3/d_m\epsilon^2)$ (Theorem 4.1). To the best of our knowledge, this is the first rate-optimal algorithm in the offline RL setting.
- Last but not least, our result can be viewed as an improved analysis of the *simulation lemma*; which demystifies the common misconception that purely model plug-in estimator is inefficient, comparing to their model-free counterpart.

To the best of our knowledge, these results are new and this is the first work that derives uniform convergence analogous to those in the statistical learning theories for offline RL.

Related work. Before formally stating our results, we briefly discuss the related literature in three categories.

1. OPE: Most existing work on OPE focuses on the *Importance Sampling* (IS) methods (Li et al., 2011; Dudík et al., 2011; Li et al., 2015; Thomas & Brunskill, 2016) or their doubly robust variants (Jiang & Li, 2016; Farajtabar et al., 2018). These methods are more generally applicable even if the the Markovian assumption is violated or the states are not observable, but has an error (or sample complexity) that depends exponential dependence in horizon H. Recently, a family of estimators based on *marginalized importance sampling* (MIS) (Liu et al., 2018; Xie et al., 2019; Kallus & Uehara, 2020, 2019; Yin & Wang, 2020) have been proposed in order to overcome the "*curse of horizon*" under the additional assumption of state observability. In the tabular setting, Yin & Wang (2020) design the Tabular-MIS estimator which matches the Cramer-Rao lower bound constructed by Jiang & Li (2016) up to a low order term for every instance (π, μ and the MDP), which translates into an $O(H^2/d_m\epsilon^2)$ episode complexity in the (pointwise) OPE problem we consider for all π. Tabular-MIS, however, is identical to the model-based plug-in estimator we use, *off-policy empirical model approximator* (OPEMA), as we discuss further in Section 2.3. These methods do not address the uniform convergence problem. The only exception is (Yin & Wang, 2020), which has a result analogous to Theorem 3.4, but for a data-splitting-type estimator.

2. Offline Learning: For the offline learning, most theoretical work consider the infinite horizon discounted setting with function approximation. Chen & Jiang (2019); Le et al. (2019) first raises the information-theoretic considerations for offline learning and uses Fitted Q-Iteration (FQI) to obtain $\epsilon V_{\max}$-optimal policy using sample complexity $\widetilde{O}((1-\gamma)^{-4}C_\mu/\epsilon^2)$ where C_μ is *concentration coefficient* (Munos, 2003) that is similar to our $1/d_m$. More recently, (Xie & Jiang, 2020b) improves the result to $\tilde{O}((1-\gamma)^{-2}C_\mu/\epsilon^2)$. However, these bounds are not tight in terms of the dependence on the effective horizon[1] $(1-\gamma)^{-1}$. More recently, Xie & Jiang (2020a); Liu et al. (2020) explore weaker settings for batch learning but with suboptimal sample complexity dependences. Our result is the first that achieves the optimal rate (despite focusing on the finite horizon episodic setting).

3. Uniform convergence in RL: There are few existing work that deals with uniform convergence in OPE. However, we notice that the celebrated simulation lemma (Kearns & Singh, 2002) is actually an uniform bound with an episode complexity of $O(H^4S^2/d_m\epsilon^2)$. Several existing work uses uniform-convergence arguments over value function classes for online RL (see, e.g., Jin et al., 2020, and the references therein). The closest to our work is perhaps (Agarwal et al., 2020), which studies model-based planning in the generative model setting. We are different in that we are in the offline learning setting. In addition, our local policy class is optimal for a larger region of ϵ_{opt} (independent to n), while their results (Lemma 10) imply optimal OPE only for empirically optimal policy with $\epsilon_{\text{opt}} \leq \sqrt{(1-\gamma)^{-5}SA/n}$. Lastly, we discovered the thesis of Tewari (2007, Ch.3 Theorem 1), which discusses the pseudo-dimension of policy classes. The setting is not compatible to ours, and does not imply a uniform OPE bound in our setting.

[1] The optimal rate should be $(1-\gamma)^{-1}C/\epsilon^2$, analogous to our $H^3/d_m\epsilon^2$ bound. The additional H^2 is due to scaling — we are obtaining ϵ-optimal policy and they obtain $\epsilon V_{\max}$-optimal policy ($V_{\max} = H$ in our case). See Table 1 for a consistent comparison.

