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

Anonymous Author(s)
Affiliation
Address
email

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. Uniform OPE $\sup_{\Pi} |Q^\pi - \hat{Q}^\pi| < \epsilon$ is a stronger measure than the point-wise OPE and ensures offline learning when Π contains all policies (the global class). In this paper, we establish an $\Omega(H^2 S/d_m \epsilon^2)$ 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 that applies to all *near-empirically optimal* policies for the MDPs with *stationary* transition. Here d_m is the minimal marginal state-action probability. Critically, 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 $\tilde{O}(H^2 S/d_m\epsilon^2)$ respectively. These results provide a unified solution for simultaneously solving different offline RL problems.

1 Introduction

Offline reinforcement learning (offline RL) targets at learning a reward-maximizing policy in an unknown *Markov Decision Process* (MDP) using a static data generated by running a behavior policy (Lange et al., 2012; Levine et al., 2020). This framework is widely applicable in applications where online exploration is demanding but historical data are plentiful. Examples include 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).

Parallel to its practical significance, recently there is a surge of theoretical investigations towards offline RL via two threads: *offline policy evaluation* (OPE), where the goal is to estimate the value of a target (fixed) policy V^π (Jiang & Li, 2016; Liu et al., 2018; Kallus & Uehara, 2020, 2019; Uehara & Jiang, 2019; Nachum et al., 2019; Xie et al., 2019; Yin & Wang, 2020; Duan et al., 2020; Wang et al., 2020b; Zhang et al., 2021) and *offline (policy) learning* which intends to output a near-optimal policy (Chen & Jiang, 2019; Le et al., 2019; Xie & Jiang, 2020a,b; Liu et al., 2020b; Hao et al., 2020; Zanette, 2020; Jin et al., 2020c; Hu et al., 2021; Yin et al., 2021a; Rashidinejad et al., 2021).

Yin et al. (2021b) 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

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

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.

To further motivate the readers why uniform OPE should be considered, we state its relation to offline learning. Indeed, 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 (please refer to Yin et al. (2021b) and the references therein for further discussions). 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). We ask the following question:

What is the statistical limit for uniform OPE and what is its connection to optimal offline learning?

We answer the first part by showing the global uniform OPE requires a lower bound of $\Omega(H^2S/d_m\epsilon^2)$[1] for the family of model-based approach and the local uniform OPE can achieve $\tilde{O}(H^2/d_m\epsilon^2)$ minimax rate by the model-based plug-in estimator and this implies optimal offline learning. Importantly, the procedure of the model-based approach via learning $\widehat{\pi}^\star$ through planning over the empirical MDP has a wider range of use in offline RL as it naturally adapts to the challenging tasks like *offline task-agnostic learning* and *offline reward-free learning*. See Section 1.2.

1.1 Related works

Offline reinforcement learning.[2] Information-theoretical considerations for offline RL are first proposed for *infinite horizon discounted setting* via Fitted Q-Iteration (FQI) type function approximation algorithms (Chen & Jiang, 2019; Le et al., 2019; Xie & Jiang, 2020a,b) which can be traced back to (Munos, 2003; Szepesvári & Munos, 2005; Antos et al., 2008a,b).

For the finite horizon case, Yin et al. (2021b) first achieves $\tilde{O}(H^3/d_m\epsilon^2)$ complexity under non-stationary transition but their results cannot further improve in the stationary setting. Recently, Yin et al. (2021a) designs the offline variance reduction algorithm for achieving the optimal $\tilde{O}(H^2/d_m\epsilon^2)$ rate. Their result is for a specific algorithm that uses data splitting while our results work for any algorithms that returns a nearly empirically optimal policy via uniform convergence. Our results on the offline task-agnostic and the reward-free settings are entirely new. Concurrently, Ren et al. (2021) considers the horizon-free setting but does not provide uniform convergence guarantee.

Model-based approaches with minimaxity. It is known model-based methods are minimax-optimal for online RL with regret $\tilde{O}(\sqrt{HSAT})$ (*e.g.* Azar et al. (2017); Efroni et al. (2019)). In the generative model setting, Agarwal et al. (2020) shows model-based approach is still minimax optimal $\tilde{O}((1-\gamma)^{-3}SA/\epsilon^2)$ by using a s-absorbing MDP construction and this model-based technique is later reused for other more general settings (*e.g.* Markov games (Zhang et al., 2020a) and linear MDPs (Cui & Yang, 2020)) and also for overcoming the sample size barrier (Li et al., 2020). In offline RL, Yin et al. (2021b) uses the model-based methods to achieve $\tilde{O}(H^3/d_m\epsilon^2)$ complexity.

Task-agnostic and Reward-free problems. The reward-free problem is initiated in the online RL (Jin et al., 2020a) where the agent needs to efficiently explore an MDP environment *without* using any reward information. It requires high probability guarantee for learning optimal policy for *any* reward function. Later, Kaufmann et al. (2020); Menard et al. (2020) establish the $\tilde{O}(H^3S^2A/\epsilon^2)$

[1] Here d_m is the minimal marginal state-action occupancy, see Assumption 2.4.

