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

Ming Yin
Department of Statistics
Department of Computer Science
UC Santa Barbara
ming_yin@ucsb.edu

Yu Bai
Salesforce Research
yu.bai@salesforce.com

Yu-Xiang Wang
Department of Computer Science
UC Santa Barbara
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 and 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 and 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 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,

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

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 $\pi^\star := \text{argmax}_\pi v^\pi$ 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* $\widehat{\pi}^\star := \text{argmax}_\pi \widehat{v}^\pi$ 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

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

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 $s_{s,a}^{(1)}, ..., s_{s,a}^{(m)}$ 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 $\widetilde{O}(H^4 S/d_m \epsilon^2)$ (Theorem 3.3) 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 $\widetilde{O}(H^3 S/d_m \epsilon^2)$ with an optimal dependence on H (Theorem 3.5).

- 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 3.7).

- We prove a information-theoretical lower bound of $\Omega(H^3/d_m \epsilon^2)$ for OPE (Theorem 3.8) 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 and Brunskill, 2016) or their doubly robust variants (Jiang and 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 and Uehara, 2020, 2019; Yin and 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 and Wang (2020) design the Tabular-MIS estimator which matches the Cramer-Rao lower bound constructed by Jiang and 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 and Wang, 2020), which has a result analogous to Theorem 3.7, 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 and 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 and 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 and 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 and 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., 2020b), 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 $\epsilon_{\text{opt}} \le \sqrt{(1-\gamma)^{-5}SA/n}$. 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.

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-

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

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 and 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: $\Pi_1 := \{\pi : s.t.\ ||\widehat{V}_t^\pi - \widehat{V}_t^{\widehat{\pi}^\star}||_\infty \le \epsilon_{\text{opt}},\ \forall t = 1, ..., H\}$ with $\epsilon_{\text{opt}} \ge 0$ 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.

[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 $i = 1, ..., n$, $0 \le r_t^{(i)} \le 1$.*

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:

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

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:

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

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

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

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

In this section, we present our results for uniform OPE problems from Section 2.1. For brevity, we use $\widehat{v}^\pi$ to denote $\widehat{v}^\pi_{\text{OPEMA}}$ in the rest of paper. Proofs of all technical results are deferred to the appendix. We start with the following Lemma:

Lemma 3.1 (martingale decomposition)**.** *For fixed π:*

$$\sum_{t=1}^{H}\langle \widehat{d}^\pi_t - d^\pi_t, r_t\rangle = \sum_{h=2}^{H}\langle V^\pi_h, (\widehat{T}_h - T_h)\widehat{d}^\pi_{h-1}\rangle + \langle V^\pi_1, \widehat{d}^\pi_1 - d^\pi_1\rangle,$$

where $T_{h+1} \in \mathbb{R}^{S\times(SA)}$ be the one step transition matrix, i.e. $T_{s_{h+1},(s_h,a_h)} = P_{h+1}(s_{h+1}|s_h,a_h)$. the inner product on the left hand side is taken w.r.t state-action and the inner product on the left hand side is taken w.r.t state only. Proof can be found in appendix (Theorem C.5).

Remark 3.2. *Note when the reward is deterministic, the left hand side is simply $\widehat{v}^\pi - v^\pi$, and the right hand side has a martingale structure which enables the applicability of concentration analysis that gives rise to the following theorems. Moreover, this decomposition is essentially "primal-dual" formulation since the LHS can be viewed as the primal form through marginal distribution representation and RHS is the dual form with value function representation.*

3.1 Uniform OPE For Global Policy Class

We present the following result Theorem 3.3 for global policy class.

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

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

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

The first term in the bound reflects the concentration of $\sup_{\pi\in\Pi}|\widehat{v}^\pi - v^\pi|$ around its mean, via McDiarmid inequality. The second term is a bound of $\mathbb{E}[\sup_{\pi\in\Pi}|\widehat{v}^\pi - v^\pi|]$. The analysis of both terms rely on the Martingale decomposition from Lemma 3.1.

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 3.8). 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

The Martingale decomposition also allows us to derive a high-probability OPE bound via a concentration argument, which complements the optimal bounds on mean square error from (Yin and Wang, 2020).

Lemma 3.4 (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:*

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

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 3.5. *Let Π consist 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:*

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

Theorem 3.5 implies an episode complexity of $\widetilde{O}(H^3S/d_m\epsilon^2)$, 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.

