Finite-Sample Analysis of Off-Policy Natural Actor-Critic Algorithm

Sajad Khodadadian[*1] **Zaiwei Chen**[*2] **Siva Theja Maguluri**[1]

Abstract

In this paper, we provide finite-sample convergence guarantees for an off-policy variant of the natural actor-critic (NAC) algorithm based on Importance Sampling. In particular, we show that the algorithm converges to a global optimal policy with a sample complexity of $\mathcal{O}(\epsilon^{-3}\log^2(1/\epsilon))$ under an appropriate choice of stepsizes. In order to overcome the issue of large variance due to Importance Sampling, we propose the Q-trace algorithm for the critic, which is inspired by the V-trace algorithm (Espeholt et al., 2018). This enables us to explicitly control the bias and variance, and characterize the trade-off between them. As an advantage of off-policy sampling, a major feature of our result is that we do not need any additional assumptions, beyond the ergodicity of the Markov chain induced by the behavior policy.

1. Introduction

Reinforcement Learning (RL) is a paradigm where an agent aims at maximizing its cumulative reward by searching for an optimal policy, in an environment modeled as a Markov Decision Process (MDP) (Sutton & Barto, 2018). RL algorithms have achieved tremendous successes in a wide range of applications such as self-driving cars with Deep Deterministic Policy Gradient (DDPG) (Lillicrap et al., 2015), and AlphaGo in the game of Go (Silver et al., 2016). The algorithms in RL can be categorized into value space methods, such as Q-learning (Watkins & Dayan, 1992), TD-learning (Sutton, 1988), and policy space methods, such as actor-critic (AC) (Konda & Tsitsiklis, 2000). Despite great empirical successes (Wang et al., 2016; Bahdanau et al., 2016), the finite-sample convergence of AC type of algorithms are not completely characterized theoretically.

An AC algorithm can be thought as a generalized policy iteration (Puterman, 1995), and consists of two phases, namely actor and critic. The objective of the actor is to improve the policy, while the critic aims at evaluating the performance of a specific policy. A step of the actor can be thought as a step of Stochastic Gradient Ascent (Bottou et al., 2018) with preconditioning. An identity pre-conditioner corresponds to regular AC, while a pre-conditioning with fisher information results in natural actor-critic (NAC) (Peters & Schaal, 2008). As for the critic, to perform a policy evaluation step, it usually uses the TD-learning method and its variants, such as TD(0), or more generally, n-step TD (Sutton, 1988). Moreover, such learning process can be done in an on or off-policy manner (Degris et al., 2012).

Off-policy Actor-Critic. In on-policy AC, the data samples are generated in an online manner, always sampling based on the current policy at hand. In contrast, in this paper, we focus on the off-policy AC, where the algorithm updates the policy based on the data collected (possibly in the past) by a fixed policy, called the *behavior policy*. Off-policy learning is inevitable in high-stakes applications such as healthcare (Dann et al., 2019), education (Mandel et al., 2014), robotics (Gu et al., 2017) and clinical trials (Liu et al., 2018; Gottesman et al., 2020). The agent there may not have direct access to the environment in order to perform online sampling, and one has to work with limited historical data that is collected under a fixed behavior policy. Moreover, off-policy AC enables off-line learning by decoupling data collection from learning, and is observed to extract the maximum possible utility out of limited available data (Levine et al., 2020).

To account for the difference between the behavior policy and the target policy (Geweke, 1989) in off-policy algorithms, a popular approach is to use Importance Sampling (IS). The IS ratio, however, can be large in some cases, which might result in high variance (Glynn & Iglehart, 1989; Precup, 2000). This phenomenon will be illustrated in detail in Section 2.3. In order to avoid such high variance, one idea is to truncate the IS ratio (Ionides, 2008), which leads to off-policy TD-learning algorithms such as Retrace(λ) (Munos et al., 2016) and V-trace (Espeholt et al., 2018).

[*] Equal contribution [1] School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA [2] PhD Program in Machine Learning, Georgia Institute of Technology, Atlanta, GA, 30332, USA. Correspondence to: Sajad Khodadadian <skhodadadian3@gatech.edu>, Zaiwei Chen <zchen458@gatech.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

Table 1. Summary of the results in the literature [1]

Algorithm	Reference	Sample Complexity [2]	Single trajectory	Comments
AC	(Wang et al., 2019)	$\tilde{\mathcal{O}}(\epsilon^{-6})$	✗	Function Approx: Sample complexity to ensure $\mathbb{E}[\|\nabla V^{\pi_t}\|^2] \leq \epsilon + \mathcal{E}_{\text{bias}}$
	(Qiu et al., 2019)	$\tilde{\mathcal{O}}(\epsilon^{-4})$	✗	
	(Kumar et al., 2019)	$\tilde{\mathcal{O}}(\epsilon^{-4})$	✗	
NAC	(Wang et al., 2019)	$\tilde{\mathcal{O}}(\epsilon^{-14})$	✗	Function Approx: Sample complexity to ensure $V^{\pi^*} - V^{\pi_t} \leq \epsilon + \mathcal{E}_{\text{bias}}$
	(Agarwal et al., 2019)	$\tilde{\mathcal{O}}(\epsilon^{-6})$	✗	
	(Khodadadian et al., 2021)	$\tilde{\mathcal{O}}(\epsilon^{-4})$	✓	Tabular RL: Convergence to global optimum $V^{\pi^*} - V^{\pi_t} \leq \epsilon$
Off-Policy NAC	Our work	$\tilde{\mathcal{O}}(\epsilon^{-3})$	✓	

[1] There are two other related works (Xu et al., 2020a) and (Xu et al., 2020b). (Xu et al., 2020a) claims a sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2})$ for NAC. (Xu et al., 2020b) claims a sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2.5})$ for AC and $\tilde{\mathcal{O}}(\epsilon^{-4})$ for NAC. In our opinion, the interpretation of the convergence results in terms of sample complexity in both papers is incorrect. In case one accepts the interpretation in (Xu et al., 2020a;b), our results imply a sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-1/N})$ for *an arbitrary* $N \in \mathbb{Z}^+$. See Appendix C.1 for a detailed explanation.
[2] In this table, $\tilde{\mathcal{O}}(\cdot)$ ignores all the logarithmic terms. See Appendix C.4 for detailed calculations and comments regarding the sample complexities presented here.

1.1. Main Contributions

In this paper, we study finite-sample convergence guarantees of an off-policy variant of the NAC algorithm. Our main contributions are threefold.

Q-Trace for Off-Policy TD-Learning: Algorithm and Finite-Sample Bounds. To estimate the Q-function for the critic, we propose an off-policy TD-learning algorithm called Q-trace. This is inspired by the V-trace algorithm (Espeholt et al., 2018) to estimate the V-function. We establish the finite-sample convergence bounds of Q-trace, and show how the truncated IS ratios can be used to explicitly trade-off the truncation bias and the variance.

Finite-Sample Bounds for Off-Policy NAC. Based on the Q-trace algorithm for the critic, we propose an off-policy NAC algorithm, which uses only a *single trajectory* of samples. To the best of our knowledge, we establish the first known finite-sample convergence guarantees of an *off-policy* NAC algorithm. Based on that, we show that in order to obtain an ϵ-optimal policy, the amount of samples required is of the size $\mathcal{O}(\epsilon^{-3}\log^2(1/\epsilon))$. This shows that the off-policy NAC outperforms even the best known theoretical convergence bounds of *on-policy* NAC algorithms. See Table 1 for more details.

