Optimal Uniform OPE and Model-based Offline Reinforcement Learning in Time-Homogeneous, Reward-Free and Task-Agnostic Settings

Ming Yin [1 2] **Yu-Xiang Wang** [1]

Abstract

This work studies the statistical limits of uniform convergence for offline policy evaluation (OPE) problems with model-based methods (for episodic MDP) and provides a unified framework towards optimal learning for several well-motivated offline tasks. We establish an $\Omega(H^2S/d_m\epsilon^2)$ lower bound (over model-based family) for the global uniform OPE and our main result establishes an upper bound of $\tilde{O}(H^2/d_m\epsilon^2)$ for the *local* uniform convergence. The highlight in achieving the optimal rate $\tilde{O}(H^2/d_m\epsilon^2)$ is our design of *singleton absorbing MDP*, which is a new sharp analysis tool that works with the model-based approach. We generalize such a model-based framework to the new settings: offline task-agnostic and the offline reward-free with optimal complexity $\tilde{O}(H^2\log(K)/d_m\epsilon^2)$ (K is the number of tasks) and $\tilde{O}(H^2S/d_m\epsilon^2)$ respectively. These results provide a unified solution for simultaneously solving different offline RL problems.

1. Introduction

Offline reinforcement learning is widely applicable in applications where online exploration is demanding but historical data are plentiful. Examples includes medicine (Liu et al., 2017) (safety concerns limit the applicability of unproven treatments but electronic records are abundant) and autonomous driving (Codevilla et al., 2018) (building infrastructure for testing new policy is expensive while collecting data from current setting is almost free).

Yin et al. (2021a) initiates the studies for offline RL from the new perspective of *uniform convergence* in OPE (uniform OPE for short) which unifies OPE and offline learning tasks. Generally speaking, given a policy class Π and offline data with n episodes, uniform OPE seeks to coming up with OPE estimators $\widehat{V}_1^\pi$ and $\widehat{Q}_1^\pi$ satisfy $\sup_{\pi\in\Pi}||\widehat{Q}_1^\pi - Q_1^\pi||_\infty < \epsilon$. The task is to achieve this with the optimal episode complexity: the "minimal" number of episodes n needed as a function of ϵ, failure probability δ, the parameters of the MDP as well as the behavior policy μ in the minimax sense.

Uniform OPE to RL is analogous of uniform convergence of empirical risk in statistical learning (Vapnik, 2013). In supervised learning, it has been proven that almost all learnable problems are learned by an (asymptotic) *empirical risk minimizer* (ERM) (Shalev-Shwartz et al., 2010). In offline RL, the natural counterpart is the *empirical optimal policy* $\widehat{\pi}^\star := \operatorname{argmax}_\pi \widehat{V}_1^\pi$ and with uniform OPE it further ensures $\widehat{\pi}^\star$ is a near-optimal policy for the offline learning via:

$$0 \le Q_1^{\pi^\star} - Q_1^{\widehat{\pi}^\star} = Q_1^{\pi^\star} - \widehat{Q}_1^{\pi^\star} + \widehat{Q}_1^{\pi^\star} - \widehat{Q}_1^{\widehat{\pi}^\star} + \widehat{Q}_1^{\widehat{\pi}^\star} - Q_1^{\widehat{\pi}^\star} \le 2\sup_\pi |Q_1^\pi - \widehat{Q}_1^\pi|. \tag{1}$$

On the *policy evaluation* side, there is often a need to evaluate the performance of a *data-dependent* policy. Uniform OPE suffices for this purpose since 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. In this paper, we study the uniform OPE problem under the *finite horizon stationary MDPs* and focus on the model-based approaches. Specifically, we consider two representative class: global policy class Π_g (contains all (deterministic) policies) and local policy class Π_l (contains policies near the empirical optimal one, see Section 2.1).