Remark 3.6. *The similar high-probability OPE bound in Lemma 3.4 was proven before by (Yin and Wang, 2020) through the data-splitting type estimator. However their theory does not imply efficient offline learning, see Section K in appendix for discussion.*

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 3.7. *Suppose* $\epsilon_{opt} \le \sqrt{H}/S$ *and* $\Pi_1 := \{\pi : s.t. \ ||\widehat{V}_t^{\pi} - \widehat{V}_t^{\widehat{\pi}^\star}||_\infty \le \epsilon_{opt}, \ \forall t = 1, ..., H\}$. *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* $1-\delta$,

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

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 3.7 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., 2020b) 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 fine-grained sample complexity lower bound of the uniform OPE problem that captures the dependence of all parameters including d_m.

Theorem 3.8 (Minimax lower bound for uniform OPE)**.** *For all* $0 < d_m \le \frac{1}{SA}$. *Let the class of problems be*

$$\mathcal{M}_{d_m} := \left\{ (\mu, M) \ \middle| \ \min_{t,s_t,a_t} d_t^{\mu}(s_t, a_t) \ge d_m \right\}.$$

There exist universal constants c_1, c_2, c_3, p *(with* $H, S, A \ge c_1$ *and* $0 < \epsilon < c_2$*) such that*

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

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

The proof uses a reduction argument that shows if a stronger uniform OPE bound exists, then it implies an algorithm that breaks an offline learning lower bound (Theorem G.2), which itself is proven by embedding many stochastic multi-armed bandits problems in a family of hard MDPs. Our construction is inspired by the MDPs in (Jiang et al., 2017) and a personal communication with Christopher Dann but involve substantial modifications to account for the differences in the assumption about rewards. The part in which we obtain explicit dependence on d_m is new and it certifies that the offline learning (and thus uniform OPE) problem strictly more difficult than their online counterpart.

On Optimality. The above result provides the minimax lower bound of complexity $\Omega(H^3/d_m \epsilon^2)$. As a comparison, Theorem 3.5 gives $\widetilde{O}(H^3 S/d_m \epsilon^2)$ is a factor of S away from the lower bound and Theorem 3.7 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* $\hat{\pi}^* = \text{argmax}_\pi \hat{v}^\pi$ — *the empirically optimal policy. Let* $\hat{\pi}$ *be any data-dependent choice of policy such that* $\hat{v}^{\hat{\pi}^*} - \hat{v}^{\hat{\pi}} \le \epsilon_{opt}$, *then. There is a universal constant* c *such that w.p.* $\ge 1-\delta$

1. $v^{\pi^*} - v^{\hat{\pi}} \le c\sqrt{\frac{H^4 S \log(HSA/\delta)}{d_m \cdot n}} + \epsilon_{opt}$.
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* $\epsilon_{opt} \le \sqrt{H}/S$ *and that* $||\widehat{V}_t^{\hat{\pi}} - \widehat{V}_t^{\widehat{\pi}^\star}||_\infty \le$

ϵ_{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., 2020b, 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.

Sparse MDP Estimate. We highlight that the result does not require the estimated MDP to be an accurate approximation in any sense. Recall that the true MDP has $O(S^2)$ parameters (ignoring the dependence on H, A and logarithmic terms), but our result is valid provided that $n = \tilde{\Omega}(1/d_m)$ which is $\Omega(S)$. This suggests that we may not even exhaustively visit all pairs to state-transitions and that the estimator of $\hat{P}_t$ is allowed to be zero in many coordinates.

Optimal Computational Complexity. Lastly, from the computational perspective, we can leverage the best existing solutions for solving optimization $\widehat{\pi}^\star := \text{argmax}_{\pi\in\Pi}\widehat{v}^\pi$. For example, with $\epsilon_{\text{opt}} > 0$, as explained by Agarwal et al. (2020b), value iteration ends in $O(H \log \epsilon_{\text{opt}}^{-1})$ iteration and takes at most $O(HSA)$ time after the model has been estimated with one pass of the data ($O(nH)$ time). We have a total computational complexity of $O(H^4/(d_m\epsilon^2) + H^2SA\log(1/\epsilon))$ time algorithm for obtaining the ϵ-suboptimal policy using $n = O(H^4/(d_m\epsilon^2)$ episodes. This is essentially optimal because the leading term H^4SA/ϵ^2 is required even to just process the data needed for the result to be information-theoretically possible. In comparison, the algorithm that obtains an exact empirical optimal policy $\widehat{\pi}^\star$, the SIMPLEX policy iteration runs in time $O(\text{poly}(H, S, A, n))$ (Ye, 2011).

