Sample Complexity of Policy-Based Methods under Off-Policy Sampling and Linear Function Approximation

Zaiwei Chen
Georgia Institute of Technology

Siva Theja Maguluri
Georgia Institute of Technology

Abstract

In this work, we study policy-based methods for solving the reinforcement learning problem, where off-policy sampling and linear function approximation are employed for policy evaluation, and various policy update rules (including natural policy gradient) are considered for policy improvement. To solve the policy evaluation sub-problem in the presence of the deadly triad, we propose a generic algorithm framework of multi-step TD-learning with generalized importance sampling ratios, which includes two specific algorithms: the λ-averaged Q-trace and the two-sided Q-trace. The generic algorithm is single time-scale, has provable finite-sample guarantees, and overcomes the high variance issue in off-policy learning. As for the policy improvement, we provide a universal analysis that establishes geometric convergence of various policy update rules, which leads to an overall $\tilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity.

1 INTRODUCTION

Policy-based methods including approximate policy iteration and various policy gradient methods are popular approaches to solve the reinforcement learning (RL) problem (Sutton and Barto, 2018). Two key ideas that are behind these successes are *function approximation* and *off-policy sampling*. Since we usually have to deal with extremely large or even continuous state and action spaces, function approximation enables the agent to overcome the curse of dimensionality so that RL is computationally tractable. On the other hand, sampling can be of high risk and/or expensive in many realistic RL problems such as clinical trials (Zhao et al., 2009) and power systems (Glavic et al., 2017). Off-policy sampling overcomes this challenge because the agent can learn in an off-line manner using historical data.

On the theoretical side, there is an increasing interest in understanding the finite-sample convergence behavior of policy-based methods. However, policy-based methods under off-policy sampling and function approximation are in general not well understood. This leads to our main contributions in the following.

Off-Policy TD-Learning under Linear Function Approximation. To solve the policy evaluation sub-problem in general policy-based approaches, we propose a generic single time-scale algorithm design of multi-step TD-learning with generalized importance sampling ratios, including two specific algorithms: the λ-averaged Q-trace algorithm and the two-sided Q-trace algorithm. We establish their finite-sample convergence guarantees, characterize the limit points as solutions to generalized multi-step projected Bellman equations (PBEs), and provide performance bounds on the limit points in terms of the error compared to the true value function.

The $\tilde{\mathcal{O}}(\epsilon^{-2})$ Sample Complexity of Policy-Based Methods with Various Policy Update Rules. Consider the general policy-based framework where the policy evaluation is solved with our proposed off-policy TD-learning algorithm, and the policy improvement uses various policy update rules, including $1/\beta_1$-greedy update, β_2-softmax update, and β_3-natural policy gradient update (see Section 3). We provide a unified approach to show that the overall sample complexity for all these algorithms is $\tilde{\mathcal{O}}(\epsilon^{-2})$. In the case of off-policy natural actor-critic, this improves the previous state-of-the-art result in Chen et al. (2021a) by a factor of ϵ^{-1}.

1.1 Related Literature

At a high level, RL algorithms can be divided into two categories: value-based methods and policy-based

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

methods. Policy-based methods include actor-critic (Konda and Tsitsiklis, 2000), its variant natural actor-critic (Kakade, 2001), and approximate policy iteration (Bertsekas, 2011), and is usually consisted of policy evaluation and policy improvement.

TD-Learning. The TD-learning method is used to solve the policy evaluation sub-problem, and is usually used in policy-based methods to ultimately find an optimal policy. The asymptotic convergence of TD-learning was established in Tsitsiklis (1994); Dayan and Sejnowski (1994); Bertsekas and Yu (2009). Finite-sample analysis of variants of TD-learning algorithms using on-policy sampling was performed in Chen et al. (2021b), and using off-policy sampling was performed in Khodadadian et al. (2021b); Chen et al. (2021c). In the function approximation setting, TD-learning with linear function approximation was studied in Tsitsiklis and Van Roy (1997); Lazaric et al. (2012); Srikant and Ying (2019); Bhandari et al. (2018) when using on-policy sampling. In the off-policy linear function approximation setting, due to the presence of the deadly triad, TD-learning algorithms can diverge (Sutton and Barto, 2018). Variants of TD-learning algorithms such as TDC (Sutton et al., 2009), GTD (Sutton et al., 2008), emphathic TD (Sutton et al., 2016), and n-step TD (with a large enough n) (Chen et al., 2021a) were used to resolve the divergence issue, and the finite-sample bounds were studied in Ma et al. (2020); Wang et al. (2021); Chen et al. (2021a). Note that TDC, GTD, and emphathic TD are two time-scale algorithms, while n-step TD is single time-scale, it suffers from a high variance due to the cumulative product of the importance sampling ratios. See Appendix D of this work for a detailed discussion.

(Natural) Actor-Critic. The asymptotic convergence of on-policy actor-critic was established in Williams and Baird (1990); Borkar (2009); Borkar and Konda (1997) when using a tabular representation, and in Konda and Tsitsiklis (2000); Bhatnagar et al. (2009) when using function approximation. In recent years, there has been an increasing interest in understanding the finite-sample behavior of (natural) actor-critic algorithms. Here is a non-exhaustive list: Lan (2021); Khodadadian et al. (2021b); Zhang et al. (2019); Qiu et al. (2019); Kumar et al. (2019); Liu et al. (2019); Wang et al. (2019); Liu et al. (2020); Wu et al. (2020); Cayci et al. (2021). The state-of-the-art sample complexity of on-policy natural actor-critic is $\tilde{\mathcal{O}}(\epsilon^{-2})$ (Lan, 2021). However, only tabular RL was considered in Lan (2021). In the off-policy setting, finite-sample analysis of natural actor-critic was studied in Khodadadian et al. (2021a) in the tabular setting, and in Chen et al. (2021a) in the linear function approximation, and the sample complexity in both cases is $\tilde{\mathcal{O}}(\epsilon^{-3})$.

1.2 Background on Reinforcement Learning

Consider modeling the RL problem as a *finite Markov decision process* (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is the state-space, $\mathcal{A}$ is the action-space, $\mathcal{P} = \{P_a \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|} \mid a \in \mathcal{A}\}$ is the set of unknown action-dependent transition probability matrices, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is the unknown reward function, and $\gamma \in (0, 1)$ is the discount factor. We assume without loss of generality that $\max_{s,a} |\mathcal{R}(s, a)| \leq 1$. The goal is to find an optimal policy π^* of selecting actions so that the long term reward is maximized. Formally, define the state-action value function associated with a policy π at state-action pair (s, a) by $Q^\pi(s, a) = \mathbb{E}_\pi[\sum_{k=0}^\infty \gamma^k \mathcal{R}(S_k, A_k) \mid S_0 = s, A_0 = a]$, where we use the notation $\mathbb{E}_\pi[\,\cdot\,]$ to indicate that the actions are chosen according to the policy π. Then the goal is to find an optimal policy π^* such that $Q^* := Q^{\pi^*}$ is maximized uniformly for all (s, a). A popular approach to solve the RL problem is to use the policy-based methods. In every iteration of general policy-based algorithms, the agent first performs a policy evaluation step to estimate the value function of the current policy iterate, which is then followed by a policy improvement step to update the policy.

2 OFF-POLICY TD-LEARNING WITH LINEAR FUNCTION APPROXIMATION

This section is dedicated to solving the policy evaluation sub-problem within general policy-based methods. Policy evaluation refers to estimating the Q-function Q^π of a given target policy π, and is usually solved with the TD-learning algorithm. Depending on whether the policy π_b used to collect samples (called the behavior policy) is equal to the target policy π or not, there are on-policy TD-learning (i.e., $\pi_b = \pi$) and off-policy TD-learning (i.e., $\pi_b \neq \pi$).

TD-learning becomes computationally intractable when the size of the state-action space is large. This motivates the use of function approximation. In linear function approximation, we choose a set of basis vectors $\phi_i \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $1 \leq i \leq d$. Let $\Phi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times d}$ be a matrix defined by $\Phi = [\phi_1, \cdots, \phi_d]$. Then, the goal is to find from the linear sub-space $\mathcal{Q} = \{\tilde{Q}_w = \Phi w \mid w \in \mathbb{R}^d\}$ the "best" approximation of the Q-function Q^π, where $w \in \mathbb{R}^d$ is the weight vector. Let $\phi(s, a) = [\phi_1(s, a), \phi_2(s, a), \cdots, \phi_d(s, a)]^\top \in \mathbb{R}^d$ be the feature vector associated with the pair (s, a).

When TD-learning is used along with off-policy sampling and linear function approximation, the deadly

Algorithm 1 A Generic Multi-Step Off-Policy TD-Learning with Linear Function Approximation

1: **Input**: Integer K, bootstrapping parameter n, stepsize sequence $\{\alpha_k\}$, initialization w_0, target policy π, behavior policy π_b, generalized importance sampling ratios $c, \rho : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}_+$, and a single trajectory of samples $\{(S_k, A_k)\}_{0 \le k \le K+n-1}$ generated by the behavior policy π_b.
2: **for** $k = 0, 1, \cdots, K-1$ **do**
3: $\quad \Delta_i(w_k) = \mathcal{R}(S_i, A_i) + \gamma \rho(S_{i+1}, A_{i+1}) \phi(S_{i+1}, A_{i+1})^\top w_k - \phi(S_i, A_i)^\top w_k,\ i \in \{k, k+1, \cdots, k+n-1\}$
4: $\quad w_{k+1} = w_k + \alpha_k \phi(S_k, A_k) \sum_{i=k}^{k+n-1} \gamma^{i-k} \prod_{j=k+1}^{i} c(S_j, A_j) \Delta_i(w_k)$
5: **end for**
6: **Output:** w_K

triad is formed and the algorithm can be unstable. We next propose a generic framework of TD-learning algorithms (including two specific algorithms: the λ-averaged Q-trace and the two-sided Q-trace), which provably converge in the presence of the deadly triad, and do not suffer from the high variance issue in off-policy learning. Throughout this paper, we impose the following assumption on the basis vectors. Such assumption is indeed without loss of generality.

Assumption 2.1. The matrix Φ has full column-rank, and satisfies $\|\Phi\|_\infty \le 1$.

2.1 Algorithm Design

We present in Algorithm 1 a generic TD-learning algorithm using off-policy sampling and linear function approximation. In Algorithm 1, the choice of the generalized importance sampling ratios $c(\cdot, \cdot)$ and $\rho(\cdot, \cdot)$ is of vital importance. We next present two specific choices, resulting in two novel algorithms called λ-averaged Q-trace and two-sided Q-trace.

The λ-Averaged Q-Trace Algorithm. Let $\lambda \in \mathbb{R}^{|\mathcal{S}|}$ be a vector-valued tunable parameter satisfying $\lambda \in [\mathbf{0}, \mathbf{1}]$. Then the generalized importance sampling ratios are chosen as $c(s,a) = \rho(s,a) = \lambda(s) \frac{\pi(a|s)}{\pi_b(a|s)} + 1 - \lambda(s)$ for all (s,a). Observe that when $\lambda = \mathbf{1}$, we have $c(s,a) = \rho(s,a) = \frac{\pi(a|s)}{\pi_b(a|s)}$, and Algorithm 1 reduces to the convergent multi-step off-policy TD-learning algorithm presented in Chen et al. (2021a), which however suffers from an exponential large variance due to the cumulative product of the importance sampling ratios. See Appendix D for more details. On the other hand, when $\lambda = \mathbf{0}$, we have $c(s,a) = \rho(s,a) = 1$, and hence the product of the generalized importance sampling ratios is deterministically equal to one, resulting in no variance at all. However, in this case, we are essentially performing policy evaluation of the behavior policy π_b instead of the target policy π, hence there will be a bias in the limit of Algorithm 1. More generally, when $\lambda \in (\mathbf{0}, \mathbf{1})$, there is a trade-off between the variance and the bias in the limit point. Such trade-off will be studied quantitatively in Section 2.3.

The Two-Sided Q-Trace Algorithm. To introduce the algorithm, we first define the two-sided truncation function. Given upper and lower truncation levels $a, b \in \mathbb{R}$, define $g_{a,b} : \mathbb{R} \mapsto \mathbb{R}$ by $g_{a,b}(x) = a$ when $x < a$, $g_{a,b}(x) = x$ when $a \le x \le b$, and $g_{a,b}(x) = b$ when $x > b$. Let $\ell, u \in \mathbb{R}^{|\mathcal{S}|}$ be two vector-valued tunable parameters satisfying $\mathbf{0} \le \ell \le \mathbf{1} \le u$. Then, for the two-sided Q-trace algorithm, the generalized importance sampling ratios are chosen as $c(s,a) = \rho(s,a) = g_{\ell(s),u(s)}\left(\pi(a|s)/\pi_b(a|s)\right)$ for all (s,a). The idea of truncating the importance sampling ratios from above was already employed in existing algorithms such as Retrace(λ) (Munos et al., 2016), V-trace (Espeholt et al., 2018), and Q-trace (Khodadadian et al., 2021b), and is used to control the high variance in off-policy learning. However, none of them were shown to converge in the function approximation setting. Introducing the lower truncation level is crucial to ensure the convergence of the two-sided Q-trace algorithm in the presence of the deadly triad. This will be illustrated in detail in Section 2.3.

2.2 The Generalized PBE

We next theoretically analyze Algorithm 1. Specifically, in this section, we formulate Algorithm 1 as a stochastic approximation algorithm for solving a generalized PBE and study its properties. We begin by stating our assumption.

Assumption 2.2. The behavior policy π_b satisfies $\pi_b(a|s) > 0$ for all (s,a), and induces an irreducible and aperiodic Markov chain $\{S_k\}$.

Assumption 2.2 implies that the Markov chain $\{S_k\}$ induced by π_b has a unique stationary distribution $\mu \in \Delta^{|\mathcal{S}|}$. Moreover, there exist $C \ge 1$ and $\sigma \in (0,1)$ such that $\max_{s \in \mathcal{S}} \|P_{\pi_b}^k(s, \cdot) - \mu(\cdot)\|_{\text{TV}} \le C\sigma^k$ for all $k \ge 0$, where P_{π_b} is the transition probability matrix of the Markov chain $\{S_k\}$ under π_b (Levin and Peres, 2017).