[2] We only provide a short discussion of the most related works due to the space constraint. A detailed discussion can be found in Appendix A.

complexity and Zhang et al. (2020d) further tightens the dependence to $\tilde{O}(H^2S^2A/\epsilon^2)$. Recently, Zhang et al. (2020c) proposes the task-agnostic setting where one needs to use exploration data to simultaneously learn K tasks and proves an upper bound $\tilde{O}(H^5SA\log(K)/\epsilon^2)$. However, although these settings remain critical in the offline regime, no statistical result has been derived so far.

1.2 Our contribution

Optimal local uniform OPE. First and foremost, 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. (2021b) ($\tilde{O}(H^3/d_m\epsilon^2)$) non-trivially through our new *singleton-absorbing MDP* technique.

Information-theoretical characterization of the global 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. (2021b) that the global uniform OPE is generically harder than the local uniform OPE / offline learning by a factor of S, such a difference will dominate when the state space is exponentially large.

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.

Singleton-absorbing MDP: a sharp analysis tool for episodic stationary transition case. On the technical end, our major contribution is the novel design of *singleton-absorbing MDP* which handles the data-dependence hurdle encountered in the stationary MDPs. To decouple the data-dependence between $\widehat{P}_{s,a}$ and $\widehat{V}$, Agarwal et al. (2020) uses a s-absorbing MDP $\widehat{V}_s$ (in lieu of $\widehat{V}$) of each state for the independence. To control the error propagation between $\widehat{V}_s$ and $\widehat{V}$, they use the ϵ-net covering such that the value of $\widehat{V}_s$ traverse the evenly-spaced grids in $[0,(1-\gamma)^{-1}]$. However, when applied to finite horizon case, the complexity increases as there are H different quantities $(V_1,...,V_H)$ and the ϵ-nets need to cover the H-dimensional space $[0,H]^H$. This result in a exponential-H covering number and the metric entropy blows up by a factor H, which yields suboptimal result. In contrast, the *singleton-absorbing MDP* technique designs a single absorbing MDP that can also control the error propagation sufficiently well. This sharp analysis tool negates the conjecture of Cui & Yang (2020) that absorbing MDP is not well suitable for finite horizon stationary MDP.

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 the identical learning pattern. Considering these tasks were originally proposed in the online regime under different contexts, such a unified view from the model-based perspective offers a new angle for understanding offline RL.

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: $s_1\sim d_1, a_t\sim\pi(\cdot|s_t), r_t=r(s_t,a_t), s_{t+1}\sim P(\cdot|s_t,a_t), \forall t\in[H]$.

Figure 1: Related comparisons of sample complexities for offline RL

Result/Method	Setting	Type	Complexity	Uniform guarantee?
Le et al. (2019)	∞-horizon	FQI variants	$\widetilde{O}((1-\gamma)^{-6}\beta_\mu/\epsilon^2)$	No
FQI (Chen & Jiang, 2019)	∞-horizon	FQI variants	$\widetilde{O}((1-\gamma)^{-6}C/\epsilon^2)$	No
MSBO/MABO (Xie & Jiang, 2020b)	∞-horizon	FQI variants	$\widetilde{O}((1-\gamma)^{-4}C_\mu/\epsilon^2)$	No
OPEMA (Yin et al., 2021b)	H-horizon	Non-splitting	$\widetilde{O}(H^3/d_m\epsilon^2)$	$\sqrt{H}/S$-local uniform
OPDVR (Yin et al., 2021a)	H-horizon	Data splitting	$\widetilde{O}(H^2/d_m\epsilon^2)$	No
Model-based Plug-in (Corollary 4.2)	H-horizon	Non-splitting	$\widetilde{O}(H^2/d_m\epsilon^2)$	$\sqrt{H/S}$-local uniform
Task-Agnostic (Theorem 5.3)	H-horizon	Non-splitting	$\widetilde{O}(H^2\log(K)/d_m\epsilon^2)$	—
Reward-Free (Theorem 5.4)	H-horizon	Non-splitting	$\widetilde{O}(H^2S/d_m\epsilon^2)$	—

* K is the number of tasks for Task-agnostic setting and β_μ, C and $1/d_m$ are data coverage parameters that measure the state-action dependence and are qualitative similar under their respective assumptions.

