

## 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  $\epsilon$ -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 partial 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 Bellman 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 realizability 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  $\epsilon$ -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 independent random variables. If  $X_i \leq b$  for all  $i$ , then*

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

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

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

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

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

## 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.,  $\mathcal{P}_*^\pi = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (\mathcal{P}^\pi)^i$ . (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*

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

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

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

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

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

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

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

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

This implies

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

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

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

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

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

□

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  $\mu$  and  $\rho$  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),

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

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

*Proof.* Under Assumption 6, for any non negative integer  $n$ , we have  $\rho^\top (\mathcal{P}^{\pi_*})^n \mathcal{P}^{\pi_1} \mathcal{P}^{\pi_2} \dots \mathcal{P}^{\pi_k} (\cdot) \leq C_{\mu, \rho} \mu$ . This implies  $\rho^\top \mathcal{P}_*^{\pi_*} \mathcal{P}^{\pi_1} \mathcal{P}^{\pi_2} \dots \mathcal{P}^{\pi_k} (\cdot) \leq C_{\mu, \rho} \mu$  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  $\mu$  and  $\rho$  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,

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

Let  $\mathcal{P}_l^Q = \mathcal{P}_*^{\pi_K} \Pi_{i=l+1}^K \mathcal{P}^{\pi_i} \left( \sum_{m=0}^l \Pi_{i=m+1}^l \lambda_i (1 - \lambda_m) (\mathcal{P}^{\pi_*})^{l+1-m} + I \right) / 2$  and  $\mathcal{P}_l^\epsilon = \mathcal{P}_*^{\pi_K} \sum_{m=l}^K (\lambda_{m+1} - \lambda_m) \Pi_{i=m+1}^K \mathcal{P}^{\pi_i} (\mathcal{P}^{\pi_*})^{m+1-l} + \Pi_{i=l+1}^K \mathcal{P}^{\pi_i} (\lambda_l \mathcal{P}^{\pi_l} + I) / 2$ . 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

$$\begin{aligned} \|g^{\pi^*} - g^{\pi_K}\|_{p,\rho} &\leq 2 \sum_{l=0}^K \Pi_{i=l+1}^K \lambda_i (\lambda_{l+1} - \lambda_l) \|\mathcal{P}_l | Q^0 - Q^{\pi^*}\|_{p,\rho} + 2 \sum_{l=1}^K \Pi_{i=l}^K \lambda_i \|\mathcal{P}_l^\epsilon | \epsilon_l\|_{p,\rho} \\ &\leq 2C_\mu^{1/p} \sum_{l=0}^K \Pi_{i=l+1}^K \lambda_i (\lambda_{l+1} - \lambda_l) \|Q^0 - Q^{\pi^*}\|_{p,\mu} + 2C_\mu^{1/p} \sum_{l=1}^K \Pi_{i=l}^K \lambda_i \|\epsilon_l\|_{p,\mu}, \end{aligned}$$

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  $\rho$ , then choosing  $\rho$  to be a Dirac distribution at each state proves the case of Assumption 5 which implies first inequality of Proposition 1.  $\square$

## D Omitted proofs in Section 4

### D.1 Proof of Lemma 1

*Proof of Lemma 1.* Let  $\mathcal{F} \subset \{f : \mathcal{S} \times \mathcal{A} \rightarrow [-f_{\max}, f_{\max}] \mid f \in B(S \times A)\}$  and  $\mathcal{G} \subset \{f : \mathcal{S} \times \mathcal{A} \rightarrow [-g_{\max}, g_{\max}] \mid f \in B(S \times A)\}$ . Let  $f_1, \dots, f_N$  cover the  $\mathcal{F}$  and  $g_1, \dots, g_{N'}$  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  $\{f_i, g_j\}_{i=1, \dots, N, j=1, \dots, N'}$ ,

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

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

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

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

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

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

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 (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  $\delta$ , we get desired result.  $\square$

## 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

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

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

where

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

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

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

with sample complexity

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

where

$$N_\epsilon = \max_{k=1,\dots,K} N_{k,\epsilon}, \quad N_{k,\epsilon} = \mathcal{N}\left(\frac{\epsilon^4}{10^6 k C_\mu^2 \|Q^{\pi^*}\|_{2,\mu}^2 R}; \mathcal{F}_k, \|\cdot\|_\infty\right), \quad \text{for } k = 1, \dots, K.$$

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

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

with sample complexity

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

where

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

□

### 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 covering number depends on number of samples, but it can be bounded by pseudo dimension which depends on only function space and  $\epsilon$  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  $\epsilon > 0$ ,*

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

Define  $L(g, f) = \mathbb{E}_{s_i, a_i \sim \mu} [\text{Var}_{s'_i \sim P(\cdot | s_i, a_i)} (r(s_i, a_i) + \max_a f(s'_i, a)) + \|g - Tf\|_{2,\mu}^2]$  where  $\text{Var}$  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.**  $\|\hat{T}f - Tf\|_{2,\mu}^2 - \inf_{g \in \mathcal{G}} \|g - Tf\|_{2,\mu}^2 \leq 2 \sup_{g \in \mathcal{G}} |L(g, f) - \hat{L}(g, f)|$ .