For simplicity, denote $c_{i,j} = \prod_{k=i}^{j} c(S_k, A_k)$. The target equation Algorithm 1 aims at solving is:

$$\mathbb{E}_{S_0 \sim \mu}\left[\phi(S_0, A_0) \sum_{i=0}^{n-1} \gamma^i c_{1,i} \Delta_i(w)\right] = 0, \qquad (1)$$

where $A_i \sim \pi_b(\cdot|S_i)$ and $S_{i+1} \sim P_{A_i}(S_i,\cdot)$. The following lemma formulates Eq. (1) in the form of a generalized PBE. To present the lemma, we first introduce some notation. Let $\mathcal{K}_{SA} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ be a diagonal matrix with diagonal entries $\{\mu(s)\pi_b(a|s)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$, and let $\mathcal{K}_{SA,\min}$ be the minimal diagonal entry. Let $\|\cdot\|_{\mathcal{K}_{SA}}$ be the weighted ℓ_2-norm with weights $\{\mu(s)\pi_b(a|s)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$, and denote $\text{Proj}_{\mathcal{Q}}$ as the projection operator onto the linear sub-space $\mathcal{Q}$ with respect to $\|\cdot\|_{\mathcal{K}_{SA}}$. Let $\mathcal{T}_c,\mathcal{H}_\rho : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \mapsto \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be two operators defined by $[\mathcal{T}_c(Q)](s,a) = \sum_{i=0}^{n-1}\mathbb{E}_{\pi_b}[\gamma^i c_{1,i} Q(S_i,A_i) \mid S_0=s, A_0=a]$ and $[\mathcal{H}_\rho(Q)](s,a) = \mathcal{R}(s,a) + \gamma\mathbb{E}_{\pi_b}[\rho(S_{k+1},A_{k+1})Q(S_{k+1},A_{k+1}) \mid S_k=s, A_k=a]$ for any $Q\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and state-action pair (s,a).

Lemma 2.1. *Eq. (1) is equivalent to:*

$$\Phi w = \text{Proj}_{\mathcal{Q}}\mathcal{B}_{c,\rho}(\Phi w), \tag{2}$$

where $\mathcal{B}_{c,\rho}(\cdot)$ is the generalized Bellman operator defined by $\mathcal{B}_{c,\rho}(Q) = \mathcal{T}_c(\mathcal{H}_\rho(Q)-Q)+Q$.

The generalized Bellman operator $\mathcal{B}_{c,\rho}(\cdot)$ was previously introduced in Chen et al. (2021c) to study off-policy TD-learning algorithms in the *tabular* setting (i.e., $\Phi = I_{SA}$), where the contraction property of $\mathcal{B}_{c,\rho}(\cdot)$ was shown. However, $\mathcal{B}_{c,\rho}(\cdot)$ alone being a contraction is not enough to guarantee the convergence of Algorithm 1 because of function approximation, which introduces an additional projection operator $\text{Proj}_{\mathcal{Q}}$. What we truly need is that (1) the composed operator $\text{Proj}_{\mathcal{Q}}\mathcal{B}_{c,\rho}(\cdot)$ is a contraction mapping, and (2) the solution $w_{c,\rho}^\pi$ of Eq. (2) is such that $\Phi w_{c,\rho}^\pi$ is an approximation of the Q-function Q^π. We next provide sufficient conditions on the choices of the generalized importance sampling ratios $c(\cdot,\cdot)$ and $\rho(\cdot,\cdot)$, and the bootstrapping parameter n so that the above two requirements are satisfied.

Let $D_c, D_\rho \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ be two diagonal matrices such that $D_c((s,a),(s,a)) = \sum_{a'\in\mathcal{A}}\pi_b(a'|s)c(s,a')$ and $D_\rho((s,a),(s,a)) = \sum_{a'\in\mathcal{A}}\pi_b(a'|s)\rho(s,a')$ for all (s,a). Let $D_{c,\max}$ and $D_{\rho,\max}$ ($D_{c,\min}$ and $D_{\rho,\min}$) be the maximam (minimum) diagonal entries of the matrices D_c and D_ρ respectively.

Condition 2.1. The generalized importance sampling ratios $c(\cdot,\cdot),\rho(\cdot,\cdot)$ satisfy (1) $c(s,a)\le\rho(s,a)$, $\forall\,(s,a)$, (2) $D_{\rho,\max} < 1/\gamma$, and (3) $\frac{\gamma(D_{\rho,\max}-D_{c,\min})}{(1-\gamma D_{c,\min})\sqrt{\mathcal{K}_{SA,\min}}} < 1$.

Condition 2.1 (1) and (2) were previously introduced in Chen et al. (2021c), and were used to show the contraction property of the operator $\mathcal{B}_{c,\rho}(\cdot)$. In particular, it was shown that the generalized Bellman operator $\mathcal{B}_{c,\rho}(\cdot)$ is a contraction mapping with respect to $\|\cdot\|_\infty$, with contraction factor $\tilde{\gamma}(n) = 1 - f_n(\gamma D_{c,\min})(1-\gamma D_{\rho,\max})$, where $f_n:\mathbb{R}\mapsto\mathbb{R}$ is defined by $f_n(x) = \sum_{i=0}^{n-1}x^i$ for any x. It is clear that $\tilde{\gamma}(n)\in(0,1)$, and is a decreasing function of n.

As illustrated earlier, $\mathcal{B}_{c,\rho}(\cdot)$ being a contraction mapping is not sufficient to guarantee the stability of Algorithm 1. We require the composed operator $\text{Proj}_{\mathcal{Q}}\mathcal{B}_{c,\rho}(\cdot)$ to be contraction mapping with appropriate choice of n. This is guaranteed by Condition 2.1 (3). To see this, first note that we have the following lemma, which is obtained by using the contraction property of $\mathcal{B}_{c,\rho}(\cdot)$ and the "equivalence" between norms in finite-dimensional spaces.

Lemma 2.2. *Under Condition 2.1, it holds for any $Q_1,Q_2 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ that $\|\text{Proj}_{\mathcal{Q}}\mathcal{B}_{c,\rho}(Q_1) - \text{Proj}_{\mathcal{Q}}\mathcal{B}_{c,\rho}(Q_2)\|_{\mathcal{K}_{SA}} \le \frac{\tilde{\gamma}(n)}{\sqrt{\mathcal{K}_{SA,\min}}}\|Q_1-Q_2\|_{\mathcal{K}_{SA}}$.*

In view of Lemma 2.2, the composed operator $\text{Proj}_{\mathcal{Q}}\mathcal{B}_{c,\rho}(\cdot)$ is a contraction mapping as long as $\lim_{n\to\infty}\tilde{\gamma}(n)/\sqrt{\mathcal{K}_{SA,\min}} < 1$, which after straightforward algebra is equivalent to Condition 2.1 (3).

To satisfy Condition 2.1 (3), intuitively we should make $D_{\rho,\max}$ and $D_{c,\min}$ arbitrarily close to each other. It is not clear if this is possible for existing off-policy TD-learning algorithms such as Retrace(λ) (Munos et al., 2016), $Q^\pi(\lambda)$ (Harutyunyan et al., 2016), V-trace (Espeholt et al., 2018), and Q-trace (Khodadadian et al., 2021a). That is the reason why none of them were shown to converge in the function approximation setting. In contrast, consider the λ-averaged Q-trace algorithm. Both D_c and D_ρ are identity matrices (which implies $D_{\rho,\max} = D_{c,\min} = 1$), hence Condition 2.1 (3) is always satisfied. Similarly, in the two-sided Q-trace algorithm, for any choice of the upper truncation level $u \ge \mathbf{1}$, we can always choose the lower truncation level $\mathbf{0}\le\ell\le\mathbf{1}$ appropriately to satisfy Condition 2.1 (3). Specifically, for any $s\in\mathcal{S}$ and $u(s)\ge 1$, choosing $\ell(s)\le 1$ such that $\sum_{a\in\mathcal{A}}\pi_b(a|s)g_{\ell(s),u(s)}(\pi(a|s)/\pi_b(a|s)) = 1$ satisfies Condition 2.1 (3). Therefore, compared to V-trace, Retrace(λ), and Q-trace, where the importance sampling ratios were only truncated above, the primary reason for introducing the lower truncation level is to satisfy Condition 2.1 (3), thereby ensuring convergence of the resulting two-sided Q-trace algorithm.

In the next lemma, we show that under Condition 2.1, with properly chosen n, the composed operator $\text{Proj}_{\mathcal{Q}}\mathcal{B}_{c,\rho}(\cdot)$ is a contraction mapping, which ensures that Eq. (2) has a unique solution, denoted by $w_{c,\rho}^\pi$. Moreover, we provide performance guarantees on the solution $w_{c,\rho}^\pi$ in terms of an upper bound on the difference between Q^π and $\Phi w_{c,\rho}^\pi$. Let $Q_{c,\rho}^\pi$ be the solution of generalized Bellman equation $Q = \mathcal{B}_{c,\rho}(Q)$, which is guaranteed to exist and is unique since $\mathcal{B}_{c,\rho}(\cdot)$ itself is a contraction mapping under Condition 2.1 (1) and (2) (Chen et al., 2021c).

Lemma 2.3. *Under Condition 2.1, suppose that the parameter n is chosen such that $\gamma_c := \tilde{\gamma}(n)/\sqrt{\mathcal{K}_{SA,\min}} < 1$. Then the composed operator $Proj_{\mathcal{Q}}\mathcal{B}_{c,\rho}(\cdot)$ is a γ_c-contraction mapping with respect to $\|\cdot\|_{\mathcal{K}_{SA}}$. In this case, the unique solution $w_{c,\rho}^\pi$ of the generalized PBE (cf. Eq. (2)) satisfies*

$$\|Q^\pi - \Phi w_{c,\rho}^\pi\|_{\mathcal{K}_{SA}} \leq \frac{1}{\sqrt{1-\gamma_c^2}}\|Q_{c,\rho}^\pi - Proj_{\mathcal{Q}} Q_{c,\rho}^\pi\|_{\mathcal{K}_{SA}} + \frac{\gamma \max_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} |\pi(a|s) - \pi_b(a|s)\rho(s,a)|}{(1-\gamma)(1-\gamma D_{\rho,\max})}. \quad (3)$$

The first term on the RHS of Eq. (3) captures the error due to function approximation, which is in the same spirit to Theorem 1 (4) of the seminal paper Tsitsiklis and Van Roy (1997), and vanishes in the tabular setting. The second term on the RHS of Eq. (3) arises because of the use of generalized importance sampling ratios, which is introduced to overcome the high variance in off-policy learning. Note that the second term vanishes when $\rho(s,a) = \pi(a|s)/\pi_b(a|s)$ for all (s,a), which corresponds to choosing $\lambda = \mathbf{1}$ in λ-averaged Q-trace and choosing $\ell(s) \leq \min_{s,a} \pi(a|s)/\pi_b(a|s)$ and $u(s) \geq \max_{s,a} \pi(a|s)/\pi_b(a|s)$ for all s in two-sided Q-trace. However, in these cases, the cumulative product of importance sampling ratios leads to a high variance in Algorithm 1. The trade-off between the variance and the bias in $w_{c,\rho}^\pi$ (i.e., second term on the RHS of Eq. (3)) will be elaborated in detail in the next subsection.

2.3 Finite-Sample Analysis

With the contraction property of the generalized PBE established, the almost sure convergence of Algorithm 1 under mild conditions directly follows from standard stochastic approximation results in the literature (Bertsekas and Tsitsiklis, 1996; Borkar, 2009). In this section, we take a step further and perform finite-sample analysis of Algorithm 1. For ease of exposition, we here only present the finite-sample bounds of λ-averaged Q-trace and two-sided Q-trace, where $c(\cdot,\cdot)$ and $\rho(\cdot,\cdot)$ are explicitly specified.

For any $\delta > 0$, let $t_\delta = \min\{k \geq 0 : \max_{s\in\mathcal{S}} \|P^k(s,\cdot) - \mu(\cdot)\|_{\text{TV}} \leq \delta\}$ be the mixing time of the Markov chain $\{S_k\}$ under π_b with precision δ. Note that Assumption 2.2 implies that $t_\delta = \mathcal{O}(\log(1/\delta))$. Let $\lambda_{\min}$ be the mininum eigenvalue of the positive definite matrix $\Phi^\top \mathcal{K}_{SA}\Phi$. Let $L = 1 + (\gamma\rho_{\max})^n$, where $\rho_{\max} = \max_{s,a} \rho(s,a)$.

We next present finite-sample guarantees of the λ-averaged Q-trace algorithm when using constant stepsize (i.e., $\alpha_k \equiv \alpha$). The results for using diminishing stepsizes are trivial extensions (Chen et al., 2019).

Theorem 2.1. *Consider $\{w_k\}$ of the λ-averaged Q-trace Algorithm. Suppose that (1) Assumptions 2.1 and 2.2 are satisfied, (2) $\lambda \in [\mathbf{0}, \mathbf{1}]$, (3) the parameter n is chosen such that $\gamma_c := \gamma^n/\sqrt{\mathcal{K}_{SA,\min}} < 1$, and (4) the stepsize α is chosen such that $\alpha(t_\alpha + n + 1) \leq \frac{(1-\gamma_c)\lambda_{\min}}{130L^2}$. Then, we have for all $k \geq t_\alpha + n + 1$ that*

$$\mathbb{E}[\|w_k - w_{c,\rho}^\pi\|_2^2] \leq c_1(1-(1-\gamma_c)\lambda_{\min}\alpha)^{k-(t_\alpha+n+1)} + c_2\frac{\alpha L^2(t_\alpha+n+1)}{(1-\gamma_c)\lambda_{\min}}, \quad (4)$$

where $c_1 = (\|w_0\|_2 + \|w_0 - w_{c,\rho}^\pi\|_2 + 1)^2$ and $c_2 = 130(\|w_{c,\rho}^\pi\|_2 + 1)^2$. Moreover, we have

$$\|Q^\pi - \Phi w_{c,\rho}^\pi\|_{\mathcal{K}_{SA}} \leq \frac{1}{\sqrt{1-\gamma_c^2}}\|Q_{c,\rho}^\pi - Proj_{\mathcal{Q}} Q_{c,\rho}^\pi\|_{\mathcal{K}_{SA}} + \frac{\gamma \max_{s\in\mathcal{S}}(1-\lambda(s))\|\pi(\cdot|s) - \pi_b(\cdot|s)\|_1}{(1-\gamma)^2}. \quad (5)$$

Using the common terminology in stochastic approximation literature, we call the first term on the RHS of Eq. (4) *convergence bias*, and the second term *variance*. When constant stepsize is used, the convergence bias goes to zero at a geometric rate while the variance is a constant roughly proportional to αt_α. Since $\lim_{\alpha\to 0} \alpha t_\alpha = 0$ under Assumption 2.2, the variance can be made arbitrarily small by using small α.