2 Problem setup and method

RL environment is usually modeled as a *Markov Decision Process* (MDP) which is denoted by $M = (\mathcal{S}, \mathcal{A}, r, P, d_1, H)$. The MDP consists of a state space $\mathcal{S}$, an action space $\mathcal{A}$ and a transition kernel $P_t : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$ with $P_t(s'|s, a)$ representing the probability transition from state s, action a to next state s' at time t. In particular here we consider non-stationary transition dynamics so P_t varies over time t. Besides, $r_t : \mathcal{S} \times A \mapsto \mathbb{R}$ is the expected reward function and given (s_t, a_t), $r_t(s_t, a_t)$ specifies the average reward obtained at time t. d_1 is the initial state distribution and H is the horizon. Moreover, we focus on the case where state space $\mathcal{S}$ and the action space $\mathcal{A}$ are finite, *i.e.* $S := |\mathcal{S}| < \infty, A := |\mathcal{A}| < \infty$. A (non-stationary) policy is formulated by $\pi := (\pi_1, \pi_2, ..., \pi_H)$, where π_t assigns each state $s_t \in \mathcal{S}$ a probability distribution over actions at each time t. Any fixed policy π together with MDP M induce a distribution over trajectories of the form $(s_1, a_1, r_1, s_2, ..., s_H, a_H, r_H, s_{H+1})$ where $s_1 \sim d_1$, $a_t \sim \pi_t(\cdot|s_t)$, $s_{t+1} \sim P_t(\cdot|s_t, a_t)$ and r_t has mean $r_t(s_t, a_t)$ for $t = 1, ..., H$.[2]

In addition, we denote $d_t^\pi(s_t, a_t)$ the induced marginal state-action distribution and $d_t^\pi(s_t)$ the marginal state distribution, satisfying $d_t^\pi(s_t, a_t) = d_t^\pi(s_t) \cdot \pi(a_t|s_t)$. Moreover, $d_1^\pi = d_1\ \forall\pi$. We use the notation $P_t^\pi \in \mathbb{R}^{S\cdot A \times S\cdot A}$ to represent the state-action transition $(P_t^\pi)_{(s,a),(s',a')} := P_t(s'|s, a)\pi_t(a'|s')$, then the marginal state-action vector $d_t^\pi(\cdot, \cdot) \in \mathbb{R}^{S\times A}$ satisfies the expression $d_{t+1}^\pi = P_{t+1}^\pi d_t^\pi$. We define the quantity $V_t^\pi(s) = \mathbb{E}_\pi[\sum_{t'=t}^H r_{t'}|s_t = s]$ and the Q-function $Q_t^\pi(s, a) = \mathbb{E}_\pi[\sum_{t'=t}^H r_{t'}|s_t = s, a_t = a]$ for all $t = 1, ..., H$. The ultimate measure of the performance of policy π is the value function:

$$v^\pi = \mathbb{E}_\pi\left[\sum_{t=1}^H r_t\right].$$

Lastly, for the standard OPE problem, the goal is to estimate v^π for a given π while assuming that n episodic data $\mathcal{D} = \left\{(s_t^{(i)}, a_t^{(i)}, r_t^{(i)}, s_{t+1}^{(i)})\right\}_{i\in[n]}^{t\in[H]}$ are rolling from a *different* behavior policy μ.

2.1 Uniform convergence problems

Uniform OPE extends the pointwise OPE to a family of policies. Specifically, for an policy class Π of interest, we aim at showing that $\sup_{\pi\in\Pi} |\widehat{v}^\pi - v^\pi| < \epsilon$ with high probability with optimal dependence in all parameters. In this paper, we consider three policy classes.

The global policy class. The policy class Π we considered here consists of all the non-stationary policies, deterministic or stochastic. This is the largest possible class we can consider and hence the hardest one.

The global deterministic policy class. Here class consists of all the non-stationary deterministic policies. By the standard results in reinforcement learning, there exists at least one deterministic policy that is optimal (Sutton & Barto, 2018). Therefore, the deterministic policy class is rich enough for evaluating any learning algorithm (*e.g.* Q-value iteration in Sidford et al. (2018)) that wants to learn to the optimal policy.

The local policy class: in the neighborhood of empirical optimal policy. Given empirical MDP $\widehat{M}$ (*i.e.* the transition kernel is replaced by $\widehat{P}_t(s_{t+1}|s_t, a_t) := n_{s_{t+1},s_t,a_t}/n_{s_t,a_t}$ if $n_{s_t,a_t} > 0$ and 0 otherwise, where n_{s_t,a_t} is the number of visitations to (s_t, a_t) among all n episodes[3]), it is convenient to learn the empirical optimal policy $\widehat{\pi}^\star := \text{argmax}_\pi\, \widehat{v}^\pi$ since the full empirical transition $\widehat{P}$ is known. Standard methods like Policy Iteration (PI) and Value Iteration (VI) can be leveraged for finding $\widehat{\pi}^\star$. This observation allows us to consider the following interesting policy class: $\Pi_1 := \{\pi : s.t.\ ||\widehat{V}_t^\pi - \widehat{V}_t^{\widehat{\pi}^\star}||_\infty \leq \epsilon_{\text{opt}},\ \forall t = 1, ..., H\}$ with $\epsilon_{\text{opt}} \geq 0$ a parameter. Here we consider $\widehat{\pi}^\star$ (instead of $\pi^\star$) since by defining with empirical optimal policy, we can use data $\mathcal{D}$ to really check class Π_1, therefore this definition is more practical.