Exploration through Off-Policy Sampling. While off-policy learning is primarily motivated by practical constraints, in this paper, we demonstrate that off-policy sampling leads to natural exploration. By exploiting off-policy sampling, we do not require either hard-to-verify assumptions made in the literature to ensure exploration (Xu et al., 2020a; Wu et al., 2020), or additional exploration steps in the algorithm that slow down the convergence (Khodadadian et al., 2021).

1.2. Related Work

Two popular algorithms for finding the optimal policy of an MDP are value iteration and policy iteration, which corresponds to Q-learning and AC in Reinforcement Learning when the underlying model is unknown.

The Q-learning algorithm, first proposed in (Watkins & Dayan, 1992) is one of the most celebrated value space methods for solving the RL problem (Sutton & Barto, 2018). Since the proposal, there has been a long line of work to establish the convergence properties of Q-learning. In particular, (Tsitsiklis, 1994; Jaakkola et al., 1994; Bertsekas & Tsitsiklis, 1996; Borkar & Meyn, 2000; Borkar, 2009) characterize the asymptotic convergence of Q-learning, (Beck & Srikant, 2012; 2013; Wainwright, 2019; Chen et al., 2020; 2021) study the finite-sample convergence bound in the mean-square sense, and (Even-Dar & Mansour, 2003; Li et al., 2020; Qu & Wierman, 2020) study the high-probability convergence bounds.

In AC framework, usually the actor uses Policy Gradient (PG) to perform policy update, and the critic uses TD-learning method to perform policy evaluation.

The PG method was shown to converge in (Sutton et al., 1999; Baxter & Bartlett, 2001; Pirotta et al., 2015; Haarnoja et al., 2017). Natural PG, which is a PG method with preconditioning, was proposed in (Kakade, 2001). More recently, there has been a line of work to establish finite-sample convergence bound of (natural) PG algorithm (Even-Dar et al., 2009; Azar et al., 2012; Geist et al., 2019; Agarwal et al., 2019; Wang et al., 2019; Liu et al., 2019; Shani et al., 2020; Mei et al., 2020; Cen et al., 2020; Bhandari & Russo, 2020).

TD-learning method, originally proposed in (Sutton, 1988), represents a family of policy evaluation algorithms in RL. The asymptotic convergence of TD-learning has been established in (Tsitsiklis, 1994; Jaakkola et al., 1994; Borkar & Meyn, 2000). Furthermore, the finite-sample convergence bounds of TD-learning have been studied in (Dalal et al., 2018; Lakshminarayanan & Szepesvari, 2018; Bhandari et al., 2018; Srikant & Ying, 2019) in the on-policy setting. Off-policy variants of TD-learning such as Retrace(λ), Tree-backup, and V-trace were studied in (Munos et al., 2016; Precup, 2000; Espeholt et al., 2018) respectively. Finite-sample bounds for V-trace are quantified in (Chen et al., 2020; 2021).

Actor-critic, as a stochastic variant of policy iteration, was proposed in (Barto et al., 1983; Borkar & Konda, 1997), and later it has extended to function approximation setting (Konda & Tsitsiklis, 2000) and NAC (Peters & Schaal, 2008; Morimura et al., 2009; Thomas et al., 2013; Bhatnagar et al., 2009). Asymptotic convergence of AC algorithms was studied in (Williams & Baird, 1990; Konda & Tsitsiklis, 2000; Borkar & Konda, 1997; Borkar, 2009; Maei, 2018; Zhang et al., 2019; 2020). Furthermore, there has been a flurry of recent work studying the finite-sample convergence of AC and NAC (Qiu et al., 2019; Kumar et al., 2019; Shani et al., 2020; Wang et al., 2019; Xu et al., 2020b;a; Wu et al., 2020; Khodadadian et al., 2021). The results are summarized in Table 1. Concurrent work (Lan, 2021) studies a variant of NAC with on-policy sampling and time-varying inverse temperature, and obtains an $\tilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity.

The rest of this paper is organized as follows. In Section 2, we first present the Q-trace algorithm for off-policy TD-learning. We then use it with the Natural Policy Gradient to get the off-policy NAC algorithm, and present the finite-sample convergence bounds and sample complexity analysis. In Section 3, we present the proof sketch of our main results, and conclude in Section 4.

2. Off-Policy Natural Actor-Critic: Algorithm and Finite-Sample Bounds

2.1. Background on Reinforcement Learning

We model our RL problem with an MDP which consists of a tuple of 5 elements $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma)$. Here $\mathcal{S}$ and $\mathcal{A}$ are finite sets of states and actions, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta^{|\mathcal{S}|}$ (where $\Delta^{|\mathcal{S}|}$ is the probability simplex on $\mathbb{R}^{|\mathcal{S}|}$) is the collection of transition probabilities that are unknown, and $\gamma \in (0, 1)$ is the discount factor.

The dynamics of an MDP is as follows. At each time step k, the system is at some state S_k of the environment. The agent chooses an action A_k based on a policy π at hand, $A_k \sim \pi(\cdot|S_k)$, and the system moves to a new state based on the transition probabilities $\mathbb{P}(S_{k+1} = \cdot|S_k, A_k)$, and induces a one-step reward $\mathcal{R}(S_k, A_k)$. The goal of the agent to find an optimal policy which maximizes the cumulative reward. Specifically, the value function of a policy π is defined by $V^\pi(\mu) = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k \mathcal{R}(S_k, A_k)|S_0 \sim \mu, A_k \sim \pi(\cdot|S_k)]$, where μ is an initial distribution over states. Then the goal is to find an optimal policy π^* such that

$$\pi^* \in \arg\max_{\pi \in \Pi} V^\pi(\mu), \tag{1}$$

where Π represents the set of all policies.

2.2. Natural Policy Gradient

Policy gradient algorithms aim at solving the optimization problem (1) by using gradient ascent or its variants in the policy space. In particular, a Mirror Descent (MD) (Nemirovskij & Yudin, 1983) update of policy with stepsize β reads as:

$$\pi_{t+1} = \arg\max_{\pi \in \Pi} \left\{ \beta \langle \nabla V^{\pi_t}(\mu), \pi - \pi_t \rangle - B(\pi, \pi_t) \right\}, \tag{2}$$

where $B(\cdot, \cdot)$ is an appropriately chosen Bregman divergence between two policies. If we replace the Bregman divergence with $B(\pi, \pi_t) = \sum_s d_\mu^{\pi_t}(s) \mathcal{KL}(\pi(\cdot|s)|\pi_t(\cdot|s))$ in Eq. (2), we get the Natural Policy Gradient (NPG) algorithm for MDPs. Here $d_\mu^\pi(s) = (1 - \gamma) \sum_{j=0}^{\infty} \gamma^j \mathbb{P}^\pi(S_j = s \mid S_0 \sim \mu)$ is the discounted state visitation distribution (Agarwal et al., 2019), and $\mathcal{KL}(\cdot \mid \cdot)$ stands for the KL-Divergence (Cover, 1999). It has been shown in (Agarwal et al., 2019) that the update equation (2) can be equivalently written as

$$\pi_{t+1}(a|s) = \frac{\pi_t(a|s) \exp(\beta Q^{\pi_t}(s, a))}{\sum_{a'} \pi_t(a'|s) \exp(\beta Q^{\pi_t}(s, a'))}, \forall\, s, a, \tag{3}$$

where $Q^\pi(s, a) = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k \mathcal{R}(S_k, A_k)|S_0 = s, A_0 = a]$ is the Q-function for policy π (Puterman, 1995). The update rule (3) can be equivalently derived using the preconditioned gradient ascent (with the Moore–Penrose inverse of the Fisher information matrix as the pre-conditioner) on the dual space of the policy π. This interpretation of (3) was presented in (Kakade, 2001; Agarwal et al., 2019). Furthermore, an interpretation of (3) in terms of Mirror Descent Modified Policy Iteration (MD-MPI) was presented in (Geist et al., 2019). An important result about the NPG is that, although the objective function of (1) is not concave, it has been shown in (Agarwal et al., 2019) that the policies achieved by the MD update of (3) converges to an optimal policy with rate $\mathcal{O}(1/t)$.