The parameter $L = 1 + (\gamma\rho_{\max})^n$ plays an important role in the finite-sample bound. In fact, L appears quadratically in the variance term of Eq. (4), and captures the impact of the cumulative product of the importance sampling ratios. To overcome the high variance in off-policy learning (i.e., to make sure that the parameter $L = 1 + (\gamma\rho_{\max})^n$ does not grow exponentially fast with respect to n), we choose $\lambda \in \mathbb{R}^{|\mathcal{S}|}$ such that $\rho_{\max} = \max_s \lambda(s)(\max_a \pi(a|s)/\pi_b(a|s) - 1) + 1 \leq 1/\gamma$. However, as long as $\lambda \neq \mathbf{1}$, the limit point of the λ-averaged Q-trace algorithm involves an additional bias term (i.e., the second term on the RHS of Eq. (5)) that does not vanish even in the tabular setting.

In light of the discussion above, it is clear that there is a trade-off between the variance (cf. second term on the RHS of Eq. (4)) and the bias in the limit point (cf. the second term on the RHS of Eq. (3)) in choosing the parameter λ. Specifically, large λ leads to large $\rho_{\max}$ and hence large L and large variance, but in this case the second term on the RHS of Eq. (3) is smaller, implying that we have a smaller bias in the limit point.

Next, we present the finite-sample bounds of the two-sided Q-trace algorithm.

Theorem 2.2. *Consider $\{w_k\}$ of the two-sided Q-trace Algorithm. Suppose that (1) Assumptions 2.1 and 2.2 are satisfied, (2) the upper and lower truncation levels $\ell, u \in \mathbb{R}^{|\mathcal{S}|}$ are chosen such that $\sum_{a\in\mathcal{A}} \pi_b(a|s) g_{\ell(s),u(s)}(\pi(a|s)/\pi_b(a|s)) = 1$ for all s,* *(3) the parameter n is chosen such that $\gamma_c := \gamma^n/\sqrt{\mathcal{K}_{SA,\min}} < 1$, and (4) the stepsize α is chosen such that $\alpha(t_\alpha + n + 1) \leq \frac{(1-\gamma_c)\lambda_{\min}}{130L^2}$. Then, we have for all $k \geq t_\alpha + n + 1$ that*

$$\mathbb{E}[\|w_k - w_{c,\rho}^\pi\|_2^2] \leq c_1(1-(1-\gamma_c)\lambda_{\min}\alpha)^{k-(t_\alpha+n+1)} + c_2\frac{\alpha L^2(t_\alpha+n+1)}{(1-\gamma_c)\lambda_{\min}}, \quad (6)$$

where $c_1 = (\|w_0\|_2 + \|w_0 - w_{c,\rho}^\pi\|_2 + 1)^2$ and $c_2 = 130(\|w_{c,\rho}^\pi\|_2+1)^2$. Moreover, we have

$$\|Q^\pi - \Phi w_{c,\rho}^\pi\|_{\mathcal{K}_{SA}} \leq \frac{1}{\sqrt{1-\gamma_c^2}}\|Q_{c,\rho}^\pi - \Phi w_{c,\rho}^\pi\|_{\mathcal{K}_{SA}} + \frac{\gamma \max_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}(u_{\pi,\pi_b}(s,a) - \ell_{\pi,\pi_b}(s,a))}{(1-\gamma)^2}, \quad (7)$$

where $u_{\pi,\pi_b}(s,a) = \max(\pi(a|s) - \pi_b(a|s)u(s), 0)$ and $\ell_{\pi,\pi_b}(s,a) = \min(\pi(a|s) - \pi_b(a|s)\ell(s), 0)$ for all (s,a).

The finite-sample bound of the two-sided Q-trace algorithm is qualitatively similar to that of the λ-averaged Q-trace algorithm. To overcome the high variance issue in off-policy learning, we choose the upper truncation level such that $\gamma u(s) \leq 1$ for all s, which ensures that the parameter $L = 1 + (\gamma\rho_{\max})^n \leq 1 + (\gamma \max_s u(s))^n$ does not grow exponentially with respect to n. Then we choose the lower truncation level accordingly to satisfy requirement (2) stated in Theorem 2.2. However, as long as there exists $s \in \mathcal{S}$ such that $u(s) < \max_{s,a}\pi(a|s)/\pi_b(a|s)$ or $\ell(s) > \min_{s,a}\pi(a|s)/\pi_b(a|s)$, the second term on the RHS of Eq. (7) is in general non-zero, hence adding an additional bias term to the limit point even in the tabular setting. As a result, the trade-off between the variance and the bias in the limit point is also present in the two-sided Q-trace algorithm.

In view of Theorems 2.1 and 2.2, one limitation of this work is that the choice of n to make $\gamma_c < 1$ depends on the unknown parameter $\mathcal{K}_{SA,\min}$ of the problem. In practice, one can start with a specific choice of n and then gradually tune n to achieve the convergence of the λ-averaged Q-trace algorithm or the two-sided Q-trace algorithm.

3 POLICY-BASED METHODS

In this section we study various policy-based algorithms and establish their finite-sample convergence guarantees. The policy evaluation sub-problem is solved with Algorithm 1.

3.1 Policy Update Rules

We begin by presenting a generic policy-based algorithm in the following. For simplicity of notation, for a given target policy π, behavior policy π_b, constant stepsize α, initialization w_0, and samples $\{(S_k, A_k)\}_{0\leq k\leq K+n-1}$, we denote the output of Algorithm 1 after K iterations by $w = \text{ALG}(w_0, \pi, \pi_b, \alpha, K, \{(S_k, A_k)\}_{0\leq k\leq K+n-1})$.

Algorithm 2 A Generic Policy-Based Algorithm

1: **Input:** Integers T, K, initial policy π_0, sample trajectory $\{(S_t, A_t)\}_{0\leq t\leq T(K+n)}$ collected under the behavior policy π_b.
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: dataset $= \{(S_k, A_k)\}_{t(K+n)\leq k\leq (t+1)(K+n)-1}$
4: $w_t = \text{ALG}(\mathbf{0}, \pi_t, \pi_b, \alpha, K, \text{dataset})$
5: $\pi_{t+1} = G(\Phi w_t, \pi_t)$
6: **end for**
7: **Output:** π_T

Although Algorithm 2 is presented with a fixed behavior policy π_b, our results can be easily generalized to the case where the behavior policy is updated across t. The only requirement on the behavior policy is that it should enable the agent to sufficiently explore the state-action space. In Algorithm 2 line 5, the function $G(\cdot,\cdot)$ represents the policy update rule, which takes the current policy iterate π_t and the Q-function estimate Φw_t as inputs. Many existing policy update rules fit into this framework, as elaborated below.

$1/\beta_1$-Greedy Update. Let $\beta_1 \in [1,\infty]$ be a tunable parameter. For any $t \geq 0$ and state-action pair (s,a), we update the policy by $\pi_{t+1}(a|s) = \frac{1}{\beta_1|\mathcal{A}|}$ when $a \neq \arg\max_{a'\in\mathcal{A}}\phi(s,a')^\top w_t$, and $\pi_{t+1}(a|s) = \frac{1}{\beta_1|\mathcal{A}|} + 1 - \frac{1}{\beta_1}$ when $a = \arg\max_{a'\in\mathcal{A}}\phi(s,a')^\top w_t$. In this work, whenever the arg max is not unique, we break tie arbitrarily. More generally, we allow the tunable parameter β_1 to be time-dependent (i.e., β_1 is a function of the iteration index t) and/or state-dependent (i.e., β_1 is a function of the state s).

β_2-Softmax Update. Let $\beta_2 > 0$ be a tunable parameter, which is allowed to be time varying and state-dependent. Then the policy is updated by

$$\pi_{t+1}(a|s) = \frac{\exp(\beta_2\phi(s,a)^\top w_t)}{\sum_{a'\in\mathcal{A}}\exp(\beta_2\phi(s,a')^\top w_t)}, \ \forall\ (s,a).$$

In $1/\beta_1$-greedy update or β_2-softmax update, there is no need to parametrize the policy because it is uniquely determined by the estimate of the Q-function, which already uses linear function approximation.

At a first glance of Algorithm 2 line 5, it seems that we need to work with $|\mathcal{S}||\mathcal{A}|$-dimensional objects to update the policy at each state-action pair, which contradicts to the motivation of using function approximation. However, there is an equivalent way of implementing Algorithm 2 without explicitly executing line 5. To see this, first note that the target policy π_t in each iteration is only used in the policy evaluation step (Algorithm 2 line 4). To view of our policy evaluation algorithm (cf. Algorithm 1), we only need to compute the policy value of π_t at state-action pairs that are visited by the sample trajectory $\{(S_k, A_k)\}$.

When using $1/\beta_1$-greedy update or β_2-softmax update, Algorithm 2 subsumes the popular value-based method SARSA (Bertsekas and Tsitsiklis, 1996) as its special case. To see this, suppose that we are in the on-policy setting (i.e., $\pi = \pi_b$), and the inner-loop iteration number K is set to 1. Then Algorithm 2 corresponds to SARSA with $1/\beta_1$-greedy exploration policy or Boltzmann exploration policy. However, we need to point out that our result does NOT imply finite-sample bounds for SARSA since we need a relatively large K to provide a sufficiently accurate estimate of the value function before using it in policy improvement.

β_3-NPG Update. Unlike $1/\beta_1$-greedy update or β_2-softmax update, where we need only the estimate of the Q-function to perform to update, in NPG, to update the policy, we need both the current policy and the estimate of its Q-function. Therefore, to keep track of the policy, in this case we also need to parametrize the policy using softmax parametrization and compatible linear function approximation. Specifically, with parameter $\theta \in \mathbb{R}^d$, the policy π associated with parameter θ is given by $\pi_\theta(a|s) = \frac{\exp(\phi(s,a)^\top \theta)}{\sum_{a' \in \mathcal{A}} \exp(\phi(s,a')^\top \theta)}$.

Let $\beta_3 > 0$ be a tunable parameter, which is allowed to be time varying. Then NPG updates the parameter θ_t of the policy according to the formula

$$\theta_{t+1} = \theta_t + \beta_3 w_t. \tag{8}$$

See Agarwal et al. (2021) for more details about this update rule. Denote π_t as π_{θ_t} for simplicity of notation. Then the update equation can be equivalently written in terms of the policy update (and also in the form of Algorithm 2 line 5) as

$$\pi_{t+1}(a|s) = \frac{\pi_t(a|s)\exp(\beta_3 \phi(s,a)^\top w_t)}{\sum_{a' \in \mathcal{A}} \pi_t(a'|s)\exp(\beta_3 \phi(s,a')^\top w_t)}, \ \forall (s,a).$$

This enables us to use the previous equation for our analysis of NPG while using Eq. (8) for the implementation of Algorithm 2.

3.2 Finite-Sample Analysis

In this section, we present the finite-sample guarantees of Algorithm 2. For ease of exposition, we implement line 4 of Algorithm 2 with the λ-averaged Q-trace algorithm. The results for using either two-sided Q-trace algorithm or Algorithm 1 with more general choices of $c(\cdot,\cdot)$ and $\rho(\cdot,\cdot)$ (as long as Condition 2.1 is satisfied) are straightforward extensions. As for the policy improvement (cf. line 5 of Algorithm 2), we use either $1/\beta_1$-greedy policy update, or β_2-softmax policy update, or β_3-NPG policy update, with the corresponding parameters satisfying the following condition. Denote $a_{t,s} = \arg\max_{a' \in \mathcal{A}} \phi(s,a')^\top w_t$.

Condition 3.1. Let $\beta > 0$ be a tunable parameter. (1) The parameter β_1 is time-varying and state-dependent, and is chosen such that $\beta_1(t,s) \geq \frac{2\gamma}{\beta} \max_{a \in \mathcal{A}} |\phi(s,a)^\top w_t|$ for all s and t. (2) The parameter β_2 is chosen such that $\beta_2 \geq \frac{\gamma}{\beta}\log(|\mathcal{A}|)$. (3) The parameter β_3 is time-varying, and is chosen such that $\beta_3(t) \geq \frac{\gamma}{\beta}\log(1/\min_{s \in \mathcal{S}} \pi_t(a_{t,s}|s))$ for all t.

Theorem 3.1. *Consider π_t of Algorithm 2. Suppose that the assumptions for applying Theorem 2.1 are satisfied, and the choices of β_1, β_2, and β_3 satisfy Condition 3.1. Then we have for any $T \geq 0$:*

$$\begin{aligned}
&\mathbb{E}[\|Q^* - Q^{\pi_T}\|_\infty] \\
\leq\ & \underbrace{\frac{2\gamma\mathcal{E}_{approx}}{(1-\gamma)^2}}_{N_1} + \underbrace{\frac{2\gamma^2\mathcal{E}_{bias}}{(1-\gamma)^4}}_{N_2} + \underbrace{\gamma^T\|Q^* - Q^{\pi_0}\|_\infty}_{N_3:\ \text{Convergence bias in the actor}} \\
&+ \underbrace{6\tilde{c}(1-(1-\gamma_c)\lambda_{\min}\alpha)^{\frac{1}{2}[K-(t_\alpha+n+1)]}}_{N_4:\ \text{Convergence bias in the critic}} \\
&+ \underbrace{70L\tilde{c}\frac{[\alpha(t_\alpha+n+1)]^{1/2}}{\sqrt{1-\gamma_c}\sqrt{\lambda_{\min}}}}_{N_5:\ \text{Critic variance}} + \underbrace{\frac{2\gamma\beta}{(1-\gamma)^2}}_{N_6},
\end{aligned} \tag{9}$$

where $\tilde{c} = \frac{\gamma}{\sqrt{\lambda_{\min}}\sqrt{1-\gamma_c}(1-\gamma)^3}$, $\mathcal{E}_{approx} = \sup_\pi \|Q_{c,\rho}^\pi - \Phi w_{c,\rho}^\pi\|_\infty$ and $\mathcal{E}_{bias} = \max_{0 \leq t \leq T} \max_{s \in \mathcal{S}} (1 - \lambda(s))\|\pi_t(\cdot|s) - \pi_b(\cdot|s)\|_1$.

Notably on the LHS, our finite-sample guarantees are stated for the last policy iterate π_T, while in many existing literature it was stated for the best policy among $\{\pi_t\}_{0 \leq t \leq T}$ (Agarwal et al., 2021).

