Nearly Horizon-Free Offline Reinforcement Learning

Tongzheng Ren
UT Austin
tongzheng@utexas.edu

Jialian Li
Tsinghua University
lijialian7@163.com

Bo Dai
Google Brain
bodai@google.com

Simon S. Du
University of Washington
ssdu@cs.washington.edu

Sujay Sanghavi
UT Austin & Amazon Search
sanghavi@mail.utexas.edu

February 11, 2022

Abstract

We revisit offline reinforcement learning on episodic time-homogeneous Markov Decision Processes (MDP). For tabular MDP with S states and A actions, or linear MDP with anchor points and feature dimension d, given the collected K episodes data with minimum visiting probability of (anchor) state-action pairs d_m, we obtain nearly horizon H-free sample complexity bounds for offline reinforcement learning when the total reward is upper bounded by 1. Specifically:

- For offline policy evaluation, we obtain an $\tilde{O}\left(\sqrt{\frac{1}{Kd_m}}\right)$ error bound for the plug-in estimator, which matches the lower bound up to logarithmic factors and does not have additional dependency on $\mathrm{poly}\,(H, S, A, d)$ in higher-order term.
- For offline policy optimization, we obtain an $\tilde{O}\left(\sqrt{\frac{1}{Kd_m}} + \frac{\min(S,d)}{Kd_m}\right)$ sub-optimality gap for the empirical optimal policy, which approaches the lower bound up to logarithmic factors and a high-order term, improving upon the best known result by Cui and Yang [2020] that has additional $\mathrm{poly}\,(H, S, d)$ factors in the main term.

To the best of our knowledge, these are the *first* set of nearly horizon-free bounds for episodic time-homogeneous offline tabular MDP and linear MDP with anchor points. Central to our analysis is a simple yet effective recursion based method to bound a "total variance" term in the offline scenarios, which could be of individual interest.

1 Introduction

Reinforcement Learning (RL) aims to learn to make sequential decisions to maximize the long-term reward in unknown environments, and has demonstrated success in game-playing [Mnih et al., 2013, Silver et al., 2016], robotics [Andrychowicz et al., 2020], and automatic algorithm design [Dai et al., 2017]. These successes rely on being able to deploy the algorithms in a way that directly interacts with the respective environment, allowing them to improve the policies in a trial-and-error way. However, such direct interactions with real environments can be expensive or even impossible in other real-world applications, *e.g.*, education [Mandel et al., 2014], health and medicine [Murphy et al., 2001, Gottesman et al., 2019], conversational AI [Ghandeharioun et al., 2019] and recommendation systems [Chen et al., 2019]. Instead, we

are often given a collection of logged experiences generated by potentially multiple and possibly unknown policies in the past.

This lack of access to real-time interactions with an environment led to the field of *offline reinforcement learning* [Levine et al., 2020]. Within this, offline policy evaluation (OPE) focuses on evaluating a policy, and offline policy optimization (OPO) focuses on improving policies; both rely only upon the given fixed past experiences without any further interactions. OPE and OPO are in general notoriously difficult, as unbiased estimators of the policy value can suffer from *exponentially* increasing variance in terms of horizon in the worst case [Li et al., 2015, Jiang and Li, 2016].

To overcome this "curse of horizon" in OPE, [Liu et al., 2018, Xie et al., 2019] first introduced marginalized importance sampling (MIS) based estimators. They showed that if (1) all of the logged experiences are generated from the same behavior policy, and (2) the behavior policy is known, then the exponential dependency on horizon can be improved to polynomial dependency. Subsequently, [Nachum et al., 2019a, Uehara et al., 2020, Yang et al., 2020] showed that the polynomial dependency could be achieved even without assumptions (1) and (2). The basic idea for these MIS-based estimators is estimating the marginal state-action density ratio between the target policy and the empirical data, so as to adjust the distribution mismatch between them. On the algorithmic side, marginal density ratio estimation can be implemented by either plug-in estimators [Xie et al., 2019, Yin and Wang, 2020], temporal-difference updates [Hallak and Mannor, 2017, Gelada and Bellemare, 2019], or solving a $\min$-$\max$ optimization [Nachum et al., 2019a, Uehara et al., 2020, Zhang et al., 2020a, Yang et al., 2020]. These OPE estimators can also be used as one component for OPO, resulting in the algorithms in [Nachum et al., 2019b, Yin et al., 2020, Liu et al., 2020a], which also inherit the polynomial dependency on horizon.

In a different but related line of work, [Wang et al., 2020, Zhang et al., 2020b] recently showed that, for the online episodic time-homogeneous tabular Markov Decision Process (MDP) that allows for the interactions with environments, the sample complexity only has $\mathrm{poly}\log$ dependency on the horizon. This motivates us to consider the following question:

Can offline reinforcement learning escape from the polynomial dependency on the horizon?

In this paper, we provide an affirmative answer to this question. Specifically, considering the episodic time-homogeneous tabular MDP with S states, A actions and horizon H, or the linear MDP with anchor points and feature dimension d, assuming the total reward for any episode is upper bounded by 1 almost surely, we obtain the following nearly H-free bounds:

- For offline policy evaluation (OPE), we show that the plug-in estimator has a *finite-sample* error of $\widetilde{O}\left(\sqrt{\frac{1}{Kd_m}}\right)$ (Theorem 1 and 5), where K is the number of episodes and d_m is the minimum visiting probability of (anchor) state-action pairs, that matches the lower bound up to logarithmic factors. We emphasize that the bound has no additional $\mathrm{poly}\,(H, S, A, d)$ dependency in the higher order term, unlike the known results of [Yin and Wang, 2020, Yin et al., 2020].

- For offline policy optimization (OPO), we show that the policy obtained by model-based planning on empirical MDP has a sub-optimality gap of $\widetilde{O}\left(\sqrt{\frac{1}{Kd_m}} + \frac{\min\{S,d\}}{Kd_m}\right)$ (Theorem 3 and 6), which matches the lower bound up to logarithmic factors and a high-order term. This also improves upon the best known result from [Cui and Yang, 2020] by removing additional $\mathrm{poly}\,(H, S, d)$ factors in the main term.

To the best of our knowledge, these are the *first* set of nearly horizon-free bounds for both OPE and OPO on time-homogeneous tabular MDP and linear MDP with anchor points. To achieve such sharp bounds, we propose a novel recursion based method to bound a "total variance" term (introduced below) that is broadly emerged in the offline reinforcement learning, which could be of individual interest.

Technique Overview. With a sequence of fairly standard steps in the literature, we can bound the error of the plug-in estimator via terms related to the square root of the "total variance" (also known as the Cramer-Rao type lower bound illustrated in [Jiang and Li, 2016]) $\sqrt{\sum_{h\in[H]}\sum_{s,a}\xi_h^\pi(s,a)\mathrm{Var}_{P(s,a)}\left(V_{h+1}^\pi(s')\right)}$ where ξ^π is the reaching probability, P is the transition and V^π is the value function under policy π. (For a more formal definition, see Section 3.). An improper bound of this term will introduce unnecessary dependency on the horizon, either in the main term or in the higher-order term. We instead bound this term with a recursive method, by observing that this "total variance" term can be approximately upper bounded by the square root of the "total variance of value square" $\sum_{h\in[H]}\sum_{s,a}\xi_h^\pi(s,a)\mathrm{Var}_{P(s,a)}\left[\left(V_{h+1}^\pi(s')\right)^2\right]$ and some remaining term (see Lemma 1 and Lemma 3 for the detail). Applying this argument recursively, we can finally obtain a $\mathrm{poly}\log H$ upper bound on the total variance, which eventually gets rid of the polynomial dependency on H.

The idea of higher order expansion has been investigated in [Li et al., 2020, Zhang et al., 2020b]. We notice that the recursion introduced in [Li et al., 2020] was designed for the infinite horizon setting, and how to generalize their technique to finite horizon setting is still unclear. Our recursion is conceptually more similar to the recursion in [Zhang et al., 2020b]. However, [Zhang et al., 2020b] considered the online setting, where we only need to bound the error on the visited state-action pairs. For the offline setting, we need to bound the error on every state-action pair that can be touched by the policy π. This introduces the reaching probability ξ^π in the "total variance" term, which we need to deal with using the MDP structure. As a result, our recursion is significantly different from their counterpart, especially in the case of linear MDP.

Organization. Our paper is organized as follows: in Section 2, we review the related literature on offline reinforcement learning, and then briefly introduce the problem we consider in Section 3. In Section 4 and Section 5, we show our results of offline policy evaluation and offline policy optimization on tabular MDP correspondingly, and in Section 6, we show how to generalize our results to linear MDP with anchor points. We finally conclude and discuss our results in Section 7.

2 Related Work

In this section, we briefly discuss the related literature in three categories, *i.e.*, offline policy evaluation, offline policy optimization, and horizon-free online reinforcement learning. Notice that, for the setting that assumes an additional generative model, typical model-based algorithms first query equal number of data from each state-action pair, then perform offline policy evaluation/optimization based on the queried data. Thus we view the reinforcement learning with generative model as a special instance of offline reinforcement learning. To make the comparison fair, for method and analysis that do not assume Assumption 1, we scale the error and sample complexity, by assuming per-step reward is upper bounded by $1-\gamma$ and $1/H$ under infinite-horizon and finite-horizon setting correspondingly.

Offline Policy Evaluation. For OPE in infinite horizon tabular MDP, [Li et al., 2020] showed that plug-in estimator can achieve the error of $\widetilde{O}\left(\sqrt{\frac{1}{d_m(1-\gamma)}}\right)$ under Assumption 1, which matches the lower bound in [Pananjady and Wainwright, 2020] up to logarithmic factors. For OPE in finite horizon time-inhomogeneous tabular MDP, [Yin and Wang, 2020, Yin et al., 2020] provided an error bound of $\widetilde{O}\left(\sqrt{\frac{1}{Kd_m}}+\frac{\sqrt{SA}}{Kd_m}\right)$ under the uniform reward assumption, which matches the lower bound [Jiang and Li, 2016] up to logarithmic fac-

Table 1: A comparison of existing offline policy evaluation results. The sample complexity in infinite horizon setting is the number of queries of transitions we need while in episodic setting is the number of episodes we need. If Non-Uniform Reward, the MDP we consider satisfies Assumption 1; otherwise, we assume the per-step reward is upper bounded by $1-\gamma$ and $1/H$ correspondingly.

Analysis	Setting	Non-Uniform Reward	Sample Complexity
[Li et al., 2020]	Infinite Horizon Tabular	Yes	$\widetilde{O}\left(\frac{1}{d_m(1-\gamma)\epsilon^2}\right)$
[Pananjady and Wainwright, 2020]	Infinite Horizon Tabular	Yes	$\Omega\left(\frac{1}{d_m(1-\gamma)\epsilon^2}\right)$
[Yin and Wang, 2020]	Finite Horizon time-inhomogeneous Tabular	No	$\widetilde{O}\left(\frac{1}{d_m\epsilon^2}+\frac{\sqrt{SA}}{d_m\epsilon}\right)$
[Jiang and Li, 2016]	Finite-Horizon time-inhomogeneous Tabular	No	$\Omega\left(\frac{1}{d_m\epsilon^2}\right)$
This work	Finite Horizon time-homogeneous Tabular/Linear	Yes	$\widetilde{O}\left(\frac{1}{d_m\epsilon^2}\right)$
Lower Bound	Finite Horizon time-homogeneous Tabular/Linear	Yes	$\Omega\left(\frac{1}{d_m\epsilon^2}\right)$

tors and an additional higher-order term. We here consider the time-homogeneous MDP, and obtain an error bound of $\tilde{O}(\sqrt{\frac{1}{Kd_m}})$, that does not have the additional $\sqrt{SA}$ in higher-order term, which is different from [Yin and Wang, 2020, Yin et al., 2020].

Beyond the tabular setting, [Duan and Wang, 2020] considered the performance of plug-in estimator with linear function approximation under the assumption of linear MDP without anchor points, and [Kallus and Uehara, 2019, 2020] provided more detailed analyses on the statistical properties of different kinds of estimators under different assumptions, which are not directly comparable to our work. Recently, there are also works [e.g. Feng et al., 2020] focusing on the interval estimation of the policy for practical application.

Table 2: A comparison of existing offline learning results. The sample complexity in infinite horizon setting is the number of queries of transitions we need while in episodic setting is the number of episodes we need. If Non-Uniform Reward, the MDP we consider satisfies Assumption 1; otherwise, we assume the per-step reward is upper bounded by $1-\gamma$ and $1/H$ correspondingly.

Analysis	Setting	Non-Uniform Reward	Sample Complexity
[Agarwal et al., 2020]	Infinite Horizon Tabular	No	$\widetilde{O}\left(\frac{1}{d_m(1-\gamma)\epsilon^2}\right)$
[Li et al., 2020]	Infinite Horizon Tabular	Yes	$\widetilde{O}\left(\frac{1}{d_m(1-\gamma)\epsilon^2}\right)$
[Yin et al., 2020]	Finite Horizon time-inhomogeneous Tabular	No	$\widetilde{O}\left(\frac{H}{d_m\epsilon^2}\right)$
[Cui and Yang, 2020]	Finite Horizon time-homogeneous Linear	No	$\widetilde{O}\left(\frac{\min\{H,S,d\}}{d_m\epsilon^2}\right)$
[Zhang et al., 2020b]	Finite Horizon time-homogeneous Tabular Online	Yes	$\widetilde{O}\left(\frac{SA}{\epsilon^2}+\frac{S^2A}{\epsilon}\right)$
This Work	Finite Horizon time-homogeneous Tabular/Linear	Yes	$\widetilde{O}\left(\frac{1}{d_m\epsilon^2}+\frac{\min\{S,d\}}{d_m\epsilon}\right)$
Lower Bound	Finite Horizon time-homogeneous Tabular/Linear	Yes	$\Omega\left(\frac{1}{d_m\epsilon^2}\right)$

Offline Policy Optimization. Offline policy optimization for infinite horizon MDP can date back to [Azar et al., 2013]. [Li et al., 2020] recently showed that a perturbed version of model-based planning can find ϵ-optimal policy within $\widetilde{O}\left(\frac{1}{d_m(1-\gamma)\epsilon^2}\right)$ queries of transitions in infinite horizon tabular MDP when the total reward is upper bounded by 1, that matches the lower bound up to logarithmic factors. For the finite

horizon time-inhomogeneous tabular MDP setting, [Yin et al., 2020] showed that model-based planning can identify ϵ-optimal policy with $\widetilde{O}\left(\frac{H}{d_m\epsilon^2}\right)$ episodes, that matches the lower bound for time-inhomogeneous MDP up to logarithmic factors. When it comes to finite horizon time-homogeneous tabular MDP and linrar MDP with anchor points, [Cui and Yang, 2020] provided a $\widetilde{O}\left(\frac{\min\{H,S,d\}}{d_m\epsilon^2}\right)$ episode complexity for model-based planning, which is $\min\{H, S, d\}$ away from the lower bound. We provide a $\widetilde{O}\left(\frac{1}{d_m\epsilon^2}+\frac{\min\{S,d\}}{d_m\epsilon}\right)$ episode complexity, that matches the lower bound up to logarithm factors and a higher order term.

