Finite-Sample Analysis of Off-Policy Natural Actor-Critic With Linear Function Approximation

Zaiwei Chen∗ 1 Sajad Khodadadian∗ 2 Siva Theja Maguluri 2

Abstract
In this paper, we develop a novel variant of off-policy natural actor-critic algorithm with linear function approximation and we establish a sample complexity of $\tilde{O}(\epsilon^{-3})$, outperforming all the previously known convergence bounds of such algorithms. In order to overcome the divergence due to deadly triad in off-policy policy evaluation under function approximation, we develop a critic that employs $n$-step TD-learning algorithm with a properly chosen $n$. We derive our sample complexity bounds solely based on the assumption that the behavior policy sufficiently explores all the states and actions, which is a much lighter assumption compared to the related literature.

1. Introduction
Reinforcement learning (RL) is a paradigm in which an agent aims at maximizing long term rewards via interacting with the environment. For solving the RL problem, there are value space methods such as $Q$-learning, and policy space methods such as actor-critic (AC) and its variants (e.g. natural actor critic (NAC)). In the AC framework, the actor aims at performing the policy update while the critic aims at estimating the value function of the current policy at hand. For AC type algorithms to perform well, the policy used to collect samples (called the behavior policy) must sufficiently explore the state-action space (Sutton and Barto, 2018). If the behavior policy coincides with the current policy iterate of AC, it is called on-policy sampling, otherwise it is called off-policy sampling.

In on-policy AC, the agent is restricted to use the current policy iterate to collect samples, which may not be exploratory. Moreover, on-policy sampling might be of high risk (e.g. self driving cars (Yurtsever et al., 2020)), high cost (e.g. robotics (Gu et al., 2017; Levine et al., 2020)), or might be unethical (e.g. in clinical trials (Gottesman et al., 2019; Liu et al., 2018; Gottesman et al., 2020)). Off-policy AC, on the other hand, is more practical than on-policy sampling (Levine et al., 2020). Specifically, off-policy sampling enables the agent to learn using the historical data, hence decouples the sampling process and the learning process. This allows the agent to learn in an off-line manner, and makes RL applicable in high-stake problems mentioned earlier. In addition, it is empirically observed that by using a suitable behavior policy, one can rectify the exploration issue in on-policy AC. As a result, off-policy learning successfully solved many practical problems in different areas, such as board game (Silver et al., 2017), city navigation (Mirowski et al., 2018), education (Mandel et al., 2014), and healthcare (Dann et al., 2019).

In practice, AC algorithms are usually used along with function approximation to overcome the curse of dimensionality in RL (Bellman, 1957). However, it has been observed that the combination of function approximation, off-policy sampling, and bootstrapping (also known as the deadly triad (Sutton and Barto, 2018)) can result in instability or even divergence (Sutton and Barto, 2018; Baird, 1995). In this work, we develop a variant of off-policy NAC with function approximation, and we establish its finite-sample convergence guarantee in the presence of the deadly triad.

1.1. Main Contributions
The main contributions of this paper are fourfold.

Finite-Sample Bounds of Off-Policy NAC. We develop a variant of NAC with off-policy sampling, where both the actor and the critic use linear function approximation, and the critic uses off-policy sampling. We establish finite-sample mean square bound of our proposed algorithm. Our result implies an $\tilde{O}(\epsilon^{-3})$ sample complexity, which is the best known convergence bound in the literature for AC algorithms with function approximation.

Novelty in the Critic. Off-policy TD with function approximation is famously (Sutton and Barto, 2018) known to diverge due to deadly triad. To overcome this difficulty, we...
employ \(n\)-step TD-learning, and show that a proper choice of \(n\) naturally achieves convergence. To the best of our knowledge, we are the first to design a single time-scale off-policy TD with function approximation with provable finite-sample bounds.