The Terms N_1 and N_2. The term N_1 represents the function approximation bias, and is present in all existing literature that study policy-based methods under function approximation (Agarwal et al., 2021). Note that $N_1 = 0$ when we use a complete basis. The term N_2 represents the bias introduced to the algorithm by using generalized importance sampling ratios $c(\cdot,\cdot)$ and $\rho(\cdot,\cdot)$. Note that we have $N_2 = 0$ when $c(s,a) = \rho(s,a) = \pi(a|s)/\pi_b(a|s)$, which corresponds to using $\lambda = \mathbf{1}$ in the λ-averaged Q-trace algorithm, and using $u(s) \geq \max_{s,a} \pi(a|s)/\pi_b(a|s)$ and $\ell(s) \leq \min_{s,a} \pi(a|s)/\pi_b(a|s)$ for all s in the two-sided Q-trace algorithm. However, this choice of λ (or u and ℓ) might lead to a high variance. In particular, the parameter L within the term N_5 could be large.

The terms N_3 and N_4. The term N_3 represents the convergence bias in the actor, and goes to zero geometrically fast as the outer loop iteration number T goes to infinity. Such geometric convergence is the main reason why we obtain improved sample complexity of β_3-NPG compared to Chen et al. (2021a), where the convergence rate of the actor is $\mathcal{O}(1/T)$. The term N_4 represents the convergence bias in the critic, and goes to zero geometrically fast as the inner loop iteration number K goes to infinity.

The terms N_5 and N_6. The term N_5 represents the variance in the critic, and is proportional to $\sqrt{\alpha t_\alpha} = \mathcal{O}(\sqrt{\alpha \log(1/\alpha)})$. Therefore, N_5 can be made arbitrarily small by using small enough stepsize α. The term N_6 captures the error introduced to the algorithm by the policy update rule $G(\cdot,\cdot)$. To elaborate, consider the following example. Suppose that the underlying MDP model has a unique optimal policy, and suppose we use $1/\beta_1$-greedy update (with a fixed β_1) in Algorithm 2 line 5. Then as long as β_1 is finite, we can never truly find the optimal policy π^* because of the deterministic nature of π^*. As a result, the difference between Q^* and Q^{π_t} will always be above some threshold, which depends on the choice of β_1, and is captured by N_6. Observe that N_6 can be made arbitrarily small by using small enough β.

Based on Theorem 3.1, we next derive the sample complexity of Algorithm 2. To enable fair comparison with existing literature, we choose $\lambda = \mathbf{1}$ to eliminate the error due to using generalized importance sampling ratios. Note that $\lambda = \mathbf{1}$ implies $\mathcal{E}_{\text{bias}} = 0$ (and hence $N_2 = 0$) in Theorem 3.1.

Corollary 3.1.1. *For a given accuracy level $\epsilon > 0$, to achieve $\mathbb{E}[\|Q^* - Q^{\pi_T}\|_\infty] \leq \epsilon + N_1$, the number of samples (e.g. the integer TK) required is of the size* $\mathcal{O}\left(\frac{\log^3(1/\epsilon)}{\epsilon^2}\right)\tilde{\mathcal{O}}\left(\frac{L^2 n}{(1-\gamma)^7(1-\gamma_c)^3\lambda_{\min}^3}\right)$.

Notably, we obtain $\tilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity for policy-based methods, which matches with the sample complexity of value-based algorithms such as Q-learning (Li et al., 2020). In the case of β_3-NPG update, to our knowledge, Cayci et al. (2021); Lan (2021) establishes the $\tilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity of on-policy NAC under regularization, and Chen et al. (2021a) establishes the $\tilde{\mathcal{O}}(\epsilon^{-3})$ sample complexity of a variant of off-policy NAC (where the infamous deadly triad is present). We improve the sample complexity in Chen et al. (2021a) by a factor of ϵ^{-1}, and we do not use regularization.

In addition to the dependence on ϵ, the dependence on $1/(1-\gamma)$ (which is usually called the effective horizon) is also improved by a factor of $1/(1-\gamma)$ compared to existing work (Agarwal et al., 2021; Chen et al., 2021a). The bootstrapping parameter n appears linearly in our sample complexity bound. This matches with the results for n-step TD-learning in the on-policy tabular setting (Chen et al., 2021b).

4 OUTLINE OF THE PROOF

In this section, we provide the proof sketch of Theorems 2.1 and 2.2, and Theorem 3.1.

4.1 Policy Evaluation

Instead of proving Theorems 2.1 and 2.2, we will state and prove finite-sample bounds for Algorithm 1 with $c(\cdot,\cdot)$ and $\rho(\cdot,\cdot)$ satisfying Condition 2.1, which subsumes Theorems 2.1 and 2.2 as its special cases. In this more general setup where we do not have $c(\cdot,\cdot) = \rho(\cdot,\cdot)$, we define the constant parameter L as

$$L = \begin{cases} (1 + (\gamma\rho_{\max})^n), & c(\cdot,\cdot) = \rho(\cdot,\cdot), \\ (1+\gamma\rho_{\max}) f_n(\gamma c_{\max}), & c(\cdot,\cdot) \neq \rho(\cdot,\cdot), \end{cases} \tag{10}$$

where $c_{\max} = \max_{s,a} c(s,a)$ and $\rho_{\max} = \max_{s,a}\rho(s,a)$.

Theorem 4.1. *Consider $\{w_k\}$ of Algorithm 1. Suppose that (1) Assumptions 2.1 and 2.2 are satisfied, (2) the generalized importance sampling ratios satisfy Condition 2.1, (3) the parameter n is chosen such that $\gamma_c := \tilde{\gamma}(n)/\sqrt{\mathcal{K}_{SA,\min}} < 1$, and (4) the constant stepsize α is chosen such that $\alpha(t_\alpha + n + 1) \leq \frac{(1-\gamma_c)\lambda_{\min}}{130L^2}$. Then, we have for all $k \geq t_\alpha + n + 1$:*

$$\begin{aligned}\mathbb{E}[\|w_k - w_{c,\rho}^\pi\|_2^2] \leq{}& c_1(1-(1-\gamma_c)\lambda_{\min}\alpha)^{k-(t_\alpha+n+1)} \\ &+ c_2 L^2\frac{\alpha(t_\alpha+n+1)}{(1-\gamma_c)\lambda_{\min}},\end{aligned} \tag{11}$$

where $c_1 = (\|w_0\|_2 + \|w_0 - w_{c,\rho}^\pi\|_2 + 1)^2$ and $c_2 = 130(\|w_{c,\rho}^\pi\|_2 + 1)^2$.

To prove Theorem 4.1, we first rewrite Algorithm 1 as a stochastic approximation algorithm. Let $\{X_k\}$ be a finite-state Markov chain defined by $X_k = (S_k, A_k, ..., S_{k+n}, A_{k+n})$ for any $k \geq 0$. Denote the state-space of $\{X_k\}$ by $\mathcal{X}$. It is clear that under Assumption 2.2, the Markov chain $\{X_k\}$ also admits a unique stationary distribution, which we denote by $\nu \in \Delta^{|\mathcal{X}|}$. Let $F : \mathbb{R}^d \times \mathcal{X} \mapsto \mathbb{R}^d$ be an operator defined by $F(w,x) = \phi(s_0,a_0)\sum_{i=0}^{n-1}\gamma^i c_{1,i}\Delta_i(w)$ for any $w \in \mathbb{R}^d$ and $x = (s_0, a_0, ..., s_n, a_n) \in \mathcal{X}$. Let $\bar{F} : \mathbb{R}^d \mapsto \mathbb{R}^d$ be the "expected" operator of $F(\cdot,\cdot)$ defined by $\bar{F}(w) = \mathbb{E}_{X\sim\nu}[F(w,X)]$. Using the notation above, the update equation (line 4) of Algorithm 1 can be compactly written as

$$w_{k+1} = w_k + \alpha_k F(w_k, X_k), \tag{12}$$

which is a stochastic approximation algorithm for solving the equation $\bar{F}(w) = 0$ with Markovian noise. Note that $\bar{F}(w) = 0$ is equivalent to the generalized PBE (2) (cf. Lemma 2.1). We next establish the properties of the operators $F(\cdot,\cdot)$, $\bar{F}(\cdot)$, and the Markov chain $\{X_k\}$

in the following proposition, which enables us to use standard stochastic approximation results in the literature to derive finite-sample bounds of Algorithm 1.

Proposition 4.1. *The following statements hold: (1)* $\|F(w_1,x)-F(w_2,x)\|_2 \leq L\|w_1-w_2\|_2$ *for any* $w_1,w_2\in\mathbb{R}^d$ *and* $x\in\mathcal{X}$, *and* $\|F(\mathbf{0},x)\|_2\leq f_n(\gamma c_{\max})$ *for any* $x\in\mathcal{X}$, *(2)* $\max_{x\in\mathcal{X}}\left\|P_X^{k+n+1}(x,\cdot)-\nu(\cdot)\right\|_{TV}\leq C\sigma^k$ *for all* $k\geq 0$, *where* P_X *is the transition probability matrix of the Markov chain* $\{X_k\}$, *and (3)* $(w-w_{c,\rho}^\pi)^\top\bar{F}(w)\leq -(1-\gamma_c)\lambda_{\min}\|w-w_{c,\rho}^\pi\|_2^2$ *for any* $w\in\mathbb{R}^d$.

Proposition 4.1 (1) establishes the Lipschitz continuity of the operator $F(\cdot,\cdot)$, Proposition 4.1 (2) establishes the geometric mixing of the auxiliary Markov chain $\{X_k\}$, and Proposition 4.1 (3) essentially guarantees that the ODE $\dot{x}(t)=\bar{F}(x(t))$ associated with stochastic approximation algorithm (12) is globally geometrically stable. The rest of the proof follows by applying Theorem 2.1 of Chen et al. (2019) to Algorithm 1, and is presented in detail in the Appendix.

4.2 Policy Improvement

We first introduce some notation. Let $\mathcal{H}:\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}\mapsto\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be the Bellman optimality operator defined by $[\mathcal{H}(Q)](s,a)=\mathcal{R}(s,a)+\gamma\mathbb{E}[\max_{a'\in\mathcal{A}}Q(S_{k+1},a')\mid S_k=s,A_k=a]$ for all (s,a), and let $\mathcal{H}_\pi:\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}\mapsto\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be the Bellman operator associated with policy π defined by $[\mathcal{H}_\pi(Q)](s,a)=\mathcal{R}(s,a)+\gamma\mathbb{E}_\pi[Q(S_{k+1},A_{k+1})\mid S_k=s,A_k=a]$ for all (s,a).

The key to prove Theorem 3.1 is the following proposition.

Proposition 4.2. *Consider* $\{\pi_T\}$ *of Algorithm 2. The following inequality holds for any* $T\geq 0$:

$$\begin{aligned}\mathbb{E}[\|Q^*-Q^{\pi_T}\|_\infty] \;\leq\; & \gamma^T\|Q^*-Q^{\pi_0}\|_\infty && A_1\\ &+\frac{2\gamma}{1-\gamma}\sum_{i=0}^{T-1}\gamma^{T-1-i}\mathbb{E}[\|Q^{\pi_i}-\Phi w_i\|_\infty] && A_2\\ &+\frac{2\gamma}{1-\gamma}\sum_{i=0}^{T-1}\gamma^{T-1-i}\mathbb{E}[\|\mathcal{H}_{\pi_{i+1}}(\Phi w_i)-\mathcal{H}(\Phi w_i)\|_\infty]. && A_3\end{aligned}$$

In light of Proposition 4.2, to proceed and establish finite-sample bound of Algorithm 2, it remains to control the terms A_2 and A_3 when the policy evaluation algorithm and the policy update rule are specified. Specifically, we control A_2 by using Theorem 2.1, and control A_3 by using Condition 3.1 on the parameters β_1, β_2, and β_3 for various policy update rules. See Appendix B.1 for more details.

Before we present the key steps to prove Proposition 4.2, consider a special case of tabular RL, and choosing $c(s,a)=\rho(s,a)=\frac{\pi(a|s)}{\pi_b(a|s)}$ in Algorithm 1. Note that the term A_2 vanishes. Since the term A_3 can be made arbitrarily small by using large enough β, Proposition 4.2 implies *geometric* convergence for NPG. The geometric convergence of NPG was previously established in Cayci et al. (2021); Lan (2021) under regularization, and in Khodadadian et al. (2021c) in the asymptotic region. We do not require regularization to establish the result, and our result holds for all $T\geq 0$.

Next we present the proof sketch of Proposition 4.2. In most of the existing literature, for policy-based type of algorithms, the analysis is usually based on the mirror descent analysis in optimization (Lan, 2020), where the $\mathcal{KL}$-divergence was chosen as a potential/Lyapunov function, and the performance difference lemma was extensively used (Agarwal et al., 2021; Cayci et al., 2021). To establish Proposition 4.2, we use a completely different approach, where we only exploit the contraction and the monotonicity of the Bellman operators $\mathcal{H}_\pi(\cdot)$ and $\mathcal{H}(\cdot)$. Such proof technique was inspired by Bertsekas and Tsitsiklis (1996) Section 6.2. However, only asymptotic error bound of approximate policy iteration was established in Bertsekas and Tsitsiklis (1996), while we establish finite-sample bounds for various policy update rules. Proposition 4.2 builds on the following two lemmas.

Lemma 4.1. *It holds for all* $t\geq 0$ *that*

$$\begin{aligned}&\max_{s,a}(Q^{\pi_t}(s,a)-Q^{\pi_{t+1}}(s,a))\\ &\leq\frac{2\gamma\|Q^{\pi_t}-\Phi w_t\|_\infty+\|\mathcal{H}_{\pi_{t+1}}(\Phi w_t)-\mathcal{H}(\Phi w_t)\|_\infty}{1-\gamma}.\end{aligned}$$

Lemma 4.2. *It holds for all* $t\geq 0$ *that*

$$\begin{aligned}&\|Q^*-Q^{\pi_{t+1}}\|_\infty\leq\gamma\|Q^*-Q^{\pi_t}\|_\infty\\ &+\frac{2\gamma\|Q^{\pi_t}-\Phi w_t\|_\infty+\|\mathcal{H}_{\pi_{t+1}}(\Phi w_t)-\mathcal{H}(\Phi w_t)\|_\infty}{1-\gamma}.\end{aligned}$$

Proposition 4.2 then follows by repeatedly using Lemma 4.2 and then taking expectation on both sides of the resulting inequality.

5 CONCLUSION

In this work, we study finite-sample guarantees of general policy-based algorithms under off-policy sampling and linear function approximation. To overcome the deadly triad and the high variance in policy evaluation, we design a convergent framework of TD-learning algorithms, including two specific algorithms called λ-averaged Q-trace and two-sided Q-trace. The resulting overall sample complexity bound is $\tilde{\mathcal{O}}(\epsilon^{-2})$, which matches with typical value-based algorithms such as Q-learning. In the case of natural actor-critic with function approximation, this advances the existing state-of-the-art result by a factor of ϵ^{-1}.