A recent work [Yin et al., 2021] considered solving the offline policy optimization with model-free Q-learning with variance reduction. Although their algorithm can match the sample complexity lower bound $\Omega\left(\frac{H^2}{d_m\epsilon^2}\right)$ when per-step reward is upper bounded by 1, model-free algorithms generally need at least $\Omega(\frac{H^2}{d_m})$ episodes of data to finish the algorithm (a.k.a the sample size barrier in [Li et al., 2020]) and it is unclear how to translate their results to the setting with total reward upper bounded by 1.

Horizon-Free Online Reinforcement Learning. There are several works obtained nearly horizon-free sample complexity bounds for online reinforcement learning. In the time-homogeneous setting, whether the sample complexity needs to scale polynomially with H was an open problem raised by [Jiang and Agarwal, 2018]. The problem was first addressed by [Wang et al., 2020] who proposed an ϵ-net over the optimal policies and a simulation-based algorithms to obtain a sample complexity that only scales logarithmically with H, though their dependency on S, A and ϵ is not optimal and their algorithm is not computationally efficient. The sample complexity bound was later substantially improved by [Zhang et al., 2020b] who obtained an $\widetilde{O}\left(\frac{SA}{\epsilon^2}+\frac{S^2A}{\epsilon}\right)$ bound, which nearly matches the contextual bandits (tabular MDP with $H = 1$) lower bound $\Omega\left(\frac{SA}{\epsilon^2}\right)$ up to an $\frac{S^2A}{\epsilon}$ factor [Lattimore and Szepesvári, 2020]. Their key ideas are (1) a new bonus function and (2) a recursion-based approach to obtain a tight bound on the sample complexity. Such kind of recursion-based approach cannot be directly applied to the offline setting, and we develop a novel recursion method to suit the offline scenarios.

3 Problem Setup

Notation: Throughout this paper, we use $[N]$ to denote the set $\{1, 2, \cdots, N\}$ for $N \in \mathbb{Z}^+$, $\Delta(E)$ to denote the set of the probability measure over the event set E. Moreover, for simplicity, we use ι to denote $\mathrm{polylog}\,(S, A, H, d, 1/\delta)$ factors (that can be changed in the context), where δ is the failure probability. We use $\widetilde{O}$ and $\widetilde{\Omega}$ to denote the upper bound and lower bound up to logarithm factors.

3.1 Markov Decision Process

Markov Decision Process (MDP) is one of the most standard models studied in the reinforcement learning, usually denoted as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, R, P, \mu)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $R : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R}^+)$ is the reward, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition, and μ is the initial state distribution. We additionally define $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^+$ to denote the expected reward.

We focus on the episodic MDP with the horizon[1] $H \in \mathbb{Z}^+$ and time-homogeneous setting[2] that P

[1] A common belief is that we can always reproduce the results between episodic time-homogeneous setting and infinite horizon setting via substitute the horizon H in episodic setting with the "effective horizon" $\frac{1}{1-\gamma}$ in episodic setting. However, this argument does not always hold, for example the dependency decouple technique used in [Agarwal et al., 2020, Li et al., 2020] cannot be directly applied in the episodic setting.

[2] Some previous work consider time-inhomogeneous setting [e.g. Jin et al., 2018], where P and R can be varied for different

and R do not depend on the level $h \in [H]$. A (potentially non-stationary) policy π is defined as $\pi = (\pi_1, \pi_2, \cdots \pi_H)$, where $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$, $\forall h \in [H]$. Following the standard definition, we define the value function $V_h^\pi(s) := \mathbb{E}_{P,\pi}[\sum_{t=h}^H R(s_t, a_t)|s_h = s]$ and the action-value function (i.e. the Q-function) $Q_h^\pi(s,a) := \mathbb{E}_{P,\pi}[\sum_{t=h}^H R(s_t, a_t)|(s_h, a_h) = (s,a)]$, which are the expected cumulative rewards under the transition P and policy π starting from $s_h = s$ and $(s_h, a_h) = (s,a)$. Notice that, even though P and R keep invariant under the change of h, V and Q always depend on h in the episodic setting, which introduces additional technical difficulties compared with the infinite horizon setting.

The expected cumulative reward under policy π is defined as: $v^\pi := \mathbb{E}_\mu [V_1^\pi(s)]$, and our ultimate goal is finding the optimal policy π^* of $\mathcal{M}$, which can be written as: $\pi^* = \arg\max_\pi (v^\pi)$. We additionally define the reaching probabilities $\xi_h^\pi(s) = \mathbb{P}_{\mu,P,\pi}(s_h = s), \xi_h^\pi(s,a) = \mathbb{P}_{\mu,P,\pi}(s_h = s, a_h = a)$, that represent the probability of reaching state s and state-action pair (s,a) at time step h. Obviously we have $\sum_s \xi_h^\pi(s) = 1, \sum_{s,a} \xi_h^\pi(s,a) = 1, \forall h \in [H]$, and we also have the following relations between $\xi_h^\pi(s)$ and $\xi_h^\pi(s,a)$: $\xi_h^\pi(s,a) = \xi_h^\pi(s)\pi(a|s)$, $\xi_{h+1}^\pi(s') = \sum_{s,a} \xi_h^\pi(s,a) P(s'|s,a)$. With ξ_h^π at hand, we can write v^π in an equivalent way: $v^\pi = \sum_{s,a} \left(\left(\sum_{h\in[H]} \xi_h^\pi(s,a) \right) r(s,a) \right)$, which provides a dual perspective on the policy evaluation [Yang et al., 2020].

3.2 Offline Reinforcement Learning

Generally, P and r are not revealed to the learner, which means we can only learn about $\mathcal{M}$ and identify the optimal policy π^* with data from different kinds of sources. In offline reinforcement learning, the learner can only have access to a collection of data $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i\in[n]}$ where $r_i \sim R(s_i, a_i)$ and $s_i' \sim P(\cdot|s_i, a_i)$, that is collected in K episodes with (known or unknown) behavior policy (so that $n = KH$), For simplicity, define $n(s,a)$ as the number of data that $(s_i, a_i) = (s,a)$, while $n(s,a,s')$ is the number of data that $(s_i, a_i, s_i') = (s,a,s')$.

With $\mathcal{D}$, the learner is generally asked to do two kinds of tasks. The first one is the offline policy evaluation (a.k.a off-policy evaluation), that aims to estimate v^π for the given π. The second one is the offline policy optimization, that aims to find the $\hat{\pi}^*$ that can perform well on $\mathcal{M}$. We are interested in the statistical limit due to the limited number of data, and how to approach this statistical limit with simple and computationally efficient algorithms.

4 Offline Policy Evaluation

In this section, we first consider the offline policy evaluation for tabular MDP with number of state $S = |\mathcal{S}| < \infty$, number of action $A = |\mathcal{A}| < \infty$, which is the basis for the more general settings. We first introduce the plug-in estimator we consider, which is equivalent to different kinds of estimators that are widely used in practice. Then we describe the assumptions we use, show the error bound of the plug-in estimator, and provide the proof sketch of the error bound.

4.1 The Plug-in Estimator

Here we introduce the plug-in estimator. We first build the empirical MDP $\widehat{\mathcal{M}}$ with the data: $\hat{P}(s'|s,a) = \frac{n(s,a,s')}{n(s,a)}, \hat{r}(s,a) = \frac{\sum_{i\in[n]} r_i \mathbf{1}_{(s_i,a_i)=(s,a)}}{n(s,a)}$, where $\mathbf{1}$ is the indicator function. Then we correspondingly define

$h \in [H]$. It is noteworthy that we need an additional H factor in the sample complexity to identify ϵ-optimal policy for time-inhomogeneous MDP compared with time-homogeneous MDP. Transforming the improvement analysis from time-homogeneous setting to time-inhomogeneous setting is trivial (we only need to replace S with HS), but not vice-versa, as the analysis for time-inhomogeneous setting probably do not exploit the invariant transition sufficiently.

$\hat{Q}_h^\pi$, $\hat{V}_h^\pi$ and finally the estimator $\hat{v}^\pi$, $\forall h \in [H]$, by substituting the P and r in Q_h^π, V_h^π and v^π with $\hat{P}$ and $\hat{r}$. Such computation can be efficiently implemented with dynamic programming. We also introduce the reaching probabilities $\hat{\xi}_h^\pi(s) = \mathbb{P}_{\mu,\hat{P},\pi}(s_h = s), \hat{\xi}_h^\pi(s,a) = \mathbb{P}_{\mu,\hat{P},\pi}(s_h = s, a_h = a)$ in the empirical MDP $\widehat{\mathcal{M}}$, which will be helpful in our analysis.

The plug-in estimator has been studied in [Duan and Wang, 2020] under the assumption of linear transition. It's known that the plug-in estimator is equivalent to the MIS estimator proposed in [Yin and Wang, 2020] and a certain version of DualDICE estimator with batch update is proposed in [Nachum et al., 2019a], due to the observation that $\hat{v}^\pi = \sum_{s,a} \left(\left(\sum_{h\in[H]} \hat{\xi}_h^\pi(s,a) \right) \hat{r}(s,a) \right)$.

4.2 Theoretical Guarantee

Here we first summarize the assumptions we use for the tabular MDP.

Assumption 1 (Bounded Total Reward). *$\forall \pi$, we have $\sum_{h\in[H]} r_h \leq 1$ almost surely, where $s_1 \sim \mu$, $a_h \sim \pi(\cdot|s_h)$, $r_h \sim R(s_h, a_h)$ and $s_{h+1} \sim P(\cdot|s_h, a_h)$, $\forall h \in [H]$. This also means $\mathbb{P}(r \sim R(s,a)|r > 1) = 0$, $\forall (s,a)$.*

This is the key assumption used in [Wang et al., 2020, Zhang et al., 2020b] to escape the polynomial dependence of horizon in episodic setting. As discussed in Jiang and Agarwal [2018], Wang et al. [2020], Zhang et al. [2020b], this assumption is also more general than the *uniformly bounded reward* assumption: $\forall (s,a), r(s,a) \leq 1/H$. Thus, all of the results in this paper can be generalized to the uniformly bounded reward with a proper scaling of the bounded total reward.

Assumption 2 (Data Coverage). $\forall (s,a)$, $n(s,a) \geq n d_m$.

This assumption has been used in [Yin and Wang, 2020, Yin et al., 2020] and is similar to *concentration coefficient* assumption originated from [Munos, 2003]. Intuitively, the performance of the offline reinforcement learning should depend on d_m, since the state-action pair with less visitation will introduce more uncertainty.

Notice that, $d_m \in \left(0, (SA)^{-1}\right]$. Assuming access to the generative model, we can query equal number of samples from each state action pair, where $d_m = (SA)^{-1}$. For the offline data sampled with a fixed behavior policy π_{BEH}, we can view $d_m \approx \frac{1}{H} \min_{s,a} \sum_{h\in[H]} \xi_h^{\pi_{\mathrm{BEH}}}(s,a)$, which measures the quality of exploration for π_{BEH}. When the number of episodes $K = \widetilde{\Omega}\left(1/d_m\right)$, by standard concentration, we know $\min_{s,a} n(s,a) \approx n d_m = HK d_m$.

We remark that, some of the recent Liu et al. [2019], Yin et al. [2021] define the data coverage via the coverage on the visitation of the optimal policy when considering offline policy optimization. This kind of data coverage can be covered by our Assumption 2, by considering the sub-MDP which only consists of the state-action pair that can be visited by the optimal policy. As we already know that the optimal policy will not leave this sub-MDP, any near-optimal policy on this sub-MDP will be near-optimal on the original MDP, and hence our results can be directly applied under this alternative definition of data coverage.

With these assumptions at hand, we can present our main results.

Theorem 1. *Under Assumption 1 and Assumption 2, suppose $K = \widetilde{\Omega}\left(1/d_m\right)$, then*

$$|v^\pi - \hat{v}^\pi| \leq \sqrt{\frac{\iota}{K d_m}}$$

holds with probability at least $1-\delta$, where K is the number of episodes, d_m is the minimum visiting probability and ι absorbs the poly log *factors.*

To demonstrate the tightness of our upper bound, we provide a minimax lower bound in Theorem 2[3].

Theorem 2. *There exists a pair of MDPs $\mathcal{M}_1$ and $\mathcal{M}_2$, and offline data $\mathcal{D}$ with $|\mathcal{D}| = KH$ and minimum visiting probability d_m, such that for some absolute constant c_0, we have*

$$\inf_{\hat{v}^\pi} \sup_{\mathcal{M}_i \in \{\mathcal{M}_1, \mathcal{M}_2\}} \mathbb{P}_{\mathcal{M}_i} \left(|\hat{v}^\pi(\mathcal{D}) - v^\pi| > \frac{c_0}{\sqrt{Kd_m}} \right) > 0.25,$$

where $\hat{v}^\pi$ is any estimator that takes $\mathcal{D}$ as input.