**Novelty in the Actor.** NAC under function approximation was developed in (Agarwal et al., 2019) by projecting the \(Q\)-values (gradients) to the lower dimensional space, and this involves the use of the discounted state visitation distribution, which is hard to estimate. We develop a new NAC algorithm for the function approximation setting that is instead based on the solution of a projected Bellman equation (Tsitsiklis and Van Roy, 1997), which our critic is designed to solve.

**Exploration through Off-Policy Sampling.** We establish the convergence bounds under the minimum set of assumptions, viz., ergodicity under the behavior policy, which ensures sufficient exploration, and thus resolving challenges faced in on-policy sampling. As a result, learning can be done using a single trajectory of samples generated by the behavior policy, and we do not require constant reset of the system that was introduced in on-policy AC algorithms (Agarwal et al., 2019; Wang et al., 2019) to ensure exploration. A similar observation about employing off-policy sampling to ensure exploration has been made in the tabular setting in (Khodadadian et al., 2021).

### 2. Main Results

In this section, we present our main results. Specifically, in Section 2.1 we briefly cover the background of RL and AC. In Section 2.2, we present our algorithm design for the critic, which uses off-policy sampling with linear function approximation. In section 2.3, we combine the critic with our actor update to form a variant of off-policy NAC with linear function approximation, and we present our finite-sample guarantees and sample complexity bounds.

#### 2.1. Preliminaries

Consider modelling the RL problem as an infinite horizon MDP, which consists of a finite set of states \(S\), a finite set of actions \(A\), a set of unknown transition probability matrices \(\mathcal{P} = \{ P_a \in \mathbb{R}^{||S|| \times |S|} \mid a \in A \}\), an unknown reward function \(R : S \times A \rightarrow \mathbb{R}\), and a discount factor \(\gamma \in (0, 1)\). Without loss of generality we assume that \(\max_{s,a} |R(s, a)| \leq 1\). For a given policy \(\pi\), its state value function is defined by \(V^\pi(s) = \mathbb{E}_\pi [\sum_{k=0}^{\infty} \gamma^k R(S_k, A_k) \mid S_0 = s] \) for all \(s \in S\), and its state-action value function is defined by \(Q^\pi(s, a) = \mathbb{E}_\pi [\sum_{k=0}^{\infty} \gamma^k R(S_k, A_k) \mid S_0 = s, A_0 = a] \) for all \((s, a) \in S \times A\). The goal of RL is to find an optimal policy \(\pi^*\) which maximizes \(V^\pi(\mu) = \sum_s \mu(s) V^\pi(s)\), where \(\mu\) is an arbitrary fixed initial distribution over the state space. It was shown in the literature that the optimal policy is in fact independent of the initial distribution. See (Bertsekas and Tsitsiklis, 1996; Puterman, 1995; Sutton and Barto, 2018) for more details for the MDP model of the RL problem.

To solve the RL problem, a popular approach is to use the AC framework (Konda and Tsitsiklis, 2000). In AC algorithm, we iteratively perform the policy evaluation and the policy improvement until an optimal policy is obtained. Specifically, in each iteration, we first estimate the \(Q\)-function (or the advantage function) of the current policy at hand, which is related to the policy gradient. Then we update the policy using gradient ascent over the space of the policies. NAC is a variant of AC where the gradient ascent step is performed with a properly chosen pre-conditioner. See (Agarwal et al., 2019) for more details about AC and NAC.