Acknowledgements

This work was partially supported by NSF Grant CMMI- 2112533, NSF Grant CCF-1740776, and an award from Raytheon Technologies. We would like to thank anonymous reviewers from AISTATS for their detailed and insightful comments.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2021). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76.

Bertsekas, D. P. (2011). Approximate policy iteration: A survey and some new methods. *Journal of Control Theory and Applications*, 9(3):310–335.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Athena Scientific.

Bertsekas, D. P. and Yu, H. (2009). Projected equation methods for approximate solution of large linear systems. *Journal of Computational and Applied Mathematics*, 227(1):27–50.

Bhandari, J., Russo, D., and Singal, R. (2018). A Finite Time Analysis of Temporal Difference Learning With Linear Function Approximation. In *Conference On Learning Theory*, pages 1691–1692.

Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., and Lee, M. (2009). Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482.

Borkar, V. S. (2009). *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer.

Borkar, V. S. and Konda, V. R. (1997). The actor-critic algorithm as multi-time-scale stochastic approximation. *Sadhana*, 22(4):525–543.

Cayci, S., He, N., and Srikant, R. (2021). Linear convergence of entropy-regularized natural policy gradient with linear function approximation. *Preprint arXiv:2106.04096*.

Chen, Z., Khodadadian, S., and Maguluri, S. T. (2021a). Finite-Sample Analysis of Off-Policy Natural Actor-Critic with Linear Function Approximation. *Preprint arXiv:2105.12540*.

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. (2021b). A Lyapunov Theory for Finite-Sample Guarantees of Asynchronous Q-Learning and TD-Learning Variants. *Preprint arXiv:2102.01567*.

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. (2021c). Finite-Sample Analysis of Off-Policy TD-Learning via Generalized Bellman Operators. *Advances in Neural Information Processing Systems*, 34.

Chen, Z., Zhang, S., Doan, T. T., Clarke, J.-P., and Maguluri, S. T. (2019). Finite-Sample Analysis of Nonlinear Stochastic Approximation with Applications in Reinforcement Learning. *Preprint arXiv:1905.11425*.

Dayan, P. and Sejnowski, T. J. (1994). TD(λ) converges with probability 1. *Machine Learning*, 14(3):295–301.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., et al. (2018). IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures. In *International Conference on Machine Learning*, pages 1407–1416.

Glavic, M., Fonteneau, R., and Ernst, D. (2017). Reinforcement learning for electric power system decision and control: Past considerations and perspectives. *IFAC-PapersOnLine*, 50(1):6918–6927.

Harutyunyan, A., Bellemare, M. G., Stepleton, T., and Munos, R. (2016). $Q(\lambda)$ with Off-Policy Corrections. In *International Conference on Algorithmic Learning Theory*, pages 305–320. Springer.

Kakade, S. M. (2001). A natural policy gradient. *Advances in neural information processing systems*, 14.

Khodadadian, S., Chen, Z., and Maguluri, S. T. (2021a). Finite-Sample Analysis of Off-Policy Natural Actor-Critic Algorithm. *The 38th International Conference on Machine Learning*, 139:5420–5431.

Khodadadian, S., Doan, T. T., Maguluri, S. T., and Romberg, J. (2021b). Finite Sample Analysis of Two-Time-Scale Natural Actor-Critic Algorithm. *Preprint arXiv:2101.10506*.

Khodadadian, S., Jhunjhunwala, P. R., Varma, S. M., and Maguluri, S. T. (2021c). On the linear convergence of natural policy gradient algorithm. *Preprint arXiv:2105.01424*.

Konda, V. R. and Tsitsiklis, J. N. (2000). Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014. Citeseer.

Kumar, H., Koppel, A., and Ribeiro, A. (2019). On the Sample Complexity of Actor-Critic Method for Reinforcement Learning with Function Approximation. *Preprint arXiv:1910.08412*.

Lan, G. (2020). *First-order and Stochastic Optimization Methods for Machine Learning*. Springer.

Lan, G. (2021). Policy Mirror Descent for Reinforcement Learning: Linear Convergence, New Sampling Complexity, and Generalized Problem Classes. *Mathematical programming.*

Lazaric, A., Ghavamzadeh, M., and Munos, R. (2012). Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research*, 13:3041–3074.

Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*, volume 107. American Mathematical Soc.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2020). Sample Complexity of Asynchronous Q-Learning: Sharper Analysis and Variance Reduction. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7031–7043. Curran Associates, Inc.

Liu, B., Cai, Q., Yang, Z., and Wang, Z. (2019). Neural proximal/trust region policy optimization attains globally optimal policy. *Advances in Neural Information Processing Systems*, 32.

Liu, Y., Zhang, K., Basar, T., and Yin, W. (2020). An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33.

Ma, S., Zhou, Y., and Zou, S. (2020). Variance-reduced off-policy TDC learning: Non-asymptotic convergence analysis. *Advances in Neural Information Processing Systems*, 33:14796–14806.

Munos, R., Stepleton, T., Harutyunyan, A., and Bellemare, M. G. (2016). Safe and efficient off-policy reinforcement learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1054–1062.

Qiu, S., Yang, Z., Ye, J., and Wang, Z. (2019). On the finite-time convergence of actor-critic algorithm. In *Optimization Foundations for Reinforcement Learning Workshop at Advances in Neural Information Processing Systems (NeurIPS)*.

Srikant, R. and Ying, L. (2019). Finite-Time Error Bounds For Linear Stochastic Approximation and TD Learning. In *Conference on Learning Theory*, pages 2803–2830.

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

Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, C., and Wiewiora, E. (2009). Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 993–1000.

Sutton, R. S., Mahmood, A. R., and White, M. (2016). An emphatic approach to the problem of off-policy temporal-difference learning. *The Journal of Machine Learning Research*, 17(1):2603–2631.

Sutton, R. S., Szepesvári, C., and Maei, H. R. (2008). A convergent $\mathcal{O}(n)$ algorithm for off-policy temporal-difference learning with linear function approximation. *Advances in neural information processing systems*, 21(21):1609–1616.

Tsitsiklis, J. N. (1994). Asynchronous stochastic approximation and Q-learning. *Machine learning*, 16(3):185–202.

Tsitsiklis, J. N. and Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE transactions on automatic control*, 42(5):674–690.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. (2019). Neural Policy Gradient Methods: Global Optimality and Rates of Convergence. In *International Conference on Learning Representations*.

Wang, Y., Zou, S., and Zhou, Y. (2021). Finite-Sample Analysis for Two Time-scale Non-linear TDC with General Smooth Function Approximation. *Advances in Neural Information Processing Systems*, 34.

Williams, R. J. and Baird, L. C. (1990). A mathematical analysis of actor-critic architectures for learning optimal controls through incremental dynamic programming. In *Proceedings of the Sixth Yale Workshop on Adaptive and Learning Systems*, pages 96–101. Citeseer.

Wu, Y. F., Zhang, W., Xu, P., and Gu, Q. (2020). A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17617–17628.

Zhang, K., Koppel, A., Zhu, H., and Başar, T. (2019). Convergence and iteration complexity of policy gradient method for infinite-horizon reinforcement learning. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 7415–7422. IEEE.

Zhao, Y., Kosorok, M. R., and Zeng, D. (2009). Reinforcement learning design for cancer clinical trials. *Statistics in medicine*, 28(26):3294–3315.

Supplementary Material: Sample Complexity of Policy-Based Methods under Off-Policy Sampling and Linear Function Approximation

A Proof of All Technical Results in Section 2

A.1 Proof of Lemma 2.1

We begin by introducing some notation. Let π_c and π_ρ be two policies defined by

$$\pi_c(a|s) = \frac{\pi_b(a|s)c(s,a)}{\sum_{a'\in\mathcal{A}}\pi_b(a'|s)c(s,a')}, \quad \text{and} \quad \pi_\rho(a|s) = \frac{\pi_b(a|s)\rho(s,a)}{\sum_{a'\in\mathcal{A}}\pi_b(a'|s)\rho(s,a')}, \; \forall\,(s,a).$$

Let P_{π_c} and P_{π_ρ} be the transition probability matrices of the Markov chain $\{S_k\}$ induced by the policies π_c and π_ρ, respectively. Then, Eq. (1) can be compactly written in vector form as

$$\Phi^\top \mathcal{K}_{SA}\sum_{i=0}^{n-1}(\gamma P_{\pi_c}D_c)^i(R+\gamma P_{\pi_\rho}D_\rho\Phi w-\Phi w)=0,$$

where $R\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is defined by $R(s,a)=\mathcal{R}(s,a)$ for all (s,a). Observe that the above equation is further equivalent to

$$\Phi(\Phi^\top\mathcal{K}_{SA}\Phi)^{-1}\Phi^\top \mathcal{K}_{SA}\sum_{i=0}^{n-1}(\gamma P_{\pi_c}D_c)^i(R+\gamma P_{\pi_\rho}D_\rho\Phi w-\Phi w)=0. \tag{13}$$

To see this, since the matrix Φ has full column-rank, and the matrix $\Phi^\top\mathcal{K}_{SA}\Phi$ is positive definite and hence invertible, we have $x=0$ if and only if $\Phi(\Phi^\top\mathcal{K}_{SA}\Phi)^{-1}x=0$.

To rewrite Eq. (13) in the desired form of the generalized PBE (2), we use the following three observations.

(1) The projection operator $\text{Proj}_{\mathcal{Q}}(\cdot)$ is explicitly given by $\text{Proj}_{\mathcal{Q}}(\cdot)=\Phi(\Phi^\top\mathcal{K}_{SA}\Phi)^{-1}\Phi^\top\mathcal{K}_{SA}(\cdot)$,

(2) The operator $\mathcal{T}_c(\cdot)$ is explicitly given by $\mathcal{T}_c(\cdot)=\sum_{i=0}^{n-1}(\gamma P_{\pi_c}D_c)^i(\cdot)$,

(3) The operator $\mathcal{H}_\rho(\cdot)$ is explicitly given by $\mathcal{H}_\rho(\cdot)=R+\gamma P_{\pi_\rho}D_\rho(\cdot)$.

Therefore, Eq. (13) is equivalent to

$$\text{Proj}_{\mathcal{Q}}[\mathcal{T}_c(\mathcal{H}_\rho(\Phi w)-\Phi w)]=0. \tag{14}$$

Finally, adding and subtracting Φw on both sides of the previous inequality and we obtain the desired generalized PBE:

$$\begin{aligned}\Phi w &= \text{Proj}_{\mathcal{Q}}[\mathcal{T}_c(\mathcal{H}_\rho(\Phi w)-\Phi w)]+\Phi w\\ &= \text{Proj}_{\mathcal{Q}}[\mathcal{T}_c(\mathcal{H}_\rho(\Phi w)-\Phi w)+\Phi w]\\ &= \text{Proj}_{\mathcal{Q}}\mathcal{B}_{c,\rho}(\Phi w),\end{aligned}$$

where the second equality follows from (1) $\Phi w\in\mathcal{Q}$ and (2) $\text{Proj}_{\mathcal{Q}}(\cdot)$ is a linear operator.

A.2 Proof of Lemma 2.2

For any $Q_1,Q_2\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, we have

$$\begin{aligned}\|\text{Proj}_{\mathcal{Q}}\mathcal{B}_{c,\rho}(Q_1)-\text{Proj}_{\mathcal{Q}}\mathcal{B}_{c,\rho}(Q_2)\|_{\mathcal{K}_{SA}} &\le \|\mathcal{B}_{c,\rho}(Q_1)-\mathcal{B}_{c,\rho}(Q_2)\|_{\mathcal{K}_{SA}}\\ &\le \|\mathcal{B}_{c,\rho}(Q_1)-\mathcal{B}_{c,\rho}(Q_2)\|_\infty && (\|\cdot\|_{\mathcal{K}_{SA}}\le\|\cdot\|_\infty)\end{aligned}$$

$$\begin{aligned} &\leq \tilde{\gamma}(n)\|Q_1 - Q_2\|_\infty \\ &\leq \frac{\tilde{\gamma}(n)}{\sqrt{\mathcal{K}_{SA,\min}}}\|Q_1 - Q_2\|_{\mathcal{K}_{SA}}, \qquad (\|\cdot\|_\infty \leq \tfrac{1}{\sqrt{\mathcal{K}_{SA,\min}}}\|\cdot\|_{\mathcal{K}_{SA}}) \end{aligned}$$

where the first inequality follows from $\text{Proj}_{\mathcal{Q}}$ being non-expansive with respect to $\|\cdot\|_{\mathcal{K}_{SA}}$, and the third inequality follows from $\mathcal{B}_{c,\rho}(\cdot)$ being a $\tilde{\gamma}(n)$-contraction operator with respect to $\|\cdot\|_\infty$ (Chen et al., 2021c) [1].

A.3 Proof of Lemma 2.3

We first show that under Condition 2.1 (3), we have $\lim_{n\to\infty}\tilde{\gamma}(n)/\sqrt{\mathcal{K}_{SA,\min}}<1$. Using the explicit expression of $\tilde{\gamma}(n)$, we have

$$\begin{aligned} \lim_{n\to\infty}\frac{\tilde{\gamma}(n)}{\sqrt{\mathcal{K}_{SA,\min}}} &= \lim_{n\to\infty}\frac{1-f_n(\gamma D_{c,\min})(1-\gamma D_{\rho,\max})}{\sqrt{\mathcal{K}_{SA,\min}}} \\ &= \lim_{n\to\infty}\frac{1-\frac{1-(\gamma D_{c,\min})^n}{1-\gamma D_{c,\min}}(1-\gamma D_{\rho,\max})}{\sqrt{\mathcal{K}_{SA,\min}}} \qquad (f_n(x)=\textstyle\sum_{i=0}^{n-1}x^i \text{ and } \gamma D_{c,\min}<1) \\ &= \frac{\gamma(D_{\rho,\max}-D_{c,\min})}{(1-\gamma D_{c,\min})\sqrt{\mathcal{K}_{SA,\min}}} \\ &< 1. \qquad \text{(Condition 2.1 (3))} \end{aligned}$$

Therefore, when n is chosen such that $\gamma_c=\frac{\tilde{\gamma}(n)}{\sqrt{\mathcal{K}_{SA,\min}}}<1$, we have by Lemma 2.2 that

$$\|\text{Proj}_{\mathcal{Q}}\mathcal{B}_{c,\rho}(Q_1)-\text{Proj}_{\mathcal{Q}}\leq \gamma_c\|Q_1-Q_2\|_{\mathcal{K}_{SA}},\ \forall\, Q_1,Q_2\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}.$$

It follows that the composed operator $\text{Proj}_{\mathcal{Q}}\mathcal{B}_{c,\rho}(\cdot)$ is a contraction mapping with respect to $\|\cdot\|_{\mathcal{K}_{SA}}$, with contraction factor γ_c.