5 PROOF OVERVIEW

Our uniform convergence analysis in Section 3.1, relies on creating an unbiased version of $\widehat{v}_{\text{OPEMA}}$ (which we call it $\widetilde{v}_{\text{OPEMA}}$) artificially and use concentration (Lemma C.1) to guarantee $\widehat{v}_{\text{OPEMA}}$ is identical to $\widetilde{v}_{\text{OPEMA}}$ in most situations. By doing so we can reduce our analysis from $\sup_{\pi\in\Pi}|\widehat{v}^\pi - v^\pi|$ to $\sup_{\pi\in\Pi}|\widetilde{v}^\pi - v^\pi|$. Specifically, $\widetilde{v}^\pi$ replaces $\widehat{P}_t$, $\widehat{r}_t$ in $\widehat{v}^\pi$ by its fictitious counterparts $\widetilde{P}_t$, $\widetilde{r}_t$, defined as:

$$\widetilde{r}_t(s_t, a_t) = \widehat{r}_t(s_t, a_t)\mathbf{1}(E_t) + r_t(s_t, a_t)\mathbf{1}(E_t^c),$$
$$\widetilde{P}_{t+1}(\cdot|s_t, a_t) = \widehat{P}_{t+1}(\cdot|s_t, a_t)\mathbf{1}(E_t) + P_{t+1}(\cdot|s_t, a_t)\mathbf{1}(E_t^c).$$

where E_t denotes the event $\{n_{s_t,a_t} \geq nd_t^\mu(s_t, a_t)/2\}$. This is saying, if observation n_{s_t,a_t} is large enough (E_t is true), we use $\widehat{P}$; otherwise we directly use P instead. This track helps dealing with out-of-sample state-action pairs. The next key is the martingale decomposition (Lemma 3.1). On one hand, by using the structure of $\sup_{\pi\in\Pi}\langle V_h^\pi, (\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi\rangle$ we can relax it into a "Rademacher-type complexity" which corresponds to $\tilde{O}(\sqrt{H^4S/d_m n})$ term in Theorem 3.3. On the other hand, this decomposition has a natural martingale structure so martingale concentration inequalities can be appropriately applied, *i.e.* Theorem 3.4. In addition, each term $\langle V_h^\pi, (\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi\rangle$ separates the non-stationary policy into two parts with empirical distribution only depends on $\pi_{1:h-1}$ that governs how the data "roll in" and the long term value function V_h^π only depends on $\pi_{h:H}$ that governs how the reward "roll out".

For local uniform convergence, by Bellman equations we can obtain a similar decomposition on Q-function:

$$\widehat{Q}_t^\pi - Q_t^\pi = \sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^\pi(\widehat{P}_h - P_h)\widehat{V}_h^\pi,$$

where $\Gamma_{t:h}^\pi = \prod_{i=t}^h P_i^\pi$ is the multi-step state-action transition and $\Gamma_{t+1:t}^\pi := I$. Since π is any policy in Π_1 which may dependent on $\mathcal{D}'$ so we cannot directly apply concentration inequalities on $(\widehat{P}_h - P_h)\widehat{V}_h^\pi$. Instead, we overcome this hurdle by doing concentration on $(\widehat{P}_h - P_h)\widehat{V}_h^{\widehat{\pi}^*}$ since $\widehat{V}_h^{\widehat{\pi}^*}$ and $\widehat{P}_h$ are independent, and we connect $\widehat{V}_h^{\widehat{\pi}^*}$ back to $\widehat{V}_h^\pi$ by using they are ϵ_{opt} close (Theorem 3.7). This idea helps avoiding the technicality of absorbing MDP used in Agarwal et al. (2020b) for infinite horizon case because of our non-stationary transition setting. For the uniform convergence lower bound, our analysis relies on reducing the problem to identifying ϵ-optimal policy and proving any algorithm that learns a ϵ-optimal policy requires at least $\Omega(H^3/d_m\epsilon^2)$ episodes in the non-stationary episodic setting. Previously, Jiang et al. (2017) proves the $\Omega(HSA/\epsilon^2)$ lower bound with assumption $\sum_{i=1}^H r_i \leq 1$. Our proof uses a modified version of their hard-to-learn MDP instance to achieve the desired result. To produce extra H^2 dependence, we leverage the Assumption 2.1 that $\sum_{i=1}^H r_i$ may be of order $O(H)$. We only present the high-level

Table 1: A comparison of related offline policy learning results. Results shown in this color are new to this paper.

