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 $\mathcal{O}(\epsilon^{-3})$, 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 $\tilde{\mathcal{O}}(\epsilon^{-3})$ 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

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

ICML Workshop on Reinforcement Learning Theory , 2021.

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 $\mathcal{S}$, a finite set of actions $\mathcal{A}$, a set of unknown transition probability matrices $\mathcal{P} = \{P_a \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|} \mid a \in \mathcal{A}\}$, an unknown reward function $\mathcal{R} : \mathcal{S}\times\mathcal{A} \mapsto \mathbb{R}$, and a discount factor $\gamma \in (0, 1)$. Without loss of generality we assume that $\max_{s,a} |\mathcal{R}(s, a)| \leq 1$. For a given policy π, its state value function is defined by $V^\pi(s) = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k \mathcal{R}(S_k, A_k) \mid S_0 = s]$ for all $s \in \mathcal{S}$, and its state-action value function is defined by $Q^\pi(s, a) = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k \mathcal{R}(S_k, A_k) \mid S_0 = s, A_0 = a]$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. The goal of RL is to find an optimal policy π^* which maximizes $V^\pi(\mu) = \sum_s \mu(s)V^\pi(s)$, where μ 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 $|\mathcal{S}||\mathcal{A}|$ 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 $\phi_i \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ for all i. Without loss of generality, we assume that ϕ_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., $\pi_\theta(a|s) = \frac{\exp(\phi(s,a)^\top\theta)}{\sum_{a'\in\mathcal{A}}\exp(\phi(s,a')^\top\theta)}$ 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 $\mathcal{Q} = \{Q_w = \Phi w \mid w \in \mathbb{R}^d\}$, 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 θ for the policy), where d is usually chosen to be much smaller than $|\mathcal{S}||\mathcal{A}|$.

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 π be the target policy we aim to evaluate, and let π_b be the behavior policy we used to collect samples. For any state-action pairs (s, a), let $\rho(s, a) = \frac{\pi(a|s)}{\pi_b(a|s)}$, which is called the importance sampling ratio between π and π_b at (s, a). For any positive integer n, Algorithm 2.1 presents the off-policy n-step TD-learning algorithm for estimating Q^π.

Algorithm 2.1 Off-Policy n-Step TD-Learning with Linear Function Approximation

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 π_b)
for $k = 0, 1, \cdots, K-1$ **do**
  $\delta_{k,i} = \mathcal{R}(S_i, A_i) + \gamma\rho(S_{i+1}, A_{i+1})\phi(S_{i+1}, A_{i+1})^\top w_k - \phi(S_i, A_i)^\top w_k$
  $\Delta_{k,n} = \sum_{i=k}^{k+n-1}\gamma^{i-k}\prod_{j=i+1}^{k+n-1}\rho(S_j, A_j)\delta_{k,i}$
  $w_{k+1} = w_k + \alpha\phi(S_k, A_k)\Delta_{k,n}$
end for
Output: w_K

In Algorithm 2.1, we employ the importance sampling ratio to account for the discrepancy between the target policy π and the behavior policy π_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 $Q^{\pi_{\theta_t}}$ as Q^{π_t}. Also, with input K, α, w_0, π, π_b, and samples $\{(S_k, A_k)\}_{0\leq k\leq K+n}$, we denote the output of Algorithm 2.1 as CRITIC$(K, \alpha, w_0, \pi, \pi_b, \{S_k, A_k\}_{0\leq k\leq K+n})$.

In each iteration of the off-policy NAC algorithm 2.2, the critic first estimates the Q-function Q^{π_t} using Φw_t. Then, the actor updates the parameter θ_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