Remark Theorem 1 and 2 show that, even with the simplest plug-in estimator, we can match the minimax lower bound $\Omega\left(\sqrt{\frac{1}{Kd_m}}\right)$ for offline policy evaluation in time-homogeneous MDP up to logarithmic factors. As a result, we can conclude that time-homogeneous MDP is not harder than the bandits in offline policy evaluation.

Remark The assumption that $K = \widetilde{\Omega}\left(1/d_m\right)$ is a mild and necessary assumption, as with only $o\left(1/d_m\right)$ episodes, there can exist some under-explored state-action pair which unavoidably leads to constant error (this is also how we construct the hard instance for the minimax lower bound).

Remark Central to our analysis is a recursion based upper bound on the "total variance" term (see Lemma 1), which enables us to have a sharper bound for the plug-in estimator compared with previous work [e.g. Yin and Wang, 2020, Yin et al., 2020] that include the unnecessary $\mathrm{poly}(S, A)$ factors in the higher-order term.

4.3 Proof Sketch

The detailed proof can be found in Appendix A, and here we sketch our proof in short. With the value difference lemma (see Lemma 4 in the Appendix A), we need to focus on bounding the term:

$$\textstyle\sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^\pi(s,a) \left[\sum_{s'} \left(P(s'|s,a) - \hat{P}(s'|s,a) \right) V_{h+1}^\pi(s') \right],$$

which, by Bernstein's inequality and Cauchy-Schwartz inequality associated with the Assumption 1 and Assumption 2, can be upper bounded by the following term with high probability:

$$\sqrt{\tfrac{\iota}{Kd_m}} \sqrt{\textstyle\sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^\pi(s,a) \mathrm{Var}_{P(s,a)}(V_{h+1}^\pi(s'))} + \tfrac{\iota}{Kd_m}$$

To bound the "total variance" term $\sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^\pi(s,a) \mathrm{Var}_{P(s,a)}[V_{h+1}^\pi(s')]$, we use a novel recursion-based method based on the following observation:

Lemma 1 (Variance Recursion For Evaluation)**.** *For the value function $V_h^\pi(s)$ induced by any π, we have that*

$$\begin{aligned}
&\sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^\pi(s,a) \mathrm{Var}_{P(s,a)} \left[V_{h+1}(s')^{2^i} \right] \\
&\leq \sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^\pi(s,a) \left[\sum_{s'} \left(P(s'|s,a) - \hat{P}(s'|s,a) \right) V_{h+1}(s')^{2^{i+1}} \right] + 2^{i+1} \qquad (1)
\end{aligned}$$

[3] To the best of our knowledge, no lower bound has been provided for finite horizon time-homogeneous setting, so here we provide a minimax lower bound, and the proof can be found in Appendix B.

Here the term in (11) can be bounded with $\sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^\pi(s,a)\mathrm{Var}_{P(s,a)}\left[V_{h+1}(s')^{2^{i+1}}\right]$, the total variance of higher-order value function via Bernstein's inequality. Thus Lemma 11 can be applied iteratively to obtain a tight bound for the "total variance" term. Define

$$\Delta_1(i) = \left|\sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^\pi(s,a)\left[\sum_{s'}\left(P(s'|s,a)-\hat{P}(s'|s,a)\right)V_{h+1}^\pi(s')^{2^i}\right]\right|.$$

Applying Lemma 11 with $V_h^\pi(s)$, we can write the following recursion:

$$\Delta_1(i) \le \sqrt{\tfrac{\iota}{Kd_m}\left(\Delta_1(i+1)+2^{i+1}\right)} + \tfrac{\iota}{Kd_m}.$$

Solve this recursion, and we can obtain the results in Theorem 11.

5 Offline Policy Optimization

In this section, we further consider the offline policy optimization for tabular MDP, which is the ultimate goal for offline reinforcement learning. We first introduce the model-based planning algorithm, which is probably the simplest algorithm for offline policy optimization. Then we analyze the performance gap between the policy obtained by model-based planning and the optimal policy.

5.1 Model-Based Planning

We consider the optimal policy on the empirical MDP $\widehat{\mathcal{M}}$, which can be defined as

$$\hat{\pi}^* = \arg\max_\pi \hat{v}^\pi. \tag{2}$$

Here $\hat{\pi}^*$ can be obtained by dynamic programming with the empirical MDP, which is also known as model-based planning. We remark that our analysis is independent to the algorithm used for solving (2). In other words, the result also applies to the optimization-based planning with the empirical MDPs [Puterman, 2014, Mohri et al., 2012], as long as it solves (2).

5.2 Theoretical Guarantee

Theorem 3 provides an upper bound on the sub-optimality of $\hat{\pi}^*$.

Theorem 3. *Under Assumption 1 and Assumption 2, suppose that $K = \widetilde{\Omega}\left(1/d_m\right)$, and then*

$$\left|v^{\pi^*} - v^{\hat{\pi}^*}\right| \le \sqrt{\tfrac{\iota}{Kd_m}} + \tfrac{S\iota}{Kd_m},$$

holds with probability at least $1-\delta$, where K is the number of episodes, d_m is the minimum visiting probability and ι absorbs the poly log *factors.*

We also provide a minimax lower bound under the finite horizon time-homogeneous setting.

Theorem 4. *There exists a pair of MDPs $\mathcal{M}_1$ and $\mathcal{M}_2$, and offline data $\mathcal{D}$ with $|\mathcal{D}| = KH$ and minimum state-action pair visiting frequency d_m, such that for some absolute constant c_0, we have*

$$\inf_{\hat{v}^\pi}\sup_{\mathcal{M}_i\in\{\mathcal{M}_1,\mathcal{M}_2\}}\mathbb{P}_{\mathcal{M}_i}\left(\left|v^{\hat{\pi}(\mathcal{D})} - v^{\pi^*}\right| > \tfrac{c_0}{\sqrt{Kd_m}}\right) > 0.25,$$

where $\hat{\pi}$ is any planning algorithm that takes $\mathcal{D}$ as input.

Remark Theorem 3 provides a bound approaching the minimax lower bound in Theorem 4 up to logarithmic factors and a higher-order term, which shows that the error of offline policy optimization does not scale polynomially on the horizon. Notice that if $d_m = \Omega\left(\frac{1}{SA}\right)$, we can obtain an error bound of $\widetilde{O}\left(\sqrt{\frac{SA}{K}} + \frac{S^2A}{K}\right)$, which can be translated to sample complexity $\widetilde{O}\left(\frac{SA}{\epsilon^2} + \frac{S^2A}{\epsilon}\right)$ that matches the best known result of sample complexity for online finite-horizon time-homogeneous setting [Zhang et al., 2020b]. We conjecture that the additional S factor in the higher-order term is only an artifact (see Lemma 2) and can be eliminated with more delicate analysis. We leave this as an open problem.

Remark There are also works considering local policy optimization [e.g. Kakade and Langford, 2002, Liu et al., 2020b, Kumar et al., 2020] when the offline data are not sufficient exploratory. We want to emphasize that, as Theorem 4 suggests, to identify the global optimal policy, we need the offline data sufficient exploratory.

5.3 Proof Sketch

The detailed proof can be found in Appendix A, and here we sketch our proof in short. Notice that

$$\begin{aligned} v^{\pi^*} - v^{\hat{\pi}^*} =& v^{\pi^*} - \hat{v}^{\pi^*} + \underbrace{\hat{v}^{\pi^*} - \hat{v}^{\hat{\pi}^*}}_{\leq 0} + \hat{v}^{\hat{\pi}^*} - v^{\hat{\pi}^*} \\ \leq & \underbrace{v^{\pi^*} - \hat{v}^{\pi^*}}_{\text{Error on Fixed Policy}} + \underbrace{\hat{v}^{\hat{\pi}^*} - v^{\hat{\pi}^*}}_{\text{Error on Data-Dependent Policy}} . \end{aligned} \tag{3}$$

We can directly apply Theorem 1 to bound the error on fixed policy. For the error on data-dependent policy, since the policy $\hat{\pi}^*$ depends on data, we need to consider

$$\sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^{\pi}(s,a) \left[\sum_{s'} \left(P(s'|s,a) - \hat{P}(s'|s,a) \right) V_{h+1}^{\hat{\pi}^*}(s') \right]$$

$$= \sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^{\pi}(s,a) \left[\sum_{s'} \left(P(s'|s,a) - \hat{P}(s'|s,a) \right) V_{h+1}^{\pi^*}(s') \right] \tag{4}$$

$$+ \sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^{\pi}(s,a) \left[\sum_{s'} \left(P(s'|s,a) - \hat{P}(s'|s,a) \right) \left(V_{h+1}^{\hat{\pi}^*}(s') - V_{h+1}^{\hat{\pi}^*}(s') \right) \right] . \tag{5}$$

As π^* is independent of $\hat{P}$, the term in (4) can be similarly handled with the techniques used in offline policy evaluation. However, due to the dependency of $\hat{\pi}^*$ and $\hat{P}$, we need to deal with the term in (5) more carefully. To decouple the dependency of $\hat{\pi}^*$ and $\hat{P}$, we first introduce the following lemma:

Lemma 2. *$\forall V_h(s) \in [0,1]$, $\forall h \in [H]$, $s \in \mathcal{S}$, then we have that with high probability,*

$$\left| \sum_{s'} \left(P(s'|s,a) - \hat{P}(s'|s,a) \right) V_h(s') \right| \leq \sqrt{\frac{S \cdot \mathrm{Var}_{P(s,a)}[V_h(s')]\iota}{n(s,a)}} + \frac{S\iota}{n(s,a)}.$$

Remark Lemma 2 has been widely used in the design and analysis of online reinforcement learning algorithms, [e.g. Zhang et al., 2020b]. It holds even $V_h(s)$ depends on $\hat{P}(s'|s,a)$, however, at the cost of an additional S factor. This is the source of the additional S factor in the higher-order term, and we believe a more fine-grained analysis can help avoid this additional S factor.

We also have the following recursion for (4) and (5):

Lemma 3 (Variance Recursion For Optimization). *For $V_h(s) = V_h^{\pi^*}(s)$ and $V_h(s) = V_h^{\pi^*}(s) - V_h^{\hat{\pi}^*}(s)$, we have that*

$$\sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^{\hat{\pi}^*}(s,a)\mathrm{Var}_{P(s,a)}\left[V_{h+1}(s')^{2^i}\right]$$

$$\leq \sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^{\hat{\pi}^*}(s,a)\left[\sum_{s'}\left(P(s'|s,a)-\hat{P}(s'|s,a)\right)V_{h+1}(s')^{2^{i+1}}\right] \tag{6}$$

$$+ 2^{i+1}\left[\sum_s \mu(s)V_1(s) + \sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^{\hat{\pi}^*}(s,a)\sum_{s'}\left[P(s'|s,a)-\hat{P}(s'|s,a)\right]V_{h+1}(s')\right]. \tag{7}$$

Denote

$$\Delta_2(i) = \left|\sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^{\pi}(s,a)\left[\sum_{s'}\left(P(s'|s,a)-\hat{P}(s'|s,a)\right)V_{h+1}^{\pi^*}(s')^{2^i}\right]\right|,$$

$$\Delta_3(i) = \left|\sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^{\pi}(s,a)\left[\sum_{s'}\left(P(s'|s,a)-\hat{P}(s'|s,a)\right)\left(V_{h+1}^{\pi^*}(s')-V_{h+1}^{\hat{\pi}^*}(s')\right)^{2^i}\right]\right|.$$

Applying Bernstein inequality or Lemma 2, and then Lemma 3, we have that:

$$\Delta_2(i) \leq \sqrt{\frac{\iota}{Kd_m}\left(\Delta_2(i+1) + 2^{i+1}(v^{\pi^*}+\Delta_2(0))\right)} + \frac{\iota}{Kd_m}$$

$$\Delta_3(i) \leq \sqrt{\frac{S\iota}{Kd_m}\left(\Delta_2(i+1) + 2^{i+1}(v^{\pi^*}-v^{\hat{\pi}^*}+\Delta_3(0))\right)} + \frac{S\iota}{Kd_m}.$$

With Assumption 1, we know $v^{\pi^*} \leq 1$. Also, with (3) and (4), we have that $v^{\pi^*} - v^{\hat{\pi}^*} \leq \Delta_3(0) + O\left(\sqrt{\frac{\iota}{Kd_m}}\right)$. Solve this recursion, then we can obtain the results in Theorem 3.

6 Extensions to Linear MDP with Anchor Points

In this section, we first introduce the definition of the linear MDP with anchor points [Yang and Wang, 2019, Cui and Yang, 2020], and then generalize our results of offline policy evaluation and optimization to this setting.