In AC framework, since we need to work with the \(Q\)-function and the policy, which are \(|S||A|\) dimensional objects, the algorithm becomes intractable when the size of the state-action space is large (Bellman, 1957). To overcome this difficulty, in this work we consider using linear function approximation for both the policy and the \(Q\)-function. Specifically, let \(\{ \phi_i \}_{1 \leq i \leq d}\) be a set of basis functions, where \(\phi_i \in \mathbb{R}^{||S|| \times |A|}\) for all \(i\). Without loss of generality, we assume that \(\phi_i, 1 \leq i \leq d\), are linearly independent and are normalized so that \(|\phi(s, a)||_1 \leq 1\) for all \((s, a)\), where \(\phi(s, a) = [\phi_1(s, a), \cdots, \phi_d(s, a)]\) is the feature associated with state-action pair \((s, a)\). Let \(\Phi = [\phi_1, \cdots, \phi_d]\) be the feature matrix. We parameterize the policy and the \(Q\)-function using compatible function approximation (Sutton et al., 1999). In particular, we use softmax parametrization for the policy, i.e., \(\pi^\theta(a|s) = \frac{\exp(\phi(s, a)^\top \theta)}{\sum_{a' \in A} \exp(\phi(s, a')^\top \theta)}\) for all \((s, a)\), where \(\theta \in \mathbb{R}^d\) is the parameter. As for the \(Q\)-function, we approximate it from the linear sub-space given by \(Q = \{ Q_w = \Phi w \mid w \in \mathbb{R}^d\}\), where \(w \in \mathbb{R}^d\) is the corresponding parameter. By doing this, we now only need to work with \(d\)-dimensional objects (i.e., \(w\) for the \(Q\)-function and \(\theta\) for the policy), where \(d\) is usually chosen to be much smaller than \(|S||A|\).

#### 2.2. Off-Policy Multi-Step TD-learning with Linear Function Approximation

In this section, we present the \(n\)-step off-policy TD-learning algorithm under linear function approximation (Sutton and Barto, 2018), which is used for solving the policy evaluation (critic) sub-problem in our AC framework. Let \(\pi\) be the target policy we aim to evaluate, and let \(\pi_0\) be the behavior policy we used to collect samples. For any state-action pairs \((s, a)\), let \(\rho(s, a) = \frac{\pi(a|s)}{\pi_0(a|s)}\), which is called the importance sampling ratio between \(\pi\) and \(\pi_0\) at \((s, a)\). For any positive integer \(n\), Algorithm 2.1 presents the off-policy \(n\)-step TD-
Algorithm 2.1 Off-Policy n-Step TD-Learning with Linear Function Approximation

1: **Input:** $K$, $\alpha$, $w_0$, $\pi$, $\pi_b$, and $\{(S_k, A_k)\}_{0 \leq k \leq (K+n)}$ (a single trajectory generated by the behavior policy $\pi_b$)
2: for $k = 0, 1, \cdots, K - 1$
3: \hspace{1em} $\delta_{k,i} = R(S_i, A_i) + \gamma \rho(S_{i+1}, A_{i+1}) \phi(S_{i+1}, A_{i+1})^T w_k - \phi(S_i, A_i)^T w_k$
4: \hspace{1em} $\Delta_{k,i} = \sum_{i=k}^{k+n-1} \gamma^{i-k} \prod_{j=i+1}^{n-1} \rho(S_j, A_j) \delta_{k,i}$
5: \hspace{1em} $w_{k+1} = w_k + \alpha \phi(S_k, A_k) \Delta_{k,n}$
6: end for
7: **Output:** $w_K$

In Algorithm 2.1, we employ the importance sampling ratio to account for the discrepancy between the target policy $\pi$ and the behavior policy $\pi_b$. Although all the three elements of the deadly triad (bootstrapping, function approximation, and off-policy sampling) (Sutton and Barto, 2018) are present, we show that by choosing $n$ appropriately, Algorithm 2.1 has provable finite-sample convergence guarantee. The detailed statement of the result is presented in Section 2.4.

In existing literature, to achieve stability in the presence of the deadly triad, algorithms such as gradient TD-learning (GTD) (Sutton et al., 2008), TD-learning with gradient correction (TDC) (Sutton et al., 2009), and emphatic TD-learning (Sutton et al., 2016) all require to maintain two iterates. Such two time-scale algorithms are in general harder to implement, and in addition, even if convergence is guaranteed, there is no characterization on the limit point. However, Algorithm 2.1 naturally achieves convergence, requires to maintain only one iterate, and has a limit point that can be characterized as the solution of a projected Bellman equation.

