Finite-Time Bounds for Average-Reward Fitted Q-Iteration

Jongmin Lee
Seoul National University
Department of Mathematical Sciences
dlwhd2000@snu.ac.kr

Ernest K. Ryu
UCLA
Department of Mathematics
eryu@math.ucla.edu

Abstract

Although there is an extensive body of work characterizing the sample complexity of discounted-return offline RL with function approximations, prior work on the average-reward setting has received significantly less attention, and existing approaches rely on restrictive assumptions, such as ergodicity or linearity of the MDP. In this work, we establish the first sample complexity results for average-reward offline RL with function approximation for weakly communicating MDPs, a much milder assumption. To this end, we introduce Anchored Fitted Q-Iteration, which combines the standard Fitted Q-Iteration with an anchor mechanism. We show that the anchor, which can be interpreted as a form of weight decay, is crucial for enabling finite-time analysis in the average-reward setting. We also extend our finite-time analysis to the setup where the dataset is generated from a single-trajectory rather than IID transitions, again leveraging the anchor mechanism.

1 Introduction

The goal of offline Reinforcement Learning (RL) is to find a near-optimal policy using a precollected dataset without any direct interaction with the environment. Characterizing the sample complexity for finding an ϵ-optimal policy using function approximation under assumptions that the offline data has sufficient coverage over the whole state-action space has been an active area of theoretical RL research. However, more prior work focuses on the discounted cumulative reward setup, and research on obtaining sample complexity in the average reward has been limited due to the absence of the discount factor and the complexity of the Bellman equation. Specifically, all prior works with function approximation rely on restrictive assumptions such as ergodicity or linearity of the MDP.

Although theoretical RL research often focuses on the discounted return setup due to the theoretical convenience offered by the discount factor and the simpler Bellman equation, many practical scenarios are more naturally modeled as agents aim to maximize the average reward. In fact, many practical RL applications do not use discounting at all. These considerations make the sample complexity of average-reward RL relevant, despite the additional technical challenges this setting presents.

Contribution. In this work, we introduce the Anchored Fitted Q-Iteration and establish the sample complexity on average reward MDPs with general function approximation for weakly communicating MDPs for the first time. We consider the cases with IID data and with single-trajectory data. Then, we show that by using the relative normalization mechanism from the classical relative value iteration, we can further improve the sample complexity.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Prior works	MDP class	dataset	Coverage coefficient
Ozdaglar et al. [63]	ergodic*	IID samples	partial
Gabbianelli et al. [30]	unichain* (+ linear)	IID samples	partial
Our work	weakly communicating*	IID samples	full
Our work	weakly communicating*	β-mixing single-trajectory	full

Table 1: Comparison of analyses of offline average-reward MDPs. Our work, which assumes the MDP is weakly communicating, significantly relaxes the structural assumption on the MDP compared to prior work. (Clarification*: Ergodic, unichain, and weakly communicating are respectively the standard MDP classes for which the results of [63], [30], and our work apply. However, the precise conditions are slightly more general in each case. See Section 1.2 for detailed definitions.)

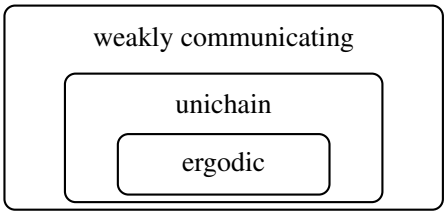

Figure 1: The MDP classes satisfy the inclusion: ergodic $\subset$ unichain $\subset$ weakly communicating

1.1 Preliminaries and notations

We briefly review the basic notions of average-reward Markov decision processes (MDPs) and reinforcement learning (RL) and refer the readers to standard references for further details [66, 7, 78].

Average-reward MDP. Let $\mathcal{M}(\mathcal{X})$ be the space of probability distributions over $\mathcal{X}$ and $\mathcal{F}(\mathcal{X})$ be the space of bounded real-valued functions over $\mathcal{X}$. Write $(\mathcal{S}, \mathcal{A}, P, r)$ to denote an infinite-horizon undiscounted MDP with finite state space $\mathcal{S}$, finite action space $\mathcal{A}$, transition matrix $P\colon \mathcal{S} \times \mathcal{A} \to \mathcal{M}(\mathcal{S})$, bounded reward $r\colon \mathcal{S} \times \mathcal{A} \to [-R, R]$. Denote $\pi\colon \mathcal{S} \to \mathcal{M}(\mathcal{A})$ for a policy, $g^\pi(s,a) = \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}_\pi \left[\sum_{t=1}^T r(s_t, a_t) \,|\, s_0 = s, a_0 = a\right]$ for the average-reward of a policy π given an initial state-action pair (s,a), where $\mathbb{E}_\pi$ denotes the expectation over all trajectories $(s_0, a_0, s_1, a_1, \ldots, s_T, a_T)$ induced by P and π.

We say $\pi_\star$ is an optimal policy if $g^{\pi_\star}(s,a) = \max_\pi g^\pi(s,a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, and we say $g^{\pi_\star}$ is the optimal average reward. (The optimal policy and optimal average reward exists for finite state-action space [66, Theorem 9.1.8].) We say π is an ϵ-optimal policy if $\|g^{\pi_\star} - g^\pi\|_\infty \le \epsilon$. Define $\mathcal{P}^\pi$ as

$$\mathcal{P}^\pi((s,a) \to (s',a')) = \mathrm{Prob}((s,a) \to (s',a') \,|\, s' \sim P(\cdot \,|\, s,a), a' \sim \pi(\cdot \,|\, s')),$$

the transition matrix induced by policy π. Then, $(\mathcal{P}^\pi Q)(s,a) = \mathbb{E}_{a'\sim\pi(\cdot\,|\,s'), s'\sim P(\cdot\,|\,s,a)}[Q(s',a')]$ for $Q \in \mathcal{F}(\mathcal{S}\times\mathcal{A})$. Define the weighted L_p-norm of $Q \in \mathcal{F}(\mathcal{S}\times\mathcal{A})$ under state-action distribution ρ as $\|Q\|_{p,\rho} = [\mathbb{E}_{(s,a)\sim\rho}|Q(s,a)|^p]^{1/p}$ for $p \ge 1$.

Coverage coefficient. A coverage coefficient quantifies the shift between the distribution of the offline data and the distribution induced by policies [61, 18]. Loosely speaking, the *full coverage* assumption, as stated in Table 1, assumes that the offline data sufficiently explores the whole state-action space regardless of policy [4, 92], while *partial coverage* only requires the offline data to sufficiently explore the state-action pairs that an optimal policy would visit [94, 42]. These types of assumptions are fundamentally necessary for the complexity analysis of offline RL [18], and different works use different types of coverage coefficients (cf. [80, 71]). The coverage coefficient we use is defined in Section 3.

Value Iteration. Given an undiscounted MDP $(\mathcal{S}, \mathcal{A}, P, r)$, the Bellman optimality operator T is

$$TQ(s,a) = r(s,a) + \mathbb{E}_{s' \sim P(\cdot \mid s,a)} \left[\max_{a' \in \mathcal{A}} Q(s', a') \right]$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. We define the standard Value Iteration (VI) as

$$Q^k = TQ^{k-1} \qquad \text{for } k = 1, 2, \ldots, K,$$

where Q^0 is an initial point.

MDP classes. MDPs are classified according to the structure of the transition matrices. (For definitions on irreducible classes, recurrent classes, transient states, and aperiodicity of transition matrices, refer to [66, Appendix A.2].) An MDP is *ergodic* if the transition matrices induced by every policy π has a single recurrent class and is aperiodic. An MDP is *unichain* if the transition matrices induced by every policy π has a single recurrent class plus a possibly empty set of transient states. An MDP is *weakly communicating* if there is a set of states where every state in the set is accessible from every other state in that set under some policy, plus a possibly empty set of states that are transient for all policies. Otherwise, in general, an MDP is *multichain*. We note that classification of MDPs is crucial in the analyses of average-reward MDPs [96, 97, 88, 50].

1.2 Conditions of Prior works

In Table 1, we remarked that the precise conditions on the MDPs are slightly general than ergodic, unichain, and weakly communicating. In this section, we state the precise conditions.

Uniform mixing [63]. The *uniform mixing* condition assumes that there exist positive $t_{mix} \in \mathbb{N}$ (which does not depend on π and ρ) such that $\|\rho^\top (\mathcal{P}^\pi)^t - \nu^\pi\|_1 \leq 1/2$ for all $t \geq t_{max}$ for any policy π and initial distribution ρ, where ν^π is the stationary distribution of $\mathcal{P}^\pi$. (This condition requires that the stationary distribution ν^π is unique for all π.) The prior work [63] uses this assumption for its analysis of average-reward MDPs. Ergodic MDPs satisfy the uniform mixing condition [13, 52], but unichain MDPs do not [66, Example 8.2.1].

All-policy Bellman equation and linear MDPs [30]. The *all-policy Bellman equation* states that for any policy π, the average reward g^π does not depend on (s, a) and there exist a $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that

$$r(s,a) + \mathbb{E}_{a' \sim \pi(\cdot \mid s'), s' \sim P(\cdot \mid s,a)} \left[Q^\pi(s', a') \right] = Q^\pi(s,a) + g^\pi$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. The prior work [30] uses this assumption for its analysis of average-reward MDP. Unichain MDPs satisfy the all-policy Bellman equation while weakly communicating MDPs, in general, do not [66, Section 8.4]. Also, the uniform mixing condition implies the all-policy Bellman equation [88, Lemma 6].

An MDP is *linear* if there exist $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, $\psi : \mathcal{S} \to \mathbb{R}^d$, and $w \in \mathbb{R}^d$ such that

$$r(s,a) = \langle \phi(s,a), w \rangle, \quad P(s' \mid s,a) = \langle \phi(s,a), \psi(s') \rangle.$$

The linear MDP assumption is often used for theoretical analyses [39, 30], but it requires knowledge of the mapping ϕ and ψ and often fails to hold in practice [32, 82].

Bellman optimality equation (our work).

Assumption 1 (Bellman optimality equation)**.** *The optimal average reward $g^{\pi_\star}$ does not depend on (s, a) and there exist a $Q^{\pi_\star} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that*

$$r(s,a) + \mathbb{E}_{s' \sim P(\cdot \mid s,a)} \left[\max_{a'} Q^{\pi_\star}(s', a') \right] = Q^{\pi_\star}(s,a) + g^{\pi_\star},$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

A policy $\pi_\star$ satisfying the Bellman optimality equation is an optimal policy [88, Section 2]. The all-policy Bellman equation implies the Bellman optimality equation, and the weakly communicating condition of MDPs also implies the Bellman optimality equation [66, Theorem 8.3.2, 8.4.1].

2 Anchored Fitted Q-Iteration

Consider the offline RL setup with precollected dataset $D = \{s_i, a_i, r_i, s_i'\}_{i=1}^n$, where $r_i = r(s_i, a_i)$ and $s_i' \sim P(\cdot \mid s_i, a_i)$. Let $\mathcal{F}$ be a nonempty function space to approximate Q-value. We now introduce our novel algorithm, *Anchored Fitted Q-Iteration (Anc-F-QI)*.

Algorithm 1 Anchored Fitted Q-Iteration $(D, K, \{\mathcal{F}_i\}_{i=1}^K \{\lambda_i\}_{i=1}^K)$

Input: $D = \{s_i, a_i, r_i, s_i'\}_{i=1}^n$, $f_0 = 0$, $K \geq 1$, $\{\lambda_i\}_{i=1}^K \subset (0, 1)$
for $k = 0, 1, \ldots, K-1$ **do**
  $\hat{T} f_k = \operatorname{argmin}_{f \in \mathcal{F}_{k+1}} \sum_{i=1}^n \left(f(s_i, a_i) - r_i - \max_{a \in \mathcal{A}} f_k(s_i', a)\right)^2$
  $f_{k+1} = (1 - \lambda_{k+1}) f_0 + \lambda_{k+1} \hat{T} f_k$ ▷ With $f_0 = 0$, this is weight decay
end for
$\pi(a \mid s) = \operatorname{argmax}_{a \in \mathcal{A}} f_K(s, a)$
Output π, f_K

In Section 4, we present our sample complexity results for Anc-F-QI. Roughly speaking, Theorem 1 establishes $\tilde{\mathcal{O}}(1/\epsilon^6)$ sample complexity with IID data and Theorem 2 establishes $\tilde{\mathcal{O}}(1/\epsilon^{12})$ sample complexity with β-mixing single-trajectory data. In Section 5, to further improve the sample complexity with the *Relative Anchored* Fitted-Q Iteration, establishing $\tilde{\mathcal{O}}(1/\epsilon^4)$ and $\tilde{\mathcal{O}}(1/\epsilon^8)$ sample complexities for IID and single-trajectory data, respectively.

2.1 The anchor mechanism and weight decay

Our method Anc-F-QI stated as Algorithm 1 consists of two main components. The first component is the first line of the for-loop, the classical Fitted Q-Iteration [25, 61] step without discount factor. Its goal is to find the function $\hat{T} f_k \approx T f_k$, where T is the Bellman operator. However, unlike Fitted Q-Iteration in the discounted cumulative reward case, $\hat{T} f_k \approx T f_k$ is not enough to establish a finite sample complexity in the average-reward setup. In the tabular case where the Fitted Q-Iteration reduces to Value Iteration (VI), it is known that VI might not converge. Specifically, there exists an average-reward MDP such that the policy error of VI does not converge to zero [22, Example 4]. Even if an aperiodicity condition is assumed, VI guarantees only asymptotic convergence without any known explicit convergence rate in the average-reward setup [66, Theorem 9.4.5].

Recently, Anchored Value Iteration (Anc-VI) was proposed to obtain finite-time bounds of policy error for average-reward MDPs [14, 50, 48]. Particularly, the *Anchored Q-Value Iteration* is

$$Q^k = (1 - \lambda_k) Q^0 + \lambda_k T Q^{k-1} \qquad \text{for } k = 1, 2, \ldots. \tag{Anc-QI}$$

where λ_k parameter is to be chosen. Compared to the standard VI, Anc-QI obtains the next iterate as a convex combination between the output of T and the *starting point* Q^0. We call the $(1 - \lambda_k) Q_0$ term the *anchor term* since it serves to pull the iterates back toward the starting point Q_0. With this Anc-VI, [50] establish non-asymptotic convergence in the average-reward setup. Specifically, Anc-VI exhibits the $O(1/k)$-rate in terms of policy error [50][Theorem 2 and Corollary 2] without any restrictions on the MDP.

This anchoring mechanism, classically also known as the Halpern iteration [33], has been widely studied in minimax optimization and fixed-point problems [70, 55, 65, 20, 93]. In the context of reinforcement learning, [49, 50] applied the anchoring mechanism to VIs for cumulative-return and average-reward MDPs under the tabular setting, and [14, 48] applied the anchoring mechanism to Q-Value Iteration for cumulative-return and average reward MDPs under the generative model setting.

In this work, we combine Fitted Q-Iteration with anchoring, as shown in the second line of the for-loop of Algorithm 1, and establish finite-time bounds on the sample complexity.

2.2 Assumptions on the function space $\mathcal{F}$

Assumption 2 (existence of argmin)**.** *In Anc-F-QI, the argmin defining $\hat{T} f_k$ exist for $k = 0, \ldots, K-1$.*

This assumption is needed for the regression step of Algorithm 1 to be well defined.

Assumption 3 (star-shaped function space)**.** *If $f \in \mathcal{F}, \eta f \in \mathcal{F}$ for all $\eta \in [0, 1]$.*

This assumption implies that the anchor step of Anc-F-QI to be well defined. Star-shaped function space is a classical notion that relaxes convexity [31, 34, 51], and if $\mathcal{F}$ corresponds to a parametrized neural network with a linear layer as the output layer, $\mathcal{F}$ is star-shaped.