Definition 1 (Linear MDP with Anchor Points [Yang and Wang, 2019, Cui and Yang, 2020]). *For the MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, R, P, \mu)$, assume there is a feature map $\phi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$, such that r and P admits a linear representation:*

$$r(s,a) = \langle\phi(s,a), \theta_r\rangle \quad P(s'|s,a) = \langle\phi(s,a), \mu(s')\rangle,$$

where μ is an unknown (signed) measure of $\mathcal{S}$. Furthermore, we assume there exists a set of anchor state-action pairs $\mathcal{K}$, such that $\forall(s,a) \in \mathcal{S} \times \mathcal{A}$, $\exists\{\lambda_k^{s,a}\}_{k\in\mathcal{K}}$,

$$\phi(s,a) = \sum_{k\in\mathcal{K}} \lambda_k^{s,a}\phi(s_k,a_k), \quad \sum_{k\in\mathcal{K}} \lambda_k^{s,a} = 1, \quad \lambda_k^{s,a} \geq 0, \forall k \in \mathcal{K}.$$

With the definition of Linear MDP as well as the anchor points assumption, we can find that $r(s,a) = \sum_{k\in\mathcal{K}} \lambda_k^{s,a} r(s_k,a_k)$, $P(s'|s,a) = \sum_{k\in\mathcal{K}} \lambda_k^{s,a} P(s'|s_k,a_k)$, which can lead to an empirical estimation of $\hat{P}$ and $\hat{r}$ by replacing $\{r(s_k,a_k)\}_{k\in\mathcal{K}}$, $\{P(s'|s_k,a_k)\}_{k\in\mathcal{K}}$ with the empirical counterpart $\{\hat{r}(s_k,a_k)\}_{k\in\mathcal{K}}$, $\{\hat{P}(s'|s_k,a_k)\}_{k\in\mathcal{K}}$ estimated from the offline data. Following [Yang and Wang, 2019, Cui and Yang, 2020], we additionally make the following assumption on the offline data:

Assumption 3 (Anchor Point Data [Yang and Wang, 2019, Cui and Yang, 2020])**.** *Assume $|\mathcal{K}| = d$. For the offline data $\mathcal{D} = \{(s_i,a_i,r_i,s_i')\}_{i\in[n]}$, $(s_i,a_i) \in \{(s_k,a_k)\}_{k\in\mathcal{K}}$, $\forall i \in [n]$. Furthermore, $\forall k \in \mathcal{K}$, $n(s_k,a_k) \geq nd_m$.*

We now present the main theorem on the offline policy evaluation and offline policy improvement on the linear MDP with anchor points.

Theorem 5. *Under Assumption 1 and Assumption 3, suppose $n = \widetilde{\Omega}(H/d_m)$, and then for a given policy π, the plug-in estimator $\hat{v}^\pi$ satisfies*

$$|v^\pi - \hat{v}^\pi| \leq \sqrt{\frac{H\iota}{nd_m}}$$

with probability at least $1-\delta$, where n is the number of offline data, d_m is the minimum visiting probability of anchor points and ι absorbs the polylog *factors.*

Theorem 6. *Under Assumption 1 and Assumption 3, suppose that $n = \widetilde{\Omega}(H/d_m)$, and then for the $\hat{\pi}^*$ obtained by model-based planning,*

$$\left|v^{\pi^*} - v^{\hat{\pi}^*}\right| \leq \sqrt{\frac{H\iota}{nd_m}} + \frac{dH\iota}{nd_m},$$

holds with probability at least $1-\delta$, where n is the number of offline data, d is the feature dimension, d_m is the minimum visiting probability of anchor points and ι absorbs the polylog *factors.*

Proof of both theorems can be found in Appendix C. From a high-level perspective, we observe that for the unseen state-action pair, we still have Bernstein-type concentration bound and a lemma similar to Lemma 2. Hence we can apply our recursion introduced in Lemma 1 and use the similar techniques for tabular MDP to obtain the desired results. We want to remark that such results demonstrate the broad applicability of our recursion-based analysis in different kinds of offline scenarios.

Remark Compared with the results in [Cui and Yang, 2020], we remove the additional dependency of $\min\{H,|\mathcal{S}|,d\}$ in the main term and approach the optimal complexity shown in the [Yang and Wang, 2019] up to logarithmic factors, which shows that model-based planning is minimax optimal for Linear MDP with anchor points and demonstrates the effectiveness of our recursion-based analysis.

Remark Here we do not directly replace the term n/H with K, as we make relatively strong assumption on all of the offline data are sampled from the anchor points. One potential question is how general the anchor point assumption is. We notice that, as [Duan and Wang, 2020] showed, the popular FQI algorithm for linear MDP provides an estimate of $\hat{P}$ as

$$\hat{P}(s'|s,a) = \phi(s,a)^\top \hat{\Lambda}^{-1} \sum_{i\in[n]} \phi(s_i,a_i)\mathbf{1}_{s'=s'_i},$$

where $\hat{\Lambda} = \sum_{i\in[n]} \phi(s_i,a_i)\phi(s_i,a_i)^\top$. As $\sum_{s'} P(s'|s,a) = 1$, by Definition 1,

$$1 = \sum_{s'} P(s'|s,a) = \phi(s,a)^\top \sum_{s'} \mu(s') = \phi(s,a)^\top \hat{\Lambda}^{-1} \left(\sum_{i\in[n]} \phi(s_i,a_i)\right).$$

Thus, if $\phi(s,a)\hat{\Lambda}^{-1}\phi(s_i,a_i) \geq 0$, $\forall (s,a), i$, then it forms a linear MDP with anchor points with $\mathcal{K} = \{(s_i,a_i)\}_{i\in[n]}$, $\lambda_k^{s,a} = \phi(s,a)\hat{\Lambda}^{-1}\phi(s_k,a_k)$ and our analysis can be simply adapted to this case where nd_m is replaced with $\lambda_{\min}(\hat{\Lambda})$ using Freedman's inequality [Freedman, 1975], and $\lambda_{\min}(\hat{\Lambda})$ denotes the minimum eigenvalue of $\hat{\Lambda}$. We notice that if $\phi(s,a)\hat{\Lambda}^{-1}\phi(s_i,a_i) < 0$, and s'_i are distinct for each i (which is probable for exponential large $|\mathcal{S}|$), then $\hat{P}(s'_i|s,a) < 0$, which can be pathological for the algorithm and analysis. We expect a well-behaved transition estimation (*i.e.*$\hat{P}$) for linear MDP shares the similar results, even without anchor point assumptions. We leave this as an open problem.

7 Conclusion

In this paper, we revisit the offline reinforcement learning on episodic time-homogeneous MDP. We show that, if the total reward is properly normalized, offline reinforcement learning is not harder than the offline bandits counterparts. Specifically, we provide performance guarantee for algorithms based on empirical MDPs, that match the lower bound up to logarithmic factors for offline policy evaluation, and up to logarithmic factors and a higher-order term for offline policy optimization, and both do not have polynomial dependency on H. There are still several open problems. For example, can we provide a sharper analysis for the policy obtained by model-based planning without any additional factors on higher-order term? Can we extend to MDP with more general assumption? We leave these problems as future work.

Acknowledgement

The authors thank the anonymous reviewer for their constructive feedback. TR would like to thank for the helpful discussion with Ming Yin and Yu Bai. SS gratefully acknowledges the funding from NSF grants 1564000 and 1934932, and SSD gratefully acknowledges the funding from NSF Award's IIS-2110170 and DMS-2134106.

References

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020.

OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349, 2013.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1042–1051. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/chen19e.html.

Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 456–464, 2019.

Qiwen Cui and Lin F Yang. Is plug-in solver sample-efficient for feature-based reinforcement learning? *arXiv preprint arXiv:2010.05673*, 2020.

Hanjun Dai, Elias B Khalil, Yuyu Zhang, Bistra Dilkina, and Le Song. Learning combinatorial optimization algorithms over graphs. *arXiv preprint arXiv:1704.01665*, 2017.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: uniform pac bounds for episodic reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5717–5727, 2017.

Yaqi Duan and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. *arXiv preprint arXiv:2002.09516*, 2020.

Yihao Feng, Tongzheng Ren, Ziyang Tang, and Qiang Liu. Accountable off-policy evaluation with kernel bellman statistics. In *International Conference on Machine Learning*, pages 3102–3111. PMLR, 2020.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

Carles Gelada and Marc G Bellemare. Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3647–3655, 2019.

Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. Approximating interactive human evaluation with self-play for open-domain dialog systems. *arXiv preprint arXiv:1906.09308*, 2019.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25 (1):16–18, 2019.

Assaf Hallak and Shie Mannor. Consistent on-line off-policy evaluation. In *International Conference on Machine Learning*, pages 1372–1383. PMLR, 2017.

Nan Jiang and Alekh Agarwal. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pages 3395–3398. PMLR, 2018.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR, 2016.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 4863–4873. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/d3b1fb02964aa64e257f9f26a31f72cf-Paper.pdf.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

Nathan Kallus and Masatoshi Uehara. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *arXiv preprint arXiv:1909.05850*, 2019.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*, 2020.

Tor Lattimore and Marcus Hutter. Near-optimal pac bounds for discounted mdps. *Theoretical Computer Science*, 558:125–143, 2014.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *arXiv preprint arXiv:2005.12900*, 2020.

Lihong Li, Rémi Munos, and Csaba Szepesvári. Toward minimax off-policy value estimation. In *Artificial Intelligence and Statistics*, pages 608–616. PMLR, 2015.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *arXiv preprint arXiv:1810.12429*, 2018.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Off-policy policy gradient with state distribution correction. *arXiv preprint arXiv:1904.08473*, 2019.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Off-policy policy gradient with stationary distribution correction. In *Uncertainty in Artificial Intelligence*, pages 1180–1190. PMLR, 2020a.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*, 2020b.

Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. In *AAMAS*, pages 1077–1084, 2014.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012. ISBN 026201825X.

Rémi Munos. Error bounds for approximate policy iteration. In *ICML*, volume 3, pages 560–567, 2003.

Susan A Murphy, Mark J van der Laan, James M Robins, and Conduct Problems Prevention Research Group. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456): 1410–1423, 2001.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, pages 2318–2328, 2019a.

Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019b.

Ashwin Pananjady and Martin J Wainwright. Instance-dependent l-infinity-bounds for policy evaluation in tabular reinforcement learning. *IEEE Transactions on Information Theory*, 67(1):566–585, 2020.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR, 2020.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Ruosong Wang, Simon S Du, Lin F Yang, and Sham M Kakade. Is long horizon reinforcement learning more difficult than short horizon reinforcement learning? *arXiv preprint arXiv:2005.00527*, 2020.

Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In Jonas Peters and David Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 550–559. PMLR, 03–06 Aug 2020a. URL http://proceedings.mlr.press/v124/xie20a.html.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. *arXiv preprint arXiv:2008.04990*, 2020b.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *arXiv preprint arXiv:1906.03393*, 2019.

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

Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-policy evaluation via the regularized lagrangian. *arXiv preprint arXiv:2007.03438*, 2020.

Ming Yin and Yu-Xiang Wang. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3948–3958. PMLR, 26–28 Aug 2020.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near optimal provable uniform convergence in off-policy evaluation for reinforcement learning. *arXiv preprint arXiv:2007.03760*, 2020.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal offline reinforcement learning via double variance reduction, 2021.

Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. Gendice: Generalized offline estimation of stationary values. *arXiv preprint arXiv:2002.09072*, 2020a.

Zihan Zhang, Xiangyang Ji, and Simon S Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. *arXiv preprint arXiv:2009.13503*, 2020b.

A Proof of the Main Theorems

A.1 Proof for Offline Policy Evaluation

The proof of all of the technical lemmas can be found in Appendix E. Our proof is organized as follows: we first decompose the estimation error to the errors introduced by reward estimation $\hat{r}$ and transition estimation $\hat{P}$ with Lemma 4, then Lemma 5 provides an upper bound for the error introduced by $\hat{r}$. For error introduced by $\hat{P}$, we first show it can be upper bounded by a "total variance" term (in (9)). A naive bound for this "total variance" term will introduce an additional H factors, thus we apply a recursion-based method to upper bound this "total variance" term (see Lemma 1). Solve the recursion in Lemma 6, and put everything together, we eventually obtain the bound in Theorem 1.

Lemma 4 (Value Difference Lemma)**.**

$$v^{\pi} - \hat{v}^{\pi} = \sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^{\pi}(s,a) \left[r(s,a) - \hat{r}(s,a) + \sum_{s'} \left[\left(P(s'|s,a) - \hat{P}(s'|s,a) \right) V_{h+1}^{\pi}(s') \right] \right]. \tag{8}$$

The following lemma provides an upper bound on the estimation error introduced by $\hat{r}$, *i.e.*, the first term in (8).

Lemma 5 (Error from Reward Estimation)**.** *Suppose Assumption 1 holds, then we have that*

$$\left| \sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^{\pi}(s,a) \left[r(s,a) - \hat{r}(s,a) \right] \right| \leq \sqrt{\frac{\iota}{Kd_m}} + \frac{\iota}{Kd_m},$$

holds with probability at least $1-\delta$.

We then use a recursive method to bound the error introduced by $\hat{P}$, *i.e.*, the second term in (8). Denote

$$\Delta_1 := \sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^{\pi}(s,a) \left[\sum_{s'} \left(P(s'|s,a) - \hat{P}(s'|s,a) \right) V_{h+1}^{\pi}(s') \right].$$

With Bernstein's inequality, we know that with high probability

$$\begin{aligned} \Delta_1 &\leq \sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^{\pi}(s,a) \left[\sqrt{\frac{\mathrm{Var}_{P(s,a)}(V_{h+1}^{\pi}(s'))\iota}{n(s,a)}} + \frac{\iota}{n(s,a)} \right] \\ &\leq \sqrt{\frac{\iota}{Kd_m}} \sqrt{\sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^{\pi}(s,a) \mathrm{Var}_{P(s,a)}(V_{h+1}^{\pi}(s'))} + \frac{\iota}{Kd_m}, \end{aligned} \tag{9}$$

where the second inequality is due to Cauchy-Schwartz inequality associated with the Assumption 2.