2.3. Off-Policy Variant of NAC with Linear Function Approximation

In this section, we combine the off-policy TD-learning with linear function approximation algorithm in the previous section, with our variant of NPG update to form the off-policy variant of NAC algorithm. For simplicity of notation, we denote $Q^{\pi_0}$ as $Q^\pi$. Also, with input $K$, $\alpha$, $w_0$, $\pi$, $\pi_b$, and samples $\{(S_k, A_k)\}_{0 \leq k \leq K+n}$, we denote the output of Algorithm 2.1 as CRITIC($K$, $\alpha$, $w_0$, $\pi_b$, $\{S_k, A_k\}_{0 \leq k \leq K+n}$).

In each iteration of the off-policy NAC algorithm 2.2, the critic first estimates the $Q$-function $Q^\pi$ using $\Phi w_t$. Then, the actor updates the parameter $\theta_t$ of the current policy. Note that unlike the on-policy NAC where the algorithm usually needs to be constantly reset to a specific state of the environment, which is impractical, off-policy sampling enables us to use a single sample trajectory collected under the behavior policy.

Algorithm 2.2 Off-Policy Natural Actor-Critic Algorithm with Linear Function Approximation

1: **Input:** $T$, $K$, $\alpha$, $\beta$, $\theta_0$, $\pi$, $\pi_b$, and $\{(S_k, A_k)\}_{0 \leq k \leq (K+n)}$ (a single trajectory generated by the behavior policy $\pi_b$)
2: for $t = 0, 1, \cdots, T - 1$
3: \hspace{1em} $w_t = \text{CRITIC}(K, \alpha, 0, \pi_t, \pi_b, \{(S_k, A_k)\}_{t \leq K+n})$
4: \hspace{1em} $\theta_{t+1} = \theta_t + \beta w_t$
5: end for
6: **Output:** $\theta_T$, where $\hat{T}$ is uniformly sampled from $[0, T - 1]$.

2.4. Finite-Sample Convergence Guarantees

In this section, we present the finite-sample convergence bounds of Algorithms 2.1 and 2.2. We begin by stating our one and only assumption.

Assumption 2.1. The behavior policy $\pi_b$ satisfies $\pi_b(a|s) > 0$ for all $(s, a)$ and the Markov chain $\{S_k\}$ induced by the behavior policy is irreducible and aperiodic.

Assumption 2.1 is standard in studying off-policy TD-learning algorithms (Maei, 2018; Zhang et al., 2020). Since we work with finite state and action spaces, under Assumption 2.1, the Markov chain $\{S_k\}$ admits a unique stationary distribution, denoted by $\mu_0 \in \Delta^{[S]}$ (Levin and Peres, 2017). In addition, we have $\|\mu_k(s) - \mu_0(s)\|_{TV} \leq C \sigma^k$ for any $k \geq 0$, where $C > 0$, $\sigma \in (0, 1)$ are constants, and $\|\cdot\|_{TV}$ stands for the total variation distance between probability distributions (Levin and Peres, 2017). Note that in this case the random process $\{(S_k, A_k)\}$ is also a Markov chain with a unique stationary distribution, which we have denoted by $\pi_0 \in \Delta^{[S] \times [A]}$, and $\kappa_0(s, a) = \mu_0(s)\pi_0(a|s)$ for all $(s, a)$.

In the existing literature, where on-policy NAC was studied, it is typically required that all the policies achieved in the iterations of the NAC induce ergodic Markov chains over the state-action space (Qiu et al., 2019; Wu et al., 2020). Such a requirement is strong and not possible to satisfy in an MDP where the optimal policy is a unique deterministic policy. Off-policy sampling enables us to relax such an unrealistic requirement while also ensuring exploration.