[2] Here r_t without any argument is random reward and $\mathbb{E}[r_t|s_t, a_t] = r_t(s_t, a_t)$.

[3] Similar definition holds for n_{s_{t+1},s_t,a_t}.

2.2 Assumptions

Next we present some mild necessary regularity assumptions for uniform convergence OPE problem.

Assumption 2.1 (Bounded rewards). *$\forall\, t = 1, ..., H$ and $i = 1, ..., n$, $0 \leq r_t^{(i)} \leq 1$.*

Assumption 2.2 (Exploration requirement). *Logging policy μ obeys that $\min_{t,s_t} d_t^\mu(s_t) > 0$, for any state s_t that is "accessible". Moreover, we define quantity $d_m := \min\{d_t^\mu(s_t, a_t) : d_t^\mu(s_t, a_t) > 0\}$.*

State s_t is "accessible" means there exists a policy π so that $d_t^\pi(s_t) > 0$. If for any policy π we always have $d_t^\pi(s_t) = 0$, then state s_t can never be visited in the given MDP. Assumption 2.2 simply says μ have the right to explore all "accessible" states. This assumption is required for the consistency of uniform convergence estimator since we have "$\sup_{\pi\in\Pi}$" and is similar to the standard *concentration coefficient* assumption made by Munos (2003); Le et al. (2019). As a short comparison, offline learning problems (*e.g.* offline policy optimization in Liu et al. (2019)) only require $d_t^\mu(s_t) > 0$ for any state s_t satisfies $d_t^{\pi^\star}(s_t) > 0$. Last but not least, even though our target policy class is deterministic, by above assumptions μ is always stochastic.

2.3 Method: Offline Policy Empirical Model Approximator

The method we use for doing OPE in uniform convergence is the *offline policy empirical model approximator* (OPEMA). OPEMA uses off-policy data to build the empirical estimators for both the transition dynamic and the expected reward and then substitute the related components in real value function by its empirical counterparts. First recall for any target policy π, by definition: $v^\pi = \sum_{t=1}^H \sum_{s_t,a_t} d_t^\pi(s_t, a_t) r_t(s_t, a_t)$, where the marginal state-action transitions satisfy $d_{t+1}^\pi = P_{t+1}^\pi d_t^\pi$. OPEMA then directly construct empirical estimates for $\widehat{P}_{t+1}(s_{t+1}|s_t, a_t)$ and $\widehat{r}_t(s_t, a_t)$ as:

$$\widehat{P}_{t+1}(s_{t+1}|s_t, a_t) = \frac{\sum_{i=1}^n \mathbf{1}[(s_{t+1}^{(i)}, a_t^{(i)}, s_t^{(i)}) = (s_{t+1}, s_t, a_t)]}{n_{s_t,a_t}}, \quad \widehat{r}_t(s_t, a_t) \quad = \frac{\sum_{i=1}^n r_t^{(i)} \mathbf{1}[(s_t^{(i)}, a_t^{(i)}) = (s_t, a_t)]}{n_{s_t,a_t}}.$$

and $\widehat{P}_{t+1}(s_{t+1}|s_t, a_t) = 0$ and $\widehat{r}_t(s_t, a_t) = 0$ if $n_{s_t,a_t} = 0$ (recall n_{s_t,a_t} is the visitation frequency to (s_t, a_t) at time t) and then the estimates for state-action transition $\widehat{P}_t^\pi$ is defined as: $\widehat{P}_t^\pi(s_{t+1}, a_{t+1}|s_t, a_t) = \widehat{P}_t(s_{t+1}|s_t, a_t)\pi(a_{t+1}|s_{t+1})$.The initial distribution is also constructed using empirical estimator $\widehat{d}_1^\pi(s_1) = n_{s_1}/n$. Based on the construction, the empirical marginal state-action transition follows $\widehat{d}_{t+1}^\pi = \widehat{P}_{t+1}^\pi \widehat{d}_t^\pi$ and the final estimator for v^π is:

$$\widehat{v}^\pi_{\text{OPEMA}} = \sum_{t=1}^H \sum_{s_t,a_t} \widehat{d}_t^\pi(s_t, a_t) \widehat{r}_t(s_t, a_t). \tag{1}$$