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

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

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

Define the function  $l_{f,g} : \mathcal{S} \times \mathcal{A} \times [-R, R] \times \mathcal{S} \rightarrow \mathbb{R}$  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  $\mathcal{L}_{\mathcal{F}, \mathcal{G}} = \{l_{f,g} \mid f \in \mathcal{F}, g \in \mathcal{G}\}$  and  $\mathcal{G}_{max} = \{\max_a g(s, a) \mid g \in \mathcal{G}\}$ . The pseudo dimension of  $\mathcal{G}_{max}$  could be bounded by following Lemma.

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

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

Now, we are ready to prove Lemma 2.

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

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

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

Since

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

this implies that

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

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

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

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  $\epsilon$  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  $\epsilon = \sqrt{\frac{c_0(\max\{c_0/b, 1\})^{1/\kappa}}{c_2 n}}$ ,

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

where  $c_0 = (V_{\mathcal{F}} + V_{\mathcal{G}_{max}})/2 \log n + \log(e/\delta) + \log(\max(c_1 c_2^{(V_{\mathcal{F}} + V_{\mathcal{G}_{max}})/2}, \bar{\beta}, 1))$ ,  $c_1 = 16e^2(V_{\mathcal{F}} + 1)(V_{\mathcal{G}_{max}} + 1)(128(2f_{max} + R)e^2)^{V_{\mathcal{F}} + V_{\mathcal{G}_{max}}}$ ,  $c_2 = \frac{1}{512(2f_{max} + R)^4}$ ,  $V_{\mathcal{G}_{max}} = 2|\mathcal{A}|V_{\mathcal{G}} \log(3|\mathcal{A}|)$ . 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  $\delta$ , we get desired result.  $\square$

#### 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

$$\begin{aligned} \|g^{\pi^*} - g^{\pi_K}\|_{\infty} &\leq C_{\mu}^{1/2} \frac{8\|Q^{\pi^*}\|_{2, \mu}}{K+2} \\ &\quad + C_{\mu}^{1/2} \frac{2K}{3} \left( \left( \frac{c_{0,K}(\max\{c_{0,K}/b, 1\})^{1/\kappa}}{c_{2,K} n} \right)^{1/4} + \max_{k=0, \dots, K-1} \sqrt{\epsilon_B(\mathcal{F}_k, \mathcal{F}_{k+1})} \right), \end{aligned}$$

$$\begin{aligned} \|g^{\pi^*} - g^{\pi_K}\|_{2, \rho} &\leq C_{\mu, \rho}^{1/2} \frac{8\|Q^{\pi^*}\|_{2, \mu}}{K+2} \\ &\quad + C_{\mu, \rho}^{1/2} \frac{2K}{3} \left( \left( \frac{c_{0,K}(\max\{c_{0,K}/b, 1\})^{1/\kappa}}{c_{2,K} n} \right)^{1/4} + \max_{k=0, \dots, K-1} \sqrt{\epsilon_B(\mathcal{F}_k, \mathcal{F}_{k+1})} \right), \end{aligned}$$

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

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

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

with sample complexity

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

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  $\tilde{O}$  ignores all logarithmic factors.

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

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

with sample complexity

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

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.  $\square$

## 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  $\mu$  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*

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

where

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

*Proof.* The proof basically follows from the proof of Lemma 1.  $\square$

Now, we prove Theorem 3.

*Proof of Theorem 3.* Consider Apporximate Relative Anchored Value Iteration

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

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

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

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 definition 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,

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

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

where

$$N_{\epsilon'} = \mathcal{N}\left(\frac{\epsilon'}{108(R+4\|Q^{\pi^*}\|_\infty)}; \mathcal{F}, \|\cdot\|_\infty\right).$$

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

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

with sample complexity

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

where

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

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

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

with sample complexity

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

where

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

□

## 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,  $\beta$ -mixing single-trajectory). Let  $\mu$  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  $R$ -Anc-F-QI satisfies

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

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

*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

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

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

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

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

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

with sample complexity

$$n = \tilde{\mathcal{O}} \left( \frac{b^{-1/\kappa}(c'_0)^{\frac{1+\kappa}{\kappa}}(R + \|Q^{\pi_*}\|_{\infty})^4\|Q^{\pi_*}\|_{\infty}^4 C_{\mu}^4}{\epsilon^8} \right)$$

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  $K = \lceil 9C_{\mu,\rho}^{1/2}\|Q^{\pi_*}\|_{2,\mu}/\epsilon \rceil$ . Then, by direct calculation, we derive that

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

with sample complexity

$$n = \tilde{\mathcal{O}} \left( \frac{b^{-1/\kappa}(c'_0)^{\frac{1+\kappa}{\kappa}}(R + \|Q^{\pi_*}\|_{\infty})^4\|Q^{\pi_*}\|_{\infty}^4 C_{\mu,\rho}^4}{\epsilon^8} \right)$$

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.  $\square$