Next consider the difference between Q^π and $\Phi w_{c,\rho}^\pi$. First of all, we have by triangle inequality that

$$\begin{aligned} \|Q^\pi-\Phi w_{c,\rho}^\pi\|_{\mathcal{K}_{SA}} &= \|Q^\pi-Q_{c,\rho}^\pi+Q_{c,\rho}^\pi-\Phi w_{c,\rho}^\pi\|_{\mathcal{K}_{SA}} \\ &\leq \|Q^\pi-Q_{c,\rho}^\pi\|_{\mathcal{K}_{SA}}+\|Q_{c,\rho}^\pi-\Phi w_{c,\rho}^\pi\|_{\mathcal{K}_{SA}}. \end{aligned} \tag{15}$$

We next bound each term on the RHS of the previous inequality. For the first term, it was already established in Proposition 2.1 of Chen et al. (2021c) that

$$\|Q^\pi-Q_{c,\rho}^\pi\|_{\mathcal{K}_{SA}}\leq\|Q^\pi-Q_{c,\rho}^\pi\|_\infty\leq\frac{\gamma\max_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}|\pi(a|s)-\pi_b(a|s)\rho(s,a)|}{(1-\gamma)(1-\gamma D_{\rho,\max})}. \tag{16}$$

Now consider the second term on the RHS of Eq. (15). First note that

$$\begin{aligned} \|Q_{c,\rho}^\pi-\Phi w_{c,\rho}^\pi\|_{\mathcal{K}_{SA}}^2 &= \|Q_{c,\rho}^\pi-\text{Proj}_{\mathcal{Q}}Q_{c,\rho}^\pi+\text{Proj}_{\mathcal{Q}}Q_{c,\rho}^\pi-\Phi w_{c,\rho}^\pi\|_{\mathcal{K}_{SA}}^2 \\ &= \|Q_{c,\rho}^\pi-\text{Proj}_{\mathcal{Q}}Q_{c,\rho}^\pi\|_{\mathcal{K}_{SA}}^2+\|\text{Proj}_{\mathcal{Q}}Q_{c,\rho}^\pi-\Phi w_{c,\rho}^\pi\|_{\mathcal{K}_{SA}}^2 \qquad (*) \\ &= \|Q_{c,\rho}^\pi-\text{Proj}_{\mathcal{Q}}Q_{c,\rho}^\pi\|_{\mathcal{K}_{SA}}^2+\|\text{Proj}_{\mathcal{Q}}\mathcal{B}_{c,\rho}(Q_{c,\rho}^\pi)-\text{Proj}_{\mathcal{Q}}\mathcal{B}_{c,\rho}(\Phi w_{c,\rho}^\pi)\|_{\mathcal{K}_{SA}}^2 \\ &\leq \|Q_{c,\rho}^\pi-\text{Proj}_{\mathcal{Q}}Q_{c,\rho}^\pi\|_{\mathcal{K}_{SA}}^2+\gamma_c^2\|Q_{c,\rho}^\pi-\Phi w_{c,\rho}^\pi\|_{\mathcal{K}_{SA}}^2, \end{aligned}$$

where Eq. $(*)$ follows from the Babylonian–Pythagorean theorem (i.e., $Q_{c,\rho}^\pi-\text{Proj}_{\mathcal{Q}}Q_{c,\rho}^\pi\perp\mathcal{Q}$ and $\text{Proj}_{\mathcal{Q}}Q_{c,\rho}^\pi-\Phi w_{c,\rho}^\pi\in\mathcal{Q}$). Rearrange the previous inequality and we have

$$\|Q_{c,\rho}^\pi-\Phi w_{c,\rho}^\pi\|_{\mathcal{K}_{SA}}\leq\frac{1}{\sqrt{1-\gamma_c^2}}\|Q_{c,\rho}^\pi-\text{Proj}_{\mathcal{Q}}Q_{c,\rho}^\pi\|_{\mathcal{K}_{SA}}. \tag{17}$$

Substituting Eqs. (16) and (17) into the RHS of Eq. (15) and we finally obtain

$$\|Q^\pi-\Phi w_{c,\rho}^\pi\|_{\mathcal{K}_{SA}}\leq\frac{\gamma\max_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}|\pi(a|s)-\pi_b(a|s)\rho(s,a)|}{(1-\gamma)(1-\gamma D_{\rho,\max})}+\frac{1}{\sqrt{1-\gamma_c^2}}\|Q_{c,\rho}^\pi-\text{Proj}_{\mathcal{Q}}Q_{c,\rho}^\pi\|_{\mathcal{K}_{SA}}.$$

[1] Chen et al. (2021c) works with an asynchronous variant of the generalized Bellman operator, which is shown to be a contraction mapping with respect to $\|\cdot\|_\infty$ with contraction factor $1-\mathcal{K}_{SA,\min}f_n(\gamma D_{c,\min})(1-\gamma D_{\rho,\max})$. In this paper we work with the synchronous generalized Bellman operator $\mathcal{B}_{c,\rho}(\cdot)$. In this case, one can easily verify that the corresponding contraction factor can be obtained by simply dropping the factor $\mathcal{K}_{SA,\min}$.

A.4 Proof of Theorem 2.1

The finite-sample bound (i.e., Eq. (4)) follows directly from Theorem 4.1. To show the performance bound (5) on the limit point $w_{c,\rho}^\pi$, we apply Lemma 2.3 to the λ-averaged Q-trace algorithm. Note that when $c(s,a) = \rho(s,a) = \lambda(s)\frac{\pi(a|s)}{\pi_b(a|s)} + 1 - \lambda(s)$ for all (s,a), we have for any $s \in \mathcal{S}$ that

$$\sum_{a\in\mathcal{A}} |\pi(a|s) - \pi_b(a|s)\rho(s,a)| = (1-\lambda(s)) \sum_{a\in\mathcal{A}} |\pi(a|s) - \pi_b(a|s)| = (1-\lambda(s))\|\pi(\cdot|s) - \pi_b(\cdot|s)\|_1.$$

This proves the result.

A.5 Proof of Theorem 2.2

The finite-sample bound (i.e., Eq. (6)) follows directly from Theorem 4.1. To show the performance bound (7) on the limit point $w_{c,\rho}^\pi$, we apply Lemma 2.3 to the two-sided Q-trace algorithm. Note that when $c(s,a) = \rho(s,a) = g_{\ell(s),u(s)}(\pi(a|s)/\pi_b(a|s))$ for all (s,a), we have

$$\begin{aligned}
&\sum_{a\in\mathcal{A}} |\pi(a|s) - \pi_b(a|s)\rho(s,a)| \\
=\;& \sum_{a\in\mathcal{A}} |(\pi(a|s) - \pi_b(a|s)\ell(s))\mathbb{I}\{\pi(a|s) < \ell(s)\pi_b(a|s)\} + (\pi(a|s) - \pi_b(a|s)u(s))\mathbb{I}\{\pi(a|s) > u(s)\pi_b(a|s)\}| \\
\leq\;& \sum_{a\in\mathcal{A}} |(\pi(a|s) - \pi_b(a|s)\ell(s))\mathbb{I}\{\pi(a|s) < \ell(s)\pi_b(a|s)\}| + \sum_{a\in\mathcal{A}} |(\pi(a|s) - \pi_b(a|s)u(s))\mathbb{I}\{\pi(a|s) > u(s)\pi_b(a|s)\}| \\
=\;& \sum_{a\in\mathcal{A}} \max(\pi(a|s) - \pi_b(a|s)u(s), 0) - \min(\pi(a|s) - \pi_b(a|s)\ell(s), 0) \\
=\;& \sum_{a\in\mathcal{A}} (u_{\pi,\pi_b}(s,a) - \ell_{\pi,\pi_b}(s,a)).
\end{aligned}$$

This proves the result.

B Proof of All Technical Results in Section 3

B.1 Proof of Theorem 3.1

We begin with the result of Proposition 4.2:

$$\begin{aligned}
\mathbb{E}[\|Q^* - Q^{\pi_T}\|_\infty] \leq \gamma^T \|Q^* - Q^{\pi_0}\|_\infty + &\underbrace{\frac{2\gamma}{1-\gamma}\sum_{i=0}^{T-1}\gamma^{T-1-i}\mathbb{E}[\|Q^{\pi_i} - \Phi w_i\|_\infty]}_{A_2} \\
&+ \underbrace{\frac{2\gamma}{1-\gamma}\sum_{i=0}^{T-1}\gamma^{T-1-i}\mathbb{E}[\|\mathcal{H}_{\pi_{i+1}}(\Phi w_i) - \mathcal{H}(\Phi w_i)\|_\infty]}_{A_3}.
\end{aligned} \tag{18}$$

B.1.1 The Term A_2

To control the term A_2, using triangle inequality and we have for any $0 \leq i \leq T-1$:

$$\begin{aligned}
\mathbb{E}[\|Q^{\pi_i} - \Phi w_i\|_\infty] &\leq \mathbb{E}[\|Q^{\pi_i} - \Phi w_{c,\rho}^{\pi_i} + \Phi w_{c,\rho}^{\pi_i} - \Phi w_i\|_\infty] \\
&\leq \mathbb{E}[\|Q^{\pi_i} - \Phi w_{c,\rho}^{\pi_i}\|_\infty] + \mathbb{E}[\|\Phi(w_{c,\rho}^{\pi_i} - w_i)\|_\infty] \\
&\leq \mathbb{E}[\|Q^{\pi_i} - \Phi w_{c,\rho}^{\pi_i}\|_\infty] + \|\Phi\|_\infty \mathbb{E}[\|w_{c,\rho}^{\pi_i} - w_i\|_\infty] \\
&\leq \mathbb{E}[\|Q^{\pi_i} - \Phi w_{c,\rho}^{\pi_i}\|_\infty] + \mathbb{E}[\|w_{c,\rho}^{\pi_i} - w_i\|_\infty] \\
&\leq \mathbb{E}[\|Q_{c,\rho}^{\pi_i} - \Phi w_{c,\rho}^{\pi_i}\|_\infty] + \mathbb{E}[\|Q_{c,\rho}^{\pi_i} - Q^{\pi_i}\|_\infty] + \mathbb{E}[\|w_{c,\rho}^{\pi_i} - w_i\|_\infty] && (\|\Phi\|_\infty \leq 1) \\
&\leq \mathcal{E}_{\text{approx}} + \frac{\gamma}{(1-\gamma)^2}\max_{s\in\mathcal{S}}(1-\lambda(s))\|\pi_i(\cdot|s) - \pi_b(\cdot|s)\|_1 + \mathbb{E}[\|w_{c,\rho}^{\pi_i} - w_i\|_\infty] && \text{(Apply Eq. (16))}
\end{aligned}$$

$$\leq \mathcal{E}_{\text{approx}} + \frac{\gamma}{(1-\gamma)^2}\mathcal{E}_{\text{bias}} + \mathbb{E}[\|w_{c,\rho}^{\pi_i} - w_i\|_\infty]. \tag{19}$$

It remains to control $\mathbb{E}[\|w_{c,\rho}^{\pi_i} - w_i\|_\infty]$. For any $0 \leq i \leq T-1$, we have by Theorem 2.1 that

$$\begin{aligned}
\mathbb{E}[\|w_{c,\rho}^{\pi_i} - w_i\|_\infty] &\leq \mathbb{E}[\|w_{c,\rho}^{\pi_i} - w_i\|_2] \\
&\leq (\mathbb{E}[\|w_{c,\rho}^{\pi_i} - w_i\|_2^2])^{1/2} \quad \text{(Jensen's Inequality)} \\
&\leq c_{1,i}(1-(1-\gamma_c)\lambda_{\min}\alpha)^{\frac{1}{2}[K-(t_\alpha+n+1)]} + c_{2,i}\frac{[\alpha(t_\alpha+n+1)]^{1/2}}{\sqrt{1-\gamma_c}\sqrt{\lambda_{\min}}},
\end{aligned}$$

where the last line follows from $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$, and $c_{1,i} = \|w_{c,\rho}^{\pi_i}\|_2 + 1$ and $c_{2,i} = 11.5L(\|w_{c,\rho}^{\pi_i}\|_2 + 1)$. To further control the constants $c_{1,i}$ and $c_{2,i}$, note that we have for any policy π that

$$\begin{aligned}
\|w_{c,\rho}^{\pi}\|_2 &\leq \frac{1}{\sqrt{\lambda_{\min}}}\|\Phi w_{c,\rho}^{\pi}\|_{\mathcal{K}_{SA}} \\
&\leq \frac{1}{\sqrt{\lambda_{\min}}}\left(\|Q_{c,\rho}^{\pi}\|_{\mathcal{K}_{SA}} + \frac{1}{\sqrt{1-\gamma_c^2}(1-\gamma)}\right) \quad \text{(Eq. (17))} \\
&\leq \frac{1}{\sqrt{\lambda_{\min}}}\left(\frac{1}{1-\gamma} + \frac{1}{\sqrt{1-\gamma_c^2}(1-\gamma)}\right) \\
&\leq \frac{2}{\sqrt{\lambda_{\min}}(1-\gamma)\sqrt{1-\gamma_c}}.
\end{aligned}$$

Therefore we have $c_{1,i} \leq \frac{3}{\sqrt{\lambda_{\min}}(1-\gamma)\sqrt{1-\gamma_c}}$ and $c_{2,i} \leq \frac{35L}{\sqrt{\lambda_{\min}}(1-\gamma)\sqrt{1-\gamma_c}}$ for any $0 \leq i \leq T-1$. Substituting the upper bound we obtained for $\mathbb{E}[\|w_{c,\rho}^{\pi_i} - w_i\|_\infty]$ into Eq. (19) and we have for any $0 \leq i \leq T-1$:

$$\begin{aligned}
\mathbb{E}[\|Q^{\pi_i} - \Phi w_i\|_\infty] \leq{}& \mathcal{E}_{\text{approx}} + \frac{\gamma}{(1-\gamma)^2}\mathcal{E}_{\text{bias}} + \frac{3}{\sqrt{\lambda_{\min}}(1-\gamma)\sqrt{1-\gamma_c}}(1-(1-\gamma_c)\lambda_{\min}\alpha)^{\frac{1}{2}[K-(t_\alpha+n+1)]} \\
&+ \frac{35L[\alpha(t_\alpha+n+1)]^{1/2}}{(1-\gamma)(1-\gamma_c)\lambda_{\min}}.
\end{aligned}$$

Finally, using the previous inequality and we obtain the following bound on the term A_2:

$$\begin{aligned}
A_2 &= \frac{2\gamma}{1-\gamma}\sum_{i=0}^{T-1}\gamma^{T-1-i}\mathbb{E}[\|Q^{\pi_i} - \Phi w_i\|_\infty] \\
&\leq \frac{2\gamma\mathcal{E}_{\text{approx}}}{(1-\gamma)^2} + \frac{2\gamma^2\mathcal{E}_{\text{bias}}}{(1-\gamma)^4} + 6\tilde{c}(1-(1-\gamma_c)\lambda_{\min}\alpha)^{\frac{1}{2}[K-(t_\alpha+n+1)]} + \frac{70\tilde{c}L[\alpha(t_\alpha+n+1)]^{1/2}}{\sqrt{\lambda_{\min}}\sqrt{1-\gamma_c}},
\end{aligned}$$

where $\tilde{c} = \frac{\gamma}{\sqrt{\lambda_{\min}}\sqrt{1-\gamma_c}(1-\gamma)^3}$.