1.1. Our contribution

Optimal local uniform OPE. We derive the $\tilde{O}(H^2/d_m\epsilon^2)$ optimal episode complexity for local uniform OPE (Theorem 4.1) via the model-based method and this implies optimal offline learning with the same rate (Corollary 4.2); this result strictly improves upon (Yin et al., 2021a) ($\tilde{O}(H^3/d_m\epsilon^2)$) non-trivially through our new *singleton-absorbing MDP* technique.

Information-theoretical characterization of the global

[1]Department of Computer Science, University of California, Santa Barbara, USA [2]Department of Statistics and Applied Probability, University of California, Santa Barbara, USA. Correspondence to: Ming Yin <ming_yin@ucsb.edu>.

Reinforcement Learning Theory Workshop at *International Conference on Machine Learning*, 2021. The full version of the paper can be found in https://arxiv.org/abs/2105.06029

uniform OPE. We characterize the statistical limit for the global uniform convergence by proving a minimax lower bound $\Omega(H^2S/d_m\epsilon^2)$ (over all model-based approaches) (Theorem 3.1). This result answers the question left by Yin et al. (2021a) that the global uniform OPE is generically harder than the local uniform OPE / offline learning.

Generalize to the new offline settings. Critically, our model-based frameworks naturally generalize to the more challenging settings like task-agnostic and reward-free settings. In particular, we establish the $\tilde{O}(H^2\log(K)/d_m\epsilon^2)$ (Theorem 5.3) and $\tilde{O}(H^2S/d_m\epsilon^2)$ (Theorem 5.4) complexities for *offline task-agnostic learning* and *offline reward-free learning*. Both results are new and optimal.

Significance: Unifying different offline settings Beyond the study of statistical limit in uniform OPE, this work solves the sample optimality problems for the local uniform OPE, offline task-agnostic and offline reward-free problems. If we take a deeper look, the algorithmic frameworks utilized are all based on the model-based empirical MDP construction and planning. Therefore, as long as we can analyze such framework sharply (*e.g.* via novel absorbing-MDP technique), then it is hopeful that our techniques can be generalized to tackle more sophisticated settings. On the other hand, things could be more tricky for online RL since the exploration phases need to be specifically designed for each settings and there may not be one general algorithmic pattern that dominates. Our findings reveal the model-based framework is fundamental for offline RL as it subsumes settings like local uniform OPE, offline task-agnostic and offline reward-free learning into identical learning pattern.

2. Problem setup

Episodic stationary reinforcement learning. A finite-horizon *Markov Decision Process* (MDP) is denoted by a tuple $M = (\mathcal{S}, \mathcal{A}, P, r, H, d_1)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state action spaces with $S := |\mathcal{S}|, A := |\mathcal{A}|$. A stationary (time-invariant) transition kernel has the form $P : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \mapsto [0,1]$ with $P(s'|s,a)$ representing the probability transition from state s, action a to next state s'. Besides, $r : \mathcal{S}\times A \mapsto \mathbb{R}$ is the expected reward function and given (s,a) which satisfies $0 \le r \le 1$ and assumed known. d_1 is the initial state distribution and H is the horizon. At time t, a policy $\pi = (\pi_1, ..., \pi_H)$ assigns each state $s \in \mathcal{S}$ a probability distribution $\pi_t(s)$ over actions. For a policy π, a random trajectory $s_1, a_1, r_1, \ldots, s_H, a_H, r_H, s_{H+1}$ is generated as follows: $s_1 \sim d_1, a_t \sim \pi(\cdot|s_t), r_t = r(s_t, a_t), s_{t+1} \sim P(\cdot|s_t, a_t), \forall t \in [H]$. In particular, we denote the average marginal state-action occupancy $d^\pi(s,a)$ as: $d^\pi(s,a) := \frac{1}{H}\sum_{t=1}^H \mathbb{P}[s_t = s|s_1 \sim d_1, \pi]\cdot\pi_t(a|s)$.

Offline setting. The offline RL assumes that episodes $\mathcal{D} = \left\{\left(s_t^{(i)}, a_t^{(i)}, r_t^{(i)}, s_{t+1}^{(i)}\right)\right\}_{i\in[n]}^{t\in[H]}$ are rolling from some behavior policy μ a priori. In particular, we do not assume the knowledge of μ.