Definition 1 (Inherent Bellman error)**.** *Define $\epsilon_B(\mathcal{F}, \mathcal{F}') = \max_{f \in \mathcal{F}} \min_{f' \in \mathcal{F}'} \|f' - Tf\|$ as the inherent Bellman error with respect to the norm $\|\cdot\|$.*

The inherent Bellman error ϵ_B quantifies the error due to the function spaces $\mathcal{F}, \mathcal{F}'$ in approximating the output of the Bellman operator [61, 3, 18]. Note that if the function spaces $\mathcal{F}, \mathcal{F}'$ are bounded (in the $\|\cdot\|_\infty$-norm), then ϵ_B is also bounded.

Assumption 4 (Bellman completeness)**.** *$\epsilon_B(\mathcal{F}, \mathcal{F}') = 0$, where ϵ_B the is inherent Bellman error.*

Bellman completeness states that if $f \in \mathcal{F}$, then $Tf \in \mathcal{F}'$. I.e., $\mathcal{F}, \mathcal{F}'$ are closed under the Bellman operator. Although the Bellman completeness assumption is seemingly strong, it is often considered in sample complexity analyses in the offline RL literature [18, 26]. In fact, the Bellman completeness condition is fundamental in the sense that the prior work [28] showed that a polynomial sample complexity cannot be established without Bellman completeness assumption.

3 Approximate Anchored Q-Value Iteration

In this section, we conduct an L_p bound analysis that will later be used to establish the main sample complexity results of Sections 4 and 5. Define the *Approximate Anchored Q-Value Iteration* as

$$Q^k = (1 - \lambda_k)Q^0 + \lambda_k(TQ^{k-1} + \epsilon_k) \tag{Apx-Anc-QI}$$

for $k = 1, 2, \ldots, K$, where T is the Bellman operator, $Q^0 \in \mathbb{R}^n$ is a starting point, and ϵ_k represents the evaluation error of TQ^{k-1}. We choose $\lambda_k = \frac{k}{k+2}$ for $k = 1, \ldots, K$, motivated by [70, 20].

We now establish a convergence analysis of Apx-Anc-QI based on L_p bounds of ϵ_k. Similar to the prior work [59, 60, 26], we assume the following coverage coefficient for our analysis.

Assumption 5 (uniform stochastic transition)**.** *For a given distribution μ on $\mathcal{S} \times \mathcal{A}$,*

$$C_\mu \stackrel{\text{def}}{=} \sup_{s,a,\pi} \left\| \frac{\mathcal{P}^\pi(\cdot \mid s, a)}{\mu(\cdot)} \right\|_\infty < \infty.$$

Assumption 6 (uniform future state distribution)**.** *For given distributions μ and ρ on $\mathcal{S} \times \mathcal{A}$,*

$$C_{\mu,\rho} \stackrel{\text{def}}{=} \sup_{\pi_1, \pi_2, \ldots \pi_k} \left\| \frac{\rho^\top \mathcal{P}^{\pi_1} \mathcal{P}^{\pi_2} \ldots \mathcal{P}^{\pi_k}(\cdot)}{\mu(\cdot)} \right\|_\infty < \infty,$$

where $\pi_1, \pi_2, \ldots \pi_k$ represents an arbitrary sequence of policies.

The coverage coefficients measure the mismatch between the distribution of offline data and the distribution induced by the transition matrices and initial distributions. We note that Assumption 5 implies Assumption 6 with $C_{\mu,\rho} \leq C_\mu$ [61, Section 5].

Proposition 1. *Let $p \in [1, \infty]$, and let μ and ρ be distributions on $\mathcal{S} \times \mathcal{A}$. Under Assumption 1 and 5 (Bellman optimality equation, uniform stochastic transition), the policy error of Apx-Anc-QI with $\lambda_k = \frac{k}{k+2}$ satisfies*

$$\|g^{\pi_\star} - g^{\pi_K}\|_\infty \leq C_\mu^{1/p} \frac{8}{K+2} \|Q^{\pi_\star} - Q^0\|_{p,\mu} + C_\mu^{1/p} \frac{2K}{3} \max_{1 \leq k \leq K} \|\epsilon_k\|_{p,\mu}.$$

Similarly, under Assumption 1 and 6 (Bellman optimality equation, uniform future state distribution), the policy error of Apx-Anc-QI with $\lambda_k = \frac{k}{k+2}$ satisfies

$$\|g^{\pi_\star} - g^{\pi_K}\|_{p,\rho} \leq C_{\mu,\rho}^{1/p} \frac{8}{K+2} \|Q^{\pi_\star} - Q^0\|_{p,\mu} + C_{\mu,\rho}^{1/p} \frac{2K}{3} \max_{1 \leq k \leq K} \|\epsilon_k\|_{p,\mu}.$$

To clarify, the g-, Q-, and ϵ-terms in Proposition 1 are functions of (s, a) and the norms $\|\cdot\|_{p,\rho}$ and $\|\cdot\|_{p,\mu}$ are taking expectations with respect to the distributions ρ and μ.

The bounds of Proposition 1 serve as the technical crux of our sample complexity results later presented in Theorems 1, 2, 3, and 4. In the bound, the first term decreases with order $\mathcal{O}(1/K)$ but the second error term increases with order $\Theta(K)$. Therefore, our subsequent arguments will ensure $\|\epsilon_k\|_{p,\mu} = \mathcal{O}(1/K^2)$ by using sufficient offline samples.

4 Sample complexity of Anchored Fitted Q-Iteration

We now present sample complexity analyses of Anc-F-QI with IID and single-trajectory data.

4.1 Range of function space $\mathcal{F}$

Before analyzing the complexity of those, we explain our issue on range of function space in average-reward setup and our choice of function space.

When considering Fitted Q-Iteration in the discounted reward setup, the functions are often assumed to be bounded by $\left\|Q_\gamma^\star\right\|_\infty$ [61, 18], where $Q_\gamma^\star$ is optimal state-action function with discount factor γ, since $\|f\|_\infty \leq \left\|Q_\gamma^\star\right\|_\infty$ implies $\|Tf\|_\infty \leq \left\|Q_\gamma^\star\right\|_\infty$. In the average-reward setup (without discounting), this property does not hold, and the Fitted Q-Iteration is expected to produce an unbounded sequence of functions. To address this issue, we allow the range of the function space to increase with each iteration.

Assumption 7 (increasing function range)**.** *Let $\mathcal{F}_0 = \{0\}$ and $\mathcal{F}_k \subset \{f : \mathcal{S} \times \mathcal{A} \to [-kR, kR] : f \in B(S \times A)\}$ and $f_k \in \mathcal{F}_k$ in Anc-F-QI for all k .*

Roughly speaking,

$$\|f_k\| \sim \|Tf_{k-1}\|_\infty = \left\| r + P \max_{a \in \mathcal{A}} f_{k-1} \right\|_\infty \lesssim R + \|f_{k-1}\|_\infty \lesssim kR + \|f_0\|_\infty ,$$

so we increase the function bound as kR.

4.2 IID dataset

In this subsection, we study sample complexity with IID dataset.

Assumption 8 (IID dataset)**.** *There is a distribution μ such that the dataset is $D = \{s_i, a_i, r_i, s_i'\}_{i=1}^n$ generated IID with $(s_i, a_i) \sim \mu$ and $s_i' \sim P(\cdot \mid s_i, a_i)$ for $i = 1, \ldots, n$.*

Since we consider possibly infinite function space, as measurement of the capacity of function space, we use covering number [21, 83].

Definition 2. *An ϵ-cover of set S with respect to metric d is a set $\{\theta_i\}_{i=1}^N \subset S$ such that for all $\theta \in S$, there is an $i \in \{1, \ldots, N\}$ such that $d(\theta, \theta^i) \leq \epsilon$. The covering number $\mathcal{N}(\epsilon; S, d)$ is the cardinality of the smallest ϵ-cover. By convention, we define $\mathcal{N}(+\infty; S, d) = 1$.*

We now present lemma which bounds approximation error of Anc-F-QI for IID dataset.

Lemma 1. *Assume Assumptions 1, 2, 3, 7, and 8 (Bellman optimality equation, existence of argmin, star-shaped function space, increasing function range, IID dataset). Let μ be the distribution generating the dataset. Let $\epsilon > 0$ and $\delta > 0$. With probability $1 - \delta$, $\{f_k, \hat{T} f_k\}_{k=0}^{K-1}$ of Anc-F-QI with $\lambda_k = \frac{k}{k+2}$ satisfies*

$$\|Tf_k - \hat{T} f_k\|_{\mu,2}^2 \leq \frac{60(k+2)^2 R^2 \ln(2K N_{k,\epsilon} N_{k+1,\epsilon}/\delta)}{n} + 3\epsilon + 13\epsilon_B(\mathcal{F}_k, \mathcal{F}_{k+1}),$$

where

$$N_{k,\epsilon} = \mathcal{N}(\tfrac{\epsilon}{108(2k+1)R}; \mathcal{F}_k, \|\cdot\|_\infty), \quad \text{for } k = 0, 1, \ldots, K-1.$$

We defer the proofs to Appendix D, but we quickly note that the proof is based on Bernstein inequality and is motivated by [21, 18].

Lemma 1 tells that the square of approximation error of the Bellman operator decreases sublinearly with respect to number of sample. Combining Theorem 1 and Lemma 1, we obtain following sample complexity result of Anc-F-QI with IID dataset.

Theorem 1. *Assume Assumptions 1, 2, 3, 5, 7, and 8 (Bellman optimality equation, existence of argmin, star-shaped function space, uniform stochastic transition, increasing function range, IID dataset). Let μ be the distribution generating the dataset. Let $\epsilon > 0$ and $\delta > 0$. With probability*

$1-\delta$, the policy error of Anc-F-QI with $\lambda_k = \frac{k}{k+2}$ and $K = \lceil 18C_\mu^{1/2}\|Q^{\pi_\star}\|_{2,\mu}/\epsilon \rceil$ satisfies $\|g^{\pi_\star} - g^{\pi_K}\|_\infty \le \epsilon + 3KC_\mu^{1/2} \max_{k=0,\dots,K-1} \sqrt{\epsilon_B(\mathcal{F}_k, \mathcal{F}_{k+1})}$ with sample complexity

$$n = \tilde{\mathcal{O}}\left(\frac{R^2 C_\mu^3 \|Q^{\pi_\star}\|_{2,\mu}^4 \log(N_\epsilon^2/\delta)}{\epsilon^6} \right),$$

where $\tilde{\mathcal{O}}$ ignores all logarithmic factors except the logarithmic dependence on the covering number N_ϵ defined as

$$N_\epsilon = \max_{k=1,\dots,K} N_{k,\epsilon}, \qquad N_{k,\epsilon} = \mathcal{N}\big(\tfrac{\epsilon^4}{10^6 k R C_\mu^2 \|Q^{\pi_\star}\|_{2,\mu}^2}; \mathcal{F}_k, \|\cdot\|_\infty\big), \quad \text{for } k = 1, \dots, K.$$

Alternatively assume Assumptions 1, 2, 3, 6, 7, and 8 (Bellman optimality equation, existence of argmin, star-shaped function space, uniform future state distribution, increasing function range, IID dataset). Let μ be the distribution generating the dataset and ρ be an arbitrary distribution on $\mathcal{S} \times \mathcal{A}$. Let $\epsilon > 0$ and $\delta > 0$. With probability $1-\delta$, the policy error of Anc-F-QI with $\lambda_k = \frac{k}{k+2}$ and $K = \lceil 18C_{\mu,\rho}^{1/2}\|Q^{\pi_\star}\|_{2,\mu}/\epsilon \rceil$, satisfies $\|g^{\pi_\star} - g^{\pi_K}\|_{2,\rho} \le \epsilon + 3KC_{\mu,\rho}^{1/2} \max_{k=0,\dots,K-1} \sqrt{\epsilon_B(\mathcal{F}_k, \mathcal{F}_{k+1})}$ with sample complexity

$$n = \tilde{\mathcal{O}}\left(\frac{R^2 C_{\mu,\rho}^3 \|Q^{\pi_\star}\|_{2,\mu}^4 \log(N_\epsilon^2/\delta)}{\epsilon^6} \right),$$

where $\tilde{\mathcal{O}}$ ignores all logarithmic factors except the logarithmic dependence on the covering number N_ϵ defined as

$$N_\epsilon = \max_{k=1,\dots,K} N_{k,\epsilon}, \qquad N_{k,\epsilon} = \mathcal{N}\big(\tfrac{\epsilon^4}{10^6 k R C_{\mu,\rho}^2 \|Q^{\pi_\star}\|_{2,\mu}^2}; \mathcal{F}_k, \|\cdot\|_\infty\big), \quad \text{for } k = 1, \dots, K$$

In the Appendix D, we show the full sample complexity with the logarithmic factors.

Under the additional assumption of Bellman completeness ($\epsilon_B = 0$), this theorem guarantee that Anc-F-QI produces an ϵ-optimal policy with $\tilde{\mathcal{O}}(1/\epsilon^6)$ sample complexity. To the best of our knowledge, this is the first sample complexity result only assuming the Bellman optimality equation or a weakly communicating MDP. In Section 5, we improve this sample complexity to $\tilde{\mathcal{O}}(1/\epsilon^4)$ using the relative normalization mechanism.

4.3 Single-trajectory dataset

In this subsection, we study sample complexity with single-trajectory dataset.

Assumption 9 (single-trajectory dataset)**.** *For given behavior policy π_b and initial distribution ν on $\mathcal{S}$, dataset is $D = \{s_i, a_i, r_i\}_{i=1}^n$ where $s_1 \sim \nu$, $a_i \sim \pi_b(\cdot \mid s_i)$, $s_{i+1} \sim P(\cdot \mid s_i, a_i)$.*

The main technical challenge with single-trajectory data is handling the dependency between samples. Following [4, 3], we introduce the following β-mixing condition ensuring that samples are sufficiently representative and rapidly mixing.

Definition 3 (β-mixing)**.** *Let $\{Z_t\}_{t=1}^\infty$ be a stochastic process. Denote by $Z^{1:t}$ the collection of $(Z_1, \dots, Z_t)$ where we allowed $t = \infty$. Let $\sigma(Z^{i:j})$ denote the σ-algebra generated by $Z^{i:j} (i \le j)$. The m-th β-mixing coefficient of $\{Z_t\}$ is defined as*

$$\beta_m = \sup_{t \ge 1} \mathbb{E}\left[\sup_{B \in \sigma(Z^{t+m:\infty})} \big| P(B \mid Z^{1:t}) - P(B) \big| \right].$$

$\{Z_t\}$ is said to be β-mixing if $\beta_m \to 0$ as $m \to \infty$. In particular, we say that a β-mixing process mixes at exponent rate with parameters $\bar{\beta}, b, \kappa > 0$ if $\beta_m \le \bar{\beta} exp(-b m^\kappa)$ holds for all $m \ge 0$.

Roughly speaking, the β-mixing condition ensures that future samples depend weakly on the past samples. We assume that our single-trajectory is β-mixing and the distribution is in a steady state, following [4, 3].

Assumption 10 (β-mixing single-trajectory)**.** *For single-trajectory dataset $\{s_i, a_i, r_i\}_{i=1}^n$, assume that s_i is strictly stationary with $s_i \sim \nu$ and β-mixing at exponent rate with parameters $\bar{\beta}, b, \kappa > 0$.*

Again, following [4, 3], as measurement of the capacity of function space, we use pseudo dimension which has been widely studied for complexity analyses with various function classes [2, 83].

Definition 4 (pseudo dimension)**.** *For a given function class $\mathcal{F}$ of binary-valued functions, we say the set $x_1^n = (x_1, \ldots, x_n)$ is shattered by $\mathcal{F}$ if cardinality of $\{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\}$ is 2^n. The VC-dimension $V_{\mathcal{F}}$ of $\mathcal{F}$ is defined as the largest integer n such that there exist the set x_1^n shattered by $\mathcal{F}$. For a given class $\mathcal{F}$ of real-valued functions, the pseudo-dimension $V_{\mathcal{F}}$ of is defined as the VC-dimension of the set of indicator function of the subgraphs of functions in $\mathcal{F}$.*

We now present lemma which bounds approximation error of of Anc-F-QI for single-trajectory dataset.