B.1.2 The Term A_3

Now consider the term A_3, whose upper bound depends on which policy update rule we use.

$1/\beta_1$-Greedy Update For simplicity of notation, denote $Q_t = \Phi w_t$. Then we have for any $0 \leq t \leq T-1$ and state-action pair (s, a) that

$$\begin{aligned}
0 &\leq [\mathcal{H}(Q_t)](s,a) - [\mathcal{H}_{\pi_{t+1}}(Q_t)](s,a) \\
&= \left[\mathcal{R}(s,a) + \gamma\sum_{s'\in\mathcal{S}}P_a(s,s')Q_t(s', a_{t,s'})\right] \quad \text{(Recall that } a_{t,s'} = \arg\max_{a'\in\mathcal{A}} Q_t(s',a')\text{)} \\
&\quad - \left\{\mathcal{R}(s,a) + \gamma\sum_{s'\in\mathcal{S}}P_a(s,s')\left[\left(1 - \frac{1}{\beta_1(t,s')} + \frac{1}{|\mathcal{A}|\beta_1(t,s')}\right)Q_t(s',a_{t,s'}) + \sum_{a'\neq a_{t,s'}}\frac{1}{|\mathcal{A}|\beta_1(t,s')}Q_t(s',a')\right]\right\} \\
&= \gamma\sum_{s'}P_a(s,s')\left(\left(\frac{1}{\beta_1(t,s')} - \frac{1}{|\mathcal{A}|\beta_1(t,s')}\right)Q_t(s',a_{t,s'}) - \sum_{a'\neq a_{t,s'}}\frac{1}{|\mathcal{A}|\beta_1(t,s')}Q_t(s',a')\right)
\end{aligned}$$

$$\leq \gamma \sum_{s'} P_a(s,s') \frac{2}{\beta_1(t,s')} \max_{a' \in \mathcal{A}} |Q_t(s',a')|$$
$$\leq \beta,$$

where the last line follows from $\beta_1(t,s) \geq \frac{2\gamma}{\beta} \max_{a \in \mathcal{A}} |Q_t(s,a)|$ for all $s \in \mathcal{S}$ (cf. Condition 3.1). Therefore, we have

$$A_3 \leq \frac{2\gamma}{1-\gamma} \sum_{i=0}^{T-1} \gamma^{T-1-i} \beta \leq \frac{2\gamma\beta}{(1-\gamma)^2}.$$

$\boldsymbol{\beta_2}$-Softmax Update The following lemma is needed for us to control the term A_3.

Lemma B.1. *For any $x \in \mathbb{R}^d$ and $y \in \Delta^d$ satisfying $y_i > 0$ for all i, denote $i_{\max} = \arg\max_{1 \leq i \leq d} x_i$, then the following inequality holds for any $\beta > 0$:*

$$\max_{1 \leq i \leq d} x_i - \frac{\sum_{i=1}^d x_i y_i e^{\beta x_i}}{\sum_{j=1}^d y_j e^{\beta x_j}} \leq \frac{1}{\beta} \log\left(\frac{1}{y_{i_{\max}}}\right).$$

Proof of Lemma B.1. For any $\beta > 0$, consider the function $h_\beta : \mathbb{R}^d \mapsto \mathbb{R}$ defined by

$$h_\beta(x) = \frac{1}{\beta} \log\left(\sum_{i=1}^d y_i e^{\beta x_i}\right).$$

Assume without loss of generality that $i_{\max} = 1$. Then it is clear that $h_\beta(x) \leq x_1$. On the other hand, we have

$$x_1 \leq \frac{1}{\beta} \log\left(\sum_{i=1}^d \frac{y_i}{y_1} e^{\beta x_i}\right) = h_\beta(x) + \frac{1}{\beta} \log\left(\frac{1}{y_1}\right). \tag{20}$$

Since it is well-known that $h_\beta(x)$ is a convex differentiable function, we have for any $x \in \mathbb{R}^d$ that $h_\beta(0) - h_\beta(x) \geq \langle \nabla h_\beta(x), -x \rangle$, which implies

$$\langle \nabla h_\beta(x), x \rangle = \frac{\sum_{i=1}^d x_i y_i e^{\beta x_i}}{\sum_{j=1}^d y_j e^{\beta x_j}} \geq h_\beta(x) - h_\beta(0) = h_\beta(x). \tag{21}$$

Using Eqs. (20) and (21) and we finally obtain

$$\max_{1 \leq i \leq d} x_i - \frac{\sum_{i=1}^d x_i y_i e^{\beta x_i}}{\sum_{j=1}^d y_j e^{\beta x_j}} \leq x_1 - h_\beta(x) \leq \frac{1}{\beta} \log\left(\frac{1}{y_1}\right).$$

□

We now proceed to control the term A_3 when using the β_2-softmax update. For any $0 \leq t \leq T-1$ and state-action pair (s,a), we have

$$\begin{aligned}
0 &\leq [\mathcal{H}(Q_t)](s,a) - [\mathcal{H}_{\pi_{t+1}}(Q_t)](s,a) \\
&= \gamma \sum_{s'} P_a(s,s') \left(\max_{a' \in \mathcal{A}} Q_t(s',a') - \sum_{a' \in \mathcal{A}} \frac{\exp(\beta_2 Q_t(s',a'))}{\sum_{a'' \in \mathcal{A}} \exp(\beta_2 Q_t(s',a''))} Q_t(s',a') \right) \\
&= \gamma \sum_{s'} P_a(s,s') \left(\max_{a' \in \mathcal{A}} Q_t(s',a') - \sum_{a' \in \mathcal{A}} \frac{\exp(\beta_2 Q_t(s',a'))/|\mathcal{A}|}{\sum_{a'' \in \mathcal{A}} \exp(\beta_2 Q_t(s',a''))/|\mathcal{A}|} Q_t(s',a') \right) \\
&\leq \frac{\gamma}{\beta_2} \log(|\mathcal{A}|) \qquad \text{(Lemma B.1)} \\
&\leq \beta,
\end{aligned}$$

where the last line follows from $\beta_2 \geq \frac{\gamma}{\beta} \log(|\mathcal{A}|)$. Therefore, we have

$$A_3 \leq \frac{2\gamma}{1-\gamma} \sum_{i=0}^{T-1} \gamma^{T-1-i} \beta \leq \frac{2\gamma\beta}{(1-\gamma)^2}.$$

β_3-NPG Update Recall that β_3-NPG updates the policy according to

$$\pi_{t+1}(a|s)=\frac{\pi_t(a|s)\exp(\beta_3(t)Q_t(s,a))}{\sum_{a'\in\mathcal{A}}\pi_t(a'|s)\exp(\beta_3(t)Q_t(s,a'))},\ \forall\,(s,a).$$

Therefore, for any $0\leq t\leq T-1$ and state-action pair (s,a), we have

$$\begin{aligned}
0&\leq [\mathcal{H}(Q_t)](s,a)-[\mathcal{H}_{\pi_{t+1}}(Q_t)](s,a)\\
&=\gamma\sum_{s'}P_a(s,s')\left(\max_{a'\in\mathcal{A}}Q_t(s',a')-\sum_{a'\in\mathcal{A}}\frac{\pi_t(a'|s')\exp(\beta_3(t)Q_t(s',a'))}{\sum_{a''\in\mathcal{A}}\pi_t(a''|s')\exp(\beta_3(t)Q_t(s',a''))}Q_t(s',a')\right)\\
&\leq\frac{\gamma}{\beta_3(t)}\log\left(\frac{1}{\pi_t(a_{t,s'}|s')}\right)\\
&\leq\beta,
\end{aligned}$$

where the last line follows from $\beta_3(t)\geq\frac{\gamma}{\beta}\log(1/\min_{s\in\mathcal{S}}\pi_t(a_{t,s}|s))$. Therefore, we have

$$A_3\leq\frac{2\gamma}{1-\gamma}\sum_{i=0}^{T-1}\gamma^{T-1-i}\beta\leq\frac{2\gamma\beta}{(1-\gamma)^2}.$$

B.2 Putting Together

Using the upper bounds we obtained for the terms A_2 and A_3 in Eq. (18) and we have for any $K\geq t_\alpha+n+1$ and $T\geq 0$ that

$$\begin{aligned}
\mathbb{E}[\|Q^*-Q^{\pi_T}\|_\infty]\leq{}&\gamma^T\|Q^*-Q^{\pi_0}\|_\infty+\frac{2\gamma\mathcal{E}_{\text{approx}}}{(1-\gamma)^2}+\frac{2\gamma^2\mathcal{E}_{\text{bias}}}{(1-\gamma)^4}+6\tilde{c}(1-(1-\gamma_c)\lambda_{\min}\alpha)^{\frac{1}{2}[K-(t_\alpha+n+1)]}\\
&+\frac{70\tilde{c}L[\alpha(t_\alpha+n+1)]^{1/2}}{\sqrt{\lambda_{\min}}\sqrt{1-\gamma_c}}+\frac{2\gamma\beta}{(1-\gamma)^2},
\end{aligned}$$

where $\tilde{c}=\frac{\gamma}{\sqrt{\lambda_{\min}}\sqrt{1-\gamma_c}(1-\gamma)^3}$.

C Proof of All Technical Results in Section 4

C.1 Proof of Proposition 4.1

(1) (a) We first rewrite the operator $F(\cdot,\cdot)$ in the following equivalent way. For any $w\in\mathbb{R}^d$ and $x=(s_0,a_0,...,s_n,a_n)\in\mathcal{X}$, we have

$$\begin{aligned}
F(w,x)={}&\phi(s_0,a_0)\sum_{i=0}^{n-1}\gamma^i\prod_{j=1}^{i}c(s_j,a_j)(\mathcal{R}(s_i,a_i)+\gamma\rho(s_{i+1},a_{i+1})\phi(s_{i+1},a_{i+1})^\top w-\phi(s_i,a_i)^\top w)\\
={}&\phi(s_0,a_0)\sum_{i=0}^{n-1}\gamma^i\prod_{j=1}^{i}c(s_j,a_j)\mathcal{R}(s_i,a_i)-\phi(s_0,a_0)\sum_{i=0}^{n-1}\gamma^i\prod_{j=1}^{i}c(s_j,a_j)\phi(s_i,a_i)^\top w\\
&+\phi(s_0,a_0)\sum_{i=0}^{n-1}\gamma^{i+1}\prod_{j=1}^{i}c(s_j,a_j)\rho(s_{i+1},a_{i+1})\phi(s_{i+1},a_{i+1})^\top w\\
={}&\phi(s_0,a_0)\sum_{i=0}^{n-1}\gamma^i\prod_{j=1}^{i}c(s_j,a_j)\mathcal{R}(s_i,a_i)-\phi(s_0,a_0)\sum_{i=0}^{n-1}\gamma^i\prod_{j=1}^{i}c(s_j,a_j)\phi(s_i,a_i)^\top w\\
&+\phi(s_0,a_0)\sum_{i=1}^{n}\gamma^i\prod_{j=1}^{i-1}c(s_j,a_j)\rho(s_i,a_i)\phi(s_i,a_i)^\top w\\
={}&\phi(s_0,a_0)\sum_{i=0}^{n-1}\gamma^i\prod_{j=1}^{i}c(s_j,a_j)\mathcal{R}(s_i,a_i)-\phi(s_0,a_0)\phi(s_0,a_0)^\top w
\end{aligned}$$

$$
\begin{aligned}
&+ \phi(s_0, a_0) \sum_{i=1}^{n-1} \gamma^i \prod_{j=1}^{i-1} c(s_j, a_j)(\rho(s_i, a_i) - c(s_i, a_i))\phi(s_i, a_i)^\top w \\
&+ \phi(s_0, a_0)\gamma^n \prod_{j=1}^{n-1} c(s_j, a_j)\rho(s_n, a_n)\phi(s_n, a_n)^\top w.
\end{aligned}
$$

We now proceed and show the Lipschitz property. For any $w_1, w_2 \in \mathbb{R}^d$ and $x = (s_0, a_0, ..., s_n, a_n) \in \mathcal{X}$, using the fact that $\|\phi(s, a)\|_2 \leq \|\phi(s, a)\|_1 \leq \|\Phi\|_\infty \leq 1$, we have

$$
\begin{aligned}
&\|F(w_1, x) - F(w_2, x)\|_2 \\
\leq\;& \|\phi(s_0, a_0)\phi(s_0, a_0)^\top (w_1 - w_2)\|_2 \\
&+ \left\|\phi(s_0, a_0) \sum_{i=1}^{n-1} \gamma^i \prod_{j=1}^{i-1} c(s_j, a_j)(\rho(s_i, a_i) - c(s_i, a_i))\phi(s_i, a_i)^\top (w_1 - w_2)\right\|_2 \\
&+ \left\|\phi(s_0, a_0)\gamma^n \prod_{j=1}^{n-1} c(s_j, a_j)\rho(s_n, a_n)\phi(s_n, a_n)^\top (w_1 - w_2)\right\|_2 \\
\leq\;& \|w_1 - w_2\|_2 + \sum_{i=1}^{n-1} \gamma^i c_{\max}^{i-1} \max_{s,a} |\rho(s, a) - c(s, a)| \|w_1 - w_2\|_2 + \gamma^n c_{\max}^{n-1} \rho_{\max} \|w_1 - w_2\|_2 \\
=\;& \left(1 + \gamma \max_{s,a} |\rho(s, a) - c(s, a)| \frac{1 - (\gamma c_{\max})^{n-1}}{1 - \gamma c_{\max}} + \gamma^n c_{\max}^{n-1} \rho_{\max}\right) \|w_1 - w_2\|_2 \\
\leq\;& \begin{cases} (1 + (\gamma\rho_{\max})^n)\|w_1 - w_2\|_2, & c(\cdot, \cdot) = \rho(\cdot, \cdot) \\ (1 + \gamma\rho_{\max}) f_n(\gamma c_{\max})\|w_1 - w_2\|_2, & c(\cdot, \cdot) \neq \rho(\cdot, \cdot). \end{cases}
\end{aligned}
$$