Method/Analysis	Setting	Guarantee	Sample complexity[b]
Agarwal et al. (2020b)	Generative model	$\epsilon + O(\epsilon_{\text{opt}}/(1-\gamma))$-optimal	$\widetilde{O}(SA/(1-\gamma)^3\epsilon^2)$
Le et al. (2019); Chen and Jiang (2019)	∞-horizon offline	ϵ-optimal policy	$\widetilde{O}((1-\gamma)^{-6}C_\mu/\epsilon^2)$
Xie and 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 $C_\mu \asymp 1/d_m$.

ideas here due the space constraint, detailed proofs are explicated in order in Appendix D, E, F, G.

6 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 $\hat{\pi}^\star$ being the outputs of policy planning algorithms. The sub-optimality gap is considered as a surrogate for the lower bound of $\sup_{\pi\in\Pi}|\widehat{v}^\pi - v^\pi|$. Concretely, the non-stationary MDP has 2 states s_0, s_1 and 2 actions a_1, a_2 where action a_1 has probability 1 going back the current state and for action a_2, there is one state s.t. after choosing a_2 the dynamic transitions to both states with equal probability $\frac{1}{2}$ and the other one has asymmetric probability assignment ($\frac{1}{4}$ and $\frac{3}{4}$). The transition after choosing a_2 is changing over different time steps therefore the MDP is non-stationary and the change is decided by a sequence of pseudo-random numbers (Figure 1(c) shows the transition kernel at a particular time step). Moreover, to make the learning problem non-trivial we use non-stationary rewards with 4 categories, *i.e.* $r_t(s,a) \in \{\frac{1}{4}, \frac{2}{4}, \frac{3}{4}, 1\}$ and assignment of $r_t(s,a)$ for each value is changing over time (see Section I in appendix for more details). Lastly, the logging policy in Figure 1(a) is uniform with $\mu_t(a_1|s) = \mu_t(a_2|s) = \frac{1}{2}$ for both states.

Figure 1(a) 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 3.4; 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.

Figure 1(b) show the comparison between $|\widehat{v}^{\widehat{\pi}^\star} - v^{\widehat{\pi}^\star}|$ and $|\widehat{v}^\pi - v^\pi|$ for some fixed policy π. The trend of $|\widehat{v}^{\widehat{\pi}^\star} - v^{\widehat{\pi}^\star}|$ shares a similar pattern as $|v^\star - v^{\widehat{\pi}^\star}|$ in Figure 1(a). More detailed discussions can be found in Section I in appendix.

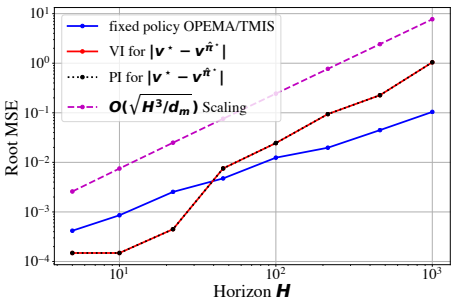
7 DISCUSSION

The Efficiency Of Model-based Methods. There had been a long-lasting debate about model-based vs model-free methods in RL. The model-based methods were considered inefficient in both space and sample complexity, due to the need to represents the transition kernel in $O(HS^2A)$. Most sample-efficient methods with the right dependence in S are model-free methods that directly represents and updates the Q-function. Our analysis reveals that direct model-based plug-in estimator is optimal in both pointwise and uniform prediction problems, which helps to correct the commonly held misunderstanding that purely model plug-in estimator is loose due to simulation lemma.