OPEMA is model-based method as it uses plug-in estimators ($\widehat{d}_t^\pi$ and $\widehat{r}_t$) for each model components (d_t^π and r_t). Traditionally, the error of OPEMA is obtained via the simulation lemma (Kearns & Singh, 2002), with $O(H^4S^2/d_m\epsilon^2)$-episode complexity. Recent work (Xie et al., 2019; Yin & Wang, 2020; Duan et al., 2020) reveals that there is an importance sampling interpretation of OPEMA

$$\widehat{v}^\pi_{\text{OPEMA}} = \frac{1}{n}\sum_{i=1}^n \sum_{t=1}^H \frac{\hat{d}^\pi(s_t^{(i)})}{\hat{d}_t^\mu(s_t^{(i)})} \hat{r}_t^\pi(s^{(i)}), \tag{2}$$

and the effectiveness of MIS of recent work partially explains why OPEMA could work, even for the Uniform OPE problem.

3 Main results for Uniform OPE

3.1 Uniform OPE for global policy class

We present the following result Theorem 3.1 for global policy class.

Theorem 3.1. *Let Π consists of all policies, then there exists an absolute constant c such that if $n > c \cdot 1/d_m \cdot \log(HSA/\delta)$, then with probability $1 - \delta$, we have:*

$$\sup_{\pi\in\Pi} |\widehat{v}^\pi - v^\pi| \leq c \left(\sqrt{\frac{H^4 \log(\frac{HSA}{\delta})}{d_m \cdot n}} + \sqrt{\frac{H^4 S \log(nHSA)}{d_m \cdot n}} \right).$$

Moreover, if failure probability $\delta < e^{-S}$*, then above can be further bounded by* $2c\sqrt{\frac{H^4}{d_m \cdot n}\log(\frac{nHSA}{\delta})}$.

Our result improves over the simulation lemma by a factor of HS but is suboptimal by another factor HS comparing to the lower bound (Theorem 3.5). In the small failure probability regime ($\delta < e^{-S}$) we can get rid of the dependence on S except for the implicit dependence through d_m. This is meaningful since we usually consider deriving results with high confidence.

3.2 Uniform OPE for deterministic policies

Lemma 3.2 (Convergence for fixed policy)**.** *Fix any policy* π*. Then there exists absolute constants* c, c_1, c_2 *such that if* $n > c \cdot 1/d_m \cdot \log(HSA/\delta)$*, then with probability* $1-\delta$*, we have:*

$$|\widehat{v}^{\pi} - v^{\pi}| \leq c_1\sqrt{\frac{H^2\log(\frac{c_2 HSA}{\delta})}{n \cdot d_m}} + \tilde{O}\left(\frac{H^2\sqrt{SA}}{n \cdot d_m}\right).$$

Note if we absorb the higher order term, our result implies sample complexity of $\widetilde{O}(H^2/d_m\epsilon^2)$ for evaluating any fixed target policy π. Notice that the total number of deterministic policies is A^{HS} in our problem, a standard union bound over all deterministic policies yields the following result.

Theorem 3.3. *Let* Π *consists of all deterministic policies, then there exists absolute constants* c, c_1, c_2 *such that if* $n > c \cdot 1/d_m \cdot \log(HSA/\delta)$*, then with probability* $1-\delta$*, we have:*

$$\sup_{\pi\in\Pi} |\widehat{v}^{\pi} - v^{\pi}| \leq c_1\sqrt{\frac{H^3 S\log(\frac{c_2 HSA}{\delta})}{n \cdot d_m}} + \tilde{O}\left(\frac{H^3 S^{1.5} A^{0.5}}{n \cdot d_m}\right).$$

Theorem 3.3 implies an episode complexity of $\widetilde{O}(H^3 S/d_m\epsilon^2)$, which is optimal in H but suboptimal by a factor of S. While the deterministic policy class seems restrictive, it could be useful in many cases because the optimal policy is deterministic, and many exploration-bonus based exploration methods use deterministic policy throughout.

3.3 Uniform OPE for the local (near *empirically optimal*) policy class

For the local (near *empirically optimal*) policy class we described in Section 2.1, the following theorem obtains the optimal episode complexity.