Lemma 2. *Assume Assumptions 1, 2, 3, 7, 9, and 10 (Bellman optimality equation, existence of argmin, star-shaped function space, increasing function range, single-trajectory dataset, β-mixing single-trajectory). Let μ be the distribution generating the dataset defined as $\mu(s,a) = \nu(s)\pi_b(a \,|\, s)$. Let $\epsilon > 0$ and $\delta > 0$. With probability $1-\delta$, $\{f_k, \hat{T} f_k\}_{k=0}^{K-1}$ of Anc-F-QI with $\lambda_k = \frac{k}{k+2}$ satisfies*

$$\|Tf_k - \hat{T} f_k\|_{\mu,2}^2 \leq \sqrt{\frac{c_{0,k}(\max\{c_{0,k}/b, 1\})^{1/\kappa}}{c_{2,k} n}} + \epsilon_B(\mathcal{F}_k, \mathcal{F}_{k+1}),$$

where $c_{0,k} = (V_{\mathcal{F}_{k+1}} + V_{(\mathcal{F}_k)_{max}}) \log n/2 + \log(e/(K\delta)) + \log(\max(c_{1,k}, \bar{\beta}))$, $c_{1,k} = 16e^2(V_{\mathcal{F}_{k+1}} + 1)(V_{(\mathcal{F}_k)_{max}} + 1)(24e)^{V_{\mathcal{F}_{k+1}} + V_{(\mathcal{F}_k)_{max}}}$, $c_{2,k} = \frac{1}{512(2k+3)^4 R^4}$, $V_{(\mathcal{F}_k)_{max}} = 2|\mathcal{A}| V_{\mathcal{F}_k} \log(3|\mathcal{A}|)$.

We defer the proofs to Appendix D, but we quickly note that the proof strategy closely follow [4, 3] and relies on the Hoeffding inequality under a mixing condition.

Rougbly speaking, Lemma 2 tells that the square of approximation error of the Bellman operator decreases at a $1/\sqrt{n}$ rate with respect to number of sample. Combining Theorem 1 and Lemma 2, we obtain following sample complexity result of Anc-F-QI with single-trajectory dataset.

Theorem 2. *Assume Assumptions 1, 2, 3, 5, 7, 9, and 10 (Bellman optimality equation, existence of argmin, star-shaped function space, uniform stochastic transition, increasing function range, single-trajectory dataset, β-mixing single-trajectory). Let μ be the distribution generating the dataset defined as $\mu(s,a) = \nu(s)\pi_b(a \,|\, s)$. Let $\epsilon > 0$ and $\delta > 0$. With $1-\delta$ probability, the policy error of Anc-F-QI with $\lambda_k = \frac{k}{k+2}$ and $K = \lceil 9C_\mu^{1/2} \|Q^{\pi_\star}\|_{2,\mu}/\epsilon \rceil$ satisfies $\|g^{\pi_\star} - g^{\pi_K}\|_\infty \leq \epsilon + KC_\mu^{1/2} \max_{k=0,\ldots,K-1} \sqrt{\epsilon_B(\mathcal{F}_k, \mathcal{F}_{k+1})}$ with sample complexity*

$$n = \tilde{\mathcal{O}}\left(1/\epsilon^{12}\right),$$

where $\tilde{\mathcal{O}}$ only shows the dependence on ϵ. Alternatively, Assume Assumptions 1, 2, 3, 6, 7, 9, and 10 (Bellman optimality equation, existence of argmin, star-shaped function space, uniform future state distribution, increasing function range, single-trajectory dataset, β-mixing single-trajectory). Let μ be the distribution generating the dataset defined as $\mu(s,a) = \nu(s)\pi_b(a \,|\, s)$ and ρ be an arbitrary distribution on $\mathcal{S} \times \mathcal{A}$. Let $\epsilon > 0$ and $\delta > 0$. With $1-\delta$ probability, the policy error of Anc-F-QI with $\lambda_k = \frac{k}{k+2}$ and $K = \lceil 9C_{\mu,\rho}^{1/2} \|Q^{\pi_\star}\|_{2,\mu}/\epsilon \rceil$ satisfies $\|g^{\pi_\star} - g^{\pi_K}\|_{2,\rho} \leq \epsilon + KC_{\mu,\rho}^{1/2} \max_{k=0,\ldots,K-1} \sqrt{\epsilon_B(\mathcal{F}_k, \mathcal{F}_{k+1})}$ with sample complexity

$$n = \tilde{\mathcal{O}}\left(1/\epsilon^{12}\right),$$

where $\tilde{\mathcal{O}}$ only shows the dependence on ϵ.

In the Appendix D, we show the full sample complexity with all of the factors.

To the best of our knowledge, this is the first sample complexity result with single-trajectory data in the average-reward setup. In Section 5, we improve this sample complexity to $\tilde{\mathcal{O}}(\epsilon^{-8})$ using the relative normalization mechanism

5 Relative Anchored Fitted Q-Iteration

In this section, we propose *Relative Anchored Fitted Q-Iteration (R-Anc-F-QI)* and improve the sample complexity. We are motivated by the classical *relative value iteration* [90]. In the tabular case, it is known that standard VI diverges in the average-reward setup [66, Theorem 9.4.1], and relative value iteration normalizes the divergent vectors [66, Section 8.5.5]. In the case of (Anchored) Fitted Q-Iteration, this normalization allows the f_k functions to be bounded and removes the inefficiency associated with the increasing function classes described in Section 4.1.

Algorithm 2 Relative Anchored Fitted Q-Iteration $(D, K, \mathcal{F}, \{\lambda_i\}_{i=1}^K)$

Input: $D = \{s_i, a_i, r_i, s_i'\}_{i=1}^n$, $f_0 = 0$, $K \geq 1$, $\{\lambda_i\}_{i=1}^K \subset (0,1)$
for $k = 0, 1, \ldots, K-1$ **do**
  $\hat{T} f_k = \operatorname{argmin}_{f \in \mathcal{F}} \sum_{i=1}^n \left(f(s_i, a_i) - r_i - \max_{a \in \mathcal{A}} f_k(s_i', a)\right)^2$
  $f_{k+1} = (1 - \lambda_{k+1}) f_0 + \lambda_{k+1}(\hat{T} f_k - \frac{\max \hat{T} f_k + \min \hat{T} f_k}{2} \mathbf{1})$
end for
$\pi(a \,|\, s) = \operatorname{argmax}_{a \in \mathcal{A}} f_K(s, a)$
Output π, f_K

The only difference with Anchored Fitted Q-Iteration is the subtraction of $\frac{\max \hat{T} f_k + \min \hat{T} f_k}{2} \mathbf{1}$ in the second line of the for-loop. By direct calculation, we can check that $\left\| f - \frac{\max f + \min f}{2} \mathbf{1} \right\|_\infty \leq \|f\|_\infty$ and subtracting a uniform constant does not effect on greedy policy due the fact that the Bellman operator satisfies $T(c\mathbf{1} + x) = c\mathbf{1} + T(x)$. Thus, we can still apply Proposition 1 to Relative Anchored Fitted Q-Iteration.

Assumption 11 (normalized function space)**.** *If $f \in \mathcal{F}$, $f - \frac{\max f + \min f}{2} \mathbf{1} \in \mathcal{F}$.*

This assumption ensures that the normalization operation is well ldefined.

Assumption 12 (range of function space)**.** *$\mathcal{F} \subset \{f : \mathcal{S} \times \mathcal{A} \to [-2\|Q^{\pi_\star}\|_\infty, 2\|Q^{\pi_\star}\|_\infty] \,|\, f \in B(S \times A)\}$, where $Q^{\pi_\star}$ is solution of Bellman optimality equation.*

Now, unlike increasing function range used for the non-relative Anchored Fitted Q-Iteration, we now have a function space bounded by $Q^{\pi_\star}$. This difference leads to improved efficiency as the following sample complexity results show.

Theorem 3. *Assume Assumptions 1, 2, 3, 5, 8, 11, and 12 (Bellman optimality equation, existence of argmin, star-shaped function space, uniform stochastic transition, normalized function space, range of function space, IID dataset). Let μ be the distribution generating the dataset. Let $\epsilon > 0$ and $\delta > 0$. With probability $1-\delta$, the policy error of R-Anc-F-QI with $\lambda_k = \frac{k}{k+2}$ and $K = \lceil 18 C_\mu^{1/2} \|Q^{\pi_\star}\|_{2,\mu} / \epsilon \rceil$ satisfies $\|g^{\pi_\star} - g^{\pi_K}\|_\infty \leq \epsilon + 3KC_\mu^{1/2} \sqrt{\epsilon_B(\mathcal{F}, \mathcal{F})}$ with sample complexity*

$$n = \tilde{\mathcal{O}}\left(\frac{(R + \|Q^{\pi_\star}\|_\infty)^2 \|Q^{\pi_\star}\|_\infty^2 C_\mu^3 \log(N_\epsilon^2 / \delta)}{\epsilon^4} \right),$$

where $\tilde{\mathcal{O}}$ ignores all logarithmic factors except the logarithmic dependence on the covering number N_ϵ defined as

$$N_\epsilon = \mathcal{N}\left(\tfrac{\epsilon^4}{10^6 C_\mu^2 (R + \|Q^{\pi_\star}\|_\infty) \|Q^{\pi_\star}\|_\infty^2}; \mathcal{F}, \|\cdot\|_\infty \right).$$

Alternatively, assume Assumptions 1, 2, 3, 6, 8, 11, and 12 (Bellman optimality equation, existence of argmin, star-shaped function space, uniform future state distribution, normalized function space, range of function space, IID dataset) Let μ be the distribution generating the dataset and ρ be an arbitrary distribution on $\mathcal{S} \times \mathcal{A}$. Let $\epsilon > 0$ and $\delta > 0$. With probability $1 - \delta$, the policy error of R-Anc-F-QI with $\lambda_k = \frac{k}{k+2}$ and $K = \lceil 18 C_{\mu,\rho}^{1/2} \|Q^{\pi_\star}\|_{2,\mu} / \epsilon \rceil$ satisfies $\|g^{\pi_\star} - g^{\pi_K}\|_{2,\rho} \leq \epsilon + 3KC_{\mu,\rho}^{1/2} \sqrt{\epsilon_B(\mathcal{F}, \mathcal{F})}$ with sample complexity

$$n = \tilde{\mathcal{O}}\left(\frac{(R + \|Q^{\pi_\star}\|_\infty)^2 \|Q^{\pi_\star}\|_\infty^2 C_{\mu,\rho}^3 \log(N_\epsilon^2 / \delta)}{\epsilon^4} \right),$$

where $\tilde{\mathcal{O}}$ ignores all logarithmic factors except the logarithmic dependence on the covering number N_ϵ defined as

$$N_\epsilon = \mathcal{N}\big(\tfrac{\epsilon^4}{10^6 C_{\mu,\rho}^2 (R+\|Q^{\pi_\star}\|_\infty)\|Q^{\pi_\star}\|_\infty^2}; \mathcal{F}, \|\cdot\|_\infty\big).$$

Theorem 4. *Assume Assumptions 1, 2, 3, 5, 9, 10, 11, and 12 (Bellman optimality equation, existence of argmin, star-shaped function space, uniform stochastic transition, normalized function space, range of function space, single-trajectory dataset, β-mixing single-trajectory). Let μ be the distribution generating the dataset defined as $\mu(s,a) = \nu(s)\pi_b(a \mid s)$. Let $\epsilon > 0$ and $\delta > 0$. With probability $1-\delta$, the policy error of Anc-F-QI with $\lambda_k = \frac{k}{k+2}$ and $K = \lceil 9C_\mu^{1/2}\|Q^{\pi_\star}\|_{2,\mu}/\epsilon \rceil$ satisfies $\|g^{\pi_\star} - g^{\pi_K}\|_\infty \le \epsilon + KC_\mu^{1/2}\sqrt{\epsilon_B(\mathcal{F},\mathcal{F})}$ with sample complexity*

$$n = \tilde{\mathcal{O}}\left(1/\epsilon^8\right),$$

where $\tilde{\mathcal{O}}$ only shows the dependence on ϵ. Alternatively, assume Assumptions 1, 2, 3, 6, 9, 10, 11, and 12 (Bellman optimality equation, existence of argmin, star-shaped function space, uniform future state distribution, normalized function space, range of function space, single-trajectory dataset, β-mixing single-trajectory). Let μ be the distribution generating the dataset defined as $\mu(s,a) = \nu(s)\pi_b(a \mid s)$ and ρ be an arbitrary distribution on $\mathcal{S}\times\mathcal{A}$. Let $\epsilon > 0$ and $\delta > 0$. With probability $1-\delta$, the policy error of Anc-F-QI with $\lambda_k = \frac{k}{k+2}$ and $K = \lceil 9C_{\mu,\rho}^{1/2}\|Q^{\pi_\star}\|_{2,\mu}/\epsilon \rceil$ satisfies $\|g^{\pi_\star} - g^{\pi_K}\|_{2,\rho} \le \epsilon + KC_{\mu,\rho}^{1/2}\sqrt{\epsilon_B(\mathcal{F},\mathcal{F})}$ with sample complexity

$$n = \tilde{\mathcal{O}}\left(1/\epsilon^8\right),$$

where $\tilde{\mathcal{O}}$ only shows the dependence on ϵ.

Indeed, with the relative normalization mechanism, we improve the sample complexities from $\tilde{\mathcal{O}}(1/\epsilon^6)$ to $\tilde{\mathcal{O}}(1/\epsilon^4)$ and $\tilde{\mathcal{O}}(1/\epsilon^{12})$ to $\tilde{\mathcal{O}}(1/\epsilon^8)$ for IID and single-trajectory data cases, respectively,

6 Conclusion

In this work, we introduced Anchored Fitted Q-Iteration (Anc-F-QI) and established new sample complexity results for the average-reward offline RL with general function approximation under the assumption of weakly communicating MDPs. Our approach combines the classical Fitted Q-Iteration with an anchoring mechanism, and the anchor mechanism is the crucial component that enables the finite-time analysis. Roughly speaking, we establish a $\tilde{\mathcal{O}}(1/\epsilon^6)$ sample complexity with IID data and $\tilde{\mathcal{O}}(1/\epsilon^{12})$ sample complexity with single-trajectory data. Then, using the relative normalization technique, we improve the sample complexity to $\tilde{\mathcal{O}}(1/\epsilon^4)$ and $\tilde{\mathcal{O}}(1/\epsilon^8)$ for IID and single-trajectory data, respectively.

One limitation of this work is the reliance on *full* coverage coefficients as described in Assumptions 5 and 6. Some prior work, such as [64] and [30], utilizes a weaker assumption that we refer to as *partial* coverage coefficients, albeit with much stronger structural assumptions on the MDP. Extending our analysis to relax the full coverage coefficient would be a worthwhile direction of future work. Another possible direction of future work is to utilize variance reduction techniques in the style of [84, 75, 48] to further improve the sample complexity.

Acknowledgments and Disclosure of Funding

This work is supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (No.RS-2024-00421203).

References

[1] R. Agarwal, D. Schuurmans, and M. Norouzi. An optimistic perspective on offline reinforcement learning. *International Conference on Machine Learning*, 2020.

[2] M. Anthony and P. L. Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.

[3] A. Antos, C. Szepesvári, and R. Munos. Fitted Q-iteration in continuous action-space MDPs. *Neural Information Processing Systems*, 2007.

[4] A. Antos, C. Szepesvári, and R. Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 2008.

[5] Q. Bai, W. U. Mondal, and V. Aggarwal. Regret analysis of policy gradient algorithm for infinite horizon average reward markov decision processes. *International Conference on Artificial Intelligence*, 2024.

[6] R. Bellman. A Markovian decision process. *Journal of Mathematics and Mechanics*, 6(5):679–684, 1957.

[7] D. P. Bertsekas. *Dynamic Programming and Optimal Control, volume II*. Athena Scientific, 4th edition, 2012.

[8] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[9] D. Blackwell. Discrete dynamic programming. *The Annals of Mathematical Statistics*, 33:719–726, 1962.

[10] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

[11] V. Boone and Z. Zhang. Achieving tractable minimax optimal regret in average reward MDPs. *Neural Information Processing Systems*, 2024.