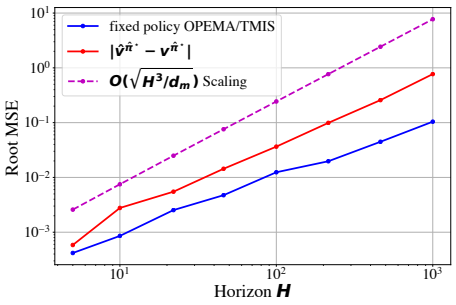
Uniform OPE Depends On π. In this paper, we primarily consider obtaining uniform bound independent to π, however, given a logging policy μ, it is often easier to evaluate some policies than others, as is revealed in the pointwise OPE bound of (Yin and Wang, 2020). Specifically, obtaining a high probability bound of the form $\sup_\pi \frac{\sqrt{n}|\hat{v}^\pi - v^\pi|}{\gamma(\pi,\mu,M,\delta)} \le C$ for some function γ and constant C would be of great interest. We could already get such a bound by applying union bound to the data-dependent high probability pointwise convergence of either (Yin and Wang, 2020) or (Duan et al., 2020) but it comes with an additional $O(S)$ factor. Characterizing the optimal per-instance OPE bound is an interesting future direction.

Simulation Lemma.[4] Our result can be viewed as a

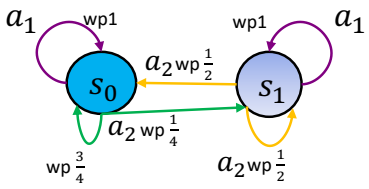
[4] There are different folklores for Simulation Lemma Agarwal et al. (2020a). We focus on the analyzing perspective, see Jiang (2018).



(a) RMSE vs. Horizon H



(b) RMSE vs. Horizon H



(c) A non-stationary MDP

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.

strengthened version of the *simulation lemma* (Kearns and 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

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

which implies an episode complexity[5] of $\widetilde{O}(H^4 S^2/d_m\epsilon^2)$. 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 3.4 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 3.7 shows that we can knock out a factor of HS^2 in the uniform convergence of *near empirically optimal* policies. Finally, there is a factor of Simprovement in the global policy class unconditionally. These savings can be used as drop-in replacements to many instances where the simulation lemma is applied to improve the parameters of the analysis therein.

[5]See Section J for more calculation details.

8 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. The work focuses on the episodic tabular MDP with nonstationary transitions. Carrying out the same analysis for the stationary transition case, infinite horizon case, as well as the linear MDP setting is highly tractable with the techniques presented. Formalizing these is left as a future work. More generally, a natural complexity measure for the policy class of RL remains elusive. We hope the work could inspire a more general statistical learning theory for RL in the near future.

Acknowledgments

The research was partially supported by NSF Awards #2007117 and #1934641. The authors thank Christopher Dann for a discussion related to sample complexity lower bounds in the pointwise bounded reward case; and Tengyang Xie and Nan Jiang for clarifying the scaling in the sample complexity of their results in (Xie and Jiang, 2020b) with us.

References

Agarwal, A., Jiang, N., Kakade, S., and Sun, W. (2020a). *Reinforcement Learning: Theory and Algorithms*.

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

Bertoluzzo, F. and 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., and 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. and 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. and Jiang, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 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. and Lu, L. (2006). Concentration inequalities and martingale inequalities: a survey. *Internet Mathematics*, 3(1):79–127.

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

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

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

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

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

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

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., and 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., and 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., and Schapire, R. E. (2017). Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning-Volume 70*, pages 1704–1713.

Jiang, N. and 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*, pages 652–661. JMLR. org.

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

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

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

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

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

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

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

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

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

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

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

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

Liu, Y., Swaminathan, A., Agarwal, A., and 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*, pages 560–567.

Quillen, D., Jang, E., Nachum, O., Finn, C., Ibarz, J., and 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)*, pages 6284–6291. IEEE.

Raghu, A., Gottesman, O., Liu, Y., Komorowski, M., Faisal, A., Doshi-Velez, F., and 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., and Ghassemi, M. (2017). Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Machine Learning for Healthcare Conference*, pages 147–163.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and 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., and 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*, pages 5186–5196.

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

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

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

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

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

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

Thomas, P. S., Theocharous, G., Ghavamzadeh, M., Durugkar, I., and 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. and Jiang, N. (2020a). Batch value-function approximation with only realizability. *arXiv preprint arXiv:2008.04990*.

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

Xie, T., Ma, Y., and Wang, Y.-X. (2019). Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems*, pages 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. and Wang, Y.-X. (2020). Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *Artificial Intelligence and Statistics*, pages 3948–3958.