Model-based RL. To make the presentation precise, we define the following:

Definition 2.1. *Model-based RL: Solving RL problems (either learning or evaluation) through learning / modeling transition dynamic P.*

The model-based approaches in general (*e.g.* Jaksch et al. (2010); Ayoub et al. (2020); Kidambi et al. (2020)) follow the procedure of modeling the full MDP $M = (\mathcal{S}, \mathcal{A}, P, r, H, d_1)$ instead of only the transition P. We (by convention) assume the mean reward function is known and the initial state distribution d_1 will not affect the choice of $\pi^\star$. Thus, Definition 2.1 suffices for our purposes.

2.1. Uniform convergence in offline RL

Recall the goal for uniform OPE is to construct estimator $\widehat{Q}_1^\pi$ such that $\sup_{\pi\in\Pi}\left\|Q_1^\pi - \widehat{Q}_1^\pi\right\| < \epsilon$. We consider two policy classes that are worth considering.

Definition 2.2 (The global (deterministic) policy class.)**.** *The global policy class Π_g consists of all the non-stationary (deterministic) policies.*

It is well-known (Sutton & Barto, 2018) there exists at least one (deterministic) optimal policy, therefore Π_g is sufficiently rich for evaluating algorithms that aim at learning the optimal policy.

Definition 2.3 (The local policy class)**.** *Given empirical MDP $\widehat{M}$ and $\widehat{V}_h^\pi$ is the value under $\widehat{M}$. Let $\widehat{\pi}^\star := \mathrm{argmax}_\pi \widehat{V}_1^\pi$ be the empirical optimal policy, then the local policy class Π_l is defined as: $\Pi_l := \left\{\pi : \text{ s.t. } \left\|\widehat{V}_h^\pi - \widehat{V}_h^{\widehat{\pi}^\star}\right\|_\infty \le \epsilon_{opt}, \forall h \in [H]\right\}$ where $\epsilon_{opt} \ge 0$ is a parameter.*

In above $\widehat{M}$ uses $\widehat{P}$ in lieu of P where $\widehat{P}(s'|s,a) = \frac{n_{s',s,a}}{n_{s,a}}$ if $n_{s,a} > 0$ and $1/S$ otherwise.[1] This class characterizes policies in the neighborhood of empirical optimal policy. Given $\widehat{P}$, it is efficient to obtain $\widehat{\pi}^\star$ using Value / Policy Iteration, therefore it is more practical to consider the neighborhood of $\widehat{\pi}^\star$ (instead of $\pi^\star$) since practitioners can use data $\mathcal{D}$ to really check Π_l whenever needed.

Assumption 2.4 (Exploration requirement)**.** *Logging policy μ obeys that $\min_s d^\mu(s) > 0$, for any state s that is "accessible". Moreover, we define the quantity $d_m := \min_{s,a}\{d^\mu(s,a) : d^\mu(s,a) > 0\}$ (recall $d^\mu(s,a)$ in Section 2) to be the minimal average marginal state-action probability.*

[1] Here $n_{s,a}$ is the number of pair (s,a) being visited among n episodes. $n_{s',s,a}$ is defined similarly.

State s is "accessible" means there exists a policy π so that $d^\pi(s) > 0$. If for any policy π we always have $d^\pi(s) = 0$, then state s can never be visited in the given MDP. Note this is weaker than (Yin et al., 2021a) since $d^\mu(s)$ is the average version of $d_t^\mu(s)$. Assumption 2.4 is the minimal assumption needed for the consistency of uniform OPE task and is qualitatively similar to the *concentrability* assumption.