Although the convergence result in (Agarwal et al., 2019) is promising to find the optimal policy in an MDP, since we

Algorithm 2.1 Q-Trace

```
**Input:** $K$, $\alpha$, $Q_0$, $\pi$, $\bar{\rho}$, and $\bar{c}$, $\{(S_k, A_k)\}_{0\leq k\leq K+n}$ (generated by the behavior policy $\pi_b$)
**for** $k = 0, 1, \cdots, K-1$ **do**
    $\alpha_k(s,a) = \alpha \mathbb{I}_{\{(s,a)=(S_k,A_k)\}}$ for all $(s,a)$
    $\Delta_{k,i} = \mathcal{R}(S_i, A_i) + \gamma \rho_\pi(S_{i+1}, A_{i+1}) Q_k(S_{i+1}, A_{i+1}) - Q_k(S_i, A_i)$ for all $k \leq i \leq k+n-1$
    $Q_{k+1}(s,a) = Q_k(s,a) + \alpha_k(s,a) \sum_{i=k}^{k+n-1} \gamma^{i-k} \prod_{j=k+1}^{i} c_\pi(S_j, A_j) \Delta_{k,i}$ for all $(s,a)$
**end for**
**Output:** $Q_K$
```

do not have access to the transition probabilities and so the Q-function in RL, we cannot update the policy according to Eq. (3). Natural Actor-Critic (NAC) algorithm, which is a sample-based variant of the update (3), proceeds as follows. In each iteration, first the critic generates an estimate Q_t of the Q-function Q^{π_t}. Then the actor updates the policy according to Eq. (3) with Q^{π_t} replaced by the estimate Q_t.

2.3. The Q-Trace Algorithm for Off-Policy Prediction

In this section, we focus on the critic sub-problem, and develop the Q-trace algorithm to estimate Q^{π_t}. Q-trace is an off-policy variant of TD-learning based on Importance Sampling. Crucially, we introduce two different truncation levels for the IS ratios in order to explicitly control the trade-off between truncation bias and the variance. This is inspired by the V-trace algorithm in (Espeholt et al., 2018).

We next introduce our notations to describe the Q-trace algorithm. Let π be the target policy (i.e., we want to evaluate Q^π) and π_b be the behavior policy (i.e., we use π_b to collect samples). We assume that the behavior policy π_b satisfies $\pi_b(a|s) > 0$ for any (s,a). This is typically necessary in off-policy setting. Let $\bar{\rho}$ and $\bar{c}$ be two truncation levels satisfying $\bar{\rho} \geq \bar{c} \geq 1$. Define $c_\pi(s,a) = \min(\bar{c}, \frac{\pi(a|s)}{\pi_b(a|s)})$ and $\rho_\pi(s,a) = \min(\bar{\rho}, \frac{\pi(a|s)}{\pi_b(a|s)})$ for all (s,a), which are the truncated IS ratios.

The off-policy Q-trace algorithm is presented in Algorithm 2.1. To better understand Algorithm 2.1, consider the following special cases. Suppose we use on-policy sampling, that is, $\pi_b = \pi$. Set $\bar{\rho} = \bar{c} = 1$. Observe that in this case we have $c_\pi(s,a) = \rho_\pi(s,a) = 1$ for all (s,a). Then Algorithm 2.1 reduces to the regular n-step TD, which is known to converge to Q^π (Tsitsiklis, 1994; Sutton & Barto, 2018).

In the off-policy setting (i.e., $\pi_b \neq \pi$), suppose we choose $\bar{\rho} = \bar{c} \geq \max_{(s,a)} \frac{\pi(a|s)}{\pi_b(a|s)}$. Then we have $c_\pi(s,a) = \rho_\pi(s,a) = \frac{\pi(a|s)}{\pi_b(a|s)}$, hence there is essentially no truncation. In this case, Algorithm 2.1 corresponds to the standard n-step TD using off-policy sampling, and therefore converges to Q^π (Precup, 2000).

A fundamental problem in off-policy TD is that the variance in the estimate can be very large or even infinity (Glynn & Iglehart, 1989; Munos et al., 2016). This is mainly because of the product of IS ratios $\prod_{j=k+1}^{i} \frac{\pi(A_j|S_j)}{\pi_b(A_j|S_j)}$. To have control on the variance of the estimate, we introduce the truncation levels $\bar{\rho}$ and $\bar{c}$. However, due to the truncation, the IS ratios are now biased, and hence the algorithm no longer converges to the target value function Q^π. In fact, Algorithm 2.1 converges to a biased limit point, denoted by $Q^{\bar{\rho},\pi}$, which need not necessarily be the value function of any policy.

Importantly, the limit point $Q^{\bar{\rho},\pi}$ depends only on the target policy π and the truncation level $\bar{\rho}$, but not on the truncation level $\bar{c}$. Therefore, we can heavily truncate the IS ratio $c_\pi(s,a)$ by using small $\bar{c}$ without affecting the limit point of the Q-trace algorithm. In fact, as we will see in Section 2.5, this is exactly what we should do. To quantify the truncation bias of Algorithm 2.1, we have the following result.

Lemma 2.1. *For any $\bar{\rho} \geq 1$ and policy π, we have*

(1) $\|Q^{\bar{\rho},\pi} - Q^\pi\|_\infty \leq \frac{\max_{(s,a)} \max(\pi(a|s) - \bar{\rho}\pi_b(a|s), 0)}{(1-\gamma)^2}$

(2) $\|Q^{\bar{\rho},\pi}\|_\infty \leq \frac{1}{1-\gamma}$

Observe from Lemma 2.1 (1) that when $\bar{\rho} \geq \max_{s,a} \frac{\pi(a|s)}{\pi_b(a|s)}$, we have $Q^{\bar{\rho},\pi} = Q^\pi$. This makes intuitive sense in that when $\bar{\rho}$ is large, there is essentially no truncation in the IS ratio $\rho_\pi(s,a)$, and we should not expect any truncation bias.

Comparison to Related Algorithms. There are two algorithms in the literature that are closely related to our Q-trace algorithm, namely the Retrace(λ) in (Munos et al., 2016) and the V-trace in (Espeholt et al., 2018). The Retrace(λ) algorithm in (Munos et al., 2016) is proposed to evaluate the Q-function, but uses a single truncation level. In contrast, we have two truncation levels $\bar{c}$ and $\bar{\rho}$, which enables us to trade-off the truncation bias and variance.

V-trace, an off-policy variant of TD to estimate the V-function, first introduced the idea of using two truncation levels. However, there are several differences between Q-trace and V-trace. First, the product of the IS ratios $c_\pi(S_j, A_j)$ starts from $j = k+1$ rather than $j = k$ in V-trace. This simple but important modification enables us to get a convergence bound in Theorem 2.1 which does not dependent on the target policy π. This is essential for us to use the Q-trace algorithm in the AC framework, as after each iteration of the actor, the critic receives a different policy π_t to evaluate. Second, as opposed to V-trace, where the limit point is a value function of some policy, the limit point $Q^{\bar{\rho},\pi}$ of Q-trace is not necessarily the Q-function of any policy. Finally, due to the structure of the Q-function, the IS ratio $\rho_\pi(S_{i+1}, A_{i+1})$ is multiplied with only one of the three terms in the temporal difference $\Delta_{k,i}$ (Algorithm 2.1 line 4), as opposed to all the three terms in V-trace.