We then upper bound $\sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^{\pi}(s,a) \mathrm{Var}_{P(s,a)}(V_{h+1}^{\pi}(s'))$ in (9) with Lemma 1. Again, by Bernstein's inequality and Cauchy-Schwarz inequality, we have the following with high probability:

$$\begin{aligned} &\sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^{\pi}(s,a) \left[\sum_{s'} \left(P(s'|s,a) - \hat{P}(s'|s,a) \right) V_{h+1}^{\pi}(s')^{2^{i+1}} \right] \\ &\leq \sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^{\pi}(s,a) \left[\sqrt{\frac{\mathrm{Var}_{P(s,a)}\left(V_{h+1}^{\pi}(s')^{2^{i+1}}\right)\iota}{n(s,a)}} + \frac{\iota}{n(s,a)} \right] \end{aligned}$$

$$\leq \sqrt{\frac{\iota}{Kd_m}} \sqrt{\sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^\pi(s,a) \mathrm{Var}_{P(s,a)} \left(V_{h+1}^\pi(s')^{2^{i+1}}\right)} + \frac{\iota}{Kd_m}. \tag{10}$$

Define

$$\mathbb{V}_1(i) := \sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^\pi(s,a) \mathrm{Var}_{P(s,a)} \left(V_{h+1}^\pi(s')^{2^{i+1}}\right).$$

Apply Lemma 1 with (10), we have the recursion as

$$\mathbb{V}_1(i) \leq \sqrt{\frac{\iota}{Kd_m} \mathbb{V}_1(i+1)} + \frac{\iota}{Kd_m} + 2^{i+1}. \tag{11}$$

Notice that $\mathbb{V}_1(i) \leq H, \forall i$. Now we can solve the recursion with the following lemma:

Lemma 6. *For the recursion formula:*

$$\mathbb{V}(i) \leq \sqrt{\lambda_1 \mathbb{V}(i+1)} + \lambda_1 + 2^{i+1}\lambda_2,$$

with $\lambda_1, \lambda_2 > 0$, if $\mathbb{V}(i) \leq H, \forall i$, then we have that

$$\mathbb{V}(0) \leq 6(\lambda_1 + \lambda_2),$$

and we need to do the recursion at most $O(\log \frac{H\lambda_1}{\lambda_2^2})$ times.

Remark We want to emphasize that, the recursion needs to be done at most $O(\log \frac{H\lambda_1}{\lambda_2})$ times. Thus, by union bound, such recursion only introduces an additional $\log\log$ factor in the error when $\lambda_1, \lambda_2 = \mathrm{poly}(S, A, H)$, that can be absorbed by ι. For simplicity, we still use ι to denote the $\mathrm{poly}\log$ factors in the following derivation.

Apply Lemma 6 with $\lambda_1 = \frac{\iota}{Kd_m}$, $\lambda_2 = 1$, we have that

$$\mathbb{V}_1(0) = O\left(\frac{\iota}{Kd_m} + 1\right),$$

and also notice in (9) that

$$\Delta_1 \leq \sqrt{\frac{\iota}{Kd_m} \mathbb{V}_1(0)} + \frac{\iota}{Kd_m}.$$

Combine this two inequality, we have that

$$\Delta_1 \leq O\left(\sqrt{\frac{\iota}{Kd_m}\left(\frac{\iota}{Kd_m} + 1\right)}\right) + \frac{\iota}{Kd_m}.$$

Suppose $K = \widetilde{\Omega}\left(\frac{1}{d_m}\right)$, we have that

$$\Delta_1 \leq O\left(\sqrt{\frac{\iota}{Kd_m}} + \frac{\iota}{Kd_m}\right). \tag{12}$$

Combined (12) with Lemma 5 and $K = \widetilde{\Omega}\left(\frac{1}{d_m}\right)$, we conclude the proof of Theorem 1.

A.2 Proof for Offline Policy Optimization

We first make the following standard decomposition:

$$\begin{aligned} v^{\pi^*} - v^{\hat{\pi}^*} =& v^{\pi^*} - \hat{v}^{\pi^*} + \underbrace{\hat{v}^{\pi^*} - \hat{v}^{\hat{\pi}^*}}_{\leq 0} + \hat{v}^{\hat{\pi}^*} - v^{\hat{\pi}^*} \\ \leq & \underbrace{v^{\pi^*} - \hat{v}^{\pi^*}}_{\text{Error on Fixed Policy}} + \underbrace{\hat{v}^{\hat{\pi}^*} - v^{\hat{\pi}^*}}_{\text{Error on Data-Dependent Policy}} . \end{aligned} \tag{13}$$

The first term characterizes the evaluation difference of optimal policy on original MDP and the empirical MDP, and the second term characterize the evaluation difference of the planning result $\hat{\pi}^*$ from the empirical MDP on original MDP and the empirical MDP.

We can directly apply Theorem 1 to bound the first term in (13). However, as $\hat{\pi}^*$ has complicated statistical dependency with $\hat{P}$, we cannot apply Theorem 1 on the second term in (13) for the evaluation error on data-dependent policy. Notice that a direct application of the absorbing MDP techniques introduced in [Agarwal et al., 2020, Li et al., 2020] for the second term will introduce additional H or S factors in the main term as shown in [Cui and Yang, 2020]. Thus, we further generalize our recursion-based method to keep the main term tight while only introduce an additional S factor at the higher-order term, which keeps the final error horizon-free.

Similar to the case in the offline evaluation, we make the following decomposition based on Lemma 4:

$$\begin{aligned} \hat{v}^{\hat{\pi}^*} - v^{\hat{\pi}^*} &= \sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^{\hat{\pi}^*}(s,a) \left[(\hat{r}(s,a) - r(s,a)) + \sum_{s'} (\hat{P}(s'|s,a) - P(s'|s,a)) V_{h+1}^{\hat{\pi}^*}(s') \right] \\ &= \sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^{\hat{\pi}^*}(s,a) \left(\Delta_r(s,a) + \Delta_P(s,a) + \Delta_{PV}(s,a) \right), \end{aligned}$$

where

$$\begin{aligned} \Delta_r(s,a) &:= \widehat{r}(s,a) - r(s,a), \\ \Delta_P(s,a) &:= \sum_{s'} (\hat{P}(s'|s,a) - P(s'|s,a)) V_{h+1}^{\pi^*}(s') \\ \Delta_{PV}(s,a) &:= \sum_{s'} \left(\hat{P}(s'|s,a) - P(s'|s,a) \right) \left(V_{h+1}^{\hat{\pi}^*}(s') - V_{h+1}^{\pi^*}(s') \right). \end{aligned}$$

For the inner product of $\hat{\xi}_h^{\hat{\pi}^*}(s,a)$ with Δ_r, as r is independent of $\hat{P}$, we can identically apply the result for offline evaluation, that leads to a $\tilde{O}\left(\sqrt{\frac{1}{Kd_m}} + \frac{1}{Kd_m}\right)$ error. We then consider the error induced by Δ_P and Δ_{PV}.

For the error introduced by Δ_P, as π^* is independent of $\hat{P}$, we can use Bernstein inequality and Cauchy-Schwartz inequality and obtain

$$\begin{aligned} \Delta_2 &:= \left| \sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^{\pi}(s,a) \Delta_P(s,a) \right| \\ &\leq \sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^{\pi}(s,a) \cdot \left[\sqrt{\frac{\cdot \mathrm{Var}_{P(s,a)}\left(V_{h+1}^{\pi^*}(s')\right) \iota}{n(s,a)}} + \frac{\iota}{n(s,a)} \right] \end{aligned}$$

$$\leq \sqrt{\frac{\iota}{Kd_m}} \sqrt{\sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^{\pi}(s,a) \mathrm{Var}_{P(s,a)} \left(V_{h+1}^{\pi^*}(s')\right)} + \frac{\iota}{Kd_m}. \tag{14}$$

Now we turn to $\sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^{\pi}(s,a) \mathrm{Var}_{P(s,a)} \left(V_{h+1}^{\pi^*}(s')\right)$. With Lemma 3, we have that

$$\begin{aligned}
&\sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^{\pi}(s,a) \mathrm{Var}_{P(s,a)} \left((V_{h+1}^{\pi^*}(s'))^{2^i}\right) \\
\leq &\sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^{\pi}(s,a) \left[\sum_{s'} \left(P(s'|s,a) - \hat{P}(s'|s,a)\right) \left(V_{h+1}^{\pi^*}(s')\right)^{2^{i+1}}\right] + 2^{i+1} \left(\Delta_2 + v^{\pi^*}\right).
\end{aligned} \tag{15}$$

We can further apply Bernstein inequality and Cauchy-Schwartz inequality to the first term in (15). Specifically, denote

$$\mathbb{V}_2(i) := \sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^{\pi}(s,a) \mathrm{Var}_{P(s,a)} \left(\left(V_{h+1}^{\pi^*}(s')\right)^{2^i}\right),$$

we have the recursion as

$$\mathbb{V}_2(i) \leq \sqrt{\frac{\iota}{Kd_m} \mathbb{V}_2(i+1)} + \frac{\iota}{Kd_m} + 2^{i+1} \left(\Delta_2 + v^{\pi^*}\right).$$

This recursion can be solved similarly as (11) by applying Lemma 6 with $\lambda_1 = \frac{\iota}{Kd_m}$ and $\lambda_2 = \Delta_2 + v^{\pi^*}$, which leads to

$$\mathbb{V}_2(0) \leq O\left(\frac{\iota}{Kd_m} + \left(v^{\pi^*} + \Delta_2\right)\right). \tag{16}$$

Meanwhile, from (14) we have that

$$\Delta_2 \leq \sqrt{\frac{\iota}{Kd_m} \mathbb{V}_2(0)} + \frac{\iota}{Kd_m}. \tag{17}$$

Combine (16) and (17), we have

$$\Delta_2 \leq O\left(\sqrt{\frac{\iota}{Kd_m}\left(\frac{\iota}{Kd_m} + v^{\pi^*} + \Delta_2\right)}\right) + \frac{\iota}{Kd_m},$$

Suppose $K = \tilde{\Omega}\left(\frac{1}{d_m}\right)$, then with Assumption 1 and the discussion in D, we have that

$$\Delta_2 \leq O\left(\sqrt{\frac{\iota}{Kd_m}} + \frac{\iota}{Kd_m}\right). \tag{18}$$

Now we turn to the error introduced by Δ_{PV}. With Lemma 2, we can again use the recursion to bound it, and finally obtain the bound in Theorem 3. By Lemma 2 and Cauchy-Schwartz inequality, we have that

$$\Delta_3 := \left|\sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^{\pi}(s,a) \Delta_{PV}(s,a)\right|$$

$$\leq \sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^{\pi}(s,a)\cdot\left[\sqrt{\frac{S\cdot \mathrm{Var}_{P(s,a)}(V_{h+1}^{\pi^*}(s')-V_{h+1}^{\hat{\pi}^*}(s'))\iota}{n(s,a)}}+\frac{S\iota}{n(s,a)}\right]$$
$$\leq \sqrt{\frac{S\iota}{Kd_m}}\sqrt{\sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^{\pi}(s,a)\mathrm{Var}_{P(s,a)}\left(V_{h+1}^{\pi^*}(s')-V_{h+1}^{\hat{\pi}^*}(s')\right)}+\frac{S\iota}{Kd_m}. \tag{19}$$

Now we turn to $\sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^{\pi}(s,a)\mathrm{Var}_{P(s,a)}(V_{h+1}^{\pi^*}(s')-V_{h+1}^{\hat{\pi}^*}(s'))$ in (19). We still bound this term with the recursive methods. With Lemma 3, we have that

$$\sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^{\pi}(s,a)\mathrm{Var}_{P(s,a)}\left((V_{h+1}^{\pi^*}(s')-V_{h+1}^{\hat{\pi}^*}(s'))^{2^i}\right)$$
$$\leq \sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^{\pi}(s,a)\left[\sum_{s'}\left(P(s'|s,a)-\hat{P}(s'|s,a)\right)\left(V_{h+1}^{\pi^*}(s')-V_{h+1}^{\hat{\pi}^*}(s')\right)^{2^{i+1}}\right]+2^{i+1}\left(\Delta_3+\left(v^{\pi^*}-v^{\hat{\pi}^*}\right)\right). \tag{20}$$

We can further apply Lemma 2 and Cauchy-Schwartz inequality to the first term in (20), which eventually lead to the recursion formula. Specifically, denote

$$\mathbb{V}_3(i):=\sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^{\pi}(s,a)\mathrm{Var}_{P(s,a)}\left(\left(V_{h+1}^{\pi^*}(s')-V_{h+1}^{\hat{\pi}^*}(s')\right)^{2^i}\right),$$

we have the recursion as

$$\mathbb{V}_3(i)\leq\sqrt{\frac{S\iota}{Kd_m}\mathbb{V}_3(i+1)}+\frac{S\iota}{Kd_m}+2^{i+1}\left(\Delta_3+v^{\pi^*}-v^{\hat{\pi}^*}\right).$$

This recursion can be solved similarly as (11) by applying Lemma 6 with $\lambda_1=\frac{S\iota}{Kd_m}$ and $\lambda_2=\Delta_3+v^{\pi^*}-v^{\hat{\pi}^*}$, which leads to

$$\mathbb{V}_3(0)\leq O\left(\frac{S\iota}{Kd_m}+\left(v^{\pi^*}-v^{\hat{\pi}^*}+\Delta_3\right)\right). \tag{21}$$

Meanwhile, from (19) we have that

$$\Delta_3\leq\sqrt{\frac{S\iota}{Kd_m}\mathbb{V}_3(0)}+\frac{S\iota}{Kd_m}. \tag{22}$$

Combine (21) and (22), we have

$$\Delta_3\leq O\left(\sqrt{\frac{S\iota}{Kd_m}\left(\frac{S\iota}{Kd_m}+v^{\pi^*}-v^{\hat{\pi}^*}+\Delta_3\right)}\right)+\frac{S\iota}{Kd_m},$$

Moreover, with (13) and (18), we have that

$$v^{\pi^*}-v^{\hat{\pi}^*}\leq O\left(\sqrt{\frac{\iota}{Kd_m}}+\frac{\iota}{Kd_m}\right)+\Delta_3.$$

Thus, with the discussion in Appendix D, we can conclude that $v^{\pi^*}-v^{\hat{\pi}^*}\leq O\left(\sqrt{\frac{\iota}{Kd_m}}+\frac{S\iota}{Kd_m}\right)$, which finishes the proof for Theorem 3.

