Global Convergence of Two-Timescale Actor-Critic for Solving Linear Quadratic Regulator

Xuyang Chen[1], **Jingliang Duan**[2], **Yingbin Liang**[3], **Lin Zhao** [1*]

[1] National University of Singapore
[2] University of Science and Technology Beijing
[3] The Ohio State University
chenxuyang@u.nus.edu, duanjl@ustb.edu.cn, liang.889@osu.edu, elezhli@nus.edu.sg

Abstract

The actor-critic (AC) reinforcement learning algorithms have been the powerhouse behind many challenging applications. Nevertheless, its convergence is fragile in general. To study its instability, existing works mostly consider the uncommon double-loop variant or basic models with finite state and action space. We investigate the more practical single-sample two-timescale AC for solving the canonical linear quadratic regulator (LQR) problem, where the actor and the critic update only once with a single sample in each iteration on an unbounded continuous state and action space. Existing analysis cannot conclude the convergence for such a challenging case. We develop a new analysis framework that allows establishing the global convergence to an epsilon-optimal solution with at most an order of epsilon to -2.5 sample complexity. To our knowledge, this is the first finite-time convergence analysis for the single sample two-timescale AC for solving LQR with global optimality. The sample complexity improves those of other variants by orders, which sheds light on the practical wisdom of single sample algorithms. We also further validate our theoretical findings via comprehensive simulation comparisons.

Introduction

The actor-critic (AC) methods (Konda and Tsitsiklis 1999) are among the most commonly used reinforcement learning (RL) algorithms, which have achieved tremendous empirical successes (Mnih et al. 2016; Silver et al. 2017). In AC methods, the actor refers to the policy and the critic characterizes the action-value function (Q-function) given the actor. In each iteration, the critic tries to approximate the Q-function by applying policy evaluation algorithms (Dann et al. 2014; Sutton and Barto 2018), while the actor typically follows policy gradient (Sutton et al. 1999; Agarwal et al. 2021) updates according to the Q-function provided by the critic. Compared with other RL algorithms, AC methods combine the advantages of both policy-based methods such as REINFORCE (Williams 1992) and value-based methods such as temporal difference (TD) learning (Sutton 1988; Bhandari, Russo, and Singal 2018) and Q-learning (Watkins and Dayan 1992). Therefore, AC methods can be naturally applied to the continuous control setting (Silver et al. 2014) and meanwhile enjoy the low variance of bootstrapping.

Despite its empirical success, theoretical guarantees of AC still lag behind. Most existing works focus exclusively on the double-loop setting, where the critic updates many steps in the inner loop, followed by an actor update in the outer loop (Yang et al. 2019; Agarwal et al. 2021; Wang et al. 2019; Abbasi-Yadkori et al. 2019; Bhandari and Russo 2021; Xu, Wang, and Liang 2020a). This setting yields accurate estimation of the Q-function and consequently the policy gradient. Therefore, double-loop setting decouples the convergence analysis of the actor and the critic, which further allows analyzing AC as a gradient method with error (Sutton et al. 1999; Kakade and Langford 2002; Shalev-Shwartz and Ben-David 2014; Ruder 2016).

A more favorable implementation of AC in practice is the single-loop two-timescale setting, where the actor and critic are updated simultaneously in each iteration with different-timescale stepsizes. Typically, the actor stepsize is smaller than that of the critic. To establish the finite-time convergence of two-timescale AC methods, most existing results either focus on the multi-sample methods (Xu, Wang, and Liang 2020b; Qiu et al. 2021) or the finite discrete action space (Wu et al. 2020). The former allows the critic to collect multiple samples to accurately estimate the Q-function given the actor, which are rarely implemented in practice. It essentially decouples actor and critic in a similar way to the double-loop setting. We study the more practical single-sample AC algorithm similar to the one considered in Wu et al. (2020), where the critic updates only once using a single sample per policy evaluation step. However, the latter only attains a stationary point under the finite-action space setting (see Table 1 for comparisons). We address the important yet more challenging question: *can the single-sample two-timescale AC find a global optimal policy on the general unbounded continuous state-action space?*

To this end, we analyze the classic single-sample AC for solving the Linear Quadratic Regulator (LQR), a fundamental control task which is commonly employed as a testbed to explore the behavior and limits of RL algorithms under continuous state-action spaces (Fazel et al. 2018; Yang et al. 2019; Tu and Recht 2018; Krauth, Tu, and Recht 2019; Duan et al. 2022). In the LQR case, the Q-function is a linear function of the quadratic form of state and action. In general, it is

*Corresponding author

Reference	Structure	Sample Complexity	Optimality
Xu, Wang, and Liang (2020b)	Multi-sample	$\mathcal{O}(\epsilon^{-2.5})$	Local
		$\mathcal{O}(\epsilon^{-4})$	Global
Wu et al. (2020)	Single-sample	$\mathcal{O}(\epsilon^{-2.5})$	Local
This paper	Single-sample	$\mathcal{O}(\epsilon^{-2.5})$	Global

Table 1: Comparison with other two-timescale actor-critic algorithms

difficult to establish the convergence of AC with linear function approximation (Bhandari, Russo, and Singal 2018). In the double-loop and multi-sample settings (Yang et al. 2019; Krauth, Tu, and Recht 2019), the Q-function of LQR can be estimated arbitrarily accurately, and the LQR cost is guaranteed to converge to the global optimum monotonically per-iteration. However, the single-sample algorithm generally does not have these nice properties. It is more challenging to control the error propagation between the actor and the critic updates over iterations and prove its global convergence.