For any policy π and any $h \in [H]$, value function $V_h^\pi(\cdot) \in \mathbb{R}^S$ and Q-value function $Q_h^\pi(\cdot,\cdot) \in \mathbb{R}^{S\times A}$ are defined as: $V_h^\pi(s) = \mathbb{E}_\pi[\sum_{t=h}^H r_t | s_h = s]$, $Q_h^\pi(s,a) = \mathbb{E}_\pi[\sum_{t=h}^H r_t | s_h, a_h = s, a]$, $\forall s, a \in \mathcal{S}, \mathcal{A}$. The goal of RL is to find a policy $\pi^\star$ such that $v^\pi := \mathbb{E}_\pi\left[\sum_{t=1}^H r_t\right]$ is maximized, which is equivalent to simultaneously maximize $V_1^\pi(s)$ (or $Q_1^\pi(s,a)$) for all s (or s,a) (Sutton & Barto, 2018). Therefore, for a targeted accuracy $\epsilon > 0$ it suffices to find a policy π_{alg} such that $\left\|Q_1^\star - Q_1^{\pi_{\text{alg}}}\right\|_\infty \le \epsilon$. We denote V_h^π, Q_h^π as column vectors and $P_{s,a}$ as the row vector. 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. We focus our attention on the model-based methods, which has witnessed numerous successes and is one of the most critical components of theoretical RL as a whole (as reviewed in Section 1.1). 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.*

We emphasize that 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. Nevertheless, we (by convention) assume the mean reward function is known and the initial state distribution d_1 will not affect the choice of optimal policy $\pi^\star$. Thus, Definition 2.1 suffices for our purposes.

2.1 Uniform convergence in offline RL

We study offline RL from the uniform OPE perspective. Concretely, uniform OPE extends the point-wise (fixed target policy) OPE to a family of policies Π. The goal is to construct estimator $\widehat{Q}_1^\pi$ such that $\sup_{\pi\in\Pi}\left\|Q_1^\pi - \widehat{Q}_1^\pi\right\| < \epsilon$, which automatically ensures point-wise OPE for any $\pi \in \Pi$. More importantly, uniform OPE directly implies offline learning when Π contains optimal policies. As explained in Section 1, let $\widehat{\pi}^\star := \operatorname{argmax}_\pi \widehat{V}_1^\pi$ be the *empirical optimal policy* for some OPE estimator $\widehat{v}^\pi$, then by (1) $\widehat{\pi}^\star$ is a near-optimal policy given uniform OPE guarantee. We consider the following two policy classes that are of the interests.

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 $\widehat{\pi}^\star := \operatorname{argmax}_\pi \widehat{V}_1^\pi$ 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.[3] 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. Next we present the regularity assumption required for uniform convergence OPE problem.

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 marginal state-action probability.*

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. (2021b) 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 (Munos, 2003). This assumption can be potentially relaxed for pure offline learning problems, *e.g.* Liu et al. (2019); Rashidinejad et al. (2021), where they only require $d^\mu(s)(d^\mu(s,a)) > 0$ for any state s (s,a) satisfies $d^{\pi^\star}(s)(d^{\pi^\star}(s,a)) > 0$.

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). If this conjecture is true, then deriving sample efficient global OPE method is just as important as deriving efficient offline learning algorithm (plus the additional benefit of evaluating data-dependent algorithms)! Yin et al. (2021b) 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. This result also reveals the global uniform convergence bound in Yin et al. (2021b) ($\tilde{O}(H^3S/d_m\epsilon^2)$) is essentially minimax rate-optimal for their *non-stationary setting* and complements the story on the optimality behavior for global uniform OPE. Moreover, from the generative model view $S/d_m\epsilon^2 \approx \Theta(S^2A/\epsilon^2)$ which is linear in the model size S^2A. This means in order to achieve global uniform convergence any algorithm needs to estimate each coordinate of transition kernel $P(s'|s,a)$ accurately. We now provide the proof sketch and the full proof is in Appendix C.

Proof Sketch. We only explain the case where $H = 2$ in this proof sketch. Our proof relies on the following novel reduction to l_1 density estimation

$$\sup_{\pi\in\Pi_g}\left\|\widehat{Q}^\pi_1 - Q^\pi_1\right\|_\infty \ge \sup_{s,a}\frac{1}{2}\left\|\widehat{P}(\cdot|s,a) - P(\cdot|s,a)\right\|_1$$

and leverages the Minimax rate for estimating discrete distribution under l_1 loss is $O(\sqrt{S/n_{s,a}})$ (Han et al., 2015). Concretely, by Definition 2.1, let $\widehat{P}$ be the learned transition by any arbitrary model-based method. Since we assume r is known and by convention $Q^\pi_{H+1} = 0$ for any π, then by Bellman equation $\widehat{Q}^\pi_h = r_h + \widehat{P}^{\pi_{h+1}}\widehat{Q}^\pi_{h+1}$, $\forall h \in [H]$. In particular, $\widehat{Q}^\pi_{H+1} = Q^\pi_{H+1} =$