3. Statistical Hardness for Model-based Global Uniform OPE

From (1) and Definition 2.2, it is clear the global uniform OPE implies offline RL, therefore it is natural to wonder whether they just are *"the same task"* (their sample complexities have the same minimax rates). (Yin et al., 2021a) proves the $\tilde{O}(H^3S/d_m\epsilon^2)$ upper bound and $\Omega(H^3/d_m\epsilon^2)$ lower bound for global uniform OPE, but it is unclear whether the additional S is essential. We answer the question affirmatively by providing a tight lower bound with a concise proof to show no model-based algorithm can surpass $\Omega(H^2S/d_m\epsilon^2)$ information-theoretical limit.

Theorem 3.1 (Minimax lower bound for global uniform OPE)**.** *Let d_m be a parameter such that $0 < d_m \le \frac{1}{SA}$. Let the problem class be $\mathcal{M}_{d_m} := \{(\mu, M) \mid \min_{s,a} d^\mu(s,a) \ge d_m\}$. Then there exists universal constants $c, C, p > 0$ such that: for any $n \ge cS/d_m \cdot \log(SAp)$,*

$$\inf_{\widehat{Q}_{1,mb}} \sup_{\mathcal{M}_{d_m}} \mathbb{P}_{\mu,M}\left(\sup_{\pi\in\Pi_g}\left\|\widehat{Q}^\pi_{1,mb} - Q^\pi_1\right\|_\infty \ge C\sqrt{\frac{H^2S}{nd_m}}\right) \ge p,$$

where $\widehat{Q}_{1,mb}$ is the output of any model-based algorithm and Π_g is defined in Definition 2.2.

By setting $\epsilon := \sqrt{\frac{H^2S}{nd_m}}$, Theorem 3.1 establishes the global uniform convergence lower bound of $\Omega(H^2S/d_m\epsilon^2)$ over the model-based methods, which builds the hard statistical threshold between the global uniform OPE and the local uniform OPE tasks since the local case has achievable $\tilde{O}(H^2/d_m\epsilon^2)$ rate Theorem 4.1. The full proof is in C.

4. Optimal local uniform OPE via model-based plug-in method

Global uniform OPE is intrinsically harder than the offline learning problem due to the additional state-space dependence and such a gap will amplify when S is (exponentially) large. This motivates us to switch to the local uniform convergence regime that enables optimal learning. We design the new *singleton-absorbing MDP* to handle the challenge which avoids the exponential-H cover used in (Cui & Yang, 2020) and answers their conjecture that absorbing MDP is actually well suitable for finite horizon stationary MDP.[2]

[2]See their Section 7, first bullet point for a discussion.

Model-based Offline Plug-in Estimator Let $n_{s,a} := \sum_{i=1}^n\sum_{h=1}^H \mathbf{1}[s_h^{(i)}, a_h^{(i)} = s, a]$ be the total counts that visit (s,a) pair, then the model-based offline plug-in estimator constructs estimator $\widehat{P}$ as: $\widehat{P}(s'|s,a) = \frac{\sum_{i=1}^n\sum_{h=1}^H \mathbf{1}[(s_{h+1}^{(i)}, a_h^{(i)}, s_h^{(i)}) = (s', s, a)]}{n_{s,a}}$, if $n_{s,a} > 0$ and $\widehat{P}(s'|s,a) = \frac{1}{S}$ if $n_{s,a} = 0$. As a consequence, the estimators $\widehat{Q}^\pi_h, \widehat{V}^\pi_h$ are computed as: $\widehat{Q}^\pi_h = r + \widehat{P}^{\pi_{h+1}}\widehat{Q}^\pi_{h+1} = r + \widehat{P}\widehat{V}^\pi_{h+1}$, with the initial distribution $\widehat{d}_1(s) = n_s/n$.

Recall $\widehat{\pi}^\star := \operatorname{argmax}_\pi \widehat{V}^\pi_1$ is the empirical optimal policy and Π_l is in Definition 2.3.