[12] H. Bourel, A. Jonsson, O.-A. Maillard, and M. S. Talebi. Exploration in reward machines with low regret. *International Conference on Artificial Intelligence and Statistics*, 2023.

[13] R. C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

[14] M. Bravo and J. P. Contreras. Stochastic Halpern iteration in normed spaces and applications to reinforcement learning. *arXiv preprint arXiv:2403.12338*, 2024.

[15] A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

[16] M. Carrasco and X. Chen. Mixing and moment properties of various garch and stochastic volatility models. *Econometric Theory*, 18(1):17–39, 2002.

[17] J. Chang, M. Uehara, D. Sreenivas, R. Kidambi, and W. Sun. Mitigating covariate shift in imitation learning via offline data with partial coverage. *Neural Information Processing Systems*, 2021.

[18] J. Chen and N. Jiang. Information-theoretic considerations in batch reinforcement learning. *International Conference on Machine Learning*, 2019.

[19] M. Chen, A. Beutel, P. Covington, S. Jain, F. Belletti, and E. H. Chi. Top-k off-policy correction for a reinforce recommender system. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 456–464, 2019.

[20] J. P. Contreras and R. Cominetti. Optimal error bounds for non-expansive fixed-point iterations in normed spaces. *Mathematical Programming*, 199(1–2):343–374, 2022.

[21] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39(1):1–49, 2002.

[22] E. Della Vecchia, S. Di Marco, and A. Jean-Marie. Illustrated review of convergence conditions of the value iteration algorithm and the rolling horizon procedure for average-cost MDPs. *Annals of Operations Research*, 199:193–214, 2012.

[23] V. Dewanto, G. Dunn, A. Eshragh, M. Gallagher, and F. Roosta. Average-reward model-free reinforcement learning: a systematic review and literature mapping. *arXiv:2010.08920*, 2020.

[24] Y. Duan, Z. Jia, and M. Wang. Minimax-optimal off-policy evaluation with linear function approximation. *International Conference on Machine Learning*, 2020.

[25] D. Ernst, P. Geurts, and L. Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 2005.

[26] J. Fan, Z. Wang, Y. Xie, and Z. Yang. A theoretical analysis of deep Q-learning. *Learning for dynamics and control*, 2020.

[27] A. Federgruen, P. J. Schweitzer, and H. C. Tijms. Contraction mappings underlying undiscounted Markov decision problems. *Journal of Mathematical Analysis and Applications*, 65(3):711–730, 1978.

[28] D. J. Foster, A. Krishnamurthy, D. Simchi-Levi, and Y. Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. *arXiv preprint arXiv:2111.10919*, 2021.

[29] R. Fruit, M. Pirotta, A. Lazaric, and E. Brunskill. Regret minimization in MDPs with options without prior knowledge. *Neural Information Processing Systems*, 2017.

[30] G. Gabbianelli, G. Neu, M. Papini, and N. M. Okolo. Offline primal-dual reinforcement learning for linear MDPs. *International Conference on Artificial Intelligence and Statistics*, 2024.

[31] R. J. Gardner. *Geometric Tomography*. Cambridge University Press, 1995.

[32] A. Ghosh, S. R. Chowdhury, and A. Gopalan. Misspecified linear bandits. *The Association for the Advancement of Artificial Intelligence*, 2017.

[33] B. Halpern. Fixed points of nonexpanding maps. *Bulletin of the American Mathematical Society*, 73(6):957–961, 1967.

[34] G. Hansen, I. Herburt, H. Martini, and M. Moszyńska. Starshaped sets. *Aequationes Mathematicae*, 94:1001–1092, 2020.

[35] D. Haussler. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

[36] R. A. Howard. *Dynamic Programming and Markov Processes.* John Wiley and Sons, 1960.

[37] G. Hübner. Improved procedures for eliminating suboptimal actions in markov programming by the use of contraction properties. *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, 1977.

[38] T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010.

[39] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan. Provably efficient reinforcement learning with linear function approximation. *Conference on learning theory*, 2020.

[40] Y. Jin, R. Gummadi, Z. Zhou, and J. Blanchet. Feasible Q-learning for average reward reinforcement learning. *International Conference on Artificial Intelligence and Statistics*, 2024.

[41] Y. Jin and A. Sidford. Towards tight bounds on the sample complexity of average-reward MDPs. *International Conference on Machine Learning*, 2021.

[42] Y. Jin, Z. Yang, and Z. Wang. Is pessimism provably efficient for offline RL? *International Conference on Machine Learning*, 2021.

[43] M. R. Kosorok and E. B. Laber. Precision medicine. *Annual review of statistics and its application*, 6(1):263–286, 2019.

[44] A. Kumar, J. Fu, M. Soh, G. Tucker, and S. Levine. Stabilizing off-policy Q-learning via bootstrapping error reduction. *Neural Information Processing Systems*, 2019.

[45] A. Kumar, A. Zhou, G. Tucker, and S. Levine. Conservative Q-learning for offline reinforcement learning. *Neural Information Processing Systems*, 2020.

[46] N. Kumar, Y. Murthy, I. Shufaro, K. Y. Levy, R. Srikant, and S. Mannor. On the global convergence of policy gradient in average reward Markov decision processes. *International Conference on Learning Representations*, 2025.

[47] N. Kumar, K. Wang, K. Y. Levy, and S. Mannor. Efficient value iteration for s-rectangular robust Markov decision processes. *International Conference on Machine Learning*, 2024.

[48] J. Lee, M. Bravo, and R. Cominetti. Near-optimal sample complexity for MDPs via anchoring. *Interantional Conference on Machine Learning*, 2025.

[49] J. Lee and E. Ryu. Accelerating value iteration with anchoring. *Neural Information Processing Systems*, 2023.

[50] J. Lee and E. Ryu. Optimal non-asymptotic rates of value iteration for average-reward MDPs. *International Conference on Learning Representations*, 2025.

[51] O. Leong, E. O'Reilly, and Y. S. Soh. The star geometry of critic-based regularizer learning. *Neural Information Processing Systems*, 2024.

[52] D. A. Levin and Y. Peres. *Markov chains and mixing times*. American Mathematical Soc., 2017.

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

[54] T. Li, F. Wu, and G. Lan. Stochastic first-order methods for average-reward Markov decision processes. *Mathematics of Operations Research*, 2024.

[55] F. Lieder. On the convergence rate of the Halpern-iteration. *Optimization Letters*, 15(2):405–418, 2021.

[56] Y. Liu, A. Swaminathan, A. Agarwal, and E. Brunskill. Provably good batch off-policy reinforcement learning without great exploration. *Neural Information Processing Systems*, 2020.

[57] S. Mahadevan. Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine Learning*, 22(1):159–195, 1996.

[58] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, and et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[59] R. Munos. Error bounds for approximate value iteration. *Association for the Advancement of Artificial Intelligence*, 2005.

[60] R. Munos. Performance bounds in l_p-norm for approximate value iteration. *SIAM journal on control and optimization*, 46(2):541–561, 2007.

[61] R. Munos and C. Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

[62] Y. Murthy and R. Srikant. On the convergence of natural policy gradient and mirror descent-like policy methods for average-reward MDPs. *IEEE Conference on Decision and Control*, pages 1979–1984, 2023.

[63] A. Ozdaglar, S. Pattathil, J. Zhang, and K. Zhang. Offline reinforcement learning via linear-programming with error-bound induced constraints. *arXiv preprint arXiv:2212.13861*, 2024.

[64] A. E. Ozdaglar, S. Pattathil, J. Zhang, and K. Zhang. Revisiting the linear-programming framework for offline RL with general function approximation. *International Conference on Machine Learning*, 2023.

[65] J. Park and E. K. Ryu. Exact optimal accelerated complexity for fixed-point iterations. *International Conference on Machine Learning*, 2022.

[66] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley and Sons, 2nd edition, 2014.

[67] P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, and S. Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Neural Information Processing Systems*, 2021.

[68] A. Rosenberg and Y. Mansour. Oracle-efficient regret minimization in factored MDPs with unknown structure. *Neural Information Processing Systems*, 2021.

[69] S. Ross and J. A. Bagnell. Agnostic system identification for model-based reinforcement learning. *International Conference on Machine Learning*, 2012.

[70] S. Sabach and S. Shtern. A first order method for solving convex bilevel optimization problems. *SIAM Journal on Optimization*, 27(2):640–660, 2017.

[71] B. Scherrer. Approximate policy iteration schemes: A comparison. *International Conference on Machine Learning*, 2014.

[72] P. J. Schweitzer and A. Federgruen. The asymptotic behavior of undiscounted value iteration in Markov decision problems. *Mathematics of Operations Research*, 2(4):360–381, 1977.

[73] P. J. Schweitzer and A. Federgruen. Geometric convergence of value-iteration in multichain Markov decision problems. *Advances in Applied Probability*, 11(1):188–217, 1979.

[74] E. Seneta. *Non-Negative Matrices and Markov Chains*. Springer Science & Business Media, 3th edition, 2006.

[75] A. Sidford, M. Wang, X. Wu, and Y. Ye. Variance reduced value iteration and faster algorithms for solving Markov decision processes. *Naval Research Logistics*, 70(5):423–442, 2023.

[76] N. Y. Siegel, J. T. Springenberg, F. Berkenkamp, A. Abdolmaleki, M. Neunert, T. Lampe, R. Hafner, N. Heess, and M. Riedmiller. Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. *International Conference on Representation Learning*, 2020.

[77] R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

[78] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An introduction*. MIT press, 2nd edition, 2018.

[79] C. Szepesvári. *Algorithms for Reinforcement Learning*. Morgan Claypool Publishers, 2010.

[80] M. Uehara and W. Sun. Pessimistic model-based offline reinforcement learning under partial coverage. *International Conference on Representation Learning*, 2022.

[81] J. Van Der Wal. Stochastic dynamic programming: successive approximations and nearly optimal strategies for Markov decision processes and Markov games. 1981.

[82] D. Vial, A. Parulekar, S. Shakkottai, and R. Srikant. Improved algorithms for misspecified linear markov decision processes. *International Conference on Artificial Intelligence and Statistics*, 2022.

[83] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge university press, 2019.

[84] M. J. Wainwright. Variance-reduced Q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*, 2019.

[85] Y. Wan, A. Naik, and R. S. Sutton. Learning and planning in average-reward Markov decision processes. *International Conference on Machine Learning*, 2021.

[86] M. Wang. Primal-dual π learning: Sample complexity and sublinear run time for ergodic Markov decision problems. *arXiv:1710.06100*, 2017.

[87] R. Wang, D. P. Foster, and S. M. Kakade. What are the statistical limits of offline RL with linear function approximation? *International Conference on Representation Learning*, 2020.

[88] C.-Y. Wei, M. J. Jahromi, H. Luo, and R. Jain. Learning infinite-horizon average-reward mdps with linear function approximation. *International Conference on Artificial Intelligence and Statistics*, 2021.

[89] C.-Y. Wei, M. J. Jahromi, H. Luo, H. Sharma, and R. Jain. Model-free reinforcement learning in infinite-horizon average-reward Markov decision processes. *International Conference on Machine Learning*, 2020.

[90] D. J. White. Dynamic programming, Markov chains, and the method of successive approximations. *J. Math. Anal. Appl*, 6(3):373–376, 1963.

[91] T. Xie, C.-A. Cheng, N. Jiang, P. Mineiro, and A. Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Neural Information Processing Systems*, 2021.

[92] T. Xie and N. Jiang. Batch value-function approximation with only realizability. *International Conference on Machine Learning*, 2021.

[93] T. Yoon and E. K. Ryu. Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm. *International Conference on Machine Learning*, 2021.

[94] W. Zhan, B. Huang, A. Huang, N. Jiang, and J. Lee. Offline reinforcement learning with realizability and single-policy concentrability. *Conference on Learning Theory*, 2022.

[95] Z. Zhang and X. Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. *Neural Information Processing Systems*, 2019.

[96] Z. Zhang and Q. Xie. Sharper model-free reinforcement learning for average-reward Markov decision processes. *Conference on Learning Theory*, 2023.

[97] M. Zurek and Y. Chen. Span-based optimal sample complexity for weakly communicating and general average reward MDPs. *Neural Information Processing Systems*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes] .

   Justification: Our abstract and introduction clearly state the claims made, including the contributions. contributions made in the paper

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes] .

   Justification: Yes, we present limitation of our work as table and discuss in conclusion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes] .

Justification: Yes, we clearly state full set of assumptions and proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA] .

   Justification: Our work does not include numerical experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA] .

Justification: Our paper does not include experiments requiring code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA] .

   Justification: Our paper does not include numerical experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA] .

   Justification: Our paper does not include numerical experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA] .

   Justification: Ou paper does not include numerical experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes] .

   Justification: Our paper conforms, in every respect, with the NeurIPS Code of Ethic.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA] .

    Justification: Since our work is a theory paper, there is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: Our paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA] .

    Justification: Our paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA] .

    Justification: Our paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA] .

Justification: Our paper does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Prior works

Average-Reward MDP The setup of average reward MDPs was introduced in the dynamic programming literature by [36], and [9] established a theoretical framework for their analysis. In reinforcement learning (RL), average-reward MDP was mainly considered in the sample-based setup where the transition matrix and reward are unknown [57, 23]. For this setup, various methods were proposed: model-based methods [41, 97], Q-learning methods [89, 85], and policy gradient methods [5, 62, 46]. Sample complexity to obtain ϵ-optimal under generative model [86, 96, 48, 54, 40] and for regret minimization [15, 38, 95, 11] also have been actively studied.

Value Iteration Value iteration (VI) was first introduced in the dynamic programming literature [6] and serve as a fundamental algorithm to compute the value functions. The sample-based variants, such as TD-Learning [77], Fitted Value Iteration [25, 61], and Deep Q-Network [58] are the workhorses of modern reinforcement learning algorithms [8, 78, 79]. VI is also routinely applied in diverse settings, including factored MDPs [68], robust MDPs [47], MDPs with reward machines [12], MDPs with options [29], and generative model [84, 75, 48].

The convergence of VI in average-reward MDPs also has been extensively studied. For unichain MDPs, delta coefficient, ergodicity coefficient, and the J-stage span contraction demonstrate the linear rate of VI [74, 37, 27, 81]. When MDP is multichain, it is known that policy error of VI might not converge to zero [22, Example 4]. Even with the aperiodicity assumption, VI guarantees only asymptotic convergence. [66, Theorem 9.4.5]. [72, 73] established necessary and sufficient conditions of convergence of VI and asymptotic linear convergence on Bellman error.

Offline Reinforcement Learning In offline RL, the agent learns decision-making strategies utilizing precollected data [53]. This framework is often applied when interaction with the environment can be expensive, and the quantities of data that can be gathered online are substantially lower than the precollected dataset [19, 43, 53]. Consequently, various offline RL methods have been actively proposed [25, 76, 45, 1], and Fitted Q-Iteration is one of the representative methods based on sample-based value iteration with function approximation [25, 61].

One issue in offline RL is the distribution mismatch between the behavior policy that collected the data and the learned policy of the agent [44, 87]. For theoretical analysis, *coverage coefficient* is assumed to ensure that offline dataset sufficiently explores whole state and action space. [60, 71, 80]. Under this assumption, sample complexity of offline RL methods actively analyzed [4, 69, 18, 64], and in particular, an L_p bound of approximate value iteration was obtained, which in turn yields convergence results for Fitted Q-Iteration [60, 61]. More recently, several works succeeded relaxing the full coverage assumption to partcal coverage [56, 67, 91, 42].

Another issue in offline RL is the representation capacity of the chosen function space. To handle large state space and action spaces, many RL frameworks including offline RL use function approximation, ranging from linear functions [24] and nonlinear (general) functions such as neural networks [26] and kernel functions [17]. In offline RL, the *inherent Bellman error* measures the approximation error incurred when projecting the output of Bellman operator into chosen function space, and Bellamn completeness assumes the inherent Bellman error is zero [61, 18]. Most sample complexity analyses in offline RL rely on inherent Bellman error or Bellman completeness assumption [56, 67, 91, 42]. Recently, however, several works achieved finite sample complexity under weaker realizabiltiy assumption, which only requires that optimal function value lies within chosen function space [92, 94].