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

0, which implies $\widehat{Q}^\pi_H = Q^\pi_H = r_H$. Now, again by definition of Bellman equation $\widehat{Q}^\pi_{H-1} = r_{H-1} + \widehat{P}^{\pi_H}\widehat{Q}^\pi_H = r_{H-1} + \widehat{P}^{\pi_H} r_H$ and $Q^\pi_{H-1} = r_{H-1} + P^{\pi_H} r_H$, therefore (recall $H = 2$ and note $r_H \in \mathbb{R}^{S\cdot A}, r^{\pi_H}_H \in \mathbb{R}^S$) $\sup_{\pi\in\Pi_g} \left\|\widehat{Q}^\pi_{H-1} - Q^\pi_{H-1}\right\|_\infty = \sup_{\pi\in\Pi_g} \left\|(\widehat{P} - P) r^{\pi_H}_H\right\|_\infty \approx \sup_{r\in\{0,1\}^S} \left\|(\widehat{P} - P) r\right\|_\infty \geq \sup_{s,a} \frac{1}{2} \left\|\widehat{P}_{s,a} - P_{s,a}\right\|_1 \geq O(\sqrt{\frac{S}{n_{s,a}}})$; Lastly, using exponential tail bound to obtain $O(\sqrt{S/n_{s,a}}) \gtrsim O(\sqrt{S/nd_m})$ with high probability. ■

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. Yin et al. (2021b) Theorem 3.7 first obtains the $\tilde{O}(H^3/d_m\epsilon^2)$ local uniform convergence for Π_l (recall Definition 2.3). Unfortunately, their technique cannot further reduces the H when the MDP is stationary. In this section, we show the model-based plug-in approach can achieve $\tilde{O}(H^2/d_m\epsilon^2)$ rate for the local uniform OPE task. To this end, we design the new *singleton-absorbing MDP* to handle the challenge in the stationary transition setting, which uses the absorbing MDP with one single H-dimensional reference point. The *singleton-absorbing MDP* technique is our major technical contribution and it 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.[4]

4.1 Model-based Offline Plug-in Estimator

Let $n_{s,a} := \sum_{i=1}^n \sum_{h=1}^H \mathbf{1}[s^{(i)}_h, a^{(i)}_h = s, a]$ be the total counts that visit (s, a) pair, then the model-based offline plug-in estimator constructs estimator $\widehat{P}$ as: $\widehat{P}(s'|s,a) = \frac{\sum_{i=1}^n \sum_{h=1}^H \mathbf{1}[(s^{(i)}_{h+1}, a^{(i)}_h, s^{(i)}_h) = (s', s, a)]}{n_{s,a}}$, 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 $\widehat{d}_1(s) = n_s/n$. Under the above setting, we have the empirical Bellman optimality equations (and its population version) as $\forall s \in \mathcal{S}, h \in [H]$: $V^\star_h(s) = \max_a \{r(s,a) + P(\cdot|s,a)V^\star_{h+1}\}$, $\widehat{V}^\star_h(s) = \max_a \left\{r(s,a) + \widehat{P}(\cdot|s,a)\widehat{V}^\star_{h+1}\right\}$.

4.2 Main results for local uniform OPE and offline learning

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

Theorem 4.1 (optimal local uniform OPE)**.** *Let $\epsilon_{opt} \leq \sqrt{H/S}$ 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. $1-\delta$,*

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

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. (2021b) (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. We explain this technique in detail in the next section.

On the other hand, characterizing policy class through the distance in value (like Π_l) is more flexible than characterizing the distance between policies themselves (*e.g.* via total variation). This is because: if two policies are "close", then their values are also similar; but the reverse may not be true since two very different policies could possibly generate similar values. Therefore the consideration of Π_l is generic and conceptually reflects the fundamental principle of RL: as long as two policies yield the same value, they are considered "equally good", no matter how different they are.[5]

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

[5] We recognize that in the specific settings (*e.g.* safe policy improvement) some of the policies that yield high values are not feasible. These considerations are beyond the scope of this paper.

Most importantly, Theorem 4.1 guarantees near-minimax optimal offline learning:

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

$$V_1^\star - V_1^{\widehat{\pi}} \leq C\left[\sqrt{\frac{H^2\iota}{nd_m}} + \frac{H^{2.5}S^{0.5}\iota}{nd_m}\right]\mathbf{1} + \epsilon_{opt}\,\mathbf{1}.$$

Corollary 4.2 first establishes the minimax rate for offline learning for any policy $\widehat{\pi}$ with the measurable gap $\epsilon_{\text{opt}} \leq \sqrt{H/S}$. 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. Besides, we leverage full data to construct empirical MDP for planning and, on the contrary, Yin et al. (2021a) uses data-splitting (split data into mini-batches and only apply each mini-batch at each specific iteration) to enable Variance Reduction technique, which could cause inefficient data use for the practical purpose. By the lower bound result from Yin et al. (2021a) (Theorem 4.2), our Corollary 4.2 is near minimax optimal.

For the rest of the section, we explain the proving ideas by introducing the *singleton-absorbing MDP* technique and the full proofs of Theorem 4.1, Corollary 4.2 can be found in Appendix B, D.