Theorem 4.1 (optimal local uniform OPE)**.** *Let $\epsilon_{opt} \le \sqrt{H/S}$ and denote $\iota = \log(HSA/\delta)$. For any $\delta \in [0,1]$, there exists universal constants c, C such that when $n > cH \cdot \iota/d_m$, w.p. $1-\delta$,*

$$\sup_{\pi\in\Pi_l}\left\|\widehat{Q}^\pi_1 - Q^\pi_1\right\|_\infty \le C\left[\sqrt{\frac{H^2\iota}{nd_m}} + \frac{H^{2.5}S^{0.5}\iota}{nd_m}\right].$$

Theorem 4.1 establishes the $\tilde{O}(H^2/d_m\epsilon^2)$ complexity bound and directly implies the upper bound for $\sup_{\pi\in\Pi_l}||\widehat{V}^\pi_1 - V^\pi_1||_\infty$ with the same rate. This result improves the local uniform convergence rate $\tilde{O}(H^3/d_m\epsilon^2)$ in Yin et al. (2021a) (Theorem 3.7) by a factor of H and is near-minimax optimal (up to the logarithmic factor). Such result is first achieved by our novel *singleton absorbing MDP* technique. Most importantly, Theorem 4.1 guarantees near-minimax optimal offline learning:

Corollary 4.2 (optimal offline learning)**.** *If $\epsilon_{opt} \le \sqrt{H/S}$ and that $\sup_t||\widehat{V}^{\widehat{\pi}}_t - \widehat{V}^{\widehat{\pi}^\star}_t||_\infty \le \epsilon_{opt}$, when $n > O(H \cdot \iota/d_m)$, then with probability $1-\delta$, element-wisely,*

$$V^\star_1 - V^{\widehat{\pi}}_1 \le C\left[\sqrt{\frac{H^2\iota}{nd_m}} + \frac{H^{2.5}S^{0.5}\iota}{nd_m}\right]\mathbf{1} + \epsilon_{opt}\mathbf{1}.$$

Corollary 4.2 first establishes the minimax rate for offline learning for any policy $\widehat{\pi}$ with the measurable gap $\epsilon_{\text{opt}} \le \sqrt{H/S}$. This extends the standard concept of offline learning by allowing any empirical planning algorithm (*e.g.* VI/PI) to find an *inexact* $\widehat{\pi}$ as an $(\tilde{O}\sqrt{H^2/nd_m} + \epsilon_{\text{opt}})$-optimal policy (instead of finding exact $\widehat{\pi}^\star$). The use of *inexact* $\widehat{\pi}$ could encourage early stopping (*e.g.* for VI/PI) therefore saves computational iterations. For the rest of the section, we brief explain the *singleton-absorbing MDP* technique and the full proofs of Theorem 4.1, Corollary 4.2 can be found in Appendix B, D.

4.1. Singleton absorbing MDP for finite horizon MDP

Essentially, the key challenge in obtaining the optimal dependence in stationary setting is the need to decouple the dependence between $P - \widehat{P}$ and $\widehat{V}^\star_h$ as we aggregate all data for constructing both $\widehat{P}$ and $\widehat{V}^\star_h$. This issue is not encountered in the non-stationary setting in general due to the

flexibility to estimate different transition P_t at each time (Yin et al., 2021a) and $\widehat{P}_t$ and $\widehat{V}^\star_{t+1}$ preserve *conditional* independence. However, when confined to stationary case, their complex $\tilde{O}(H^3/d_m\epsilon^2)$ becomes suboptimal. Moreover, the direct use of s-absorbing MDP in (Agarwal et al., 2020) does not yield tight bounds for the finite horizon stationary setting, as it requires s-absorbing MDPs with H-dimensional fine-grid cover to make sure $\widehat{V}^\star_h$ is close to one of the elements in the cover (which has size $\approx H^H$ and it is not optimal (Cui & Yang, 2020)). We overcome this hurdle by choosing *only one* delicate absorbing MDP to approximate $\widehat{V}^\star_h$ which will not incur additional dependence on horizon H caused by the union bound. We begin with the general definition of absorbing MDP.