B Proof of the Lower Bounds

B.1 Lower Bound for Offline Policy Evaluation

Our lower bound instance is adapted from the instances in [Azar et al., 2013, Lattimore and Hutter, 2014, Pananjady and Wainwright, 2020] for finite horizon time-homogeneous setting.

Proof. We consider a two-state MDP with state s_1, s_2, with an unique action a. s_1 is an absorbing state, which means $P(s_1|s_1, a) = 1$, while $P(s_2|s_2, a) = p$, $P(s_1|s_2, a) = 1 - p$. We assume the reward is deterministic with $r(s_1, a) = 0$, $r(s_2, a) = \frac{1}{H}$ that satisfies Assumption 1. Assume we want to have an accurate estimation of $V_1(s_2)$, which is equivalent to have a sufficient accurate estimation of p. With straightforward calculation, we have that

$$V(p) := V_1(s_2) = \frac{p - p^{H+1}}{1 - p} \frac{1}{H}.$$

Notice that

$$\frac{\partial V(p)}{\partial p} = \frac{1 - (1 + (1 - p)H)p^H}{(1 - p)^2} \frac{1}{H}.$$

Let $p_1 = 1 - \frac{c_1}{H}$, where c_1 is an absolute constant, we know that

$$\frac{\partial V(p_1)}{\partial p_1} = \frac{1 - c_1(1 - \frac{c_1}{H})^H}{(1 - p_1)^2} \frac{1}{H} \geq \frac{1 - c_1 e^{-c_1}}{(1 - p_1)^2} \frac{1}{H} = \frac{1 - c_1 e^{-c_1}}{c_1^2} H,$$

which is monotonically decreasing w.r.t c_1. Assume $p_2 = 1 - \frac{c_2}{H}$ where $c_2 < c_1$ is another absolute constant, we have that

$$V(p_2) - V(p_1) \geq \frac{1 - c_1 e^{-c_1}}{(1 - p_1)^2} \frac{1}{H} (p_2 - p_1) = \frac{1 - c_1 e^{-c_1}}{c_1^2} (c_1 - c_2).$$

We now use Le Cam's method to show that without sufficient number of data from $\text{Bern}(p)$, we cannot identify $p = p_1$ or $p = p_2$ with high probability, and thus cannot have ideal estimation error on both of p_1 and p_2. We start from the following lemma:

Lemma 7.

$$KL(Bern(p) \| Bern(q)) \leq \frac{(p - q)^2}{q(1 - q)}.$$

Proof.

$$\begin{aligned}
&\text{KL}(\text{Bern}(p) \| \text{Bern}(q)) \\
=&p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q} \\
\leq&p \frac{p - q}{q} + (1 - p) \log \frac{q - p}{1 - p} \\
=&\frac{(p - q)^2}{q(1 - q)},
\end{aligned}$$

where the inequality is due to the fact that $\log(1 + x) \leq x$. □

Assume $\Psi : [0,1]^{n(s_2,a)} \to \{p_1, p_2\}$ is a test with n i.i.d samples from Bern(p), and use $\mathbb{P}_1$ and $\mathbb{P}_2$ to denote the probability measure under p_1 and p_2, we have that

$$\begin{aligned}
&\inf_{\Psi} \{\mathbb{P}_1(\Psi(\mathcal{D}) \neq p_1) + \mathbb{P}_2(\Psi(\mathcal{D}) \neq p_2)\} \\
\geq & 1 - \|(\text{Bern}(p_1))^{n(s_2,a)} - (\text{Bern}(p_2))^n\|_{\text{TV}} \quad \text{(Le Cam's inequality)} \\
\geq & 1 - \sqrt{\frac{n(s_2,a)}{2} \text{KL}\left(\text{Bern}(p_1) \| \text{Bern}(p_2)\right)} \quad \text{(Pinsker's inequality)} \\
\geq & 1 - \sqrt{\frac{n(s_2,a)(p_1-p_2)^2}{p_1(1-p_1)}} \\
= & 1 - \sqrt{\frac{n(s_2,a)(c_1-c_2)^2}{c_1(H-c_1)}}.
\end{aligned}$$

Take $c_2 = c_1 - \sqrt{\frac{c_1(H-c_1)}{2n(s_2,a)}}$, we know that with probability at least 0.5 we cannot identify $p = p_1$ or $p = p_2$. And notice that

$$V(p_2) - V(p_1) \geq \frac{1 - c_1 e^{-c_1}}{c_1^2} \sqrt{\frac{c_1(H-c_1)}{2n(s_2,a)}} \geq 2c_0 \sqrt{\frac{H}{n(s_2,a)}},$$

where c_0 is an absolute constant only depends on c_1. Thus we know that, with $n(s_2,a)$ samples from $P(s_2,a)$, we must suffer from an estimation error of $\Omega\left(\sqrt{\frac{H}{n(s_2,a)}}\right)$ with probability at least 0.25. Notice that we can set $n(s_2,a) = nd_m$, thus finishes the proof. □

B.2 Lower Bound of Offline Policy Improvement

We can further show the lower bound of offline improvement for finite horizon time-homogeneous MDP, based on the hard instance we mentioned above.

Proof. We introduce additional states s_0 and s_3 in the previous hard instance, with the transition from s_0, $P(s_1|s_0,a_1) = 1$, $P(s_3|s_0,a) = 1$, $\forall a \neq a_1$, and s_3 is an absorbing state with total reward in H steps (i.e. $V_1(s_3)$) as $V(p_1) + c_0\sqrt{\frac{H}{nd_m}}$. We always start from s_0, and we need to choose the action at s_0. Notice that, if $p = p_2$, then the optimal arm is a_1, while if $p = p_1$, then the optimal arm is not a_1, both with a sub-optimal gap of at least $c_0\sqrt{\frac{H}{nd_m}}$, which finishes the proof. □

C Proof for Linear MDP with Anchor Points

C.1 Proof for Offline Policy Evaluation

Notice that, the value difference lemma holds for any kinds of MDP. Thus, if we have Bernstein-type concentration for $\hat{r}(s,a) - r(s,a)$ and $\sum_{s'}(P(s'|s,a) - \hat{P}(s'|s,a))V(s')$, we can adopt the techniques for tabular MDP and obtain the desired results.

Lemma 8. *With probability at least $1-\delta$, we have that $\forall(s,a)$,*

$$|\hat{r}(s,a) - r(s,a)| \leq \sqrt{\frac{r(s,a)}{nd_m}} + \frac{\iota}{nd_m}$$

where ι absorbs the logarithm factors $\log(\mathrm{poly}(d)/\delta)$.

Proof. Notice that

$$\begin{aligned}
&|\hat{r}(s,a) - r(s,a)| \\
\leq &\sum_{k\in\mathcal{K}} \lambda_k^{s,a} |\hat{r}(s_k,a_k) - r(s_k,a_k)| \\
\leq &\sum_{k\in\mathcal{K}} \lambda_k^{s,a} \left(\sqrt{\frac{r(s_k,a_k)}{nd_m}} + \frac{1}{nd_m} \right) \\
\leq &\sqrt{\frac{\sum_{k\in\mathcal{K}} \lambda_k^{s,a} r(s_k,a_k)}{nd_m}} + \frac{\iota}{nd_m} \\
= &\sqrt{\frac{r(s,a)}{nd_m}} + \frac{\iota}{nd_m},
\end{aligned}$$

where the second inequality is due to the Bernstein's inequality on each (s_k, a_k) with $\iota = \log(2d/\delta)$ and $r(s_k,a_k) \in [0,1]$, and the third inequality is due to Cauchy-Schwartz inequality and $\sum_{k\in\mathcal{K}} \lambda_k^{s,a} = 1$ □

Lemma 9. *Suppose $V(s')$ is independent from $\hat{P}(s'|s,a)$, then with probability at least $1-\delta$, we have that $\forall(s,a)$*

$$\left| \sum_{s'} (P(s'|s,a) - \hat{P}(s'|s,a)) V(s') \right| \leq \sqrt{\frac{\mathit{Var}_{P(s,a)} V(s') \iota}{nd_m}} + \frac{\iota}{nd_m},$$

where ι absorbs the logarithm factors $\log(\mathrm{poly}(d)/\delta)$.

Proof. First, we have that

$$\begin{aligned}
&\left| \sum_{s'} (P(s'|s,a) - \hat{P}(s'|s,a)) V(s') \right| \\
\leq &\sum_{k\in\mathcal{K}} \lambda_k^{s,a} \left| \sum_{s'} \left(P(s'|s_k,a_k) - \hat{P}(s'|s_k,a_k) \right) V(s') \right| \\
\leq &\sum_{k\in\mathcal{K}} \lambda_k^{s,a} \left(\sqrt{\frac{\mathrm{Var}_{P(s_k,a_k)} V(s') \iota}{nd_m}} + \frac{\iota}{nd_m} \right) \\
\leq &\sqrt{\frac{\sum_{k\in\mathcal{K}} \lambda_k^{s,a} \mathrm{Var}_{P(s_k,a_k)} V(s') \iota}{nd_m}} + \frac{\iota}{nd_m},
\end{aligned}$$

where the second inequality is due to the Bernstein's inequality on each (s_k, a_k) with $\iota = \log(2d/\delta)$ and the last inequality is due to Cauchy-Schwartz inequality and $\sum_{k\in\mathcal{K}} \lambda_k^{s,a} = 1$. Notice that

$$\sum_{k\in\mathcal{K}} \lambda_k^{s,a} \mathrm{Var}_{P(s_k,a_k)} V(s')$$

$$
\begin{aligned}
=&\sum_{k\in\mathcal{K}}\lambda_k^{s,a}\left(\sum_{s'}P(s'|s_k,a_k)V(s')^2-\left(\sum_{s'}P(s'|s_k,a_k)V(s')\right)^2\right)\\
\leq&\sum_{s'}P(s'|s,a)V(s')^2-\left(\sum_{k\in\mathcal{K}}\lambda_k^{s,a}P(s'|s_k,a_k)V(s')\right)^2\\
=&\mathrm{Var}_{P(s,a)}V(s'),
\end{aligned}
$$

where the inequality is due to Cauchy-Schwartz inequality and $\sum_{k\in\mathcal{K}}\lambda_k^{s,a}=1$. Substitute this term back and we conclude the proof. □

Notice that, Lemma 9 simultaneously holds for all of the (s,a) if all of the concentration on the anchor points hold. Hence we can apply the analysis for tabular MDP and obtain the desired results, which finishes the proof for offline policy evaluation on linear MDP with anchor points.

C.2 Proof for Offline Policy Optimization

Here we need a Bernstein-type concentration for $\sum_{s'}(P(s'|s,a)-\hat{P}(s'|s,a))V(s')$ when $V(s')$ and $\hat{P}(s'|s,a)$ are correlated. A naive application of Lemma 2 will introduce a $|\mathcal{S}|$ factor in the higher order term, which is not satisfactory, as $|S|$ can be exponentially large. Here use another method to replace this dependency on $|\mathcal{S}|$ with the feature dimension d.

Lemma 10. *Suppose $\tilde{V}(s')$ is independent from $\hat{P}(s'|s,a)$, then with probability $1-\delta$, we have that*

$$
\left|\sum_{s'}(P(s'|s,a)-\hat{P}(s'|s,a))V(s')\right|\leq\sqrt{\frac{\mathrm{Var}_{P(s,a)}V(s')\iota}{nd_m}}+\frac{\iota}{nd_m}+\|\tilde{V}-V\|_\infty\left(1+\sqrt{\frac{\iota}{nd_m}}\right),
$$

where ι absorbs the logarithm factors $\log(\mathrm{poly}(d)/\delta)$.

Proof. Notice that

$$
\begin{aligned}
&\left|\sum_{s'}(P(s'|s,a)-\hat{P}(s'|s,a))V(s')\right|\\
\leq&\left|\sum_{s'}(P(s'|s,a)-\hat{P}(s'|s,a))\tilde{V}(s')\right|+\left|\sum_{s'}(P(s'|s,a)-\hat{P}(s'|s,a))(V(s')-\tilde{V}(s'))\right|\\
\leq&\sqrt{\frac{\mathrm{Var}_{P(s,a)}\tilde{V}(s')\iota}{nd_m}}+\frac{\iota}{nd_m}+\|\tilde{V}-V\|_\infty\\
\leq&\sqrt{\frac{\mathrm{Var}_{P(s,a)}V(s')\iota}{nd_m}}+\sqrt{\frac{\mathrm{Var}_{P(s,a)}(\tilde{V}(s')-V(s'))\iota}{nd_m}}\frac{\iota}{nd_m}+\|\tilde{V}-V\|_\infty\\
\leq&\sqrt{\frac{\mathrm{Var}_{P(s,a)}\tilde{V}(s')\iota}{nd_m}}+\frac{\iota}{nd_m}+\|\tilde{V}-V\|_\infty\left(1+\sqrt{\frac{\iota}{nd_m}}\right),
\end{aligned}
$$

where the first inequality is due to the triangle inequality, the second inequality is due to Lemma 9 and algebra, the third inequality is due to the triangle inequality for the variance, *i.e.* $\sqrt{\mathrm{Var}(X+Y)}\leq\sqrt{\mathrm{Var}(X)}+\sqrt{\mathrm{Var}(Y)}$, and the last inequality is due to the fact that $\mathrm{Var}(V)\leq\|V\|_\infty$. □

With Lemma 10, we can construct an ϵ-net (under ℓ_∞ norm) for V to obtain Bernstein-type concentration. For tabular MDP, this ϵ-net is of size $O(\epsilon^{-|\mathcal{S}|})$, which leads to the same result of Lemma 2. However, in linear MDP, Q follows a linear form $Q(s,a) = \phi(s,a)^\top w_Q$, thus the V we consider lies in a d-dimensional manifolds, and the size of the ϵ-net we exactly need is $O(\epsilon^{-d})$. This observation leads to the following corollary:

Corollary 1. *For any $V(s)'$, with probability $1-\delta$, we have that*

$$\left|\sum_{s'}(P(s'|s,a) - \hat{P}(s'|s,a))V(s')\right| \leq \sqrt{\frac{d\mathrm{Var}_{P(s,a)}V(s')\iota}{nd_m}} + \frac{d\iota}{nd_m} + \epsilon\left(1 + \sqrt{\frac{d\iota}{nd_m}}\right),$$

where ι absorbs the logarithm factors, $\log(\mathrm{poly}(d, 1/\epsilon)/\delta)$.