Most of prior works in offline RL focused on discounted-reward setup, and to the best of our knowledge, two prior works established the finite sample complexity in the offline average-reward setup [63, 30]. Both proposed a primal-dual approach, reformulating the Bellman equation as a bilinear saddle-point problem, to obtain an ϵ-optimal policy under partial coverage. However, they imposed restrictive structural assumptions on MDP such as uniform mixing or linearity and considered only IID dataset. (See the Table 1.)

B Preliminaries

The followings are inequalities from prior works used in the proof.

Fact 1 (Bernstein inequality)**.** *Let $X_1, \dots, X_n$ are indepedent random variables. If $X_i \le b$ for all i, then*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X_i] \ge \epsilon\right) \le exp\left[-\frac{n^2\epsilon^2}{2\sum_{i=1}^{n}\mathbb{E}[X_i^2] + nb\epsilon/3}\right]$$

Furthermore, if all the $\mathbb{E}[X_i^2]$ are equal, with $1-\delta$ probability,

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X_i \le \sqrt{2\mathbb{E}[X_1^2]\ln(1/\delta)/n} + \frac{2b\ln(1/\delta)}{3n}.$$

Fact 2 ([4], Lemma 4)**.** *Suppose that $Z_1, \dots, Z_n \in \mathcal{Z}$ is a stationary β-mixing process with mixing coefficients β_m, $Z_t' \in \mathcal{Z}(t \in H)$ are the block-independent ghost samples. $H = \{2ik_N + j : 0 \le i < m_n, 1 \le j \le k_N\}$ and $\mathcal{F}$ is permissible class of $\mathcal{Z} \to [-M, M]$ functions. Then*

$$P\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{N}\sum_{n=1}^{N} f(Z_n) - \mathbb{E}[f(Z_1)]\right| > \epsilon\right) \le 16\mathbb{E}[\mathcal{N}(\epsilon/8, \mathcal{F}, l_{(Z_t')_{t\in H}})]e^{-\frac{m_N\epsilon^2}{128M^2}} + 2m_N\beta_{k_N+1}.$$

C Omitted proofs in Section 3

C.1 Proof of Proposition 1

Define the limiting matrix $\mathcal{P}_*^\pi$ as the Cesàro limit of $\mathcal{P}^\pi$, i.e., $\mathcal{P}_*^\pi = \lim \frac{1}{n}\sum_{i=1}^{n}(\mathcal{P}^\pi)^i$. (The limiting matrix always exists for finite state-action spaces [66, Appendix A.4].) Then, $\mathcal{P}_*^\pi$ is stochastic and, by definition, $g^\pi = \mathcal{P}_*^\pi r$ [66, Proposition 8.1.1].

We first prove following lemma.

Lemma 3. *Let $\lambda_{K+1} = 1$. Under Assumption 1 (Bellman optimality equation), the policy error of Apx-Anc-QI satisfies*

$$\begin{aligned}
&g^{\pi_\star} - g^{\pi_K} = \mathcal{P}_*^{\pi_K}(g^{\pi_\star} - TQ^K + Q^K)\\
&\le \mathcal{P}_*^{\pi_K}\left(\sum_{l=0}^{K}\Pi_{i=l+1}^{K}\lambda_i(\lambda_{l+1}-\lambda_l)\Pi_{i=l+1}^{K}\mathcal{P}^{\pi_i}\left(\sum_{m=0}^{l}\Pi_{i=m+1}^{l}\lambda_i(1-\lambda_m)(\mathcal{P}^{\pi_\star})^{l+1-m} - I\right)(Q^0 - Q^{\pi_\star})\right.\\
&\left.+\sum_{l=1}^{K}\Pi_{i=l}^{K}\lambda_i\left(\sum_{m=l}^{K}(\lambda_{m+1}-\lambda_m)\Pi_{i=m+1}^{K}\mathcal{P}^{\pi_i}(\mathcal{P}^{\pi_\star})^{m+1-l} + \Pi_{i=l+1}^{K}\mathcal{P}^{\pi_i}(\lambda_l\mathcal{P}^{\pi_l} - I)\right)\epsilon_l\right).
\end{aligned}$$

Proof of Lemma 3. By definition of Apx-Anc-QI, we have

$$\begin{aligned}
&TQ^K - Q^K\\
&= (1-\lambda_K)(TQ^K - Q^0) + \lambda_K(TQ^K - TQ^{K-1}) - \lambda_K\epsilon_K\\
&\ge (1-\lambda_K)(TQ^K - Q^0) + \lambda_K\mathcal{P}^{\pi_K}(Q^K - Q^{K-1}) - \lambda_K\epsilon_K\\
&\ge (1-\lambda_K)(TQ^K - Q^0) - \lambda_K\epsilon_K\\
&+ \lambda_K\mathcal{P}^{\pi_K}((\lambda_K - \lambda_{K-1})(TQ^{K-1} - Q^0) + \lambda_{K-1}(TQ^{K-1} - TQ^{K-2}) + \lambda_K\epsilon_K - \lambda_{K-1}\epsilon_{K-1})\\
&\ge \sum_{l=0}^{K}\Pi_{i=l+1}^{K}\lambda_i(\lambda_{l+1}-\lambda_l)\Pi_{i=l+1}^{K}\mathcal{P}^{\pi_i}(TQ^l - Q^0) + \sum_{l=1}^{K}\Pi_{i=l}^{K}\lambda_i\Pi_{i=l+1}^{K}\mathcal{P}^{\pi_i}(\lambda_l\mathcal{P}^{\pi_l} - I)\epsilon_l
\end{aligned}$$

where first inequality comes from greedy policy and last inequality comes from induction.

For any $0 \leq l \leq K$,

$$
\begin{aligned}
&TQ^l - Q^0 \\
&= TQ^l - Q^{\pi_\star} - (Q^0 - Q^{\pi_\star}) \\
&= TQ^l - TQ^{\pi_\star} + g^{\pi_\star} - (Q^0 - Q^{\pi_\star}) \\
&\geq \mathcal{P}^{\pi_\star}(Q^l - Q^{\pi_\star}) + g^{\pi_\star} - (Q^0 - Q^{\pi_\star}) \\
&= \mathcal{P}^{\pi_\star}(\lambda_l(TQ^{l-1} - Q^{\pi_\star}) + (1-\lambda_l)(Q^0 - Q^{\pi_\star}) + \lambda_l \epsilon_l) + g^{\pi_\star} - (Q^0 - Q^{\pi_\star}) \\
&\geq \left(\sum_{m=0}^{l} \Pi_{i=m+1}^{l} \lambda_i (\mathcal{P}^{\pi_\star})^{l+1-m}(1-\lambda_m) - I \right)(Q^0 - Q^{\pi_\star}) + \sum_{m=0}^{l} \Pi_{i=m+1}^{l} \lambda_i g^{\pi_\star} \\
&+ \sum_{m=1}^{l} \Pi_{i=m}^{l} \lambda_i (\mathcal{P}^{\pi_\star})^{l+1-m} \epsilon_m,
\end{aligned}
$$

where second equality comes from Bellman optimality equation. By combining previous two inequalities, we get

$$
\begin{aligned}
&TQ^K - Q^K \\
&\geq \sum_{l=0}^{K} \Pi_{i=l+1}^{K} \lambda_i (\lambda_{l+1} - \lambda_l) \Pi_{i=l+1}^{K} \mathcal{P}^{\pi_i} \sum_{m=0}^{l} \Pi_{i=m+1}^{l} \lambda_i g^{\pi_\star} \\
&+ \sum_{l=0}^{K} \Pi_{i=l+1}^{K} \lambda_i (\lambda_{l+1} - \lambda_l) \Pi_{i=l+1}^{k} \mathcal{P}^{\pi_i} \left(\sum_{m=0}^{l} \Pi_{i=m+1}^{l} \lambda_i (\mathcal{P}^{\pi_\star})^{l+1-m}(1-\lambda_m) - I \right)(Q^0 - Q^{\pi_\star}) \\
&+ \sum_{l=1}^{K} \Pi_{i=l}^{K} \lambda_i \Pi_{i=l+1}^{K} \mathcal{P}^{\pi_i} (\lambda_l \mathcal{P}^{\pi_l} - I) \epsilon_l \\
&+ \sum_{l=1}^{K} \sum_{m=1}^{l} \Pi_{i=l+1}^{K} \lambda_i (\lambda_{l+1} - \lambda_l) \Pi_{i=l+1}^{K} \mathcal{P}^{\pi_i} \Pi_{i=m}^{l} \lambda_i (\mathcal{P}^{\pi_\star})^{l+1-m} \epsilon_m \\
&= g^{\pi_\star} + \sum_{l=1}^{K} \Bigg(\sum_{m=l}^{k} \Pi_{i=m+1}^{K} \lambda_i (\lambda_{m+1} - \lambda_m) \Pi_{i=l}^{m} \lambda_i \Pi_{i=m+1}^{K} \mathcal{P}^{\pi_i} (\mathcal{P}^{\pi_\star})^{m+1-l} \\
&+ \Pi_{i=l}^{K} \lambda_i \Pi_{i=l+1}^{K} \mathcal{P}^{\pi_i} (\lambda_l \mathcal{P}^{\pi_l} - I) \Bigg) \epsilon_l \\
&+ \sum_{l=0}^{K} \Pi_{i=l+1}^{K} \lambda_i (\lambda_{l+1} - \lambda_l) \Pi_{i=l+1}^{K} \mathcal{P}^{\pi_i} \left(\sum_{m=0}^{l} \Pi_{i=m+1}^{l} \lambda_i (1-\lambda_m) (\mathcal{P}^{\pi_\star})^{l+1-m} - I \right)(Q^0 - Q^{\pi_\star}).
\end{aligned}
$$

This implies

$$
\begin{aligned}
&TQ^K - Q^K - g^{\pi_\star} \\
&\geq \sum_{l=0}^{K} \Pi_{i=l+1}^{K} \lambda_i (\lambda_{l+1} - \lambda_l) \Pi_{i=l+1}^{K} \mathcal{P}^{\pi_i} \left(\sum_{m=0}^{l} \Pi_{i=m+1}^{l} \lambda_i (1-\lambda_m) (\mathcal{P}^{\pi_\star})^{l+1-m} - I \right)(Q^0 - Q^{\pi_\star}) \\
&+ \sum_{l=1}^{K} \Pi_{i=l}^{K} \lambda_i \left(\sum_{m=l}^{K} (\lambda_{m+1} - \lambda_m) \Pi_{i=m+1}^{K} \mathcal{P}^{\pi_i} (\mathcal{P}^{\pi_\star})^{m+1-l} + \Pi_{i=l+1}^{K} \mathcal{P}^{\pi_i} (\lambda_l \mathcal{P}^{\pi_l} - I) \right) \epsilon_l.
\end{aligned}
$$

Finally, following the proof of [66, Theorem 8.5.5], we have

$$
\begin{aligned}
g^{\pi_\star} - g^{\pi_K} = \mathcal{P}_*^{\pi_K}(g^{\pi_\star} - r) &= \mathcal{P}_*^{\pi_K}(g^{\pi_\star} - r - \mathcal{P}^{\pi_K} Q^K + Q^K) \\
&= \mathcal{P}_*^{\pi_K}(g^{\pi_\star} - TQ^K + Q^K),
\end{aligned}
$$

where first equality comes from Bellman optimality equation and second equality comes from property of limiting matrix. This implies that

$$
\begin{aligned}
&g^{\pi_\star} - g^{\pi_K} = \mathcal{P}_*^{\pi_K}(g^{\pi_\star} - TQ^K + Q^K) \\
&\leq \mathcal{P}_*^{\pi_K}\left(\sum_{l=0}^{K}\Pi_{i=l+1}^{K}\lambda_i(\lambda_{l+1}-\lambda_l)\Pi_{i=l+1}^{K}\mathcal{P}^{\pi_i}\left(\sum_{m=0}^{l}\Pi_{i=m+1}^{l}\lambda_i(1-\lambda_m)(\mathcal{P}^{\pi_\star})^{l+1-m} - I\right)(Q^0 - Q^{\pi_\star})\right. \\
&\left. + \sum_{l=1}^{K}\Pi_{i=l}^{K}\lambda_i\left(\sum_{m=l}^{K}(\lambda_{m+1}-\lambda_m)\Pi_{i=m+1}^{K}\mathcal{P}^{\pi_i}(\mathcal{P}^{\pi_\star})^{m+1-l} + \Pi_{i=l+1}^{K}\mathcal{P}^{\pi_i}(\lambda_l\mathcal{P}^{\pi_l} - I)\right)\epsilon_l\right).
\end{aligned}
$$

□

The following are lemmas about coverage coefficient $C_{\mu,\rho}$.

Lemma 4. *If $\mathcal{P}_1$ and $\mathcal{P}_2$ are stochastic matrix satisfying $\rho^\top\mathcal{P}_i \leq C_{\mu,\rho}\mu$ for $i = 1, 2$ and given distribution μ and ρ on $\mathcal{S}\times\mathcal{A}$, then $\rho^\top(a\mathcal{P}_1 + (1-a)\mathcal{P}_2) \leq C_{\mu,\rho}\mu$ for $0 \leq a \leq 1$.*

Lemma 5. *Under Assumption 6 (uniform future state distribution),*

$$
\sup_{\pi_1,\pi_2,\ldots\pi_k}\left\|\frac{\rho^\top\mathcal{P}_*^{\pi_\star}\mathcal{P}^{\pi_1}\mathcal{P}^{\pi_2}\cdots\mathcal{P}^{\pi_k}(\cdot)}{\mu(\cdot)}\right\|_\infty \leq C_{\mu,\rho}
$$

where $\pi_\star\pi_1, \pi_2, \ldots\pi_k$ represents an arbitrary sequence of policies with optimal policy.

Proof. Under Assumption 6, for any non negative integer n, we have $\rho^\top(\mathcal{P}^{\pi_\star})^n\mathcal{P}^{\pi_1}\mathcal{P}^{\pi_2}\cdots\mathcal{P}^{\pi_k}(\cdot) \leq C_{\mu,\rho}\mu$. This implies $\rho^\top\mathcal{P}_*^{\pi_\star}\mathcal{P}^{\pi_1}\mathcal{P}^{\pi_2}\cdots\mathcal{P}^{\pi_k}(\cdot) \leq C_{\mu,\rho}\mu$ by definition of limiting matrix. □

Lemma 6. *If $\mathcal{P}$ is stochastic matrix satisfying $\rho^\top\mathcal{P} \leq C_{\mu,\rho}\mu^\top$ for given distribution μ and ρ on $\mathcal{S}\times\mathcal{A}$, then $\|\mathcal{P}Q\|_{p,\rho} \leq C_\mu^{1/p}\|Q\|_{p,\mu}$.*

Proof. Since $|\mathcal{P}Q(s,a)|^p = |\mathbb{E}_{(s',a')\sim\mathcal{P}(\cdot\,|\,s,a)}[Q(s',a')]|^p \leq \mathbb{E}_{(s',a')\sim\mathcal{P}(\cdot\,|\,s,a)}[|Q(s',a')|^p]) = \mathcal{P}|Q|^p(s,a)$ by Jensen's inequality, $\rho^\top|\mathcal{P}Q|^p \leq \rho^\top\mathcal{P}|Q|^p \leq C_{\mu,\rho}\mu^\top|Q|^p$. □

Now, we are ready to prove Proposition 1.