In summary, we propose the off-policy Q-trace algorithm to evaluate the Q-function in the critic. Moreover, the flexibility of choosing the truncation levels in Q-trace enables us to explicitly trade-off the truncation bias and the variance.

2.4. Off-Policy Natural Actor-Critic Algorithm

We are now ready to present our off-policy NAC algorithm 2.2. In iteration t, the critic first estimates the Q-function Q^{π_t} using the Q-trace algorithm, which itself runs over K iterations. Then the actor uses the estimate Q_{t+1} in Eq. (3) to perform a policy update. Thus, we have a two-loop algorithm.

Algorithm 2.2 Off-Policy Natural Actor-Critic

```
Input: $T$, $K$, $\alpha$, $\beta$, $Q_0 = \mathbf{0}$, $\pi_0$, $\bar{\rho}$, $\bar{c}$, and $\{(S_k, A_k)\}_{0\leq k\leq T(K+n)}$ (a single trajectory generated by the behavior policy $\pi_b$)
for $t = 0, 1, \cdots, T-1$ do
    Critic update:
    DataSet $= \{(S_i, A_i)\}_{t(K+n)\leq i\leq (t+1)(K+n)}$
    $Q_{t+1} = Q$-Trace$(K, \alpha, Q_0, \pi_t, \bar{c}, \bar{\rho}$, DataSet$)$
    Actor update:
    $\pi_{t+1}(a|s) = \frac{\pi_t(a|s)\exp(\beta Q_{t+1}(s,a))}{\sum_{a'}\pi_t(a'|s)\exp(\beta Q_{t+1}(s,a'))}\ \forall\,(s,a)$
end for
Output: $\{\pi_t\}_{0\leq t\leq T-1}$
```

In Algorithm 2.2, due to off-policy sampling, the sampling process and the learning process are decoupled, which allows the agent to learn in an off-line manner (Levine et al., 2020). Moreover, note that we are using a *single trajectory* of samples $\{(S_k, A_k)\}_{0\leq k\leq T(K+n)}$ to perform the update. In related literature (Xu et al., 2020b;a; Wang et al., 2019), sampling needs to be often restarted with an arbitrary initial state, which is not practical in many real-world applications. See Appendix C.2 for more details.

2.5. Finite-Sample Convergence Guarantees

In this section, we present our main results about the finite-sample convergence bounds of the Q-trace algorithm 2.1 for off-policy TD-learning, and the off-policy NAC Algorithm 2.2. We begin by stating our one and only assumption.

Assumption 2.1. The Markov chain $\{S_k\}$ induced by the behavior policy π_b is irreducible and aperiodic.

Assumption 2.1 is commonly made in related work about RL algorithms with Markovian sampling (Tsitsiklis & Van Roy, 1997; 1999; Maei, 2018; Zhang et al., 2020), and it implies that the Markov chain $\{S_k\}$ has a unique stationary distribution $\mu_b \in \Delta^{|\mathcal{S}|}$. Moreover, since the state space $\mathcal{S}$ is finite, there exist $C > 0$ and $u \in (0,1)$ such that

$$\|P^k(s,\cdot) - \mu_b(\cdot)\|_{\text{TV}} \leq Cu^k$$

for any $k \geq 0$ and $s \in \mathcal{S}$, where $\|\cdot\|_{\text{TV}}$ is the total variation distance (Levin & Peres, 2017).

A major issue in the design of AC algorithms is to ensure enough exploration to all state-action pairs (s,a). It was demonstrated in (Khodadadian et al., 2021) that the algorithm can get stuck in a local optimum if there is not enough exploration. Sampling from a fixed policy that leads to an ergodic Markov chain naturally ensures exploration, and so we do not need any additional assumptions. In contrast, prior literature on the analysis of on-policy AC either makes additional assumptions that are hard to satisfy (Xu et al., 2020a; Wu et al., 2020) or introduce an additional exploration step in the algorithm (Khodadadian et al., 2021) that slows the convergence. See Appendix C.3 for more details.

To state our result, we need the following notation. Let $\tau_\alpha = \min\{k \geq 0 : \max_{s\in\mathcal{S}} \|P^k(s,\cdot) - \mu_b(\cdot)\|_{\text{TV}} \leq \alpha\}$, where α is the constant stepsize used in the critic step of Algorithm 2.2. The quantity τ_α can be viewed as the *mixing time* of the Markov chain $\{S_k\}$ with accuracy α. Furthermore, under the geometric mixing property (implied by Assumption 2.1), the mixing time τ_α can be bounded by $L(\log(1/\alpha)+1)$ for some constant $L > 0$. Let $f(\bar{c},\gamma) = \frac{1-(\gamma\bar{c})^n}{1-\gamma\bar{c}}$ when $\gamma\bar{c} \neq 1$, and $= n$ when $\gamma\bar{c} = 1$. Suppose the constant stepsize α within the critic is properly chosen. The explicit condition is given in Appendix A.3. Then we have the following result.

Theorem 2.1. *Consider $\{Q_k\}$ of Algorithm 2.1. Suppose that (1) Assumption 2.1 is satisfied, (2) Q_0 is initiated at $\mathbf{0}$, and (3) the constant stepsize α is chosen such that $\alpha(\tau_\alpha + n + 1) \leq \min\left(\frac{1}{12(\bar{\rho}+1)f(\bar{c},\gamma)}, \frac{(1-\gamma_c)^2}{8208(\bar{\rho}+1)^2 f(\bar{c},\gamma)^2\log(|\mathcal{S}||\mathcal{A}|)}\right)$, where $\gamma_c \in (0,1)$ (defined in Proposition 3.1 (3) (b)) does not depend on the target policy π, Then we have for all $k \geq \tau_\alpha + n + 1$:*

$$\mathbb{E}[\|Q_k - Q^{\bar{\rho},\pi}\|_\infty^2] \leq \underbrace{\frac{c_1}{(1-\gamma)^2}\left(1-\frac{1-\gamma_c}{2}\alpha\right)^{k-(\tau_\alpha+n+1)}}_{T_1\text{:Convergence Bias}}$$

$$+ \underbrace{\frac{c_2\log(|\mathcal{S}||\mathcal{A}|)}{(1-\gamma_c)^2(1-\gamma)^2}(\bar{\rho}+1)^2 f(\bar{c},\gamma)^2\alpha(\tau_\alpha+n+1)}_{T_2\text{:Convergence Variance}},$$

where c_1 and c_2 are numerical constants.

Observe that the whole RHS of the convergence bound does not depend on the target policy π. This is important for us to later use Theorem 2.1 to show the finite-sample guarantees of off-policy NAC algorithm 2.2.

This result characterizes the rate of convergence of Q-trace algorithm to its stationary point, $Q^{\bar{\rho},\pi}$. The error on the RHS has two terms, which are called bias and variance respectively in the SA literature (Chen et al., 2020). To contrast this with the bias due to truncation, we call it the convergence bias. The second error term is simply called the variance. Theorem 2.1 implies that under an appropriate constant stepsize α, while the Q-trace algorithm achieves exponentially decaying convergence bias, it leads to a constant variance that cannot be eliminated, and is of the size $\mathcal{O}(\alpha\log(1/\alpha))$. The logarithmic factor is due to the mixing time τ_α, which arises as a consequence of performing Markovian sampling of $\{(S_k, A_k)\}$.