Theorem 3.4. *Suppose* $\epsilon_{opt} \leq \sqrt{H}/S$ *and* $\Pi_1 := \{\pi : s.t.\ ||\widehat{V}_t^{\pi} - \widehat{V}_t^{\widehat{\pi}^\star}||_\infty \leq \epsilon_{opt},\ \forall t = 1, ..., H\}$*. Then there exists constant* c_1, c_2 *such that for any* $0 < \delta < 1$*, when* $n > c_1 H^2 \log(HSA/\delta)/d_m$*, we have with probability* $1-\delta$*,*

$$\sup_{\pi\in\Pi_1} \left\|\widehat{Q}_1^{\pi} - Q_1^{\pi}\right\|_\infty \leq c_2\sqrt{\frac{H^3\log(HSA/\delta)}{n \cdot d_m}}.$$

This uniform convergence result is presented with l_∞ norm over (s, a). A direct corollary is $\sup_{\pi\in\Pi_1}\left\|\widehat{V}_1^{\pi} - V_1^{\pi}\right\|_\infty$ achieves the same rate. Theorem 3.4 provides the sample complexity of $O(H^3\log(HSA/\delta)/d_m\epsilon^2)$ and the dependence of all parameters are optimal up to the logarithmic term. Note that our bound does not explicitly depend on ϵ_{opt}, which is an improvement over (Agarwal et al., 2020) as they have an additional $O(\epsilon_{\text{opt}}/(1-\gamma))$ error in the infinite horizon setting. Besides, our assumption on ϵ_{opt} is mild since the required upper bound is proportional to $\sqrt{H}$. Lastly, this result implies a $O(\epsilon + \epsilon_{\text{opt}})$-optimal policy for offline/batch learning of the optimal order $O(H^3\log(HSA/\delta)/d_m\epsilon^2)$ (Theorem 4.1), which means statistical learning result enables offline learning.

3.4 Information-theoretical lower bound

Finally, we present a lower bound for the uniform OPE problem. In order to obtain a fine-grained lower bound that depends on d_m, we set up the a family of problems (μ, M) pairs that are parameterized by d_m.

Table 1: A comparison of related policy learning results.

Method/Analysis	Setting	Guarantee	Sample complexity[b]
Agarwal et al. (2020)	Generative model	$\epsilon + O(\epsilon_{\text{opt}}/(1-\gamma))$-optimal	$\widetilde{O}(SA/(1-\gamma)^3\epsilon^2)$
Le et al. (2019); Chen & Jiang (2019)	∞-horizon offline	ϵ-optimal policy	$\widetilde{O}((1-\gamma)^{-6}C_\mu/\epsilon^2)$
Xie & Jiang (2020b)	∞-horizon offline	ϵ-optimal policy	$\widetilde{O}((1-\gamma)^{-4}C_\mu/\epsilon^2)$
SIMPLEX for exact empirical optimal[a]	H-horizon offline	ϵ-optimal policy	$\widetilde{O}(H^3/d_m\epsilon^2)$
PI/VI for ϵ_{opt}-empirical optimal	H-horizon offline	$(\epsilon+\epsilon_{\text{opt}})$-optimal policy	$\widetilde{O}(H^3/d_m\epsilon^2)$
Minimax lower bound (Theorem G.2)	H-horizon offline	over class $\mathcal{M}_{d_m}$	$\Omega(H^3/d_m\epsilon^2)$

[a] PI/VI or SIMPLEX is not essential and can be replaced by any efficient empirical MDP solver.
[b] *Episode* complexity in H-horizon setting is comparable to *step* complexity in ∞-horizon setting because our finite-horizon MDP is *time-inhomogeneous*. Informally, we can just take $(1-\gamma)^{-1} \asymp H$ and $C_\mu \asymp 1/d_m$.

Theorem 3.5 (Minimax lower bound for uniform OPE). *For all $0 < d_m \leq \frac{1}{SA}$, let the class of problems be $\mathcal{M}_{d_m} := \big\{(\mu, M) \ \big| \ \min_{t,s_t,a_t} d_t^\mu(s_t,a_t) \geq d_m\big\}$. There exists universal constants c_1, c_2, c_3, p (with $H, S, A \geq c_1$ and $0 < \epsilon < c_2$) such that*

$$\inf_{\widehat{v}} \sup_{(\mu,M)\in\mathcal{M}_{d_m}} \mathbb{P}_{\mu,M}\left(\sup_{\pi\in\Pi} |\widehat{v}^\pi - v^\pi| \geq \epsilon\right) \geq p$$

if $n \leq c_3 H^3/d_m\epsilon^2$. Here Π consists of all deterministic policies.

On optimality. Above result provides the minimax lower bound of complexity $\Omega(H^3/d_m\epsilon^2)$. As a comparison, Theorem 3.3 gives $\widetilde{O}(H^3S/d_m\epsilon^2)$ is one factor away from the lower bound and Theorem 3.4 has the same rate of the lower bound up to logarithmic factor.