Proof of Proposition 1. By Lemma 3,

$$
\begin{aligned}
&g^{\pi_\star} - g^{\pi_K} \\
&\leq \mathcal{P}_*^{\pi_K}\left(\sum_{l=0}^{K}\Pi_{i=l+1}^{K}\lambda_i(\lambda_{l+1}-\lambda_l)\Pi_{i=l+1}^{K}\mathcal{P}^{\pi_i}\left(\sum_{m=0}^{l}\Pi_{i=m+1}^{l}\lambda_i(1-\lambda_m)(\mathcal{P}^{\pi_\star})^{l+1-m} - I\right)(Q^0 - Q^{\pi_\star})\right. \\
&\left. + \sum_{l=1}^{K}\Pi_{i=l}^{K}\lambda_i\left(\sum_{m=l}^{K}(\lambda_{m+1}-\lambda_m)\Pi_{i=m+1}^{K}\mathcal{P}^{\pi_i}(\mathcal{P}^{\pi_\star})^{m+1-l} + \Pi_{i=l+1}^{K}\mathcal{P}^{\pi_i}(\lambda_l\mathcal{P}^{\pi_l} - I)\right)\epsilon_l\right) \\
&\leq \mathcal{P}_*^{\pi_K}\left(\sum_{l=0}^{K}\Pi_{i=l+1}^{K}\lambda_i(\lambda_{l+1}-\lambda_l)\Pi_{i=l+1}^{K}\mathcal{P}^{\pi_i}\left(\sum_{m=0}^{l}\Pi_{i=m+1}^{l}\lambda_i(1-\lambda_m)(\mathcal{P}^{\pi_\star})^{l+1-m} + I\right)|Q^0 - Q^{\pi_\star}|\right. \\
&\left. + \sum_{l=1}^{K}\Pi_{i=l}^{K}\lambda_i\left(\sum_{m=l}^{K}(\lambda_{m+1}-\lambda_m)\Pi_{i=m+1}^{K}\mathcal{P}^{\pi_i}(\mathcal{P}^{\pi_\star})^{m+1-l} + \Pi_{i=l+1}^{K}\mathcal{P}^{\pi_i}(\lambda_l\mathcal{P}^{\pi_l} + I)\right)|\epsilon_l|\right).
\end{aligned}
$$

Let $\mathcal{P}_l^Q = \mathcal{P}_*^{\pi_K}\Pi_{i=l+1}^{K}\mathcal{P}^{\pi_i}\left(\sum_{m=0}^{l}\Pi_{i=m+1}^{l}\lambda_i(1-\lambda_m)(\mathcal{P}^{\pi_\star})^{l+1-m} + I\right)/2$ and $\mathcal{P}_l^\epsilon = \mathcal{P}_*^{\pi_K}\sum_{m=l}^{K}(\lambda_{m+1}-\lambda_m)\Pi_{i=m+1}^{K}\mathcal{P}^{\pi_i}(\mathcal{P}^{\pi_\star})^{m+1-l} + \Pi_{i=l+1}^{K}\mathcal{P}^{\pi_i}(\lambda_l\mathcal{P}^{\pi_l} + I)/2$. Then $\mathcal{P}_l^Q$ and $\mathcal{P}_l^\epsilon$ satisfying $\rho^\top\mathcal{P}_l^Q \leq C_{\mu,\rho}\mu$ and $\rho^\top\mathcal{P}_l^\epsilon \leq C_{\mu,\rho}\mu$ for all $0 \leq l \leq K$ by Lemma 4 and 5. Thus, we

have

$$\|g^{\pi_\star} - g^{\pi_K}\|_{p,\rho} \leq 2\sum_{l=0}^{K} \Pi_{i=l+1}^{K} \lambda_i (\lambda_{l+1} - \lambda_l) \|\mathcal{P}_l |Q^0 - Q^{\pi_\star}|\|_{p,\rho} + 2\sum_{l=1}^{K} \Pi_{i=l}^{K} \lambda_i \|\mathcal{P}_l^\epsilon |\epsilon_l|\|_{p,\rho}$$

$$\leq 2C_\mu^{1/p} \sum_{l=0}^{K} \Pi_{i=l+1}^{K} \lambda_i (\lambda_{l+1} - \lambda_l) \|Q^0 - Q^{\pi_\star}\|_{p,\mu} + 2C_\mu^{1/p} \sum_{l=1}^{K} \Pi_{i=l}^{K} \lambda_i \|\epsilon_l\|_{p,\mu},$$

where last inequality comes from Lemma 6. By plugging $\lambda_k = \frac{k}{k+2}$, we conclude. Note that since $C_\mu \leq C_{\mu,\rho}$ for any distribution ρ, then choosing ρ to be a Dirac distribution at each state proves the case of Assumption 5 which implies first inequality of Proposition 1. □

D Omitted proofs in Section 4

D.1 Proof of Lemma 1

Proof of Lemma 1. Let $\mathcal{F} \subset \{f : \mathcal{S} \times \mathcal{A} \to [-f_{max}, f_{max}] \,|\, f \in B(S \times A)\}$ and $\mathcal{G} \subset \{f : \mathcal{S} \times \mathcal{A} \to [-g_{max}, g_{max}] \,|\, f \in B(S \times A)\}$. Let $f_1, \ldots, f_N$ cover the $\mathcal{F}$ and $g_1, \ldots, g_{N'}$ cover the $\mathcal{G}$ where $N = \mathcal{N}(\epsilon/M; \mathcal{F}, \|\cdot\|_\infty)$, $N' = \mathcal{N}(\epsilon/M; \mathcal{G}, \|\cdot\|_\infty)$, $M = 108(R + 2f_{max})$. $\mathcal{F} \times \mathcal{G} = \cup S_{i,j}$ where $S_{i,j} = \{(f, g) : \|f - f_i\|_\infty \leq \epsilon, \|g - g_j\|_\infty \leq \epsilon\}$. Without loss of generality, suppose $g_{max} \leq f_{max}$.

First note that $\mathbb{E}_{s_i' \sim P(\cdot \,|\, s_i, a_i)}[r(s_i, a_i) + \max_a g(s_i', a)] = Tg(s_i, a_i)$, $|r_i + \max_a g(s, a)| \leq R + f_{max}$, $|Tg(s, a)| \leq R + f_{max}$.

For arbitrary $f \in \mathcal{F}, g \in \mathcal{G}$, define $X_i^{f,g} = (f(s_i, a_i) - r(s_i, a_i) - \max_a g(s_i', a))^2 - (Tg(s_i, a_i) - r(s_i, a_i) - \max_a g(s_i', a))^2$. Then, $\mathbb{E}_{s_i, a_i \sim \mu, s_i' \sim P(\cdot \,|\, s_i, a_i)}[X_i^{f,g}] = \|Tg - f\|_{\mu,2}^2$ and $\mathbb{E}[(X_i^{f,g})^2] \leq 9(R + 2f_{max})^2 \|Tg - f\|_{\mu,2}^2$ since $X_i^{f,g} = (f(s_i, a_i) - Tg(s_i, a_i))(f(s_i, a_i) + Tg(s_i, a_i) - 2r(s_i, a_i) - 2\max_a g(s_i', a))$, and $|X_i^{f,g}| \leq 3(R + 2f_{max})^2$.

By Bernstein inequality and union bound, with $1 - \delta$ probability, for all $\{f_i, g_j\}_{i=1,\ldots,N, j=1,\ldots,N'}$,

$$\|Tg_j - f_i\|_{\mu,2}^2 - \sum_{i=1}^{n} X_i^{f_i, g_j}/n \leq \sqrt{\frac{18(R + 2f_{max})^2 \|Tg_j - f_i\|_{\mu,2}^2 \ln(\mathcal{N}_{\mathcal{F},\mathcal{G}}/\delta)}{n}} + \frac{2(R + 2f_{max})^2 \ln(\mathcal{N}_{\mathcal{F},\mathcal{G}}/\delta)}{n}$$

where $\mathcal{N}_{\mathcal{F},\mathcal{G}} = \mathcal{N}(\epsilon/M; \mathcal{G}, \|\cdot\|_\infty)\mathcal{N}(\epsilon/M; \mathcal{F}, \|\cdot\|_\infty)$. Through $2\sqrt{ab} \leq a + b$, we have

$$\forall f_i \in \mathcal{F}, \forall g_i \in \mathcal{G}, \quad \|Tg_j - f_i\|_{\mu,2}^2 - 2\sum_{i=1}^{n} X_i^{f_i, g_j}/n \leq \frac{22(R + 2f_{max})^2 \ln(\mathcal{N}_{\mathcal{F},\mathcal{G}}/\delta)}{n}$$

Now, for covering number argument, we use following Lemma.

Lemma 7. *For $f \in \mathcal{F}$, $g \in \mathcal{G}$, $c > 0$, $\|Tg - f\|_{\mu,2}^2 - c\sum_{i=1}^{n} X_i^{f,g}/n$ is $(2 + 8c)(2f_{max} + R)$-Lipchitz.*

Proof. Since $\|Tg_1 - f_1\|_{\mu,2}^2 - \|Tg_2 - f_2\|_{\mu,2}^2 \leq \mathbb{E}|(Tg_1 - Tg_2 + f_2 - f_1)(Tg_1 + Tg_2 - f_2 - f_1)| \leq (\|g_1 - g_2\|_\infty + \|f_1 - f_2\|_\infty)2(R + 2f_{max})$, $\|Tg - f\|_{\mu,2}^2$ is $2(R + 2f_{max})$- Lipchitz. Also, since $|\sum_{i=1}^{n} X_i^{f_1,g_1}/n - \sum_{i=1}^{n} X_i^{f_2,g_2}/n| = \frac{1}{n}\sum_{i=1}^{n} |(\max g_2 - \max g_1 + f_1 - f_2)(f_2 + f_1 - \max g_1 - \max g_2 - 2r) - (Tg_1 - Tg_2 + \max g_2 - \max g_1)(Tg_1 + Tg_2 + \max g_2 + \max g_1 - 2r)| \leq (\|g_1 - g_2\|_\infty + \|f_1 - f_2\|_\infty)2(R + 2f_{max}) + 8\|g_1 - g_2\|_\infty (f_{max} + R) \leq (\|g_1 - g_2\|_\infty + \|f_1 - f_2\|_\infty)8(2f_{max} + R)$, $\sum_{i=1}^{n} X_i^{f_1,g_1}/n$ $8(2f_{max} + R)$-Lipchitz. By adding two Lipchitz functions, we obtain desired result. □

By Lipchitzness of $\|Tg_j - f_i\|_{\mu,2}^2 - 2\sum_{i=1}^{n} X_i^{f_i,g_j}/n$ and definition of covering number, if $f, g \in S_{i,j}$

$$\|Tg - f\|_{\mu,2}^2 - 2\sum_{i=1}^{n} X_i^{f,g}/n - (\|Tg_j - f_i\|_{\mu,2}^2 - 2\sum_{i=1}^{n} X_i^{f_i,g_j}/n) \leq \epsilon.$$

This implies that with $1-\delta$ probability,

$$\forall f \in \mathcal{F}, \forall g \in \mathcal{G} \quad \|Tg - f\|_{\mu,2}^2 \leq \epsilon + \frac{22(R+2f_{max})^2 \ln(\mathcal{N}_{\mathcal{F},\mathcal{G}}/\delta)}{n} + 2\sum_{i=1}^{n} X_i^{f,g}/n. \quad (1)$$

By other side of Bernstein's inequality and covering number, for all $\{f_i, g_j\}_{i=1,\dots,N, j=1,\dots,N'}$, we have

$$\sum_{i=1}^{n} X_i^{f_i,g_j}/n - \|Tg_j - f_i\|_{\mu,2}^2 \leq \sqrt{\frac{18(R+2f_{max})^2 \|Tg_j - f_i\|_{\mu,2}^2 \ln(\mathcal{N}_{\mathcal{F},\mathcal{G}}/\delta)}{n}} + \frac{2(R+2f_{max})^2 \ln(\mathcal{N}_{\mathcal{F},\mathcal{G}}/\delta)}{n}.$$

If $\sum_{i=1}^{n} X_i^{f_i,g_j}/n \geq \frac{4(R+2f_{max})^2 \ln(\mathcal{N}_{\mathcal{F},\mathcal{G}}/\delta)}{n}$, with $1-\delta$ probability, for all $\{f_i, g_j\}_{i=1,\dots,N, j=1,\dots,N'}$, we have

$$\sum_{i=1}^{n} X_i^{f_i,g_j}/n - \|Tg_j - f_i\|_{\mu,2}^2 \leq \sqrt{4.5\sum_{i=1}^{n} X_i^{f_i,g_j}/n \|Tg_j - f_i\|_{\mu,2}^2} + \frac{2(R+2f_{max})^2 \ln(\mathcal{N}_{\mathcal{F},\mathcal{G}}/\delta)}{n}$$

and by $2\sqrt{ab} \leq a + b$, this implies

$$\sum_{i=1}^{n} X_i^{f_i,g_j}/n - 6.5\|Tg_j - f_i\|_{\mu,2}^2 \leq \frac{4(R+2f_{max})^2 \ln(\mathcal{N}_{\mathcal{F},\mathcal{G}}/\delta)}{n}.$$

Even if $\sum_{i=1}^{n} X_i^{f_i,g_j}/n \leq \frac{4(R+2f_{max})^2 \ln(\mathcal{N}_{\mathcal{F},\mathcal{G}}/\delta)}{n}$, previous inequality still holds. Since $\sum_{i=1}^{n} X_i^{f_i,g_j}/n - 6.5\|Tg_j - f_i\|_{\mu,2}^2$ is $54(R+2f_{max})$-Lipshitz, with similar argument, we have

$$\forall f \in \mathcal{F}, g \in \mathcal{G}, \quad \sum_{i=1}^{n} X_i^{f,g}/n - 6.5\|Tg - f\|_{\mu,2}^2 \leq \epsilon + \frac{4(R+2f_{max})^2 \ln(\mathcal{N}_{\mathcal{F},\mathcal{G}}/\delta)}{n}. \quad (2)$$

Let $\tilde{T}g = \operatorname{argmin}_{f \in \mathcal{F}} \|f - Tg\|_{2,\mu}$ and $f = \tilde{T}g$ in inequality (2). Then, by definition of Inherent Bellman error,

$$\forall g \in \mathcal{G}, \quad \sum_{i=1}^{n} X_i^{\tilde{T}g,g}/n \leq \epsilon + 6.5\epsilon_B + \frac{4(R+2f_{max})^2 \ln(\mathcal{N}_{\mathcal{F},\mathcal{G}}/\delta)}{n}.$$

Also, let $f = \hat{T}g$ in inequality inequality (1). Then, by definition of $\hat{T}g$, we have $\sum_{i=1}^{n} X_i^{\hat{T}g,g} \leq \sum_{i=1}^{n} X_i^{\tilde{T}g,g}$. Combining with previous inequality, with $1-2\delta$ probability,

$$\forall g \in \mathcal{G}, \quad \|Tg - \hat{T}g\|_{\mu,2}^2 \leq 3\epsilon + 13\epsilon_B + \frac{30(R+2f_{max})^2 \ln(\mathcal{N}_{\mathcal{F},\mathcal{G}}/\delta)}{n}.$$

Finally, let $\mathcal{G} = \mathcal{F}_k, \mathcal{F} = \mathcal{F}_{k+1}$, and $g = f_k$, and by manipulating δ, we get desired result. □

D.2 Proof of Theorem 1

Proof of Theorem 1. By combining Lemma 1 and Proposition 1, we directly obtain following results. Under assumptions stated in Theorem 1, we have

$$\|g^{\pi_\star} - g^{\pi_K}\|_\infty \leq C_\mu^{1/2} \frac{8\|Q^{\pi_\star}\|_{2,\mu}}{K+2} + C_\mu^{1/2} \frac{2K}{3} \left(\sqrt{3\epsilon'} + \sqrt{\frac{60(K+1)^2 R^2 \ln(2KN_{\epsilon'}^2/\delta)}{n}} + \max_{k=0,\dots,K-1} \sqrt{13\epsilon_B(\mathcal{F}_k, \mathcal{F}_{k+1})} \right),$$

$$\|g^{\pi_\star} - g^{\pi_K}\|_{2,\rho} \le C_{\mu,\rho}^{1/2} \frac{8\|Q^{\pi_\star}\|_{2,\mu}}{K+2} + C_{\mu,\rho}^{1/2} \frac{2K}{3} \left(\sqrt{3\epsilon'} + \sqrt{\frac{60(K+1)^2 R^2 \ln(2KN_{\epsilon'}^2/\delta)}{n}} + \max_{k=0,\dots,K-1} \sqrt{13\epsilon_B(\mathcal{F}_k, \mathcal{F}_{k+1})} \right),$$

where

$$N'_\epsilon = \max_{k=1,\dots,K} N_{k,\epsilon'}, \qquad N_{k,\epsilon} = \mathcal{N}\big(\tfrac{\epsilon'}{108(2k+1)R}; \mathcal{F}_k, \|\cdot\|_\infty\big), \quad \text{for } k = 1, \dots, K.$$