4.3 Singleton absorbing MDP for finite horizon MDP

For the ease of illustration, we explain our idea via bounding $||\widehat{Q}_h^{\widehat{\pi}^\star} - Q_h^{\widehat{\pi}^\star}||_\infty$ (instead of $\sup_{\pi\in\Pi_l} ||\widehat{Q}_1^\pi - Q_1^\pi||_\infty$) and choose related quantity $\widehat{\pi}^\star$ (instead of $\widehat{\pi}$) and $\widehat{V}_h^\star$ (instead of $\widehat{V}_h^{\widehat{\pi}}$) to discuss. Essentially, the key challenge in obtaining the optimal dependence in stationary setting is the need to decouple the dependence between $P - \widehat{P}$ and $\widehat{V}_h^\star$ as we aggregate all data for constructing both $\widehat{P}$ and $\widehat{V}_h^\star$. 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., 2021b) and $\widehat{P}_t$ and $\widehat{V}_{t+1}^\star$ 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 $\widehat{V}_h^\star$ 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 $\widehat{V}_h^\star$ which will not incur additional dependence on horizon H caused by the union bound. We begin with the general definition of absorbing MDP initialized in Agarwal et al. (2020) and then introduce the *singleton 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 $V_{\{s,u_t\}}^\pi$ to denote $V_{s,M_{s,\{u_t\}_{t=1}^H}}^\pi$ and similarly for Q_t, r and transition P. Also, $V_{\{s,u_t\}}^\star$ ($Q_{\{s,u_t\}}^\star$) 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. *$V_t^\star(s) - V_{t+1}^\star(s) \geq 0$, $\forall s \in \mathcal{S}, t \in [H]$. Moreover, fix a state s. If we choose $u_t^\star := V_t^\star(s) - V_{t+1}^\star(s)$, then we have the vector form equations: $V_{h,\{s,u_t^\star\}}^\star = V_{h,M}^\star \quad \forall h \in [H]$. Similarly, if we choose $\hat{u}_t^\star := \widehat{V}_t^\star(s) - \widehat{V}_{t+1}^\star(s)$, then $\widehat{V}_{h,\{s,\hat{u}_t^\star\}}^\star = \widehat{V}_{h,M}^\star$, $\forall h \in [H]$.*

The proofs are deferred to Appendix B. Note by Lemma 4.3 the assignment of $u_t^\star (:= r_{t,\{s,u_t^\star\}})$ is well-defined. Lemma 4.3 is crucial since, under the specification of $u_t^\star$, the optimal value in $M_{s,\{u_t^\star\}_{t=1}^H}$ is identical to the optimal value in original M. Based on these, we have the definition:

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

Using Definition 4.4, for each (s,a) row the term $(\widehat{P}_{s,a} - P_{s,a})\widehat{V}_h^\star$ can be substituted by $(\widehat{P}_{s,a} - P_{s,a})\widehat{V}_{h,\{s,u_t^\star\}}^\star$, where $\widehat{P}_{s,a}$ and $\widehat{V}_{h,\{s,u_t^\star\}}^\star$ are independent by construction and Bernstein concentration applies. Furthermore, by the selection of $u_t^\star$, we can control the error of $||\widehat{V}_h^\star - \widehat{V}_{h,\{s,u_t^\star\}}^\star||_\infty$ to have

rate $O(\sqrt{\frac{1}{n}})$ which forces the term $(\widehat{P}_{s,a} - P_{s,a})(\widehat{V}^\star_h - \widehat{V}^\star_{h,\{s,u^\star_t\}})$ to have higher order error. These are the critical building blocks for bounding $||\widehat{Q}^{\widehat{\pi}^\star}_h - Q^{\widehat{\pi}^\star}_h||_\infty$.

Indeed, by Bellman equations we have the decomposition: $\widehat{Q}^{\widehat{\pi}^\star}_h - Q^{\widehat{\pi}^\star}_h = \ldots = \sum_{t=h}^H \Gamma^{\widehat{\pi}^\star}_{h+1:t}\left(\widehat{P} - P\right)\widehat{V}^\star_{t+1}$, where $\Gamma^\pi_{h+1:t} = \prod_{i=h+1}^t P^{\pi_i}$ is multi-step state-action transition and $\Gamma_{h+1:h} := I$. Then for each (s,a) row