Standard s-absorbing MDP in the finite horizon setting. The general s-absorbing MDP is defined as follows: for a fixed state s and a sequence $\{u_t\}_{t=1}^H$, MDP $M_{s,\{u_t\}_{t=1}^H}$ is identical to M for all states except s, and state s is absorbing ($P_{M_{s,\{u_t\}_{t=1}^H}}(s|s,a)=1$) for all a, and $r_t(s,a)=u_t$ for all $a, t\in\mathcal{A},[H]$. For convenience, we use $V^\pi_{\{s,u_t\}}$ to denote $V^\pi_{s,M_{s,\{u_t\}_{t=1}^H}}$ and similarly for Q_t, r and transition P. Also, $V^\star_{\{s,u_t\}}$ ($Q^\star_{\{s,u_t\}}$) is the optimal value under $M_{s,\{u_t\}_{t=1}^H}$. Before defining singleton absorbing MDP, we first present the following Lemma 4.3 which supports the our design.

Lemma 4.3. *$V^\star_t(s)-V^\star_{t+1}(s)\geq 0, \forall s\in\mathcal{S}, t\in[H]$. Moreover, fix a state s. If we choose $u^\star_t := V^\star_t(s)-V^\star_{t+1}(s)$, then we have the vector form equations: $V^\star_{h,\{s,u^\star_t\}} = V^\star_{h,M}\quad \forall h\in[H]$. Similarly, if we choose $\hat{u}^\star_t := \widehat{V}^\star_t(s)-\widehat{V}^\star_{t+1}(s)$, then $\widehat{V}^\star_{h,\{s,\hat{u}^\star_t\}}=\widehat{V}^\star_{h,M}$, $\forall h\in[H]$.*

Definition 4.4 (Singleton-absorbing MDP). *For each state s, the singleton-absorbing MDP is chosen to be $M_{s,\{u^\star_t\}_{t=1}^H}$, where $u^\star_t := V^\star_t(s)-V^\star_{t+1}(s)$ for all $t\in[H]$.*

The difference between the standard covering-based absorbing MDP and the singleton absorbing MDP is the former uses a set of MDPs $(V_{1,u_1},\ldots,V_{H,u_H})$ where $(u_1,\ldots,u_H)$ traverse all the H-dimensional grid in $[0,H]^H$ therefore the set has cardinality $O(e^H)$ but the singleton absorbing MDP only uses $(V_{1,u^\star_1},\ldots,V_{H,u^\star_H})$ therefore has the cardinality 1, which can achieve the optimality (Figure 1).

5. New settings: offline Task-agnostic and offline Reward-free learning

From Corollary 4.2, our model-based offline learning algorithm has two steps: 1. constructing offline empirical MDP $\widehat{M}$ using the offline dataset $\mathcal{D}=\{(s^{(i)}_t,a^{(i)}_t,r(s^{(i)}_t,a^{(i)}_t),s^{(i)}_{t+1})\}_{i\in[n]}^{t\in[H]}$; 2. performing any accurate black-box *planning* algorithm and returning $\widehat{\pi}^\star$(or $\widehat{\pi}$) as the final output. However, the only *effective* data (data that contains stochasticity) is $\mathcal{D}'=\{(s^{(i)}_t,a^{(i)}_t)\}_{i\in[n]}^{t\in[H]}$. This indicates we are essentially using the state-action space exploration data $\mathcal{D}'$ to solve the task-specific problem with reward r. With this perspective in mind, it is natural to ask: given only the offline exploration data $\mathcal{D}'$, can we efficiently learn a set of potentially conflicting K tasks (K rewards) simultaneously? Even more, can we efficiently learn all tasks simultaneously? This brings up the following definitions.