Given $\epsilon > 0$, for the first inequality, let $K = \lceil 18C_\mu^{1/2}\|Q^{\pi_\star}\|_{2,\mu}/\epsilon \rceil, \epsilon' = \frac{4\epsilon^2}{27K^2 C_\mu}, n = \frac{36K^2 C_\mu}{\epsilon^2} 60R^2(K+1)^2 \ln(2K\mathcal{N}_{\epsilon'}^2/\delta)$. Then, by direct calculation, we derive that

$$\|g^{\pi_\star} - g^{\pi_K}\|_\infty \le \epsilon + 3KC_\mu^{1/2} \max_{k=0,\dots,K-1} \sqrt{\epsilon_B(\mathcal{F}_k, \mathcal{F}_{k+1})}$$

with sample complexity

$$n = \mathcal{O}\left(\frac{\|Q^{\pi_\star}\|_{2,\mu}^4 C_\mu^3 R^2}{\epsilon^6} \ln(\mathcal{N}_\epsilon^2 C_\mu^{1/2}/(\delta\epsilon)) \right)$$

where

$$N_\epsilon = \max_{k=1,\dots,K} N_{k,\epsilon}, \qquad N_{k,\epsilon} = \mathcal{N}\big(\tfrac{\epsilon^4}{10^6 k C_\mu^2 \|Q^{\pi_\star}\|_{2,\mu}^2 R}; \mathcal{F}_k, \|\cdot\|_\infty\big), \quad \text{for } k = 1, \dots, K.$$

Similarly, given $\epsilon > 0$, for second inequality, let $K = \lceil 18C_{\mu,\rho}^{1/2}\|Q^{\pi_\star}\|_{2,\mu}/\epsilon \rceil, \epsilon' = \frac{4\epsilon^2}{27K^2 C_{\mu,\rho}}, n = \frac{36K^2 C_{\mu,\rho}}{\epsilon^2} 60R^2(K+1)^2 \ln(2K^2\mathcal{N}_{\epsilon'}/\delta)$, and

$$\|g^{\pi_\star} - g^{\pi_K}\|_{2,\rho} \le \epsilon + 3KC_{\mu,\rho}^{1/2} \max_{k=0,\dots,K-1} \sqrt{\epsilon_B(\mathcal{F}_k, \mathcal{F}_{k+1})}$$

with sample complexity

$$n = \mathcal{O}\left(\frac{\|Q^{\pi_\star}\|_{2,\mu}^4 C_{\mu,\rho}^3 R^2}{\epsilon^6} \ln(\mathcal{N}_\epsilon^2 C_\mu^{1/2}/(\delta\epsilon)) \right)$$

where

$$N_\epsilon = \max_{k=1,\dots,K} N_{k,\epsilon}, \qquad N_{k,\epsilon} = \mathcal{N}(\tfrac{\epsilon^4}{10^6 k C_{\mu,\rho}^2 \|Q^{\pi_\star}\|_{2,\mu}^2 R}; \mathcal{F}_k, \|\cdot\|_\infty), \quad \text{for } k = 1, \dots, K.$$

□

D.3 Proof of Lemma 2

We first introduce empirical covering number.

Definition 5 (empirical covering number)**.** *For a given function class $\mathcal{F}$ of real valued functions and set $x^{1:n} = (x_1, \dots, x_n)$, denote the covering number of $\mathcal{F}$ equipped with the empirical l_1 pseudo metric $l_{x^{1:n}}(f,g) = \frac{1}{n}\sum_{i=1}^n |f(x_i) - g(x_i)|$ by $\mathcal{N}(\epsilon, \mathcal{F}, x^{1:n})$.*

Although the empirical convering number depends on number of samples, but it can be bounded by pseudo dimension which depends on only function space and ϵ as following fact shows.

Fact 3 ([35], Corollary 3)**.** *For any $x^{1:n} = (x_1, \dots, x_n)$, any function class $\mathcal{F}$ of real-valued functions taking values in $[0, M]$ with pseudo-dimension $V_\mathcal{F} < \infty$, and any $\epsilon > 0$,*

$$\mathcal{N}(\epsilon, \mathcal{F}, l_{x^{1:N}}) \le e(V_\mathcal{F} + 1)\left(\frac{2eM}{\epsilon}\right)^{V_\mathcal{F}}.$$

Define $L(g,f) = \mathbb{E}_{s_i,a_i\sim\mu}[Var_{s'_i\sim P(\,|\,s_i,a_i)}(r(s_i,a_i) + \max f(s'_i,a))] + \|g - Tf\|_{2,\mu}^2$ where Var denotes variance with respect to s'_i, and $\hat{L}(g,f) = \frac{1}{n}\sum_{i=1}^n (g(s_i,a_i) - r(s_i,a_i) - \max_a f(s'_i,a))^2$. Then, $\mathbb{E}[\hat{L}(g,f)] = L(g,f)$ and following lemma holds.

Lemma 8. $\|\hat{T}f - Tf\|_{2,\mu}^2 - \inf_{g\in\mathcal{G}} \|g - Tf\|_{2,\mu}^2 \le 2\sup_{g\in\mathcal{G}} |L(g,f) - \hat{L}(g,f)|$.

Proof of Lemma 8. $\|\hat{T}f - Tf\|_{2,\mu}^2 - \inf_{g\in\mathcal{G}} \|g - Tf\|_{2,\mu}^2 = L(\hat{T}f, f) - \inf_{g\in\mathcal{F}} L(g,f) = L(\hat{T}f,f) - \hat{L}(\hat{T}f,f) + \hat{L}(\hat{T}f,f) - \inf_{g\in\mathcal{G}} L(g,f) \le 2\sup_{f\in\mathcal{F}} |L(g,f) - \hat{L}(g,f)|$ by definition of $\hat{T}f$. □

For $\{\hat{T}f_k, f_k\}_{k=0}^{K-1}$ of Anc-F-QI, previous lemma implies that

$$\begin{aligned}\|\hat{T}f_k - Tf_k\|_{2,\mu}^2 - \inf_{g\in\mathcal{G}} \|g - Tf_k\|_{2,\mu}^2 &\le \sup_{f\in\mathcal{F}} (\|\hat{T}f - Tf\|_{2,\mu}^2 - \inf_{g\in\mathcal{G}} \|g - Tf\|_{2,\mu}^2) \\ &\le 2\sup_{g\in\mathcal{G}, f\in\mathcal{F}} |L(g,f) - \hat{L}(g,f)|.\end{aligned}$$

Define the function $l_{f,g} : \mathcal{S}\times\mathcal{A}\times[-R,R]\times\mathcal{S} \to \mathbb{R}$ as $l_{f,g}(s_i,a_i,r_i,s_{i+1}) = (f(s_i,a_i) - r_i - \max_a g(s_{i+1},a))^2$ and the function space $\mathcal{L}_{\mathcal{F},\mathcal{G}} = \{l_{f,g} \,|\, f\in\mathcal{F}, g\in\mathcal{G}\}$ and $\mathcal{G}_{max} = \{\max_a g(s,a) \,|\, g\in\mathcal{G}\}$. The pseudo dimension of $\mathcal{G}_{max}$ could be bounded by following Lemma.

Lemma 9. *Define* $\mathcal{G}_{max} = \{\max_{a\in\mathcal{A}} g(\cdot,a) : g\in\mathcal{G}\}$. $V_{\mathcal{G}_{max}} \le 2|\mathcal{A}|V_{\mathcal{G}}\log(3|\mathcal{A}|)$.

Proof of Lemma 9. By the definition of pseudo dimension, we have $V_{\mathcal{G}} \ge V_{\mathcal{G}^i}$ where $\mathcal{G}^i = \{g(x,a_i) \,|\, g\in\mathcal{G})\}$. Since $\max_{a\in\mathcal{A}} g(\cdot,a) \le 0 \iff \forall i\ g(\cdot,a_i) \le 0$, the claim follows from Lemma 3.2.3 of [10]. □

Now, we are ready to prove Lemma 2.

Proof of Lemma 2. Let $\mathcal{F} \subset \{f : \mathcal{S}\times\mathcal{A} \to [-f_{max}, f_{max}] \,|\, f\in B(S\times A)\}$ and $\mathcal{G} \subset \{g : \mathcal{S}\times\mathcal{A} \to [-g_{max}, g_{max}] \,|\, g\in B(S\times A)\}$. Without loss of generality, $g_{max} \le f_{max}$.

By similar argument in proof of Proposition 4 of [16], $\{s_i,a_i,r_i\}$ is also β-mixing with the coefficient $\{\beta_i\}$ and this implies $\{s_i,a_i,r_i,s_{i+1}\}$ is also stationary β-mixing with coefficient $\{\beta_{i-1}\}$. By direct calculation, $|\hat{L}(f,g)| \le (2f_{max}+R)^2$. Now, we apply Fact 2 with $l(f,g)$ and $Z_i = (s_i,a_i,r_i,s_{i+1})$. Then, we get

$$P\left(\sup_{f\in\mathcal{F},g\in\mathcal{G}} \left|\hat{L}(f,g) - L(f,g)\right| > \epsilon\right) \le 16\mathbb{E}[\mathcal{N}(\epsilon/8, \mathcal{L}_{\mathcal{F},\mathcal{G}}, (Z'_t)_{t\in H})] e^{-\frac{m_N\epsilon^2}{128(2f_{max}+R)^4}} + 2m_N\beta_{k_N}.$$

Since

$$\begin{aligned}&\hat{L}(f_1,g_1) - \hat{L}(f_2,g_2) \\ &= \frac{1}{n}\left|\sum_{i=1}^n (f_1(s_i,a_i) - r(s_i,a_i) - \max_{a\in\mathcal{A}} g_1(s_{i+1},a))^2 - \sum_{i=1}^n (f_2(s_i,a_i) - r(s_i,a_i) - \max_{a\in\mathcal{A}} g_2(s_{i+1},a))^2\right| \\ &\le 2\frac{2f_{max}+R}{n}\sum_{i=1}^n (|f_1(s_i,a_i) - f_2(s_i,a_i)| + |\max_{a\in\mathcal{A}} g_1(s_{i+1},a) - \max_{a\in\mathcal{A}} g_2(s_{i+1},a)|),\end{aligned}$$

this implies that

$$\mathcal{N}(4(2f_{max}+R)\epsilon, \mathcal{L}_{\mathcal{F},\mathcal{G}}, (z^{1:n}) \le \mathcal{N}(\epsilon,\mathcal{F},s^{2:n+1})\mathcal{N}(\epsilon,\mathcal{G}_{max},(s,a)^{1:n})$$

where $z_i = (s_i,a_i,r_i,s_{i+1})$ by definition of empirical covering number. Finally, by Fact 3, we get

$$\begin{aligned}&\mathcal{N}(\epsilon/8, \mathcal{L}_{\mathcal{F},\mathcal{G}}, (Z'_t)_{t\in H}) \\ &\le e(V_{\mathcal{F}}+1)\left(\frac{128(2f_{max}+R)e}{\epsilon}\right)^{V_{\mathcal{F}}} e(V_{\mathcal{F}_{max}}+1)\left(\frac{128(2f_{max}+R)e}{\epsilon}\right)^{V_{\mathcal{G}_{max}}} \\ &= C\left(\frac{1}{\epsilon}\right)^{V_{\mathcal{F}}+V_{\mathcal{G}_{max}}}\end{aligned}$$

where $C = e^2(V_{\mathcal{F}}+1)(V_{\mathcal{G}_{max}}+1)(128(2f_{max}+R)e)^{V_{\mathcal{F}}+V_{\mathcal{G}_{max}}}$.

For calculation, we use following prior result.

Fact 4 ([4], Lemma 14). *Let* $\beta_m \leq \bar{\beta}e^{(-bm^\kappa)}, N \geq 1, k_N = \lceil (C_2 N\epsilon^2/b)^{\frac{1}{1+\kappa}} \rceil, m_N = N/(2k_N), 0 < \delta \leq 1, V \geq 2$ *and* $C_1, C_2, \bar{\beta}, b, \kappa > 0$. *Define* ϵ *and* C_0 *as*

$$\epsilon = \sqrt{\frac{C_0(\max\{C_0/b, 1\})^{1/\kappa}}{C_2 N}}$$

with $C_0 = V/2 \log N + \log(e/\delta) + \log(\max(C_1 C_2^{V/2}, \bar{\beta}, 1))$

$$C_1 \left(\frac{1}{\epsilon}\right)^V e^{-4C_2 m_N \epsilon^2} + 2m_N \beta_{k_N} \leq \delta.$$

Then, by this fact and previous arguments, for $\epsilon = \sqrt{\frac{c_0(\max\{c_0/b,1\})^{1/\kappa}}{c_2 n}}$,

$$P\left(\sup_{f\in\mathcal{F}, g\in\mathcal{G}} \left|\hat{L}(f,g) - L(f,g)\right| \leq \epsilon\right) \geq 1 - \delta$$

where $c_0 = (V_\mathcal{F} + V_{\mathcal{G}_{max}})/2 \log n + \log(e/\delta) + \log(\max(c_1 c_2^{(V_\mathcal{F}+V_{\mathcal{G}_{max}})/2}, \bar{\beta}, 1), c_1 = 16e^2(V_\mathcal{F} + 1)(V_{\mathcal{G}_{max}} + 1)(128(2f_{max} + R)e2)^{V_\mathcal{F}+V_{\mathcal{G}_{max}}}, c_2 = \frac{1}{512(2f_{max}+R)^4}, V_{\mathcal{G}_{max}} = 2|\mathcal{A}|V_\mathcal{G} \log(3|\mathcal{A}|)$. Let $\mathcal{G} = \mathcal{F}_k, \mathcal{F} = \mathcal{F}_{k+1}$ and $g = f_k$. By Lemma 8, this implies that with $1 - \delta$ probability,

$$\|Tf_k - \hat{T}f_k\|_{\mu,2}^2 \leq \epsilon_B + \sqrt{\frac{c_0(\max\{c_0/b, 1\})^{1/\kappa}}{4c_2 n}}.$$

Finally, by manipulating δ, we get desired result. □

D.4 Proof of Theorem 2

Proof of Theorem 2. By combining Lemma 2 and Proposition 1, we directly obtain following results. Under assumptions stated in Theorem 2, we have

$$\|g^{\pi_\star} - g^{\pi_K}\|_\infty \leq C_\mu^{1/2} \frac{8\|Q^{\pi_\star}\|_{2,\mu}}{K+2} + C_\mu^{1/2} \frac{2K}{3} \left(\left(\frac{c_{0,K}(\max\{c_{0,K}/b, 1\})^{1/\kappa}}{c_{2,K} n}\right)^{1/4} + \max_{k=0,\ldots,K-1} \sqrt{\epsilon_B(\mathcal{F}_k, \mathcal{F}_{k+1})}\right),$$

$$\|g^{\pi_\star} - g^{\pi_K}\|_{2,\rho} \leq C_{\mu,\rho}^{1/2} \frac{8\|Q^{\pi_\star}\|_{2,\mu}}{K+2} + C_{\mu,\rho}^{1/2} \frac{2K}{3} \left(\left(\frac{c_{0,K}(\max\{c_{0,K}/b, 1\})^{1/\kappa}}{c_{2,K} n}\right)^{1/4} + \max_{k=0,\ldots,K-1} \sqrt{\epsilon_B(\mathcal{F}_k, \mathcal{F}_{k+1})}\right),$$

where $c_{0,K} = \max_{k=0,\ldots,K-1} c_{0,k}, c_{0,k} = (V_{\mathcal{F}_{k+1}} + V_{(\mathcal{F}_k)_{max}})/2 \log n + \log(e/(K\delta)) + \log(\max(c_{1,k}, \bar{\beta}, 1)), c_{1,k} = 16e^2(V_{\mathcal{F}_{k+1}} + 1)(V_{(\mathcal{F}_k)_{max}} + 1)(24e)^{V_{\mathcal{F}_{k+1}}+V_{(\mathcal{F}_k)_{max}}}, c_{2,K} = \frac{1}{512(2K+1)^4 R^4}, V_{(\mathcal{F}_k)_{max}} = 2|\mathcal{A}|, V_{\mathcal{F}_k} \log(3|\mathcal{A}|)$.