Input: T, K, α, β, θ_0, π, π_b, and $\{(S_k, A_k)\}_{0\leq k\leq T(K+n)}$ (a *single trajectory* generated by the behavior policy π_b)
for $t = 0, 1, \ldots, T-1$ **do**
  $w_t =$ CRITIC$(K, \alpha, \mathbf{0}, \pi_t, \pi_b, \{(S_k, A_k)\}_{t(K+n)\leq k\leq (t+1)(K+n)})$
  $\theta_{t+1} = \theta_t + \beta w_t$
end for
Output: $\theta_{\hat{T}}$, 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 π_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 $\mu_b \in \Delta^{|\mathcal{S}|}$ (Levin and Peres, 2017). In addition, we have $\|P^k(s,\cdot) - \mu_b(\cdot)\|_{\text{TV}} \leq C\sigma^k$ for any $k \geq 0$, where $C > 0$, $\sigma \in (0, 1)$ are constants, and $\|\cdot\|_{\text{TV}}$ 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 $\kappa_b \in \Delta^{|\mathcal{S}||\mathcal{A}|}$, and $\kappa_b(s, a) = \mu_b(s)\pi_b(a|s)$ 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 α, let $t_\alpha = \min\{k \geq 0 : \|P^k(s,\cdot) - \mu_b(\cdot)\|_{\text{TV}} \leq \alpha\}$, 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_π as the solution of the projected Bellman equation

$$Q_w = \Pi_{\kappa_b}\mathcal{T}_\pi^n(Q_w) = \Phi(\Phi^\top \mathcal{K}\Phi)^{-1}\Phi^\top \mathcal{K}\mathcal{T}_\pi^n(Q_w), \quad (1)$$

where $Q_w = \Phi w$. Here $\mathcal{T}_\pi^n(\cdot)$ 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 ℓ_2-norm with weights $\{\kappa_b(s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ (Tsitsiklis and Van Roy, 1997). Let $\zeta_\pi = \max_{s,a}\frac{\pi(a|s)}{\pi_b(a|s)}$, which measures the mismatch between π and π_b. Let $\lambda_{\min}$ be the smallest eigenvalue of the positive definite matrix $\Phi^\top\mathcal{K}\Phi$. Let $\xi = \max_\theta \|Q^{\pi_\theta} - \Phi w_{\pi_\theta}\|_\infty$, where Q^{π_θ} is the Q-function associated with the policy π_θ. Note that the quantity ξ measures how powerful the function approximation architecture is. Let $\zeta_{\max} = \max_{s,a}\frac{1}{\pi_b(a|s)}$, which is a uniform upper bound of ζ_π for any target policy π.

Theorem 2.1. *Consider the output $\theta_{\hat{T}}$ of Algorithm 2.2. Suppose that Assumptions 2.1 is satisfied, the parameter n is chosen such that $n \geq \frac{2\log(\gamma_c)+\log(\kappa_{b,\min})}{2\log(\gamma)}$ (where $\gamma_c \in (0,1)$ is some tunable constant), and α is chosen such that $\alpha(t_\alpha+n+1) \leq \frac{1-\gamma_c}{456 f(\gamma\zeta_\pi)^2}$. For any starting distribution μ, we have for any $K \geq t_\alpha+n+1$ and $T\geq 1$:*

$$\begin{aligned} V^{\pi^*}(\mu) - \mathbb{E}\left[V^{\pi_{\hat{T}}}(\mu)\right] \leq & \underbrace{\frac{2}{(1-\gamma)^2 T}}_{A_1:\text{ convergence bias in the actor}} \\ & + \underbrace{\frac{3\xi}{(1-\gamma)^2}}_{A_2:\text{ bias due to function approximation}} \\ & + \underbrace{\frac{3}{(1-\gamma)^2}c_3(1-(1-\gamma_c)\lambda_{\min}\alpha)^{\frac{K-(t_\alpha+n+1)}{2}}}_{A_3:\text{ convergence bias in the critic}} \\ & + \underbrace{\frac{33c_3 f(\gamma\zeta_{\max})[\alpha(t_\alpha+n+1)]^{1/2}}{(1-\gamma)^2(1-\gamma_c)^{1/2}\lambda_{\min}^{1/2}}}_{A_4:\text{ variance in the Critic}}, \end{aligned}$$

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

The term A_1 represents the convergence bias of the actor, and goes to zero at a rate of $\mathcal{O}(1/T)$ 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 $\mathcal{O}(\sqrt{\alpha\log(1/\alpha)})$, which goes to zero as the inner loop stepsize α 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 $\max_\pi\|V^\pi - \Phi w_\pi\|_\infty \leq \xi$, under the API algorithm $\limsup_{k\to\infty}\|V^{\pi_k} - V^{\pi^*}\|_\infty \leq \frac{2\gamma\xi}{(1-\gamma)^2}$, 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 $V^{\pi^*}(\mu) - \mathbb{E}\left[V^{\pi_{\hat{T}}}(\mu)\right] \leq \epsilon + \frac{3\xi}{(1-\gamma)^2}$, the number of samples requires is of the size $\mathcal{O}\left(\epsilon^{-3}\log^2(1/\epsilon)\right)\tilde{\mathcal{O}}\left(f(\gamma\zeta_{\max})^2 n(1-\gamma)^{-8}(1-\gamma_c)^{-3}\lambda_{\min}^{-3}\right)$.*