The following corollary provides the error of the estimate Q_k with respect to the true Q-function Q^π.

Corollary 2.1.1. *Under the same assumptions of Theorem 2.1, we have for all $k \geq \tau_\alpha + n + 1$:*

$$\mathbb{E}\left[\|Q_k - Q^\pi\|_\infty\right] \leq \sqrt{T_1} + \sqrt{T_2} + \frac{\max(1-\bar{\rho}\min_{s,a}\pi_b(a|s),0)}{(1-\gamma)^2},$$

where the terms T_1 and T_2 are given in Theorem 2.1.

The proof of Corollary 2.1.1 immediately follows by combining Lemma 2.1 with Theorem 2.1 and using Jensen's inequality. We next present the finite-sample performance bound of the off-policy NAC algorithm 2.2.

Theorem 2.2. *Consider $\{\pi_t\}$ generated by Algorithm 2.2. Suppose that Assumption 2.1 is satisfied, and $K \geq \tau_\alpha + n + 1$. Then we have the following performance bound:*

$$\begin{aligned} & V^{\pi^*}(\mu) - \max_{0\leq t\leq T-1}\mathbb{E}\left[V^{\pi_t}(\mu)\right] \\ \leq\; & \underbrace{\frac{24}{(1-\gamma)^3}\left(1-\frac{1-\gamma_c}{2}\alpha\right)^{\frac{1}{2}(K-(\tau_\alpha+n+1))}}_{E_1:\text{Convergence bias in the Critic}} \\ & + \underbrace{\frac{1200\log^{1/2}(|\mathcal{S}||\mathcal{A}|)}{(1-\gamma)^3(1-\gamma_c)}(\bar{\rho}+1)f(\bar{c},\gamma)[\alpha(\tau_\alpha+n+1)]^{1/2}}_{E_2:\text{Variance in the Critic}} \\ & + \underbrace{\frac{4\max(0, 1-\bar{\rho}\min_{s,a}\pi_b(a|s))}{(1-\gamma)^4}}_{E_3:\text{Truncation bias}} \\ & + \underbrace{\frac{\log(e|\mathcal{A}|)}{(1-\gamma)^2\beta T}}_{E_4:\text{Convergence error in the Actor}}, \end{aligned}$$

The terms E_1 and E_2 correspond to the two terms on the RHS of the convergence bounds in Theorem 2.1, and capture the convergence bias and the variance in the critic estimate. We now focus on the terms E_3 and E_4, and the trade-off between the variance E_2 and the truncation bias E_3.

Error Due to Truncated IS Ratio. The term E_3 accounts for the error due to introducing the truncation level $\bar{\rho}$ in the critic (i.e., the Q-trace Algorithm 2.1). Recall that because of $\bar{\rho}$, the limit point of the critic is $Q^{\bar{\rho},\pi_t}$ instead of Q^{π_t}. Note that when $\bar{\rho} \geq \frac{1}{\min_{s,a}\pi_b(a|s)}$ (which implies $\bar{\rho} \geq \max_{s,a}\frac{\pi_t(a|s)}{\pi_b(a|s)}$ for any t), there is essentially no truncation in the IS ratio $\rho_{\pi_t}(s,a)$, and hence we have $E_3 = 0$, which agrees with Lemma 2.1.

Error Bound of the Actor. The term E_4 is due to the error in the actor update. That is, E_4 would be the only error term we have if we can directly use Q^{π_t} in the actor update of Algorithm 2.2. Observe that $E_4 = \mathcal{O}(\frac{1}{T})$, which agrees with results in (Agarwal et al., 2019) [Theorem 5.3].

Bias-Variance Trade-Off. Recall that the motivation for introducing the truncation levels $\bar{\rho}$ and $\bar{c}$ is to control the variance in the critic estimate. We first consider the impact of $\bar{\rho}$. Observe that the term E_3 is in favor of large $\bar{\rho}$ while the term E_2 grows linearly with respect to $\bar{\rho}$. Therefore, there is an explicit trade-off between the variance and the truncation bias in choosing $\bar{\rho}$. As a result, if we want to have convergence to the global optimal, by choosing $\bar{\rho} = \frac{1}{\min_{s,a}\pi_b(a|s)}$, we introduce an additional $\frac{1}{\min_{s,a}\pi_b(a|s)}$ factor in the variance term E_2.

The truncation level $\bar{c}$ appears only in the variance term E_2. In view of the expression of $f(\bar{c},\gamma)$ (defined before Theorem 2.1), we should choose $\bar{c}$ such that $\bar{c}\gamma < 1$ to avoid an exponential factor in the variance term. These observations are similar to (Espeholt et al., 2018; Chen et al., 2020; 2021), where the V-trace algorithm is studied.

One drawback with Theorem 2.2 is that the performance bound is stated in terms of $\max_{0\leq t\leq T-1}\mathbb{E}[V^{\pi_t}(\mu)]$, while in practice we do not know which policy among $\{\pi_t\}_{0\leq t\leq T-1}$ has the best performance. To overcome this problem, using standard techniques in optimization (Lan, 2020), we can obtain the following refined performance bound of Algorithm 2.2.

Corollary 2.2.1. *Let T' be a random sample uniformly drawn from $\{0,1,...,T-1\}$. Then we have the following performance guarantee on $\pi_{T'}$:*

$$V^{\pi^*}(\mu) - \mathbb{E}\left[V^{\pi_{T'}}(\mu)\right] \leq E_1 + E_2 + E_3 + E_4,$$

where the terms $\{E_i\}_{1\leq i\leq 4}$ are given in Theorem 2.2.

The convergence guarantee in in Corollary 2.2.1 holds for the policy attained by Algorithm 2.2 at a random point between 0 and $T-1$. However, in practice one usually takes the last policy achieved by the algorithm as the output. Numerical experiments of off-policy NAC algorithm 2.2 in Figure 1 shows that in expectation, the algorithm can converges almost monotonically. Theoretically showing a performance bound for $V^{\pi^*}(\mu) - \mathbb{E}\left[V^{\pi_{T-1}}(\mu)\right]$ is a future direction of this work.

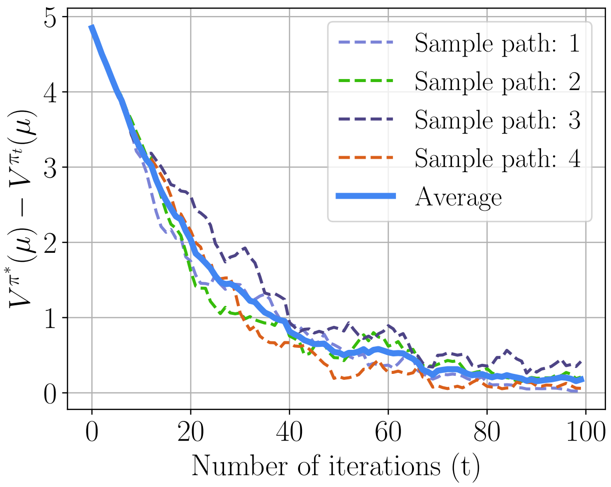

Figure 1. Convergence of Algorithm 2.2 on a 5 state, 3 action MDP. Each dashed line is for one sample path of the algorithm, and the solid line is the average of the 4 sample paths. See Appendix D for more details.