Definition 5.1 (Offline Task-agnostic Learning)**.** *Given a offline exploration datatset $\mathcal{D}'=\{(s^{(i)}_t,a^{(i)}_t)\}_{i\in[n]}^{t\in[H]}$ by μ with n episodes. Given K tasks with reward $\{r_k\}_{k=1}^K$ and the corresponding K MDPs $M_k=(\mathcal{S},\mathcal{A},P,r_k,H,d_1)$. Can we use $\mathcal{D}'$ to output $\hat{\pi}_1,\ldots,\hat{\pi}_K$ such that $\mathbb{P}\left[\forall r_k,k\in[K],\left\|V^\star_{1,M_k}-V^{\hat{\pi}_k}_{1,M_k}\right\|_\infty\leq\epsilon\right]\geq 1-\delta$?*

Definition 5.2 (Offline Reward-free Learning)**.** *Given a offline exploration datatset $\mathcal{D}'=\{(s^{(i)}_t,a^{(i)}_t)\}_{i\in[n]}^{t\in[H]}$ by μ with n episodes. For any reward r and the corresponding MDP $M=(\mathcal{S},\mathcal{A},P,r,H,d_1)$. Can we use $\mathcal{D}'$ to output $\hat{\pi}$ such that $\mathbb{P}\left[\forall r,\left\|V^\star_{1,M}-V^{\hat{\pi}}_{1,M}\right\|_\infty\leq\epsilon\right]\geq 1-\delta$?*

Our singleton absorbing MDP technique adapts to those settings and we have the following two theorems. The proofs of Theorem 5.3, 5.4 can be found in Appendix E, F.

Theorem 5.3 (optimal offline task-agnostic learning)**.** *Given $\mathcal{D}'=\{(s^{(i)}_t,a^{(i)}_t)\}_{i\in[n]}^{t\in[H]}$ by μ. Given K tasks with reward $\{r_k\}_{k=1}^K$ and the corresponding K MDPs $M_k=(\mathcal{S},\mathcal{A},P,r_k,H,d_1)$. Denote $\iota=\log(HSA/\delta)$. Let $\widehat{\pi}^\star_k := \operatorname{argmax}_\pi \widehat{V}^\pi_{1,M_k}$ $\forall k\in[K]$, when $n>O(H\cdot[\iota+\log(K)]/d_m)$, then with probability $1-\delta$, $\left\|V^\star_{1,M_k}-V^{\widehat{\pi}^\star_k}_{1,M_k}\right\|_\infty\leq O\left[\sqrt{\frac{H^2(\iota+\log(K))}{nd_m}}+\frac{H^{2.5}S^{0.5}(\iota+\log(K))}{nd_m}\right]\quad\forall k\in[K]$.*

Theorem 5.4 (optimal offline reward-free learning)**.** *Given $\mathcal{D}'=\{(s^{(i)}_t,a^{(i)}_t)\}_{i\in[n]}^{t\in[H]}$ by μ. For any reward r denote the corresponding MDP $M=(\mathcal{S},\mathcal{A},P,r,H,d_1)$. Denote $\iota=\log(HSA/\delta)$. Let $\widehat{\pi}^\star_M := \operatorname{argmax}_\pi \widehat{V}^\pi_{1,M}$ $\forall r$, when $n>O(HS\cdot\iota/d_m)$, then with probability $1-\delta$, $\left\|V^\star_{1,M}-V^{\widehat{\pi}^\star_M}_{1,M}\right\|_\infty\leq O\left[\sqrt{\frac{H^2S\cdot\iota}{nd_m}}+\frac{H^2S\cdot\iota}{nd_m}\right],\ \forall r,M$.*

6. Discussion

By a direct translation of both theorems, we have sample complexity of order $\widetilde{O}(H^2\log(K)/d_m\epsilon^2)$ and $\widetilde{O}(H^2S/d_m\epsilon^2)$. All the parameters have the optimal rates, see the lower bounds in Zhang et al. (2020b) and Jin et al. (2020a). The principle of our *Singleton absorbing MDP* technique (with model-based construction) in decoupling the dependence between $\widehat{P}_{s,a}$ and $\widehat{V}^\star$ is not confined to tabular MDPs and we have further the linear MDP with anchor points example in appendix.

References

Agarwal, A., Jiang, N., and Kakade, S. M. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 2019.

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

Antos, A., Munos, R., and Szepesvari, C. Fitted q-iteration in continuous action-space mdps. In *Advances in Neural Information Processing Systems*, pp. 9–16, 2008a.