$$(\widehat{P}_{s,a} - P_{s,a})\widehat{V}^\star_h = (\widehat{P}_{s,a} - P_{s,a})(\widehat{V}^\star_h - \widehat{V}^\star_{h,\{s,u^\star_t\}}) + (\widehat{P}_{s,a} - P_{s,a})\widehat{V}^\star_{h,\{s,u^\star_t\}} \lesssim ||\widehat{P}_{s,a} - P_{s,a}||_1$$
$$\cdot ||\widehat{V}^\star_h - \widehat{V}^\star_{h,\{s,u^\star_t\}}||_\infty + \sqrt{\frac{\mathrm{Var}_{s,a}(\widehat{V}^\star_{h,\{s,u^\star_t\}})}{n_{s,a}}} \lesssim \sqrt{\frac{S}{n_{s,a}}}\left\|\widehat{V}^\star_h - \widehat{V}^\star_{h,\{s,u^\star_t\}}\right\|_\infty + \sqrt{\frac{\mathrm{Var}_{s,a}(\widehat{V}^\star_h)}{n_{s,a}}}\ (\star) \tag{2}$$

here $(\star)$ is the place that the traditional technique applies the union bound over their *exponential large* ϵ-net and we do not have it! Next, by Lemma 4.3 and Lemma B.2 in Appendix

$$||\widehat{V}^\star_h - \widehat{V}^\star_{h,\{s,u^\star_t\}}||_\infty = ||\widehat{V}^\star_{h,\{s,\hat{u}^\star_t\}} - \widehat{V}^\star_{h,\{s,u^\star_t\}}||_\infty \leq H \max_t |\hat{u}^\star_t - u^\star_t| \leq 2H \max_t |\widehat{V}^\star_t - V^\star_t|,$$

by a crude bound (Lemma J.10), $\max_t |\widehat{V}^\star_t - V^\star_t| \lesssim H^2\sqrt{\frac{S}{n_{s,a}}}$ which makes $\sqrt{\frac{1}{n_{s,a}}}||\widehat{V}^\star_h - \widehat{V}^\star_{h,\{s,u^\star_t\}}||_\infty$ have order $1/n_{s,a}$. Finally, to reduce the horizon dependence we apply $\sum_{t=h}^H \Gamma^\pi_{h+1:t}\sqrt{\mathrm{Var}_{s,a}\left(V^\pi_{t+1}\right)} \leq \sqrt{(H-h)^3}$ for any π. This (informally) bounds $\widehat{Q}^{\widehat{\pi}^\star}_h - Q^{\widehat{\pi}^\star}_h$ by $||\widehat{Q}^{\widehat{\pi}^\star}_h - Q^{\widehat{\pi}^\star}_h||_\infty \lesssim \sqrt{\frac{H^3}{n_{s,a}}} + \frac{Poly(H,S)}{n_{s,a}}$. Lastly, use $\min_{s,a} n_{s,a} \gtrsim H \cdot d_m$ to finish the proof.

Remark 4.5. *We emphasize the appropriate selection of $M_{s,\{u^\star_t\}_{t=1}^H}$ ($\widehat{M}_{s,\{u^\star_t\}_{t=1}^H}$) is the key for achieving optimality. It guarantees two things: 1. $\widehat{V}^\star_{h,\{s,u^\star_t\}}$ approximates $\widehat{V}^\star_h$ with sufficient accuracy (has rate $\sqrt{1/n_{s,a}}$); 2. it avoids the fine-grid design with exponential union bound in the dominate term ($\sqrt{\frac{\mathrm{Var}_{s,a}(\widehat{V}^\star_h)\log(|U_{s,a}|/\delta)}{N}}$ with $|U_{s,a}|$ to be at least H^H Cui & Yang (2020).)*

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 $\mathcal{D} = \{(s^{(i)}_t, a^{(i)}_t, r(s^{(i)}_t, a^{(i)}_t), s^{(i)}_{t+1})\}^{t\in[H]}_{i\in[n]}$; 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 $\mathcal{D}' = \{(s^{(i)}_t, a^{(i)}_t)\}^{t\in[H]}_{i\in[n]}$. 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 $\mathcal{D}' = \{(s^{(i)}_t, a^{(i)}_t)\}^{t\in[H]}_{i\in[n]}$ 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 $\mathcal{D}' = \{(s^{(i)}_t, a^{(i)}_t)\}^{t\in[H]}_{i\in[n]}$ 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 $\mathbb{P}\left[\forall r, \left\|V^\star_{1,M} - V^{\hat{\pi}}_{1,M}\right\|_\infty \leq \epsilon\right] \geq 1-\delta$?*

Definition 5.1 and Definition 5.2 are the offline counterparts of Zhang et al. (2020c) and Jin et al. (2020a) in online RL. Those settings are of practical interests in the offline regime as well since in practice reward functions are often iteratively engineered to encourage desired behavior via trial and error and using one shot of offline exploration data $\mathcal{D}'$ to tackle problems with different reward functions (different tasks) could help improve sample efficiency significantly.