Remark. It was argued in (Khodadadian et al., 2021, Appendix C) that sample complexity is not well-defined when 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 ϵ is $\tilde{\mathcal{O}}(\epsilon^{-3})$. 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 ϵ^{-1}. As stated in Theorem 2.1, in order to use smaller γ_c in our analysis, we need to choose larger n in executing Algorithm 2.1. An advantage of using large n is that it leads to a lower function approximation bias ξ. To see this, consider the projected Bellman equation $Q_w = \Pi_{\kappa_b}\mathcal{T}_\pi^n(Q_w)$. When n tends to infinity, since $\lim_{n\to\infty}\mathcal{T}_\pi^n(Q_w) = Q^\pi$ due to value iteration (Banach fixed-point theorem for the operator $\mathcal{T}_\pi(\cdot)$), the solution of the projected Bellman equation coincides with the projection of Q^π 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 $\tilde{\mathcal{O}}(\epsilon^{-3})$, which advances the state-of-the-art result in the literature.

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Preprint arXiv:1908.00261*, 2019.

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pages 30–37. Elsevier, 1995.

R Bellman. Dynamic programming princeton university press princeton. *New Jersey Google Scholar*, 1957.

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

Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516. PMLR, 2019.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1): 16–18, 2019.

Omer Gottesman, Joseph Futoma, Yao Liu, Sonali Parbhoo, Leo Celi, Emma Brunskill, and Finale Doshi-Velez. Interpretable off-policy evaluation in reinforcement learning by highlighting influential transitions. In *International Conference on Machine Learning*, pages 3658–3667. PMLR, 2020.

Shixiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3389–3396. IEEE, 2017.

Sajad Khodadadian, Zaiwei Chen, and Siva Theja Maguluri. Finite-sample analysis of off-policy natural actor-critic algorithm. *Preprint arXiv:2102.09318*, 2021.

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

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

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *Preprint arXiv:2005.01643*, 2020.

Yao Liu, Omer Gottesman, Aniruddh Raghu, Matthieu Komorowski, Aldo A Faisal, Finale Doshi-Velez, and Emma Brunskill. Representation Balancing MDPs for Off-policy Policy Evaluation. *Advances in Neural Information Processing Systems*, 31:2644–2653, 2018.

Hamid Reza Maei. Convergent actor-critic algorithms under off-policy training and function approximation. *Preprint arXiv:1802.07842*, 2018.

Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. In *AAMAS*, pages 1077–1084, 2014.

Piotr Mirowski, Matt Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin, Karen Simonyan, Andrew Zisserman, Raia Hadsell, et al. Learning to navigate in cities without a map. In *Advances in Neural Information Processing Systems*, pages 2419–2430, 2018.

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

Shuang Qiu, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. 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.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.

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

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, pages 1057–1063, 1999.

Richard S Sutton, Csaba Szepesvári, and Hamid Reza Maei. 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, 2008.

Richard S Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. 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, 2009.

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

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

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural Policy Gradient Methods: Global Optimality and Rates of Convergence. In *International Conference on Learning Representations*, 2019.

Yue Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A Finite Time Analysis of Two Time-Scale Actor Critic Methods. *Preprint arXiv:2005.01350*, 2020.

Tengyu Xu, Zhuoran Yang, Zhaoran Wang, and Yingbin Liang. Doubly robust off-policy actor-critic: Convergence and optimality. *Preprint arXiv:2102.11866*, 2021.

Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 8:58443–58469, 2020.

Shangtong Zhang, Bo Liu, Hengshuai Yao, and Shimon Whiteson. Provably convergent two-timescale off-policy actor-critic with function approximation. In *International Conference on Machine Learning*, pages 11204–11213. PMLR, 2020.