2.6. Sample Complexity Analysis

With Theorem 2.2 at hand, we now analyze the sample complexity of off-policy NAC algorithm 2.2.

Sample Complexity for Global Optimum. Suppose that $\bar{\rho} \geq 1/\pi_{b,\min}$, where $\pi_{b,\min} := \min_{s,a} \pi_b(a|s)$. In this case, we have $E_3 = 0$, i.e., the bias due to truncation is eliminated, and hence we have convergence to a global optimum. Theorem 2.2 implies the following sample complexity result, whose proof is presented in Appendix B.3.

Corollary 2.2.2. *In order to obtain an ϵ-optimal policy, the total number of samples required (i.e., TK) is of the size*

$$\mathcal{O}(\epsilon^{-3}\log^2(1/\epsilon))\tilde{\mathcal{O}}((1-\gamma)^{-11}M_{\min}^{-3}\pi_{b,\min}^{-2}),$$

where in $\tilde{\mathcal{O}}(\cdot)$ we ignore all logarithmic terms, and $M_{\min} = \min_{s,a}\mu_b(s)\pi_b(a|s)$.

The $\mathcal{O}(\epsilon^{-3}\log^2(1/\epsilon))$ dependence on the accuracy ϵ advances the state of the art results in on-policy NAC. See Table 1 for more details. The dependence on the state-action space is at least $|\mathcal{S}|^3|\mathcal{A}|^5$, which is achieved when $\pi_b(a|s) = \frac{1}{|\mathcal{A}|}$ for all a and $\mu_b(s) = \frac{1}{|\mathcal{S}|}$ for all s (i.e., uniform exploration). The $\tilde{\mathcal{O}}((1-\gamma)^{-11})$ dependence on the discount factor while seemingly loose, agrees with known results about NPG in (Agarwal et al., 2019) (Corollary 6.3). See Appendix C.4 for more details about the comparison to (Agarwal et al., 2019).

Note that in off-policy TD-learning, one set of samples can be used multiple times to evaluate different policies. Therefore, it is natural to consider repeatedly using the same set of samples in the critic (the Q-trace algorithm) in the off-policy NAC algorithm. In that case, the sample complexity is reduced from $KT = \tilde{\mathcal{O}}(\epsilon^{-3})$ to only $K = \tilde{\mathcal{O}}(\epsilon^{-2})$. Although this approach seems reasonable, numerical experiments suggest that it may lead to the divergence of Algorithm 2.2. See Appendix D for more details.

3. Proof Sketch of Our Main Results

In this section, we present the key steps in proving Theorems 2.1 and 2.2.

3.1. Proof Sketch of Theorem 2.1

To prove Theorem 2.1, we begin by introducing some notations. For any $k \geq 0$, let $X_k = (S_k, A_k, ..., S_{k+n})$. It is clear that $\{X_k\}$ is a Markov chain, whose state-space is denoted by $\mathcal{X}$. Moreover, under Assumption 2.1, the Markov chain $\{X_k\}$ has a unique stationary distribution, denoted by μ_X. Let $\mathcal{T} : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \times \Pi \times \mathcal{X} \mapsto \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be an operator defined by $[\mathcal{T}(Q,\pi,x)](s,a) = [\mathcal{T}(Q,\pi,s_0,a_0,...,s_n)](s,a) = \mathbb{I}_{(s,a)=(s_0,a_0)}\sum_{i=0}^{n-1}\gamma^i\prod_{j=1}^{i}c_\pi(s_j,a_j)(\mathcal{R}(s_i,a_i) + \gamma\rho_\pi(s_{i+1},a_{i+1})Q(s_{i+1},a_{i+1}) - Q(s_i,a_i)) + Q(s,a)$ for all (s,a). We further define $\mathcal{T}_e : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \times \Pi \mapsto \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ by $\mathcal{T}_e(Q,\pi) = \mathbb{E}_{X\sim\mu_X}\mathcal{T}(Q,\pi,X)$, which can be viewed as the expected version of the operator $\mathcal{T}$.

Using the notation given above, the Q-trace update equation (Algorithm 2.1 line 5) can be equivalently written by

$$\begin{aligned} Q_{k+1} &= Q_k + \alpha(\mathcal{T}(Q_k,\pi,X_k) - Q_k) && (4)\\ &= Q_k + \alpha(\mathcal{T}_e(Q_k,\pi) - Q_k) \\ &\quad + \alpha(\mathcal{T}(Q_k,\pi,X_k) - \mathcal{T}_e(Q_k,\pi)) && (*) \end{aligned}$$

The above update equation can be viewed as a stochastic approximation algorithm for solving the fixed-point equation $\mathcal{T}_e(Q,\pi) = Q$ with Markovian noise. To see this, assume for the moment that the term $(*)$ is identically zero. Then the Algorithm is the fixed-point iteration for solving the equation $\mathcal{T}_e(Q,\pi) = Q$, and it is known to converge when the operator $\mathcal{T}_e(\cdot,\pi)$ is a *contraction mapping* (Banach, 1922). Now in the presence of the term $(*)$, the algorithm becomes a Markovian stochastic approximation algorithm for solving $\mathcal{T}_e(Q,\pi) = Q$.

Intuitively, once we show the desired contraction property of the operator $\mathcal{T}_e(\cdot,\pi)$ and have control on the error caused by the Markovian noise $(*)$, we should be able to establish the convergence bounds of Algorithm (4). In order to show such properties, we need the following notation.

(1) Let $\pi_{\bar{c}}$ and $\pi_{\bar{\rho}}$ be two policies defined by

$$\pi_{\bar{c}}(a|s) = \frac{\min(\bar{c}\pi_b(a|s), \pi(a|s))}{\sum_{a'}\min(\bar{c}\pi_b(a'|s), \pi(a'|s))} \text{ and}$$

$$\pi_{\bar{\rho}}(a|s) = \frac{\min(\bar{\rho}\pi_b(a|s), \pi(a|s))}{\sum_{a'} \min(\bar{\rho}\pi_b(a'|s), \pi(a'|s))}, \ \forall\, (s,a).$$

(2) Let $C_\pi, D_\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ be diagonal matrices s.t. $C_\pi((s,a),(s,a)) = \sum_a \min(\bar{c}\pi_b(a|s), \pi(a|s))$ and $D_\pi((s,a),(s,a)) = \sum_a \min(\bar{\rho}\pi_b(a|s), \pi(a|s))$ for all (s,a). Let $C_{\min} = \bar{c}\min_{s,a}\pi_b(a|s)$. Note that we have $C_{\min} I \leq C_\pi \leq D_\pi \leq I$ (component-wise).

(3) Let $P_\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ be a stochastic matrix defined by $P_\pi((s,a),(s',a')) = P_a(s,s')\pi(a'|s')$, i.e., the probability of transition from (s,a) to (s',a') under policy π. Let R be a vector in $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ such that $R(s,a) = \mathcal{R}(s,a)$ for all (s,a).

(4) Let $M \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ be a diagonal matrix such that $M((s,a),(s,a)) = \mu_b(s)\pi_b(a|s)$, which is the steady-state probability of visiting (s,a). Let $M_{\min} = \min_{s,a}\mu_b(s)\pi_b(a|s)$. Note that $0 < M_{\min} < 1$ under Assumption 2.1.

Now we are ready to establish the desired properties of Algorithm (4) in the following proposition, whose proof is presented in Appendix A.1.