4 Main results for offline learning

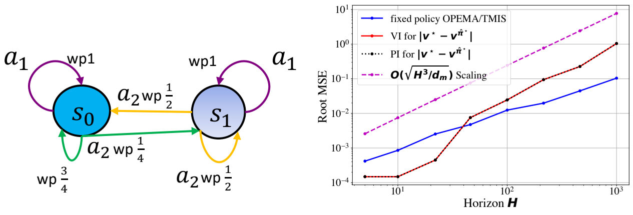
In this section we discuss the implication of our results on offline learning. As we discussed earlier in the introduction, a uniform OPE bound of ϵ implies that the corresponding ERM algorithm finds a 2ϵ-suboptimal policy. But it also implies that all other offline policy-learning algorithms that are not ERM, we could gracefully decompose their error into optimization error and statistical (generalization) error.

Theorem 4.1. *Let $\hat{\pi}^* = \operatorname{argmax}_\pi \hat{v}^\pi$ — the empirically optimal policy. Let $\hat{\pi}$ be any data-dependent choice of policy such that $\hat{v}^{\hat{\pi}^*} - \hat{v}^{\hat{\pi}} \leq \epsilon_{opt}$, then. There is a universal constant c such that w.p. $\geq 1-\delta$*

1. $v^{\pi^*} - v^{\hat{\pi}} \leq c\sqrt{\frac{H^4 S\log(HSA/\delta)}{d_m \cdot n}} + \epsilon_{opt}$.
2. *If $\delta < e^{-S}$, the bound improves to* $c\sqrt{\frac{H^4 S\log(HSA/\delta)}{d_m \cdot n}} + \epsilon_{opt}$. *And if in addition $\hat{\pi}$ is deterministic, the bound further improves to* $c\sqrt{\frac{H^3\min\{H,S\}\log(HSA/\delta)}{d_m \cdot n}} + \epsilon_{opt}$.
3. *If $\epsilon_{opt} \leq \sqrt{H}/S$ and that $||\widehat{V}_t^{\hat{\pi}} - \widehat{V}_t^{\hat{\pi}^*}||_\infty \leq \epsilon_{opt}$, $\forall t = 1, ..., H$, then* $v^{\pi^*} - v^{\hat{\pi}} \leq c\sqrt{\frac{H^3\log(HSA/\delta)}{d_m \cdot n}} + \epsilon_{opt}$.

The third statement implies that all sufficiently accurate planning algorithms based on the empirically estimated MDP are optimal. For example, we can run value iteration or policy iteration to the point that $\epsilon_{\text{opt}} \leq O(H^3/nd_m)$.

Comparing to existing work. Previously no algorithm is known to achieve the optimal sample complexity in the offline setting. Our result also applies to the related generative model setting by replacing $1/d_m$ with SA, which avoids the data-splitting procedure usually encountered by specific algorithm design (e.g., Sidford et al., 2018). The analogous policy-learning results In the generative model setting (Agarwal et al., 2020, Theorem 1) , achieves a suboptimality of $\tilde{O}((1-\gamma)^{-3}SA/n + (1-\gamma)^{-1}\epsilon_{\text{opt}})$ with no additional assumption on ϵ_{opt}. Informally, if we replace $(1-\gamma)^{-1}$ with H, then our result improves the bound from $H\epsilon_{\text{opt}}$ to just ϵ_{opt} for $\epsilon_{\text{opt}} \leq \sqrt{H}/S$. These results are summarized in Table 1.



(a) A non-stationary MDP (b) RMSE vs. Horizon H

Figure 1: Log-log plot showing the dependence on horizon of uniform OPE and pointwise OPE via learning ($|v^\star - v^{\widehat{\pi}^\star}|$) over a non-stationary MDP example.

5 Numerical simulation

In this section we use a simple simulated environment to empirically demonstrate the correct scaling in H. Direct evaluating $\sup_{\pi\in\Pi} |\widehat{v}^\pi - v^\pi|$ empirically is computationally infeasible since the policy classes we considered here contains either A^{HS} or ∞ many policies. Instead, in the experiment we will plot the sub-optimality gap $|v^\star - v^{\widehat{\pi}^\star}|$ with $\widehat{\pi}^\star$ being the outputs of policy planning algorithms. The sub-optimality gap is considered as a surrogate for the lower bound of $\sup_{\pi\in\Pi} |\widehat{v}^\pi - v^\pi|$. Details for simulations are explained in the Appendix.