Here the additional d comes from the logarithm of the size of ϵ-net. We can further choose $\epsilon = \frac{d}{nd_m}$ and absorb $\epsilon\left(1 + \sqrt{\frac{d\iota}{nd_m}}\right)$ into $\frac{d\iota}{nd_m}$, then apply the analysis for tabular MDP and replace S with d to obtain the desired result, hence conclude the proof.

D Step-by-Step Solving for Δ_2 and $v^{\pi^*} - v^{\hat{\pi}^*}$

D.1 Explicit Bound for Δ_2

Notice that, for some absolute constant c,

$$\Delta_2 \leq c\sqrt{\frac{\iota}{Kd_m}\left(\frac{\iota}{Kd_m} + \Delta_2 + v^{\pi^*}\right)} + \frac{\iota}{Kd_m}.$$

which means

$$\begin{aligned} c\Delta_2 \leq& \sqrt{\frac{c\iota}{Kd_m}\left(\frac{c\iota}{Kd_m} + c\Delta_2 + cv^{\pi^*}\right)} + \frac{c\iota}{Kd_m} \\ \leq& \frac{c\Delta_2}{2} + \frac{cv^{\pi^*}}{2} + \frac{2c\iota}{Kd_m}, \end{aligned}$$

thus we have that

$$\Delta_2 \leq v^{\pi^*} + \frac{4\iota}{Kd_m}.$$

Substitute back, suppose $N = \widetilde{\Omega}\left(\frac{1}{d_m}\right)$, then with Assumption 1, we have that

$$\Delta_2 \leq c\sqrt{\frac{\iota}{Kd_m}\left(\frac{5\iota}{Kd_m} + 2v^{\pi^*}\right)} + \frac{\iota}{Kd_m} = O\left(\sqrt{\frac{\iota}{Kd_m}} + \frac{\iota}{Kd_m}\right).$$

D.2 Explicit bound for $v^{\pi^*} - v^{\hat{\pi}^*}$

Notice that, for some absolute constant c, we have that

$$\Delta_3 \leq c\sqrt{\frac{S\iota}{Kd_m}\left(\frac{S\iota}{Kd_m} + v^{\pi^*} - v^{\hat{\pi}^*} + \Delta_3\right)} + \frac{S\iota}{Kd_m},$$

which means

$$\begin{aligned} c\Delta_3 \leq& \sqrt{\frac{cS\iota}{Kd_m}\left(\frac{cS\iota}{Kd_m} + c(v^{\pi^*} - v^{\hat{\pi}^*}) + c\Delta_3\right)} + \frac{cS\iota}{Kd_m} \\ \leq& \frac{c(v^{\pi^*} - v^{\hat{\pi}^*})}{2} + \frac{c\Delta_3}{2} + \frac{2cS\iota}{Kd_m}, \end{aligned}$$

thus we have that

$$\Delta_3 \leq (v^{\pi^*} - v^{\hat{\pi}^*}) + \frac{4S\iota}{Kd_m}.$$

We then substitute back, and know that

$$\Delta_3 \leq c\sqrt{\frac{S\iota}{Kd_m}\left(2\left(v^{\pi^*} - v^{\hat{\pi}^*}\right) + \frac{5S\iota}{Kd_m}\right)} + \frac{S\iota}{Kd_m}.$$

Furthermore, for another absolute constant c', we have that

$$\begin{aligned} v^{\pi^*} - v^{\hat{\pi}^*} \leq& c'\left(\sqrt{\frac{\iota}{Kd_m}} + \frac{\iota}{Kd_m}\right) + \Delta_3 \\ \leq& c'\sqrt{\frac{\iota}{Kd_m}} + c\sqrt{\frac{S\iota}{Kd_m}\left(2\left(v^{\pi^*} - v^{\hat{\pi}^*}\right) + \frac{5S\iota}{Kd_m}\right)} + \frac{(c'+1)S\iota}{Kd_m}, \end{aligned}$$

which means

$$\left(\sqrt{2\left(v^{\pi^*} - v^{\hat{\pi}^*}\right) + \frac{5S\iota}{Kd_m}} - c\sqrt{\frac{S\iota}{Kd_m}}\right)^2 \leq 2c'\sqrt{\frac{\iota}{Kd_m}} + \frac{(c^2 + 2c' + 2)S\iota}{Kd_m},$$

that can be translated to the bound

$$\begin{aligned} v^{\pi^*} - v^{\hat{\pi}^*} \leq& \left(\sqrt{2c'\sqrt{\frac{\iota}{Kd_m}} + \frac{(c^2 + 2c' + 2)S\iota}{Kd_m}} + c\sqrt{\frac{S\iota}{Kd_m}}\right)^2 - \frac{5S\iota}{Kd_m} \\ \leq& 2\left(\sqrt{2c'\sqrt{\frac{\iota}{Kd_m}} + \frac{(2c^2 + 2c' + 2)S\iota}{Kd_m}}\right)^2 \\ =& O\left(\sqrt{\frac{\iota}{Kd_m}} + \frac{S\iota}{Kd_m}\right) \end{aligned}$$

where we use $\sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}$.

E Proof of Technical Lemmas

E.1 Bernstein's Inequality

Lemma 11 (Bernstein's Inequality)**.** *Let $\{X_i\}_{i=1}^n$ be i.i.d random variables from X with values bounded in $[0,1]$, then with probability at least $1-\delta$, we have that*

$$\left|\sum_{i=1}^{n} X_i - \mathbb{E}[X]\right| \leq \sqrt{\frac{2Var(X)\log\frac{2}{\delta}}{n}} + \frac{\log\frac{2}{\delta}}{3n},$$

where $Var(X)$ is the variance of X.

For the proof of Bernstein's inequality, we refer the interested reader to Wainwright [2019].

E.2 Proof of Lemma 1

Proof. We have that

$$\begin{aligned}
&\sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^\pi(s,a)\mathrm{Var}_{P(s,a)}\left(V_{h+1}^\pi(s')^{2^i}\right)\\
&=\sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^\pi(s,a)\left[\sum_{s'}P(s'|s,a)V_{h+1}^\pi(s')^{2^{i+1}}-\left(\sum_{s'}P(s'|s,a)V_{h+1}^\pi(s')^{2^i}\right)^2\right]\\
&=\sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^\pi(s,a)\left[\sum_{s'}\left(P(s'|s,a)-\hat{P}(s'|s,a)\right)V_{h+1}^\pi(s')^{2^{i+1}}+V_h^\pi(s)^{2^{i+1}}-\left(\sum_{s'}P(s'|s,a)V_{h+1}^\pi(s')^{2^i}\right)^2\right]\\
&\quad-\sum_s\mu(s)V_1^\pi(s)^{2^{i+1}}\\
&\leq\sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^\pi(s,a)\left[\sum_{s'}\left(P(s'|s,a)-\hat{P}(s'|s,a)\right)V_{h+1}^\pi(s')^{2^{i+1}}\right]\\
&\quad+\sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^\pi(s,a)\left[Q_h^\pi(s,a)^{2^{i+1}}-\left(\sum_{s'}P(s'|s,a)V_{h+1}^\pi(s')\right)^{2^{i+1}}\right]\\
&\leq\sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^\pi(s,a)\left[\sum_{s'}\left(P(s'|s,a)-\hat{P}(s'|s,a)\right)V_{h+1}^\pi(s')^{2^{i+1}}\right]\\
&\quad+2^{i+1}\sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^\pi(s,a)\left[Q_h^\pi(s,a)-\left(\sum_{s'}P(s'|s,a)V_{h+1}^\pi(s')\right)\right],\\
&=\sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^\pi(s,a)\left[\sum_{s'}\left(P(s'|s,a)-\hat{P}(s'|s,a)\right)V_{h+1}^\pi(s')^{2^{i+1}}\right]+2^{i+1}\sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^\pi(s,a)r(s,a)
\end{aligned}$$

where in the second step we use the fact that

$$\begin{aligned}
&\sum_{s,a} \hat{\xi}_h^\pi(s,a) \left[\sum_{s'} \hat{P}(s'|s,a) V_{h+1}(s')^{2^{i+1}} \right] \\
=&\sum_{s'} \left[\sum_{s,a} \hat{\xi}_h^\pi(s,a) \hat{P}(s'|s,a) \right] V_{h+1}(s')^{2^{i+1}} \\
=&\sum_{s'} \hat{\xi}_{h+1}^\pi(s') V_{h+1}(s')^{2^{i+1}} \\
=&\sum_{s',a'} \hat{\xi}_{h+1}^\pi(s',a') V_{h+1}(s')^{2^{i+1}}.
\end{aligned}$$

We drop the $\sum_s \mu(s) V_1^\pi(s)^{2^{i+1}}$ and use the convexity of x^{2^i} and $V_h^\pi(s) = \mathbb{E}_\pi Q_h^\pi(s,a)$ in the third step, and the last step is indicated by the assumption that $V_h(s) \le 1, \forall h \in [H], s \in \mathcal{S}$.

With Assumption 1, we know that

$$\sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^\pi(s,a) r(s,a) \le 1,$$

as each trajectory that can generated by $\widehat{\mathcal{M}}$ can be generated by $\mathcal{M}$, thus finish the proof. □

E.3 Proof of Lemma 2

Proof.

$$\begin{aligned}
&\left| \sum_{s'} \left(P(s'|s,a) - \hat{P}(s'|s,a) \right) V_h(s') \right| \\
=&\left| \sum_{s'} \left(P(s'|s,a) - \hat{P}(s'|s,a) \right) \left(V_h(s') - \sum_{s'} P(s'|s,a) V_h(s') \right) \right| \\
\le&\sum_{s'} \sqrt{\frac{P(s'|s,a)\iota}{n(s,a)}} \left| V_h(s') - \sum_{s'} P(s'|s,a) V_h(s') \right| + \frac{S\iota}{n(s,a)} \\
\le&\sqrt{\frac{S \mathrm{Var}_{P(s,a)}\left(V_h(s')\right)\iota}{n(s,a)}} + \frac{S\iota}{n(s,a)},
\end{aligned}$$

where the first equality is due to the fact that $\sum_{s'} P(s'|s,a) = \sum_{s'} \hat{P}(s'|s,a) = 1$, the second inequality is due to Bernstein's inequality on each s' and Assumption 1, and the last inequality holds by Cauchy-Schwarz inequality. □

E.4 Proof of Lemma 3

Proof. We have that

$$\sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^{\hat{\pi}^*}(s,a) \mathrm{Var}_{P(s,a)} \left(V_{h+1}(s')^{2^i} \right)$$

$$
\begin{aligned}
&= \sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^{\hat{\pi}^*}(s,a)\left[\sum_{s'}P(s'|s,a)V_{h+1}(s')^{2^{i+1}} - \left(\sum_{s'}P(s'|s,a)V_{h+1}(s')^{2^i}\right)^2\right] \\
&= \sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^{\hat{\pi}^*}(s,a)\left[\sum_{s'}\left(P(s'|s,a)-\hat{P}(s'|s,a)\right)V_{h+1}(s')^{2^{i+1}} + V_h(s)^{2^{i+1}} - \left(\sum_{s'}P(s'|s,a)V_{h+1}(s')^{2^i}\right)^2\right] \\
&\quad - \sum_s \mu(s)V_1(s)^{2^{i+1}} \\
&\le \sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^{\hat{\pi}^*}(s,a)\left[\sum_{s'}\left(P(s'|s,a)-\hat{P}(s'|s,a)\right)V_{h+1}(s')^{2^{i+1}}\right] \\
&\quad + \sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^{\hat{\pi}^*}(s,a)\left[V_h(s)^{2^{i+1}} - \left(\sum_{s'}P(s'|s,a)V_{h+1}(s')\right)^{2^{i+1}}\right] \\
&\le \sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^{\hat{\pi}^*}(s,a)\left[\sum_{s'}\left(P(s'|s,a)-\hat{P}(s'|s,a)\right)V_{h+1}(s')^{2^{i+1}}\right] \\
&\quad + 2^{i+1}\sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^{\hat{\pi}^*}(s,a)\left[V_h(s) - \left(\sum_{s'}P(s'|s,a)V_{h+1}(s')\right)\right],
\end{aligned}
$$

where in the last step, we use the fact that $\hat{\pi}^*$ is a deterministic policy, and

$$
\begin{aligned}
V_h^{\pi^*}(s) &\ge r(s,a) + \sum_{s'}P(s'|s,a)V_{h+1}^{\pi^*}(s'), \quad \forall a\in\mathcal{A}, \\
V_h^{\hat{\pi}^*}(s) &= r(s,\hat{\pi}^*(s)) + \sum_{s'}P(s'|s,\hat{\pi}^*(s))V_{h+1}^{\pi^*}(s'),
\end{aligned}
$$

which guarantees that for $V_h(s) = V_h^{\pi^*}(s)$ and $V_h(s) = V_h^{\pi^*}(s) - V_h^{\hat{\pi}^*}(s)$,

$$
\hat{\xi}_h^{\hat{\pi}^*}(s,a)\left[V_h(s) - \left(\sum_{s'}P(s'|s,a)V_{h+1}(s')\right)\right] \ge 0.
$$

Moreover, we have that

$$
\begin{aligned}
&\sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^{\pi}(s,a)\left[V_h(s) - \sum_{s'}\hat{P}(s'|s,a)V_{h+1}(s')\right] \\
=&\sum_{h\in[H]}\sum_{s}\hat{\xi}_h^{\pi}(s)V_h(s) - \sum_{h\in[H]\setminus\{1\}}\sum_{s}\hat{\xi}_h^{\pi}(s)V_h(s) \\
=&\sum_s \mu(s)V_1(s).
\end{aligned}
$$