We next present the finite-sample convergence bound of the off-policy NAC with linear function approximation. We begin by introducing some notation. For a given stepsize $\alpha$, let $t_\alpha = \min\{k \geq 0 : \|P_k(s, \cdot) - \mu_0(\cdot)\|_{TV} \leq \alpha\}$, which represents the mixing time of the Markov chain $\{S_k\}$, and can be bounded by an affine function of $\log(1/\alpha)$ under Assumption 2.1. Let $f(x) = n + 1$ when $x = 1$ and $f(x) = \frac{1 - x^{n+1}}{1-x}$ when $x \neq 1$. Denote $w_\pi$ as the solution of
the projected Bellman equation

\[ Q_w = \Pi_{\kappa_b} T^\pi_w(Q_w) = \Phi (\Phi^T \mathcal{K} \Phi)^{-1} \Phi^T \mathcal{K} T^\pi_w(Q_w), \]

where \( Q_w = \Phi w \). Here \( T^\pi_n(\cdot) \) denotes the \( n \)-step Bellman operator, and \( \Pi_{\kappa_b}(\cdot) \) stands for the projection operator onto the linear sub-space \( \mathcal{Q} \) with respect to the weighted \( \ell_2 \)-norm with weights \( \{\kappa_b(s, a)\}_{(s, a) \in \mathcal{S} \times \mathcal{A}} \) (Tsitsiklis and Van Roy, 1997). Let \( \zeta_n = \max_{s, a} \|\pi_n(s, a)\|, \) which measures the mismatch between \( \pi \) and \( \pi_n \). Let \( \lambda_{\min} \) be the smallest eigenvalue of the positive definite matrix \( \Phi^T \mathcal{K} \Phi \). Let \( \xi = \max_{s, a} \|Q^\pi - \Phi w^\pi\|_\infty \), where \( Q^\pi \) is the \( Q \)-function associated with the policy \( \pi_\theta \). Note that the quantity \( \xi \) measures how powerful the function approximation architecture is. Let \( \zeta_{\max} = \max_{s, a} \frac{1}{\pi_n(s, a)} \), which is a uniform upper bound of \( \zeta_n \) for any target policy \( \pi \).

**Theorem 2.1.** Consider the output \( \theta^\pi_k \) of Algorithm 2.2. Suppose that Assumptions 2.1 is satisfied, the parameter \( n \) is chosen such that \( n \geq \frac{2 \log(\gamma_c) + \log(\xi)}{2 \log(\gamma)} \) (where \( \gamma_c \in (0, 1) \) is some tunable constant), and \( \alpha(s, a) \) is chosen such that \( \alpha(t_\alpha + n + 1) \leq \frac{1 - \gamma_c}{2 \log(\gamma)} \). For any starting distribution \( \mu \), we have for any \( K \geq t_\alpha + n + 1 \) and \( T \geq 1 \):

\[ V^{\pi^\pi}(\mu) - \mathbb{E}[V^{\pi^\pi}(\mu)] \leq \frac{2}{(1 - \gamma)^2 T} + \frac{3\xi}{(1 - \gamma)^2} \]

**A1:** convergence bias in the actor

\[ + \frac{3\xi}{(1 - \gamma)^2} c_3(1 - (1 - \gamma)\lambda_{\min}^2) \frac{K - t_\alpha + n + 1}{T} \]

**A2:** bias due to function approximation

\[ + \frac{3\xi}{(1 - \gamma)^2} c_4(1 - (1 - \gamma)\lambda_{\min}^2) \frac{K - t_\alpha + n + 1}{T} \]

**A3:** convergence bias in the critic

\[ + \frac{3\xi}{(1 - \gamma)^2} \frac{1}{(1 - \gamma)^2} c_3(1 - (1 - \gamma)\lambda_{\min}^2) \frac{1}{T} \]

**A4:** variance in the critic

where \( c_3 = 1 + \frac{2}{(1 - \gamma)^2(1 - \gamma)\lambda_{\min}^2} \).