Figure 1(b) use a fixed number of episodes $n = 2048$ while varying H to examine the horizon dependence for uniform OPE. We can see for fixed pointwise OPE with OPEMA (blue line), $|v^\pi - \widehat{v}^\pi|$ scales as $O(\sqrt{H^2})$ which reflects the bound of Lemma 3.2; for the model-based planning, we ran both VI and PI until they converge to the empirical optimal policy $\widehat{\pi}^\star$. The figure shows that for this MDP example $|v^\star - v^{\widehat{\pi}^\star}|$ scales as $O(\sqrt{H^3/d_m})$ for fixed n since it is parallel to the reference magenta line. This fact empirically shows $O(\sqrt{H^3/d_m})$ bound is required confirms the scaling of our theoretical results.

6 Discussion

Simulation Lemma. Our result can be viewed as a strengthened version of the *simulation lemma* (Kearns & Singh, 2002) (see also the exposition in (Jiang, 2018), which uses similar notations to us). The OPE bound that can be obtained by applying the simulation lemma is

$$|\widehat{v}^\pi - v^\pi| \leq H^2 \sup_{t,s_t,a_t} \left\| \widehat{P}(\cdot|s_t,a_t) - P(\cdot|s_t,a_t) \right\|_1 \leq \widetilde{O}\left(\sqrt{\frac{H^4S^2}{nd_m}}\right)$$

which implies an episode complexity[4] of $\widetilde{O}(H^4S^2/d_m\epsilon^2)$. The main limitation of the simulation lemma is that it does not distinguish between pointwise / uniform convergence (and their bound is in fact a uniform OPE bound), thus will suffer from a loose bound when applied to fixed policies or data-dependent policies that qualify for the smaller policy classes that we considered. For example, our Lemma 3.2 shows that for the same plug-in estimator, the bound improves to $\widetilde{O}(H^2/d_m\epsilon^2)$ for pointwise OPE and Theorem 3.4 shows that we can knock out a factor of HS^2 in the uniform convergence of *near empirically optimal* policies.

7 Conclusion

This work represents the first systematic study of uniform convergence in offline policy evaluation. We derive near optimal results for three representative policy classes. By viewing offline policy evaluation from the uniform convergence perspective, we are able to unify two central topics in offline RL, OPE and offline learning while establishing optimal rates in a subset of these settings including the first rate-optimal offline reinforcement learning method. We hope the work could inspire a more general statistical learning theory for RL in the near future.

[4] See Section J for more calculation details.

References

Agarwal, A., Kakade, S., & Yang, L. F. (2020). Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, (pp. 67–83).

Bertoluzzo, F., & Corazza, M. (2012). Testing different reinforcement learning configurations for financial trading: Introduction and applications. *Procedia Economics and Finance*, *3*, 68–77.

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., & Snelson, E. (2013). Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, *14*(1), 3207–3260.

Brafman, R. I., & Tennenholtz, M. (2002). R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, *3*(Oct), 213–231.

Chen, J., & Jiang, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, (pp. 1042–1051).

Chernoff, H., et al. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, *23*(4), 493–507.

Chung, F., & Lu, L. (2006). Concentration inequalities and martingale inequalities: a survey. *Internet Mathematics*, *3*(1), 79–127.

Dann, C., & Brunskill, E. (2015). Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, (pp. 2818–2826).

Dasari, S., Ebert, F., Tian, S., Nair, S., Bucher, B., Schmeckpeper, K., Singh, S., Levine, S., & Finn, C. (2020). Robonet: Large-scale multi-robot learning. In *Conference on Robot Learning*, (pp. 885–897).

Duan, Y., Jia, Z., & Wang, M. (2020). Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, (pp. 8334–8342).

Dudík, M., Langford, J., & Li, L. (2011). Doubly robust policy evaluation and learning. In *International Conference on Machine Learning*, (pp. 1097–1104).

Ernst, D., Stan, G.-B., Goncalves, J., & Wehenkel, L. (2006). Clinical data based optimal sti strategies for hiv: a reinforcement learning approach. In *Decision and Control, 2006 45th IEEE Conference on*, (pp. 667–672). IEEE.

Farajtabar, M., Chow, Y., & Ghavamzadeh, M. (2018). More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, (pp. 1447–1456).

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., & Celi, L. A. (2019). Guidelines for reinforcement learning in healthcare. *Nat Med*, *25*(1), 16–18.

Jaques, N., Ghandeharioun, A., Shen, J. H., Ferguson, C., Lapedriza, A., Jones, N., Gu, S., & Picard, R. (2019). Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*.

Jiang, N. (2018). Notes on tabular methods.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., & Schapire, R. E. (2017). Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning-Volume 70*, (pp. 1704–1713).