Antos, A., Szepesvári, C., and Munos, R. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008b.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 263–272. JMLR. org, 2017.

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

Chernoff, H. et al. 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, 1952.

Codevilla, F., Lopez, A. M., Koltun, V., and Dosovitskiy, A. On offline evaluation of vision-based driving models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 236–251, 2018.

Cui, Q. and Yang, L. F. Is plug-in solver sample-efficient for feature-based reinforcement learning? In *Advances in neural information processing systems*, 2020.

Ding, D., Wei, X., Yang, Z., Wang, Z., and Jovanovic, M. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 3304–3312. PMLR, 2021.

Efroni, Y., Merlis, N., Ghavamzadeh, M., and Mannor, S. Tight regret bounds for model-based reinforcement learning with greedy policies. In *Advances in Neural Information Processing Systems*, 2019.

Han, Y., Jiao, J., and Weissman, T. Minimax estimation of discrete distributions under l_1 loss. *IEEE Transactions on Information Theory*, 61(11):6343–6354, 2015.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

Jiang, N. Notes on tabular methods. 2018.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pp. 4870–4879. PMLR, 2020a.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? *arXiv preprint arXiv:2012.15085*, 2020b.

Kaufmann, E., Ménard, P., Domingues, O. D., Jonsson, A., Leurent, E., and Valko, M. Adaptive reward-free exploration. *arXiv preprint arXiv:2006.06294*, 2020.

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. Morel: Model-based offline reinforcement learning. *arXiv preprint arXiv:2005.05951*, 2020.

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

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 33, 2020.

Liu, Q., Yu, T., Bai, Y., and Jin, C. A sharp analysis of model-based reinforcement learning with self-play. *arXiv preprint arXiv:2010.01604*, 2020a.

Liu, Y., Logan, B., Liu, N., Xu, Z., Tang, J., and Wang, Y. Deep reinforcement learning for dynamic treatment regimes on medical registry data. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 380–385. IEEE, 2017.

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

Mannor, S. and Tsitsiklis, J. N. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004.

Menard, P., Domingues, O. D., Jonsson, A., Kaufmann, E., Leurent, E., and Valko, M. Fast active learning for pure exploration in reinforcement learning. *arXiv preprint arXiv:2007.13442*, 2020.

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

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *arXiv preprint arXiv:2103.12021*, 2021.

Ren, T., Li, J., Dai, B., Du, S. S., and Sanghavi, S. Nearly horizon-free offline reinforcement learning. *arXiv preprint arXiv:2103.14077*, 2021.

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

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

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

Szepesvári, C. and Munos, R. Finite time bounds for sampling based fitted value iteration. In *Proceedings of the 22nd international conference on Machine learning*, pp. 880–887, 2005.

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

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

Wang, R., Du, S. S., Yang, L. F., and Salakhutdinov, R. On reward-free reinforcement learning with linear function approximation. *arXiv preprint arXiv:2006.11274*, 2020a.

Wang, R., Foster, D. P., and Kakade, S. M. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020b.

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

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

Yang, L. and Wang, M. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pp. 6995–7004. PMLR, 2019.

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

Yin, M., Bai, Y., and Wang, Y.-X. Near-optimal offline reinforcement learning via double variance reduction. *arXiv preprint arXiv:2102.01748*, 2021b.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C., and Ma, T. Mopo: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*, 2020.

Zanette, A. Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online rl. *arXiv preprint arXiv:2012.08005*, 2020.

Zhang, K., Kakade, S. M., Başar, T., and Yang, L. F. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *arXiv preprint arXiv:2007.07461*, 2020a.

Zhang, X., Singla, A., et al. Task-agnostic exploration in reinforcement learning. *Advances in Neural Information Processing Systems*, 2020b.

Zhang, Z., Du, S. S., and Ji, X. Nearly minimax optimal reward-free reinforcement learning. *arXiv preprint arXiv:2010.05901*, 2020c.