The term \( A_1 \) represents the convergence bias of the actor, and goes to zero at a rate of \( O(1/T) \) as the outer loop iteration number \( T \) goes to infinity. The term \( A_3 \) measures the convergence bias in the critic, and goes to zero geometrically fast as the inner loop iteration number \( K \) goes to infinity. The term \( A_4 \) represents the impact of the variance in the critic, and is of the size \( O(\sqrt{\alpha \log(1/\alpha)}) \), which goes to zero as the inner loop stepsize \( \alpha \) goes to zero.

The term \( A_2 \) captures the error introduced to the system due to function approximation, and cannot be eliminated asymptotically. Moreover, known results in approximate policy iteration (API) literature suggest that the \( 1/(1 - \gamma)^2 \) coefficient inside the term \( A_2 \) is inevitable. Specifically, it is shown (Bertsekas, 2011; Bertsekas and Tsitsiklis, 1996) that when \( \max_n \|V^{\pi^\pi} - \Phi w^\pi\|_\infty \leq \xi \), under the API algorithm

\[ \lim_{n \to \infty} \|V^{\pi_k} - V^{\pi^\pi}\|_\infty \leq \frac{2\xi}{(1 - \gamma)^2}, \]

and an example is presented in (Bertsekas and Tsitsiklis, 1996, Section 6.2.3), where the inequality is tight. Since NAC algorithm can be viewed as an API algorithm with a softmax policy update (which is also weighted by the current policy), it is natural to expect a similar function approximation bias. Therefore, to improve the function approximation bias term \( A_2 \), one has to develop instance dependent bound, which is one of our future direction.

### 2.5. Sample Complexity Analysis

In this section, we derive sample complexity of off-policy NAC algorithm based on Theorem 2.1.

**Corollary 2.1.1.** In order to achieve \( V^{\pi^\pi}(\mu) - \mathbb{E}[V^{\pi^\pi}(\mu)] \leq \epsilon + \frac{3\xi}{(1 - \gamma)^2} \), the number of samples requires is of the size \( O\left(\epsilon^{-3} \log^2(1/\epsilon)\right) \) \( \mathcal{O}\left(\frac{1}{(1 - \gamma)^2}\right) \) and \( \mathcal{O}\left(\frac{1}{\gamma}\right) \). For any starting distribution \( \mu \), there is a trade-off in the choice of \( \gamma \): for \( \gamma \) close to 1, the convergence error does not go to zero. Therefore, one should not use sample complexity when we do not have global convergence due to the function approximation bias.

However, we present Corollary 2.1.1 in terms of “sample complexity” in the same sense as used in prior literature to enable a fair comparison.

In view of the sample complexity bound, the dependency on the required accuracy level \( \epsilon \) is \( \mathcal{O}(\epsilon^{-3}) \). This improves the state-of-the-art sample complexity of off-policy NAC with function approximation result in the literature in (Xu et al., 2021) by a factor of \( \epsilon^{-1} \).

To see this, consider the projected Bellman equation \( Q_w = \Pi_{\kappa_b} T^\pi_w(Q_w) \). Since \( n \) tends to infinity, since \( \lim_{n \to \infty} T^\pi_n(Q_w) = Q^\pi \) due to value iteration (Banach fixed-point theorem for the operator \( T^\pi(.) \)), the solution of the projected Bellman equation coincides with the projection of \( Q^\pi \) to the linear sub-space \( \mathcal{Q} \), which has the best function approximation bias. However, note that the parameter \( n \) also appears in the numerator of the sample complexity bound (which is due to the variance term in the critic), hence there is a trade-off in the choice of \( n \).

To summarize, increasing (decreasing) the parameter \( n \) leads to better (worse) critic convergence bias and function approximation bias, but has worse (better) critic variance.

### 3. Conclusion

In this paper, we establish finite-sample convergence guarantees of off-policy NAC with linear function approximation. To overcome the deadly triad in the critic, we use \( n \)-step TD-learning, which is a one-time scale algorithm for policy
evaluation using off-policy sampling and linear function approximation, and has provable convergence bounds. Our finite-sample bounds imply a sample complexity of $\tilde{O}(e^{-3})$, which advances the state-of-the-art result in the literature.

References


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