Jiang, N., & Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, (pp. 652–661). JMLR. org.

Jin, C., Allen-Zhu, Z., Bubeck, S., & Jordan, M. I. (2018). Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, (pp. 4863–4873).

Jin, C., Yang, Z., Wang, Z., & Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, (pp. 2137–2143).

Kallus, N., & Uehara, M. (2019). Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes. *arXiv preprint arXiv:1909.05850*.

Kallus, N., & Uehara, M. (2020). Double reinforcement learning for efficient off-policy evaluation in markov decision processes. In *International Conference on Machine Learning*, (pp. 1922–1931).

Kearns, M., & Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Machine learning*, *49*(2-3), 209–232.

Krishnamurthy, A., Agarwal, A., & Langford, J. (2016). PAC reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, (pp. 1840–1848).

Lange, S., Gabel, T., & Riedmiller, M. (2012). Batch reinforcement learning. In *Reinforcement learning*, (pp. 45–73). Springer.

Le, H., Voloshin, C., & Yue, Y. (2019). Batch policy learning under constraints. In *International Conference on Machine Learning*, (pp. 3703–3712).

Li, L., Chu, W., Langford, J., & Wang, X. (2011). Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *ACM international conference on Web search and data mining*, (pp. 297–306).

Li, L., Munos, R., & Szepesvari, C. (2015). Toward minimax off-policy value estimation. In *Artificial Intelligence and Statistics*, (pp. 608–616).

Liu, Q., Li, L., Tang, Z., & Zhou, D. (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, (pp. 5361–5371).

Liu, Y., Swaminathan, A., Agarwal, A., & Brunskill, E. (2019). Off-policy policy gradient with state distribution correction. In *Uncertainty in Artificial Intelligence*.

Liu, Y., Swaminathan, A., Agarwal, A., & Brunskill, E. (2020). Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*.

Munos, R. (2003). Error bounds for approximate policy iteration. In *International Conference on Machine Learning*, (pp. 560–567).

Quillen, D., Jang, E., Nachum, O., Finn, C., Ibarz, J., & Levine, S. (2018). Deep reinforcement learning for vision-based robotic grasping: A simulated comparative evaluation of off-policy methods. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, (pp. 6284–6291). IEEE.

Raghu, A., Gottesman, O., Liu, Y., Komorowski, M., Faisal, A., Doshi-Velez, F., & Brunskill, E. (2018). Behaviour policy estimation in off-policy policy evaluation: Calibration matters. *arXiv preprint arXiv:1807.01066*.

Raghu, A., Komorowski, M., Celi, L. A., Szolovits, P., & Ghassemi, M. (2017). Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Machine Learning for Healthcare Conference*, (pp. 147–163).

Shalev-Shwartz, S., Shamir, O., Srebro, N., & Sridharan, K. (2010). Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, *11*, 2635–2670.

Sidford, A., Wang, M., Wu, X., Yang, L., & Ye, Y. (2018). Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, (pp. 5186–5196).

Sridharan, K. (2002). A gentle introduction to concentration inequalities. *Dept. Comput. Sci., Cornell Univ., Tech. Rep*.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tang, L., Rosales, R., Singh, A., & Agarwal, D. (2013). Automatic ad format selection via contextual bandits. In *ACM international conference on information & knowledge management*, (pp. 1587–1594).

Tewari, A. (2007). *Reinforcement learning in large or unknown MDPs*. University of California, Berkeley.

Thomas, P., & Brunskill, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, (pp. 2139–2148).

Thomas, P. S. (2015). *Safe reinforcement learning*. Ph.D. thesis, University of Massachusetts Amherst.

Thomas, P. S., Theocharous, G., Ghavamzadeh, M., Durugkar, I., & Brunskill, E. (2017). Predictive off-policy policy evaluation for nonstationary decision problems, with applications to digital marketing. In *Twenty-Ninth IAAI Conference*.

Tropp, J., et al. (2011). Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, *16*, 262–270.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Xie, T., & Jiang, N. (2020a). Batch value-function approximation with only realizability. *arXiv preprint arXiv:2008.04990*.

Xie, T., & Jiang, N. (2020b). Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Uncertainty in Artificial Intelligence*, (pp. 550–559).

Xie, T., Ma, Y., & Wang, Y.-X. (2019). Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems*, (pp. 9668–9678).

Ye, Y. (2011). The simplex and policy-iteration methods are strongly polynomial for the markov decision problem with a fixed discount rate. *Mathematics of Operations Research*, *36*(4), 593–603.

Yin, M., & Wang, Y.-X. (2020). Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *Artificial Intelligence and Statistics*, (pp. 3948–3958).