Given $\epsilon > 0$, for the first inequality, let $K = \lceil 9C_\mu^{1/2} \|Q^{\pi_\star}\|_{2,\mu}/\epsilon \rceil$. Then, by direct calculation, we derive that

$$\|g^{\pi_\star} - g^{\pi_K}\|_\infty \leq \epsilon + KC_\mu^{1/2} \max_{k=0,\ldots,K-1} \sqrt{\epsilon_B(\mathcal{F}_k, \mathcal{F}_{k+1})}$$

with sample complexity

$$n = \tilde{\mathcal{O}}\left(\frac{b^{-1/\kappa}(c'_{0,K})^{\frac{1+\kappa}{\kappa}} R^4 \|Q^{\pi_\star}\|_{2,\mu}^8 C_\mu^6}{\epsilon^{12}}\right)$$

where $c'_{0,K} = \max_{k=0,\ldots,K-1} c'_{0,k}, c'_{0,k} = \log(1/\delta) + \log(\max(c_{1,k}, \bar{\beta})), c_{1,k} = 16e^2(V_{\mathcal{F}_{k+1}} + 1)(V_{(\mathcal{F}_k)_{max}} + 1)(24e)^{V_{\mathcal{F}_{k+1}}+V_{(\mathcal{F}_k)_{max}}}, V_{(\mathcal{F}_k)_{max}} = 2|\mathcal{A}|, V_{\mathcal{F}_k} \log(3|\mathcal{A}|)$, and $\tilde{\mathcal{O}}$ ignores all logarithmic factors.

Similarly, given $\epsilon > 0$, for the second inequality, let $K = \lceil 9C_{\mu,\rho}^{1/2}\|Q^{\pi_\star}\|_{2,\mu}/\epsilon \rceil$. Then, by direct calculation, we derive that

$$\|g^{\pi_\star} - g^{\pi_K}\|_\infty \leq \epsilon + KC_{\mu,\rho}^{1/2} \max_{k=0,\dots,K-1} \sqrt{\epsilon_B(\mathcal{F}_k, \mathcal{F}_{k+1})}$$

with sample complexity

$$n = \tilde{\mathcal{O}}\left(\frac{b^{-1/\kappa}(c'_{0,K})^{\frac{1+\kappa}{\kappa}} R^4 \|Q^{\pi_\star}\|_{2,\mu}^8 C_{\mu,\rho}^6}{\epsilon^{12}}\right)$$

where $c'_{0,K} = \max_{k=0,\dots,K-1} c'_{0,k}$, $c'_{0,k} = \log(1/\delta) + \log(\max(c_{1,k}, \bar{\beta}))$, $c_{1,k} = 16e^2(V_{\mathcal{F}_{k+1}} + 1)(V_{(\mathcal{F}_k)_{max}} + 1)(24e)^{V_{\mathcal{F}_{k+1}} + V_{(\mathcal{F}_k)_{max}}}$, $V_{(\mathcal{F}_k)_{max}} = 2|\mathcal{A}|, V_{\mathcal{F}_k}\log(3|\mathcal{A}|)$, and $\tilde{\mathcal{O}}$ ignores all logarithmic factors. □

E Omitted proofs in Section 5

E.1 Proof of Theorem 3

We first prove following key lemma.

Lemma 10. *Assume Assumptions 1, 2, 3, 8, 11, and 12 (Bellman optimality equation, existence of argmin, star-shaped function space, normalized function space, range of function space, IID dataset). Let μ be the distribution generating the dataset. Let $\epsilon > 0$ and $\delta > 0$. With probability $1 - \delta$, $\{f_k, \hat{T}f_k\}_{k=0}^{K-1}$ of R-Anc-F-QI satisfies*

$$\|Tf_k - \hat{T}f_k\|_{\mu,2}^2 \leq \frac{30(R + 4\|Q^{\pi_\star}\|_\infty)^2 \ln(2KN_\epsilon^2/\delta)}{n} + 3\epsilon + 13\epsilon_B(\mathcal{F}, \mathcal{F}),$$

where

$$N_\epsilon = \mathcal{N}(\tfrac{\epsilon}{108(R+4\|Q^{\pi_\star}\|_\infty)}; \mathcal{F}, \|\cdot\|_\infty).$$

Proof. The proof basically follows from the proof of Lemma 1. □

Now, we prove Theorem 3.

Proof of Theorem 3. Consider Apporximate Relative Anchored Value Iteration

$$Q_r^k = (1 - \lambda_k)Q_r^0 + \lambda_k(TQ_r^{k-1} + \epsilon_k - c_k\mathbf{1}) \tag{Apx-R-Anc-QI}$$

for $c_k \in \mathbb{R}$. Also, consider corresponding Approximate Anchored Value Iteration with same ϵ_k and starting point Q_r^0

$$Q^k = (1 - \lambda_k)Q_r^0 + \lambda_k(TQ^{k-1} + \epsilon_k). \tag{Apx-Anc-QI}$$

Since $Q^k - Q_r^k = d_k\mathbf{1}$ for some $d_k \in \mathbb{R}$, $\max_a Q^k(s,a) = \max_a Q_r^k(s,a)$ for all $s \in \mathcal{S}$ by the defintion of Bellman operator and this implies induced policies are same. Thus, Proposition 1 also holds for Apx-R-Anc-QI.

By combining Lemma 10 and Proposition 1, we directly obtain following results. Under assumptions stated in Theorem 3,

$$\begin{aligned}\|g^{\pi_\star} - g^{\pi_K}\|_\infty \leq\ & C_\mu^{1/2}\frac{8\|Q^{\pi_\star}\|_{2,\mu}}{K+2} \\ & + C_\mu^{1/2}\frac{2K}{3}\left(\sqrt{3\epsilon'} + \sqrt{\frac{30(R + 4\|Q^{\pi_\star}\|_\infty)^2 \ln(2KN_{\epsilon'}^2/\delta)}{n}} + \sqrt{13\epsilon_B(\mathcal{F}, \mathcal{F})}\right).\end{aligned}$$

$$\begin{aligned}\|g^{\pi_\star} - g^{\pi_K}\|_{2,\rho} \leq\ & C_{\mu,\rho}^{1/2}\frac{8\|Q^{\pi_\star}\|_{2,\mu}}{K+2} \\ & + C_{\mu,\rho}^{1/2}\frac{2K}{3}\left(\sqrt{3\epsilon'} + \sqrt{\frac{30(R + 4\|Q^{\pi_\star}\|_\infty)^2 \ln(2KN_{\epsilon'}^2/\delta)}{n}} + \sqrt{13\epsilon_B(\mathcal{F}, \mathcal{F})}\right),\end{aligned}$$

where

$$N_{\epsilon'} = \mathcal{N}\big(\tfrac{\epsilon'}{108(R+4\|Q^{\pi_\star}\|_\infty)}; \mathcal{F}, \|\cdot\|_\infty\big).$$

Given $\epsilon > 0$, for the first inequality, let $K = \lceil 18 C_\mu^{1/2} \|Q^{\pi_\star}\|_{2,\mu}/\epsilon \rceil, \epsilon' = \frac{4\epsilon^2}{27K^2 C_\mu}, n = \frac{36K^2 C_\mu}{\epsilon^2} 30(R + 4\|Q^{\pi_\star}\|_\infty)^2 \ln(2KN_{\epsilon'}^2/\delta)$. Then, by direct calculation, we derive that

$$\|g^{\pi_\star} - g^{\pi_K}\|_\infty \leq \epsilon + 3KC_\mu^{1/2}\sqrt{\epsilon_B(\mathcal{F}, \mathcal{F})}$$

with sample complexity

$$n = \mathcal{O}\left(\frac{(R + \|Q^{\pi_\star}\|_\infty)^2 \|Q^{\pi_\star}\|_\infty^2 C_\mu^3}{\epsilon^4} \ln(\mathcal{N}_\epsilon^2 C_\mu^{1/2}/(\delta\epsilon))\right)$$

where

$$N_\epsilon = \mathcal{N}\big(\tfrac{\epsilon^4}{10^6 C_\mu^2 (R+\|Q^{\pi_\star}\|_\infty)\|Q^{\pi_\star}\|_\infty^2}; \mathcal{F}, \|\cdot\|_\infty\big).$$

Similarly, given $\epsilon > 0$, for second inequality, let $K = \lceil 18 C_{\mu,\rho}^{1/2} \|Q^{\pi_\star}\|_{2,\mu}/\epsilon \rceil, \epsilon' = \frac{4\epsilon^2}{27K^2 C_{\mu,\rho}}, n = \frac{36K^2 C_{\mu,\rho}}{\epsilon^2} 30(R + 4\|Q^{\pi_\star}\|_\infty)^2 \ln(2K^2\mathcal{N}_{\epsilon'}/\delta)$, and

$$\|g^{\pi_\star} - g^{\pi_K}\|_{2,\rho} \leq \epsilon + 3KC_{\mu,\rho}^{1/2}\sqrt{\epsilon_B(\mathcal{F}, \mathcal{F})}$$

with sample complexity

$$n = \mathcal{O}\left(\frac{(R + \|Q^{\pi_\star}\|_\infty)^2 \|Q^{\pi_\star}\|_\infty^2 C_{\mu,\rho}^3}{\epsilon^4} \ln(\mathcal{N}_\epsilon^2 C_{\mu,\rho}^{1/2}/(\delta\epsilon))\right)$$

where

$$N_\epsilon = \mathcal{N}\big(\tfrac{\epsilon^4}{10^6 C_{\mu,\rho}^2 (R+\|Q^{\pi_\star}\|_\infty)\|Q^{\pi_\star}\|_\infty^2}; \mathcal{F}, \|\cdot\|_\infty\big).$$

□

E.2 Proof of Theorem 4

We first prove following key Lemma.

Lemma 11. *Assume Assumptions 1, 2, 3, 9, 10, 11, and 12 (Bellman optimality equation, existence of argmin, star-shaped function space, normalized function space, range of function space, single-trajectory dataset, β-mixing single-trajectory). Let μ be the distribution generating the dataset defined as $\mu(s,a) = \nu(s)\pi_b(a \mid s)$. Let $\epsilon > 0$ and $\delta > 0$. With probability $1-\delta$, $\{f_k, \hat{T} f_k\}_{k=0}^{K-1}$ of R-Anc-F-QI satisfies*

$$\|T f_k - \hat{T} f_k\|_{\mu,2}^2 \leq \epsilon_B(\mathcal{F}, \mathcal{F}) + \sqrt{\frac{c_0(\max\{c_0/b, 1\})^{1/\kappa}}{c_2 n}}$$

where $c_0 = (V_\mathcal{F} + V_{\mathcal{F}_{max}}) \log n/2 + \log(e/(K\delta)) + \log(\max(c_1, \bar{\beta}))$, $c_1 = 16e^2(V_\mathcal{F}+1)(V_{\mathcal{F}_{max}} + 1)(24e)^{V_\mathcal{F}+V_{\mathcal{F}_{max}}}$, $c_2 = \frac{1}{512(R+4\|Q^{\pi_\star}\|_\infty)^4}$, $V_{\mathcal{F}_{max}} = 2|\mathcal{A}|V_\mathcal{F} \log(3|\mathcal{A}|)$.

Proof. The proof basically follows from the proof of Lemma 2. □

Now, we prove Theorem 4.

Proof of Theorem 4. By combining Lemma 11 and Proposition 1, we directly obtain following results. Under assumptions stated in Theorem 11, we have

$$\begin{aligned}\|g^{\pi_\star} - g^{\pi_K}\|_\infty \leq\ & C_\mu^{1/2} \frac{8\|Q^{\pi_\star}\|_{2,\mu}}{K+2} \\ & + C_\mu^{1/2} \frac{2K}{3} \left(\left(\frac{c_0(\max\{c_0/b, 1\})^{1/\kappa}}{c_2 n} \right)^{1/4} + \sqrt{\epsilon_B(\mathcal{F}, \mathcal{F})} \right),\end{aligned}$$

$$\|g^{\pi_\star} - g^{\pi_K}\|_{2,\rho} \leq C_{\mu,\rho}^{1/2} \frac{8\|Q^{\pi_\star}\|_{2,\mu}}{K+2} + C_{\mu,\rho}^{1/2} \frac{2K}{3} \left(\left(\frac{c_0(\max\{c_0/b, 1\})^{1/\kappa}}{c_2 n} \right)^{1/4} + \sqrt{\epsilon_B(\mathcal{F}, \mathcal{F})} \right),$$

where $c_0 = (V_{\mathcal{F}} + V_{\mathcal{F}_{max}})/2 \log n + \log(e/(K\delta)) + \log(\max(c_1, \bar{\beta}, 1)), c_1 = 16e^2(V_{\mathcal{F}}+1)(V_{\mathcal{F}_{max}}+1)(24e)^{V_{\mathcal{F}}+V_{\mathcal{F}_{max}}}, c_2 = \frac{1}{512(R+4\|Q^{\pi_\star}\|_\infty)^4}, V_{\mathcal{F}_{max}} = 2|\mathcal{A}|V_{\mathcal{F}} \log(3|\mathcal{A}|)$.

Given $\epsilon > 0$, for the first inequality, let $K = \lceil 9C_\mu^{1/2}\|Q^{\pi_\star}\|_{2,\mu}/\epsilon \rceil$. Then, by direct calculation, we derive that

$$\|g^{\pi_\star} - g^{\pi_K}\|_\infty \leq \epsilon + KC_\mu^{1/2}\sqrt{\epsilon_B(\mathcal{F}, \mathcal{F})}$$

with sample complexity

$$n = \tilde{\mathcal{O}}\left(\frac{b^{-1/\kappa}(c_0')^{\frac{1+\kappa}{\kappa}}(R + \|Q^{\pi_\star}\|_\infty)^4 \|Q^{\pi_\star}\|_\infty^4 C_\mu^4}{\epsilon^8} \right)$$

where $c_0' = \log(1/\delta) + \log(\max(c_1, \bar{\beta})), c_1 = 16e^2(V_{\mathcal{F}}+1)(V_{\mathcal{F}_{max}}+1)(24e)^{V_{\mathcal{F}}+V_{\mathcal{F}_{max}}}, V_{\mathcal{F}_{max}} = 2|\mathcal{A}|V_{\mathcal{F}} \log(3|\mathcal{A}|)$, and $\tilde{\mathcal{O}}$ ignores all logarithmic factors.

Similarly, given $\epsilon > 0$, for the second inequality, let $K = \lceil 9C_{\mu,\rho}^{1/2}\|Q^{\pi_\star}\|_{2,\mu}/\epsilon \rceil$. Then, by direct calculation, we derive that

$$\|g^{\pi_\star} - g^{\pi_K}\|_\infty \leq \epsilon + KC_{\mu,\rho}^{1/2}\sqrt{\epsilon_B(\mathcal{F}, \mathcal{F})}$$

with sample complexity

$$n = \tilde{\mathcal{O}}\left(\frac{b^{-1/\kappa}(c_0')^{\frac{1+\kappa}{\kappa}}(R + \|Q^{\pi_\star}\|_\infty)^4 \|Q^{\pi_\star}\|_\infty^4 C_{\mu,\rho}^4}{\epsilon^8} \right)$$

where $c_0' = \log(1/\delta) + \log(\max(c_1, \bar{\beta})), c_1 = 16e^2(V_{\mathcal{F}}+1)(V_{\mathcal{F}_{max}}+1)(24e)^{V_{\mathcal{F}}+V_{\mathcal{F}_{max}}}, V_{\mathcal{F}_{max}} = 2|\mathcal{A}|V_{\mathcal{F}} \log(3|\mathcal{A}|)$, and $\tilde{\mathcal{O}}$ ignores all logarithmic factors. □