(b) For any $x = (s_0, a_0, ..., s_n, a_n) \in \mathcal{X}$, we have

$$
\|F(w, \mathbf{0})\|_2 = \left\|\phi(s_0, a_0) \sum_{i=0}^{n-1} \gamma^i \prod_{j=1}^{i} c(s_j, a_j)\mathcal{R}(s_i, a_i)\right\|_2 \leq \sum_{i=0}^{n-1} \gamma^i c_{\max}^i \leq f_n(\gamma c_{\max}).
$$

(2) It is clear that the stationary distribution ν of the Markov chain $\{X_k\}$ is given by

$$
\nu(s_0, a_0, ..., s_n, a_n) = \mu(s_0) \left(\prod_{i=0}^{n-1} \pi_b(a_i|s_i) P_{a_i}(s_i, s_{i+1})\right) \pi_b(a_n|s_n), \quad \forall\, (s_0, a_0, ..., s_n, a_n) \in \mathcal{X}.
$$

Moreover, for any $x = (s_0, a_0, ..., s_n, a_n) \in \mathcal{X}$, we have for any $k \geq 0$ that

$$
\begin{aligned}
&\left\|P_{\pi_b}^{k+n+1}(x, \cdot) - \nu(\cdot)\right\|_{\text{TV}} \\
=\;& \frac{1}{2} \sum_{s_0', a_0', \cdots, s_n', a_n'} \left|\sum_s P_{a_n}(s_n, s) P_{\pi_b}^k(s, s_0') - \mu(s_0')\right| \left[\prod_{i=0}^{n-1} \pi_b(a_i' \mid s_i') P_{a_i'}(s_i', s_{i+1}')\right] \pi_b(a_n' \mid s_n') \\
=\;& \frac{1}{2} \sum_{s_0'} \left|\sum_s P_{a_n}(s_n, s) P_{\pi_b}^k(s, s_0') - \mu(s_0')\right| \\
\leq\;& \frac{1}{2} \sum_s P_{a_n}(s_n, s) \sum_{s_0'} \left|P_{\pi_b}^k(s, s_0') - \mu(s_0')\right| \\
\leq\;& \max_{s \in \mathcal{S}} \|P_{\pi_b}^k(s, \cdot) - \mu(\cdot)\|_{\text{TV}} \\
\leq\;& C\sigma^k.
\end{aligned}
$$

Therefore, we have $\max_{x \in \mathcal{X}} \left\|P_{\pi_b}^{k+n+1}(x, \cdot) - \nu(\cdot)\right\|_{\text{TV}} \leq C\sigma^k$ for all $k \geq 0$.

(3) Using the fact that $\mathcal{B}_{c,\rho}(\cdot)$ is a linear operator (Chen et al., 2021c), we have for any $w \in \mathbb{R}^d$ that

$$
(w - w_{c,\rho}^\pi)^\top \bar{F}(w)
$$

$$
\begin{aligned}
&= (w-w_{c,\rho}^\pi)^\top\Phi^\top\mathcal{K}_{SA}\left(\mathcal{B}_{c,\rho}(\Phi w)-\Phi w\right)\\
&= (w-w_{c,\rho}^\pi)^\top\Phi^\top\mathcal{K}_{SA}\left(\mathcal{B}_{c,\rho}(\Phi w)-\mathcal{B}_{c,\rho}(\Phi w_{c,\rho}^\pi)\right)-(w-w_{c,\rho}^\pi)^\top\Phi^\top\mathcal{K}_{SA}\Phi(w-w_{c,\rho}^\pi)\\
&= (w-w_{c,\rho}^\pi)^\top\Phi^\top\mathcal{K}_{SA}\Phi(\Phi^\top\mathcal{K}_{SA}\Phi)^{-1}\Phi^\top\mathcal{K}_{SA}\mathcal{B}_{c,\rho}(\Phi(w-w_{c,\rho}^\pi))-(w-w_{c,\rho}^\pi)^\top\Phi^\top\mathcal{K}_{SA}\Phi(w-w_{c,\rho}^\pi)\\
&= (w-w_{c,\rho}^\pi)^\top\Phi^\top\mathcal{K}_{SA}\Phi(\Phi^\top\mathcal{K}_{SA}\Phi)^{-1}\Phi^\top\mathcal{K}_{SA}\mathcal{B}_{c,\rho}(\Phi(w-w_{c,\rho}^\pi))-(w-w_{c,\rho}^\pi)^\top\Phi^\top\mathcal{K}_{SA}\Phi(w-w_{c,\rho}^\pi)\\
&\leq \|\Phi(w-w_{c,\rho}^\pi)\|_{\mathcal{K}_{SA}}\|\Phi(\Phi^\top\mathcal{K}_{SA}\Phi)^{-1}\Phi^\top\mathcal{K}_{SA}\mathcal{B}_{c,\rho}(\Phi(w-w_{c,\rho}^\pi))\|_{\mathcal{K}_{SA}}-\|\Phi(w-w_{c,\rho}^\pi)\|_{\mathcal{K}_{SA}}^2\\
&= \|\Phi(w-w_{c,\rho}^\pi)\|_{\mathcal{K}_{SA}}\|\text{Proj}_{\mathcal{Q}}\mathcal{B}_{c,\rho}(\Phi(w-w_{c,\rho}^\pi))\|_{\mathcal{K}_{SA}}-\|\Phi(w-w_{c,\rho}^\pi)\|_{\mathcal{K}_{SA}}^2\\
&\leq \gamma_c\|\Phi(w-w_{c,\rho}^\pi)\|_{\mathcal{K}_{SA}}\|\Phi(w-w_{c,\rho}^\pi)\|_{\mathcal{K}_{SA}}-\|\Phi(w-w_{c,\rho}^\pi)\|_{\mathcal{K}_{SA}}^2\\
&= -(1-\gamma_c)\|\Phi(w-w_{c,\rho}^\pi)\|_{\mathcal{K}_{SA}}^2\\
&\leq -(1-\gamma_c)\lambda_{\min}\|w-w_{c,\rho}^\pi\|_2^2
\end{aligned}
$$

C.2 Proof of Lemma 4.1

For simplicity of notation, denote $\delta_t=\max_{s,a}(Q^{\pi_t}(s,a)-Q^{\pi_{t+1}}(s,a))$. Then we have by definition of δ_t that $Q^{\pi_{t+1}}\geq Q^{\pi_t}-\delta_t\mathbf{1}$. Using the monotonicity of the Bellman operator (Bertsekas and Tsitsiklis, 1996, Lemma 2.1 and Lemma 2.2) and we have

$$Q^{\pi_{t+1}}=\mathcal{H}_{\pi_{t+1}}(Q^{\pi_{t+1}})\geq\mathcal{H}_{\pi_{t+1}}(Q^{\pi_t}-\delta_t\mathbf{1})=\mathcal{H}_{\pi_{t+1}}(Q^{\pi_t})-\gamma\delta_t\mathbf{1}.$$

It follows that

$$
\begin{aligned}
&Q^{\pi_t}-Q^{\pi_{t+1}}\\
&\leq Q^{\pi_t}-\mathcal{H}_{\pi_{t+1}}(Q^{\pi_t})+\gamma\delta_t\mathbf{1}\\
&= Q^{\pi_t}-\mathcal{H}_{\pi_{t+1}}(Q^{\pi_t})+\mathcal{H}_{\pi_{t+1}}(Q_t)-\mathcal{H}_{\pi_{t+1}}(Q_t)+\mathcal{H}(Q_t)-\mathcal{H}(Q_t)+\gamma\delta_t\mathbf{1}\\
&\leq \mathcal{H}_{\pi_t}(Q^{\pi_t})-\mathcal{H}_{\pi_t}(Q_t)-\mathcal{H}_{\pi_{t+1}}(Q^{\pi_t})+\mathcal{H}_{\pi_{t+1}}(Q_t)-\mathcal{H}_{\pi_{t+1}}(Q_t)+\mathcal{H}(Q_t)+\gamma\delta_t\mathbf{1}\\
&\leq 2\gamma\|Q^{\pi_t}-Q_t\|_\infty\mathbf{1}+\|\mathcal{H}_{\pi_{t+1}}(Q_t)-\mathcal{H}(Q_t)\|_\infty\mathbf{1}+\gamma\delta_t\mathbf{1}.
\end{aligned}
$$

Therefore, we have

$$\delta_t\leq 2\gamma\|Q^{\pi_t}-Q_t\|_\infty+\|\mathcal{H}_{\pi_{t+1}}(Q_t)-\mathcal{H}(Q_t)\|_\infty+\gamma\delta_t,$$

which implies

$$\delta_t\leq\frac{2\gamma\|Q^{\pi_t}-Q_t\|_\infty+\|\mathcal{H}_{\pi_{t+1}}(Q_t)-\mathcal{H}(Q_t)\|_\infty}{1-\gamma}.$$

C.3 Proof of Lemma 4.2

For simplicity of notation, denote $\zeta_t=\max_{s,a}(Q^*(s,a)-Q^{\pi_t}(s,a))=\|Q^*-Q^{\pi_t}\|_\infty$. Then we have by definition of ζ_t that $Q^{\pi_t}\geq Q^*-\zeta_t\mathbf{1}$. Using the monotonicity of the Bellman operator and we have

$$
\begin{aligned}
Q^{\pi_{t+1}}&=\mathcal{H}_{\pi_{t+1}}(Q^{\pi_{t+1}})\\
&\geq\mathcal{H}_{\pi_{t+1}}(Q^{\pi_t}-\max_{s,a}(Q^{\pi_t}(s,a)-Q^{\pi_{t+1}}(s,a))\mathbf{1})\\
&=\mathcal{H}_{\pi_{t+1}}(Q^{\pi_t})-\gamma\max_{s,a}(Q^{\pi_t}(s,a)-Q^{\pi_{t+1}}(s,a))\mathbf{1}\\
&\geq\mathcal{H}_{\pi_{t+1}}(Q^{\pi_t})-\frac{2\gamma^2\|Q^{\pi_t}-Q_t\|_\infty+\gamma\|\mathcal{H}_{\pi_{t+1}}(Q_t)-\mathcal{H}(Q_t)\|_\infty}{1-\gamma}\mathbf{1},
\end{aligned}
\tag{22}
$$

where the last line follows from Lemma 4.1. We next control $\mathcal{H}_{\pi_{t+1}}(Q^{\pi_t})$ from below in the following. Again by monotonicity of the Bellman operator we have

$$
\begin{aligned}
\mathcal{H}_{\pi_{t+1}}(Q^{\pi_t})&\geq\mathcal{H}_{\pi_{t+1}}(Q_t-\|Q_t-Q^{\pi_t}\|_\infty\mathbf{1})\\
&=\mathcal{H}_{\pi_{t+1}}(Q_t)-\gamma\|Q_t-Q^{\pi_t}\|_\infty\mathbf{1}\\
&=\mathcal{H}_{\pi_{t+1}}(Q_t)-\mathcal{H}(Q_t)+\mathcal{H}(Q_t)-\gamma\|Q_t-Q^{\pi_t}\|_\infty\mathbf{1}\\
&\geq\mathcal{H}_{\pi_{t+1}}(Q_t)-\mathcal{H}(Q_t)+\mathcal{H}(Q^{\pi_t}-\|Q_t-Q^{\pi_t}\|_\infty\mathbf{1})-\gamma\|Q_t-Q^{\pi_t}\|_\infty\mathbf{1}\\
&=\mathcal{H}_{\pi_{t+1}}(Q_t)-\mathcal{H}(Q_t)+\mathcal{H}(Q^{\pi_t})-2\gamma\|Q_t-Q^{\pi_t}\|_\infty\mathbf{1}
\end{aligned}
$$

$$\begin{aligned}&\geq \mathcal{H}_{\pi_{t+1}}(Q_t) - \mathcal{H}(Q_t) + \mathcal{H}(Q^* - \zeta_t \mathbf{1}) - 2\gamma \|Q_t - Q^{\pi_t}\|_\infty \mathbf{1}\\&= \mathcal{H}_{\pi_{t+1}}(Q_t) - \mathcal{H}(Q_t) + \mathcal{H}(Q^*) - \gamma\zeta_t \mathbf{1} - 2\gamma \|Q_t - Q^{\pi_t}\|_\infty \mathbf{1}\\&\geq -\|\mathcal{H}_{\pi_{t+1}}(Q_t) - \mathcal{H}(Q_t)\|_\infty \mathbf{1} + Q^* - \gamma\zeta_t \mathbf{1} - 2\gamma \|Q_t - Q^{\pi_t}\|_\infty \mathbf{1}.\end{aligned}$$

Using the previous inequality in Eq. (22) and we have

$$Q^{\pi_{t+1}} - Q^* \geq -\gamma\zeta_t \mathbf{1} - \frac{2\gamma\|Q^{\pi_t} - Q_t\|_\infty + \|\mathcal{H}_{\pi_{t+1}}(Q_t) - \mathcal{H}(Q_t)\|_\infty}{1-\gamma}\mathbf{1},$$

which implies

$$\zeta_{t+1} \leq \gamma\zeta_t + \frac{2\gamma\|Q^{\pi_t} - Q_t\|_\infty + \|\mathcal{H}_{\pi_{t+1}}(Q_t) - \mathcal{H}(Q_t)\|_\infty}{1-\gamma}.$$

D The High Variance in Chen et al. (2021a)

Consider Theorem 2.1 of Chen et al. (2021a). The constant c_2 on the second term is proportional to $\sum_{i=0}^{n-1}(\gamma \max_{s,a} \frac{\pi(a|s)}{\pi_b(a|s)})^i$ (which appears as $f(\gamma\zeta_\pi)$ using the notation of Chen et al. (2021a)). When $\frac{\pi(a|s)}{\pi_b(a|s)} > 1/\gamma$ (which can usually happen in practice where γ is chosen to be close to 1), the parameter c_2 grows exponentially fast with respect to the bootstrapping parameter n. Moreover, since n needs to be chosen large enough for the results in Chen et al. (2021a) to hold, the variance term on the finite-sample bound of the n-step off-policy TD-learning algorithm with linear function approximation is exponentially large.