Proposition 3.1. *The following properties hold regarding the operators $\mathcal{T}(\cdot)$, $\mathcal{T}_e(\cdot)$, and the Markov chain $\{X_k\}$.*

(1) The operator $\mathcal{T}(\cdot)$ satisfies $\|\mathcal{T}(Q_1,\pi,x) - \mathcal{T}(Q_2,\pi,x)\|_\infty \leq 2(\bar{\rho}+1)f(\bar{c},\gamma)\|Q_1 - Q_2\|_\infty$ and $\|\mathcal{T}(\mathbf{0},\pi,x)\|_\infty \leq f(\bar{c},\gamma)$ for any $Q_1, Q_2 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $\pi \in \Pi$, and $x \in \mathcal{X}$.

(2) For all $k \geq 0$, it holds that

$$\max_{x\in\mathcal{X}} \|P^{k+n+1}(x,\cdot) - \mu_X(\cdot)\|_{TV} \leq Cu^k,$$

where $\|\cdot\|_{TV}$ is the total variation distance.

(3) The operator $\mathcal{T}_e(\cdot)$ has the following properties:

(a) $\mathcal{T}_e(\cdot,\pi)$ is a linear operator given by $\mathcal{T}_e(Q,\pi) = AQ+b$, where $A = I - \sum_{i=0}^{n-1}\gamma^i M(P_{\pi_{\bar{c}}}C_\pi)^i(I - \gamma P_{\pi_{\bar{\rho}}}D_\pi)$ and $b = \sum_{i=0}^{n-1}\gamma^i M(P_{\pi_{\bar{c}}}C_\pi)^i R$.

(b) $\mathcal{T}_e(\cdot,\pi)$ is a ***contraction mapping*** *with respect to $\|\cdot\|_\infty$, with contraction factor*

$$\gamma_c = 1 - \frac{M_{\min}(1-\gamma)(1-(\gamma C_{\min})^n)}{1-\gamma C_{\min}}.$$

(c) $\mathcal{T}_e(\cdot,\pi)$ has a unique fixed-point $Q^{\bar{\rho},\pi}$, which is the unique solution to the modified Bellman's equation $Q = R + \gamma P_{\pi_{\bar{\rho}}}D_\pi Q$.

Several remarks are in order. First, using Proposition 3.1 (1), we have by triangle inequality that

$$\|\mathcal{T}(Q,\pi,x)\|_\infty \leq 2f(\bar{c},\gamma)((\bar{\rho}+1)\|Q\|_\infty + 1)$$

for any Q, π and x. This is important to control the Markovian noise as it implies that the noisy operator $\|\mathcal{T}(Q_k,\pi,X_k)\|_\infty$ is at most an affine function of $\|Q_k\|_\infty$.

Proposition 3.1 (2) implies that the Markov chain $\{X_k\}$ mixes geometrically fast, which is also an important property we need to control the Markovian noise.

Proposition 3.1 (3) establishes all the desired properties for the expected operator $\mathcal{T}_e(\cdot)$. First of all, $\mathcal{T}_e(\cdot,\pi)$ is a contraction operator, with a contraction factor γ_c independent of the target policy π. This uniform contraction property is necessary for us to combine the critic with the actor later in Section 3.2.2, as the policy π_t is time-varying.

Note that from Proposition 3.1 (3) (c) we see that when $\bar{\rho} \geq \max_{s,a}\frac{\pi(a|s)}{\pi_b(a|s)}$, such modified Bellman's equation becomes the regular Bellman's equation for Q^π, and hence we have $Q^{\bar{\rho},\pi} = Q^\pi$, which agrees with Lemma 2.1.

The above proposition enables us to interpret Eq. (4) as a Markovian Stochastic Approximation involving a contraction mapping. Theorem 2.1 then follows from using finite-sample bounds on Markovian Stochastic Approximation established in (Chen et al., 2021). See Appendix A.3 for the detailed proof.

3.2. Proof Sketch of Theorem 2.2

The high level idea of proving Theorem 2.2 is as follows. We first analyze the iterates $\{\pi_t\}$ updated by the actor in Algorithm 2.2. The performance bound of π_t would involve the error in the critic estimate, i.e., the difference between Q_{t+1} and Q^{π_t}. We then use Corollary 2.1.1 of the Q-trace algorithm 2.1 to control the critic estimation error and finish the proof of Theorem 2.2.

3.2.1. Analysis of the Actor

By analyzing the update of the actor, we obtain the performance bound of $\{\pi_t\}$ in the following proposition.

Proposition 3.2. *Consider iterates $\{\pi_t\}$ of Algorithm 2.2. We have for any $T \geq 1$:*

$$V^{\pi^*}(\mu) - \max_{0\leq t\leq T-1}\mathbb{E}\left[V^{\pi_t}(\mu)\right]$$
$$\leq \underbrace{\frac{\log(e|\mathcal{A}|)}{(1-\gamma)^2\beta T}}_{\text{Error in the actor}} + \underbrace{\frac{4}{(1-\gamma)^2 T}\sum_{t=0}^{T-1}\mathbb{E}[\|Q^{\pi_t} - Q_{t+1}\|_\infty]}_{\text{Error in the Critic}}.$$

The proof of Proposition 3.2 is inspired by that of Theorem 5.3 in (Agarwal et al., 2019), and is presented in Appendix B.1. The main difference is that in (Agarwal et al., 2019) they assume access to the dynamics of the underlying MDP. Hence they can directly use the Q-function Q^{π_t} in the policy update. Here in the RL setting, we can only use the noisy estimate Q_t to perform the policy update. As a consequence, when compared to Theorem 5.3 of (Agarwal et al., 2019), we have the critic error term $\frac{4}{(1-\gamma)^2 T}\sum_{t=0}^{T-1}\mathbb{E}[\|Q^{\pi_t} - Q_{t+1}\|_\infty]$ on the RHS of the resulting inequality of Proposition 3.2.

3.2.2. Combining the Actor and the Critic

In view of Proposition 3.2, what remains to do in proving Theorem 2.2 is to apply Corollary 2.1.1 to control the error term $\mathbb{E}[\|Q^{\pi_t} - Q_{t+1}\|_\infty]$ for any $0 \leq t \leq T-1$. However, there is a challenge in doing this. Corollary 2.1.1 and Theorem 2.1 are stated for a fixed target policy π, while in Algorithm 2.2 the policies π_t are stochastic. We overcome this challenge by using a conditioning argument and exploiting Markovian nature of the samples. The full details are presented in Appendix B.2.

4. Conclusion and Future Work

In this work, we study the convergence bounds of NAC, where the critic uses the Q-trace algorithm to perform off-policy learning. Such off-policy NAC algorithm enables us to overcome the difficulty of exploration in on-policy NAC, and establish the convergence bounds under minimal assumptions. A future direction is to extend our results to the case where function approximation is used. Note that off-policy TD with function approximation can be unstable in general (Sutton & Barto, 2018). The first step in this direction is to modify the algorithm to achieve convergence.

Acknowledgments

This work was partially supported by NSF Grant CIF-1944993, NSF Grant CCF-1740776, and an award from Raytheon Technologies. Maguluri acknowledges seed funding from Georgia Institute of Technology.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Preprint arXiv:1908.00261*, 2019.

Azar, M. G., Gómez, V., and Kappen, H. J. Dynamic policy programming. *The Journal of Machine Learning Research*, 13(1):3207–3245, 2012.

Bahdanau, D., Brakel, P., Xu, K., Goyal, A., Lowe, R., Pineau, J., Courville, A., and Bengio, Y. An actor-critic algorithm for sequence prediction. *Preprint arXiv:1607.07086*, 2016.

Banach, S. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fund. math*, 3(1):133–181, 1922.