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* $\mathcal{D}' = \{(s_t^{(i)}, a_t^{(i)})\}_{i\in[n]}^{t\in[H]}$ *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* $\widehat{\pi}_k^\star := \operatorname{argmax}_\pi \widehat{V}_{1,M_k}^\pi \ \forall k \in [K]$*, when* $n > O(H \cdot [\iota + \log(K)]/d_m)$*, then with probability* $1-\delta$*,* $\left\|V_{1,M_k}^\star - V_{1,M_k}^{\widehat{\pi}_k^\star}\right\|_\infty \leq O\left[\sqrt{\frac{H^2(\iota+\log(K))}{nd_m}} + \frac{H^{2.5}S^{0.5}(\iota+\log(K))}{nd_m}\right] \quad \forall k \in [K]$.

Theorem 5.4 (optimal offline reward-free learning)**.** *Given* $\mathcal{D}' = \{(s_t^{(i)}, a_t^{(i)})\}_{i\in[n]}^{t\in[H]}$ *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* $\widehat{\pi}_M^\star := \operatorname{argmax}_\pi \widehat{V}_{1,M}^\pi \ \forall r$*, when* $n > O(HS \cdot \iota/d_m)$*, then with probability* $1-\delta$*,* $\left\|V_{1,M}^\star - V_{1,M}^{\widehat{\pi}_M^\star}\right\|_\infty \leq O\left[\sqrt{\frac{H^2S\cdot\iota}{nd_m}} + \frac{H^2S\cdot\iota}{nd_m}\right], \ \forall r, M$.

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. (2020c) and Jin et al. (2020a).[6] The higher order dependence in Theorem 5.4 is also tight comparing to Theorem 5.3. Such statistically optimal results reveal the model-based methods generalize well to those seemingly challenging problems in the offline regime. Changing to these harder problems would not affect the optimal statistical efficiency of the model-based approach.

Further discussion: Linear MDP with anchor representations 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 therefore it is natural to generalize such idea for the episodic stationary transition setting for other problems. As an example, we further present a sharp result for the setting of finite horizon linear MDP with anchor points. We narrate by assuming a generative oracle (that allows sampling from $s' \sim P(\cdot|s,a)$) for the ease of exposition.

Definition 5.5 (Linear MDP with anchor points (Yang & Wang, 2019; Cui & Yang, 2020))**.** *Let* $\mathcal{S}$ *be the exponential large space and* $\mathcal{A}$ *be the infinite (or even continuous) spaces. Assume there is feature map* $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^K$ *(where* $K \ll |\mathcal{S}|$*), i.e.* $\phi(s,a) = [\phi_1(s,a), \ldots, \phi_K(s,a)]$*. Transition* P *admits a linear representation:* $P(s'|s,a) = \sum_{k\in[K]} \phi_k(s,a)\psi_k(s')$ *where* $\psi_1(\cdot), \ldots, \psi_K(\cdot)$ *are unknowns. We further assume there exists a set of anchor state-action pairs* $\mathcal{K}$ *such that any* (s,a) *can be represented as a convex combination of the anchors* $\{(s_k,a_k)|k \in \mathcal{K}\}$*:* $\exists \{\lambda_k^{s,a}\} : \phi(s,a) = \sum_{k\in\mathcal{K}} \lambda_k^{s,a}\phi(s_k,a_k), \sum_{k\in\mathcal{K}} \lambda_k^{s,a} = 1, \lambda_k \geq 0, \forall k \in \mathcal{K}, (s,a) \in (\mathcal{S},\mathcal{A})$.

Under the definition, denote N be the number of samples at each anchor pairs. Then we have the following (see Appendix H for the proof):

Theorem 5.6 (Optimal sample complexity)**.** *Under Definition 5.5, let* $\widehat{\pi}^\star = \operatorname{argmax}_\pi \widehat{V}_1^\pi$*. Then if* $N \geq cH^2|\mathcal{S}|\log(KH/\delta)$*, we have with probability* $1-\delta$*,* $||Q_1^\star - Q_1^{\widehat{\pi}^\star}||_\infty \leq \widetilde{O}(\sqrt{H^3/N})$.

Comparing to Theorem 4 of Cui & Yang (2020), Theorem 5.6 removes the additional dependence $\min\{|S|, K, H\}$. In term of the total sample complexity, Theorem 5.6 gives $\tilde{O}(KH^3/\epsilon^2)$ while Cui & Yang (2020) has $\widetilde{O}(KH^4/\epsilon^2)$ (see their Section 7, first bullet point). Our result again reveals the model-based method is statistically optimal for the current setting.

6 Conclusion and Future Works

This work studies the uniform convergence problems for offline policy evaluation (OPE) and provides complete answers for their optimality behaviors. We achieve the optimal sample complexity for stationary MDPs using a novel adaptation of the absorbing MDP trick, which is more generally applicable to the new offline task-agnostic and reward-free settings combining with the model-based approach and we hope it can be applied to a broader range of future problems. We end the section by two future directions. **1.** Our main result (Theorem 4.1) has an additional $\sqrt{HS}$ dependence in the higher order error term and we cannot further remove it based on our current technique. **2.** The current study of uniform OPE derives results with expression using parameter dependence and deriving instance-dependent uniform convergence result will draw a clearer picture on the individual behaviors for each policy. We leave those as the future works.