So we can conclude that

$$
\sum_{h\in[H]}\sum_{s,a}\hat{\xi}_h^{\pi}(s,a)\left[V_h(s) - \left(\sum_{s'}P(s'|s,a)V_{h+1}(s')\right)\right]
$$

$$= \sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^\pi(s,a) \sum_{s'} \left[\hat{P}(s'|s,a) - P(s'|s,a)\right] V_{h+1}(s') + \sum_s \mu(s) V_1(s),$$

thus finish the proof. □

E.5 Proof of Lemma 4

Proof. Lemma 4 have been shown in Dann et al. [2017, Lemma E.15]. Here we include the proof for completeness.

$$\begin{aligned}
&v^\pi - \hat{v}^\pi \\
=&\sum_{s,a} \hat{\xi}_1^\pi(s,a)(Q_1^\pi(s,a) - \hat{Q}_1^\pi(s,a)) \\
=&\sum_{s,a} \left[\hat{\xi}_1^\pi(s,a) \left(r(s,a) - \hat{r}(s,a) + \sum_{s'} \left[P(s'|s,a) V_2^\pi(s')\right] - \sum_{s'} \left[\hat{P}(s'|s,a)\hat{V}_2^\pi(s'))\right]\right)\right] \\
=&\sum_{s,a} \left[\hat{\xi}_1^\pi(s,a) \left(r(s,a) - \hat{r}(s,a) + \sum_{s'} \left[(P(s'|s,a) - \hat{P}(s'|s,a)) V_2^\pi(s')\right]\right)\right] + \sum_s \hat{\xi}_2^\pi(s)(V_2^\pi(s) - \hat{V}_2^\pi(s)) \\
=&\cdots \\
=&\sum_{h\in[H]} \sum_{s,a} \left[\hat{\xi}_h^\pi(s,a) \left(r(s,a) - \hat{r}(s,a) + \sum_{s'} \left[(P(s'|s,a) - \hat{P}(s'|s,a)) V_{h+1}^\pi(s')\right]\right)\right]
\end{aligned}$$

□

E.6 Proof of Lemma 5

Proof. By Bernstein's inequality, we know that with high probability,

$$\begin{aligned}
&\left|\sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^\pi(s,a) \left[r(s,a) - \hat{r}(s,a)\right]\right| \\
\leq &\sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^\pi(s,a) \left[\sqrt{\frac{r(s,a)\iota}{n(s,a)}} + \frac{\iota}{n(s,a)}\right] \\
\leq &\sqrt{\frac{\iota}{K d_m}} \sqrt{\sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^\pi(s,a) r(s,a)} + \frac{\iota}{K d_m},
\end{aligned}$$

where the first inequality use the fact that $r(s,a) \leq 1$ and the second inequality is Cauchy-Schwarz inequality associated with Assumption 2. With Assumption 1, we know that

$$\sum_{h\in[H]} \sum_{s,a} \hat{\xi}_h^\pi(s,a) r(s,a) \leq 1,$$

as each trajectory that can generated by $\widehat{\mathcal{M}}$ can be generated by $\mathcal{M}$, which finishes the proof. □

E.7 Proof of Lemma 6

Proof. Lemma 6 is similar to the Lemma 11 in Zhang et al. [2020b], and here we provide a simplified proof. Define

$$i_0 := \max\{i \mid 2^{2i}\lambda_2^2 \le \lambda_1 \mathbb{V}(i)\} \cup \{0\},$$

which is the largest integer satisfies the inequality $2^{2i}\lambda_2^2 \le \lambda_1 \mathbb{V}_1(i)$. We will make the recursion at most i_0 times. As $\mathbb{V}(i)$ is upper bounded by H, we know $i_0 = O\left(\log \frac{H\lambda_1}{\lambda_2^2}\right)$. If $i_0 = 0$, we have that

$$\mathbb{V}(1) \le \frac{4\lambda_2^2}{\lambda_1}.$$

Otherwise,

$$\begin{aligned}
&2^{2i_0}\lambda_2^2 \\
\le&\lambda_1 \mathbb{V}(i_0) \\
\le&\lambda_1 \left(\sqrt{\lambda_1 \mathbb{V}(i_0+1)} + \lambda_1 + 2^{i_0+1}\lambda_2\right) \\
\le&\lambda_1 \left(\sqrt{2^{2i_0+2}\lambda_2^2} + \lambda_1 + 2^{i_0+1}\lambda_2\right) \\
=&\lambda_1(2^{i_0+2}\lambda_2 + \lambda_1),
\end{aligned}$$

we have that

$$(2^{i_0}\lambda_2 - 2\lambda_1)^2 \le 5\lambda_1^2,$$

which means

$$2^{i_0}\lambda_2 \le (\sqrt{5}+2)\lambda_1,$$

so $\forall i \le i_0$, we have that

$$\mathbb{V}(i) \le \sqrt{\lambda_1 \mathbb{V}(i+1)} + (\sqrt{5}+3)\lambda_1.$$

As

$$\mathbb{V}(i_0) \le 2^{i_0+2}\lambda_2 + \lambda_1 \le (4\sqrt{5}+9)\lambda_1,$$

and when $\mathbb{V}(i+1) \ge \left(\frac{1}{2} + \sqrt{\sqrt{5}+\frac{13}{4}}\right)\lambda_1$, we have $\mathbb{V}(i) \le \mathbb{V}(i+1)$, when $\mathbb{V}(i+1) \le \left(\frac{1}{2} + \sqrt{\sqrt{5}+\frac{13}{4}}\right)\lambda_1$, $\mathbb{V}(i) \le \left(\frac{1}{2} + \sqrt{\sqrt{5}+\frac{13}{4}}\right)\lambda_1$. So $\mathbb{V}(1) \le (4\sqrt{5}+9)\lambda_1$. Combine the two cases, we have that

$$\mathbb{V}(1) \le \max\left\{\frac{4\lambda_2^2}{\lambda_1}, (4\sqrt{5}+9)\lambda_1\right\},$$

which means

$$\mathbb{V}(0) \le \max\left\{4\lambda_2 + \lambda_1, \left(\sqrt{4\sqrt{5}+9}+1\right)\lambda_1 + 2\lambda_2\right\} \le 6(\lambda_1+\lambda_2),$$

which concludes the proof. □

F Further Discussion

F.1 Time-Homogenous vs. Time-Inhomogenous

F.1.1 Offline Policy Evaluation

One can notice that the lower bound of offline policy evaluation under time-homogenous setting (shown above) and time-inhomogenous setting (that can be simplified from the Cramer-Rao lower bound of [Jiang and Li, 2016]) are identical. It can be surprising at the first glance, as one common belief for reinforcement learning is that the time-homogenous MDP should be easier than the time-inhomogenous MDP. However, we remark that, this argument does not hold for the policy evaluation. This is due to the fact that in time-homogenous setting, as the error from the estimation of transition will be accumulated along the horizon, we need $\Omega(H)$ samples for each transition to make sure that the accumulated error should be $O(1)$. However, in the time-inhomogenous setting, the error from the estimation of each level transition is probably not accumulated, which means we only need $\Omega(1)$ samples for each transition at each level to make sure the final error to be $O(1)$. This can also be seen in the analysis of offline policy evaluation under time-inhomogenous setting in [Yin and Wang, 2020, Yin et al., 2020], where the authors decouple the error from each sample (s_h, a_h, s_{h+1}), which forms a martingale difference sequence with each step variance $O(1)$ that can then apply the Freedman's inequality to obtain a tight bound. Under time-inhomogenous setting, we notice that the error from each sample (s, a, s') can only form a martingale difference sequence with each step variance $O(H)$, which will not lead to a tighter bound. We want to emphasize that *the results in [Yin and Wang, 2020, Yin et al., 2020] does not directly indicate results in this paper.*

Moreover, we notice that the analysis in [Yin and Wang, 2020, Yin et al., 2020] can be translated to a value-dependent bound and thus can be horizon-free under Assumption 1, with the following lemma:

Lemma 12. *Under Assumption 1, we have that*

$$\sum_{h\in[H]}\sum_{s,a}\xi_h^\pi(s,a)Var(r(s,a)+V_{h+1}^\pi(s'))\le 3v^\pi,$$

where the term at the left hand side is exactly the variance term of MIS estimator considered in [Yin and Wang, 2020, Lemma 3.4]

Proof.

$$\begin{aligned}
&\sum_{h\in[H]}\sum_{s,a}\xi_h^\pi(s,a)\text{Var}\left(r(s,a)+V_{h+1}^\pi(s')\right)\\
\le&\sum_{h\in[H]}\sum_{s,a}\xi_h^\pi(s,a)\left[r(s,a)+\sum_{s'}P(s'|s,a)[V_{h+1}^\pi(s')]^2-\left(\sum_{s'}P(s'|s,a)V_{h+1}^\pi(s')\right)^2\right]\\
\le&\sum_{h\in[H]}\sum_{s,a}\xi_h^\pi(s,a)\left[r(s,a)+Q_h^\pi(s,a)^2-\left[\sum_{s'}P(s'|s,a)V_{h+1}^\pi(s')\right]^2\right]\\
\le&3\sum_{h\in[H]}\sum_{s,a}\xi_h^\pi(s,a)r(s,a)=3v^\pi,
\end{aligned}$$

where we use the following fact:

$$\sum_{s,a,s'}\xi_h^\pi(s,a)P(s'|s,a)\left[V_{h+1}^\pi(s')\right]^2$$

$$
\begin{aligned}
&= \sum_{s'} \xi^{\pi}_{h+1}(s') \left[V^{\pi}_{h+1}(s')\right]^2 \\
&= \sum_{s'} \xi^{\pi}_{h+1}(s') \left[\sum_{a'} \pi(a'|s') Q^{\pi}_{h+1}(s', a')\right]^2 \\
&\leq \sum_{s'} \xi^{\pi}_{h+1}(s') \sum_{a'} \pi_{h+1}(a'|s') Q^{\pi}_{h+1}(s', a')^2 \\
&= \sum_{s', a'} \xi^{\pi}_{h+1}(s', a') Q^{\pi}_{h+1}(s', a')^2.
\end{aligned}
$$

□

This shows that the offline policy evaluation under time-inhomogeneous setting is also horizon-free, which also matches the intuition that if the density ratio can be lower bounded, then we can construct a trajectory-wise importance sampling (IS) estimator that only depends on the number of episodes. Notice that, in [Yin and Wang, 2020, Yin et al., 2020], the higher order term has an additional $\sqrt{SA}$ factor. Tightening such factor is beyond the scope of this paper.

F.1.2 Offline Policy Optimization

For the offline policy optimization, it's known that time-inhomogeneous MDP cannot avoid the additional H factor in sample complexity, thus can never be horizon-free. We also want to remark that with Lemma 12, the results in [Yin et al., 2020] can be translated to a $\tilde{O}\left(\sqrt{\frac{H^2}{nd_m}}\right)$ error bound as well as a $\tilde{O}\left(\frac{H^2}{d_m\epsilon^2}\right)$ sample complexity bound that holds for all range of $\epsilon \in (0, 1]$. Our bound, however, have an additional higher order term $\tilde{O}\left(\frac{SH}{d_m\epsilon}\right)$. We would like to remark that, this is due to the time-homogeneous nature of our setting, which makes $V_h^{\hat{\pi}^*}$ heavily depends on $\hat{P}$. On the other hand, in time-inhomogeneous setting, $V_h^{\hat{\pi}^*}$ only depends on $\hat{P}_k(s, a)$ for $k > h$, thus can directly apply Bernstein's inequality when bounding the term $\left(\hat{P}_h(s, a) - P_h(s, a)\right) V_{h+1}^{\hat{\pi}^*}$, which will not introduce additional S factor. The best known result for finite horizon time-homogenous MDP [Zhang et al., 2020b] also has this additional S factor, and how to eliminate this additional S factor remains an open problem.

F.2 Finite-Horizon vs. Infinite-Horizon

F.2.1 Offline Policy Evaluation

We remark that [Li et al., 2020] provides a value-dependent bound for the policy evaluation under infinite-horizon generative model setting that accommodates full range of ϵ. We obtain the similar results in finite-horizon setting that can accommodate full range of ϵ, however, with a different and probably simpler analysis.

F.2.2 Offline Policy Optimization

[Li et al., 2020] also provides a minimax-optimal sample complexity bound up to logarithmoc factors for policy optimization with generative model that accommodates full range of ϵ. We notice that such kinds of analysis cannot be directly applied to the finite-horizon setting, as their "absorbing MDP" technique cannot be directly applied to the finite-horizon MDP, due to the difference of time-homogenous value function

in infinite-horizon setting and time-inhomogenous value function in finite-horizon setting, which has been pointed out by [Cui and Yang, 2020]. And thus most of the existing work does not provide a sample complexity bound that can match the lower bound. To the best of our knowledge, our work first provide a sample complexity bound that match the lower bound up to logarithmic factors and an high-order term.

F.3 With General Function Approximation

There are also works considering the offline policy optimization with general function approximation under different kinds of function class assumption like realizability and completeness [e.g. Chen and Jiang, 2019, Xie and Jiang, 2020a,b], which generally do not imply tight bounds under certain scenarios like e.g. tabular MDP. We leave the extension to general function approximation as future work.

F.4 Without Sufficient Exploratory Data

Recently, [Liu et al., 2020b, Kumar et al., 2020] also introduces another perspective on performing offline policy optimization within a local policy set when the offline data is not sufficient exploratory, which is different from the global policy optimization we consider here. We want to emphasize that, if we want to approach the global optimal policy, our assumption on good data coverage *i.e.*Assumption 2 is necessary. Otherwise, we will suffer from the error from under-explored state-action pair, as Theorem 2 and Theorem 4 suggests.