Barto, A. G., Sutton, R. S., and Anderson, C. W. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5):834–846, 1983. doi: 10.1109/TSMC.1983.6313077.

Baxter, J. and Bartlett, P. L. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

Beck, C. L. and Srikant, R. Error bounds for constant step-size Q-learning. *Systems & control letters*, 61(12):1203–1208, 2012.

Beck, C. L. and Srikant, R. Improved upper bounds on the expected error in constant step-size Q-learning. In *2013 American Control Conference*, pp. 1926–1931. IEEE, 2013.

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

Bhandari, J. and Russo, D. A note on the linear convergence of policy gradient methods. *Preprint arXiv:2007.11120*, 2020.

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

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

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

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

Borkar, V. S. and Meyn, S. P. The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. Fast global convergence of natural policy gradient methods with entropy regularization. *Preprint arXiv:2007.06558*, 2020.

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. Finite-Sample Analysis of Contractive Stochastic Approximation Using Smooth Convex Envelopes. *Advances in Neural Information Processing Systems*, 33, 2020.

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

Cover, T. M. *Elements of information theory*. John Wiley & Sons, 1999.

Dalal, G., Szörényi, B., Thoppe, G., and Mannor, S. Finite sample analysis for TD(0) with function approximation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

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

Degris, T., White, M., and Sutton, R. Off-Policy Actor-Critic. In *International Conference on Machine Learning*, 2012.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., et al. IMPALA: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International Conference on Machine Learning*, pp. 1407–1416, 2018.

Even-Dar, E. and Mansour, Y. Learning rates for Q-learning. *Journal of Machine Learning Research*, 5(Dec):1–25, 2003.

Even-Dar, E., Kakade, S. M., and Mansour, Y. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Geist, M., Scherrer, B., and Pietquin, O. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pp. 2160–2169. PMLR, 2019.

Geweke, J. Bayesian inference in econometric models using monte carlo integration. *Econometrica: Journal of the Econometric Society*, pp. 1317–1339, 1989.

Glynn, P. W. and Iglehart, D. L. Importance sampling for stochastic simulations. *Management science*, 35(11): 1367–1392, 1989.

Gottesman, O., Futoma, J., Liu, Y., Parbhoo, S., Celi, L., Brunskill, E., and Doshi-Velez, F. Interpretable off-policy evaluation in reinforcement learning by highlighting influential transitions. In *International Conference on Machine Learning*, pp. 3658–3667. PMLR, 2020.

Gu, S., Holly, E., Lillicrap, T., and Levine, S. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 3389–3396. IEEE, 2017.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pp. 1352–1361. PMLR, 2017.

Ionides, E. L. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

Jaakkola, T., Jordan, M. I., and Singh, S. P. Convergence of stochastic iterative dynamic programming algorithms. In *Advances in neural information processing systems*, pp. 703–710, 1994.

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

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

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

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

Lakshminarayanan, C. and Szepesvari, C. Linear Stochastic Approximation: How Far Does Constant Step-Size and Iterate Averaging Go? In *International Conference on Artificial Intelligence and Statistics*, pp. 1347–1355, 2018.

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

Lan, G. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *arXiv preprint arXiv:2102.00135*, 2021.

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

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

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. Sample Complexity of Asynchronous Q-Learning: Sharper Analysis and Variance Reduction. *Preprint arXiv:2006.03041*, 2020.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *Preprint arXiv:1509.02971*, 2015.

Liu, B., Cai, Q., Yang, Z., and Wang, Z. Neural proximal/trust region policy optimization attains globally optimal policy. *Preprint arXiv:1906.10306*, 2019.

Liu, Y., Gottesman, O., Raghu, A., Komorowski, M., Faisal, A. A., Doshi-Velez, F., and Brunskill, E. Representation Balancing MDPs for Off-policy Policy Evaluation. *Advances in Neural Information Processing Systems*, 31: 2644–2653, 2018.

Maei, H. R. Convergent actor-critic algorithms under off-policy training and function approximation. *arXiv preprint arXiv:1802.07842*, 2018.

Mandel, T., Liu, Y.-E., Levine, S., Brunskill, E., and Popovic, Z. Offline policy evaluation across representations with applications to educational games. In *AAMAS*, pp. 1077–1084, 2014.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pp. 6820–6829. PMLR, 2020.

Morimura, T., Uchibe, E., Yoshimoto, J., and Doya, K. A generalized natural actor-critic algorithm. In *Advances in neural information processing systems*, pp. 1312–1320, 2009.

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

Nemirovskij, A. S. and Yudin, D. B. Problem complexity and method efficiency in optimization. *Chichester: John Wiley*, 1983.

Peters, J. and Schaal, S. Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190, 2008.

Pirotta, M., Restelli, M., and Bascetta, L. Policy gradient in lipschitz markov decision processes. *Machine Learning*, 100(2):255–283, 2015.

Precup, D. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, pp. 80, 2000.

Puterman, M. L. Markov decision processes: Discrete stochastic dynamic programming. *Journal of the Operational Research Society*, 46(6):792–792, 1995.

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

Qu, G. and Wierman, A. Finite-Time Analysis of Asynchronous Stochastic Approximation and Q-Learning. In *Conference on Learning Theory*, pp. 3185–3205. PMLR, 2020.

Shani, L., Efroni, Y., and Mannor, S. Adaptive Trust Region Policy Optimization: Global Convergence and Faster Rates for Regularized MDPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5668–5675, 2020.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

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

Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

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

Sutton, R. S., McAllester, D. A., Singh, S. P., Mansour, Y., et al. Policy gradient methods for reinforcement learning with function approximation. In *NIPs*, volume 99, pp. 1057–1063. Citeseer, 1999.

Thomas, P. S., Dabney, W., Mahadevan, S., and Giguere, S. Projected natural actor-critic. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, pp. 2337–2345, 2013.

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

Tsitsiklis, J. N. and Van Roy, B. Analysis of temporal-difference learning with function approximation. In *Advances in neural information processing systems*, pp. 1075–1081, 1997.

Tsitsiklis, J. N. and Van Roy, B. Average cost temporal-difference learning. *Automatica*, 35(11):1799–1808, 1999.

Wainwright, M. J. Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning. *Preprint arXiv:1905.06265*, 2019.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. Neural policy gradient methods: Global optimality and rates of convergence. *Preprint arXiv:1909.01150*, 2019.

Wang, Z., Bapst, V., Heess, N., Mnih, V., Munos, R., Kavukcuoglu, K., and de Freitas, N. Sample efficient actor-critic with experience replay. *Preprint arXiv:1611.01224*, 2016.

Watkins, C. J. and Dayan, P. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Williams, R. J. and Baird, L. C. 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*, pp. 96–101. Citeseer, 1990.

Wu, Y., Zhang, W., Xu, P., and Gu, Q. A Finite Time Analysis of Two Time-Scale Actor Critic Methods. *Preprint arXiv:2005.01350*, 2020.

Xu, T., Wang, Z., and Liang, Y. Improving sample complexity bounds for (natural) actor-critic algorithms. *Advances in Neural Information Processing Systems*, 33, 2020a.

Xu, T., Wang, Z., and Liang, Y. Non-asymptotic Convergence Analysis of Two Time-scale (Natural) Actor-Critic Algorithms. *Preprint arXiv:2005.03557*, 2020b.

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

Zhang, S., Liu, B., Yao, H., and Whiteson, S. Provably convergent two-timescale off-policy actor-critic with function approximation. In *International Conference on Machine Learning*, pp. 11204–11213. PMLR, 2020.