[6] We add a discussion in Appendix G to explain more clearly why our rates are optimal for these problems.

References

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

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).

Ameko, M. K., Beltzer, M. L., Cai, L., Boukhechba, M., Teachman, B. A., & Barnes, L. E. (2020). Offline contextual multi-armed bandits for mobile health interventions: A case study on emotion regulation. In *Fourteenth ACM Conference on Recommender Systems*, (pp. 249–258).

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

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

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

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

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.

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

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

Dann, C., Li, L., Wei, W., & Brunskill, E. (2019). Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, (pp. 1507–1516). PMLR.

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).

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

Feng, Y., Li, L., & Liu, Q. (2019). A kernel loss for solving the bellman equation. *arXiv preprint arXiv:1905.10506*.

Feng, Y., Ren, T., Tang, Z., & Liu, Q. (2020). Accountable off-policy evaluation with kernel bellman statistics. In *International Conference on Machine Learning*, (pp. 3102–3111). PMLR.

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

Hao, B., Duan, Y., Lattimore, T., Szepesvári, C., & Wang, M. (2020). Sparse feature selection makes batch reinforcement learning more sample efficient. *arXiv preprint arXiv:2011.04019*.

Hao, B., Duan, Y., Lu, H., Szepesvári, C., Wang, M., et al. (2021). Bootstrapping statistical inference for off-policy evaluation. *arXiv preprint arXiv:2102.03607*.

Hu, Y., Kallus, N., & Uehara, M. (2021). Fast rates for the regret of offline reinforcement learning. *arXiv preprint arXiv:2102.00479*.

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

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

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

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

Jin, Y., Yang, Z., & Wang, Z. (2020c). Is pessimism provably efficient for offline rl? *arXiv preprint arXiv:2012.15085*.

Kallus, N., & Uehara, M. (2019). Intrinsically efficient, stable, and bounded off-policy evaluation for reinforcement learning. In *Advances in Neural Information Processing Systems*, (pp. 3325–3334).

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).

Kato, M., Uehara, M., & Yasui, S. (2020). Off-policy evaluation and learning for external validity under a covariate shift. In *Advances in neural information processing systems*.

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

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

Kuzborskij, I., Vernade, C., Gyorgy, A., & Szepesvari, C. (2020). Confident off-policy evaluation and selection through self-normalized importance weighting. *arXiv preprint arXiv:2006.10460*.

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

Lattimore, T., & Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

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

Levine, S., Kumar, A., Tucker, G., & Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

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

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

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

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. (2020b). Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*.

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

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

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

Nachum, O., Chow, Y., Dai, B., & Li, L. (2019). Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *arXiv preprint arXiv:1906.04733*.

Nguyen-Tang, T., Gupta, S., Tran-The, H., & Venkatesh, S. (2021). On finite-sample analysis of offline reinforcement learning with deep relu networks. *arXiv preprint arXiv:2103.06671*.

Osinski, B., Jakubowski, A., Ziecina, P., Milos, P., Galias, C., Homoceanu, S., & Michalewski, H. (2020). Simulation-based reinforcement learning for real-world autonomous driving. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, (pp. 6411–6418). IEEE.

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

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

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

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.

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

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

Uehara, M., Imaizumi, M., Jiang, N., Kallus, N., Sun, W., & Xie, T. (2021). Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*.

Uehara, M., & Jiang, N. (2019). Minimax weight and q-function learning for off-policy evaluation. *arXiv preprint arXiv:1910.12809*.

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

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

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

Wang, Y.-X., Agarwal, A., & Dudık, M. (2017). Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, (pp. 3589–3597). PMLR.

Xiao, C., Wu, Y., Lattimore, T., Dai, B., Mei, J., Li, L., Szepesvari, C., & Schuurmans, D. (2021). On the optimality of batch policy optimization algorithms. *arXiv preprint arXiv:2104.02293*.

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).

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

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

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

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

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

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

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

Zhang, R., Dai, B., Li, L., & Schuurmans, D. (2020b). Gendice: Generalized offline estimation of stationary values. *arXiv preprint arXiv:2002.09072*.

Zhang, S., Wan, Y., Sutton, R. S., & Whiteson, S. (2021). Average-reward off-policy policy evaluation with function approximation. *arXiv preprint arXiv:2101.02808*.

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

Zhang, Z., Du, S. S., & Ji, X. (2020d). Nearly minimax optimal reward-free reinforcement learning. *arXiv preprint arXiv:2010.05901*.

Zhang, Z., Ji, X., & Du, S. S. (2020e). Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. *arXiv preprint arXiv:2009.13503*.

Zhou, D., Gu, Q., & Szepesvari, C. (2020). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. *arXiv preprint arXiv:2012.08507*.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] , see Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We do not have any negative social impact since this is a theoretical work.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]