We distinguish our work from other model-free RL algorithms for solving LQR in Table 2. The zeroth-order methods and the policy iteration method are included for completeness. In particular, we note that Zhou and Lu (2022) analyzed the finite-time convergence under a single-timescale stepsize and multi-sample setting. The analysis requires the strong assumption on the uniform boundedness of the critic parameters. In comparison, our analysis does not require this assumption and considers the more challenging single-sample setting.

Within the literature of two-timescale AC for solving general MDP problems (see Table 1), we note that the analysis of Wu et al. (2020) depends critically on the assumptions of finite action space, bounded reward function, and bounded feature functions. However, these fundamental assumptions do not hold in the LQR case, making its analysis more challenging.

Main Contribution

Our main contributions are summarized as follows:

• Our work contributes to the theoretical understanding of AC methods. We for the first time show that the classic single-sample two-timescale AC can provably find the ϵ-accurate global optimum with a sample complexity of $\mathcal{O}(\epsilon^{-2.5})$, under the continuous state-action space with linear function approximation. This is the same complexity order as those in Wu et al. (2020); Xu, Wang, and Liang (2020b), but the latter only attain the local optimum.

We also add to the work of RL on continuous control tasks. It is novel that even without completely solving the policy evaluation sub-problem and only updating the actor with a biased policy gradient, the two-timescale AC algorithm can still find the global optimal policy for LQR, under common assumptions. Our work may serve as the first step towards understanding the limits of single-sample two-timescale AC on continuous control tasks.

Compared with the state-of-the-art work of double-loop AC for solving LQR (Yang et al. 2019), we show the practical wisdom single-sample AC enjoys a lower sample complexity than $\mathcal{O}(\epsilon^{-5})$ of the latter. We also show the algorithm is much more sample-efficient empirically compared to a few classic works.

• Technical-wise, despite the non-convexity of the LQR problem, we still find the global optimal policy under the single-sample update by exploiting the gradient domination property (Polyak 1963; Nesterov and Polyak 2006; Fazel et al. 2018). Existing popular analysis (Fazel et al. 2018; Yang et al. 2019) relies on the contraction of the cost learning errors. This nevertheless does not hold in the single-sample case. We alternatively establish the global convergence by showing the natural gradient of the objective function converges to zero and then using the gradient domination. Our work provides a more general proof framework for finding the optimal policy of LQR using various RL algorithms.

Related Work

Due to the extensive studies on AC methods, we hereby review only those works that are mostly relevant to our study.

Actor-Critic methods. The AC algorithm was proposed in Witten (1977); Sutton (1984); Konda and Tsitsiklis (1999). Kakade (2001) extended it to the natural AC algorithm. The asymptotic convergence of AC algorithms has been well established in Kakade (2001); Bhatnagar et al. (2009); Castro and Meir (2010); Zhang et al. (2020). Many recent works focused on the finite-time convergence of AC methods. Under the double-loop setting, Yang et al. (2019) established the global convergence of AC methods for solving LQR. Wang et al. (2019) studied the global convergence of AC methods with both the actor and the critic being parameterized by neural networks. Kumar, Koppel, and Ribeiro (2019) studied the finite-time local convergence of a few AC variants with linear function approximation, where the number of inner loop iterations grows linearly with the outer loop counting number.

Under the two-timescale AC setting, Khodadadian et al. (2022); Hu, Ji, and Telgarsky (2021) studied its finite-time convergence in tabular (finite state-action) case. For two-timescale AC with linear function approximation (see Table 1 for a summary), Wu et al. (2020) established the finite-time local convergence to a stationary point at a sample complexity of $\mathcal{O}(\epsilon^{-2.5})$. Xu, Wang, and Liang (2020b) studied both local convergence and global convergence for two-timescale (natural) AC, with $\mathcal{O}(\epsilon^{-2.5})$ and $\mathcal{O}(\epsilon^{-4})$ sample complexity, respectively, under the discounted accumulated reward. The algorithm collects multiple samples to update the critic.

Under the single-timescale setting, Fu, Yang, and Wang (2020) considered the regularized least-square temporal difference (LSTD) update for critic and established the finite-

Reference	Algorithm	Structure	
Fazel et al. (2018)	Zeroth-order	Double-loop	
Malik et al. (2019)	Zeroth-order		
Yang et al. (2019)	Actor-Critic		
Krauth, Tu, and Recht (2019)	Policy Iteration	Single-loop	Multi-sample
Zhou and Lu (2022)	Actor-Critic		Multi-sample (Single-timescale)
This paper	Actor-Critic		Single-sample (Two-timescale)

Table 2: Comparison with other model-free RL algorithms for solving LQR.

time convergence for both linear function approximation and nonlinear function approximation using neural networks.

RL algorithms for LQR. RL algorithms in the context of LQR have seen increased interest in the recent years. These works can be mainly divided into two categories: model-based methods (Dean et al. 2020; Mania, Tu, and Recht 2019; Cohen, Koren, and Mansour 2019; Dean et al. 2018) and model-free methods. Our main interest lies in the model-free methods. Notably, Fazel et al. (2018) established the first global convergence result for LQR under the policy gradient method using derivative-free (one-point gradient estimator based zeroth-order) optimization at a sample complexity of $\mathcal{O}(\epsilon^{-4})$. Malik et al. (2019) employed two-point gradient estimator based zeroth-order optimization methods for solving LQR and improved the sample complexity to $\mathcal{O}(\epsilon^{-1})$. Tu and Recht (2019) characterized the sample complexity gap between model-based and model-free methods from an asymptotic viewpoint where their model-free algorithm is based on REINFORCE.

Apart from policy gradient methods, Tu and Recht (2018) studied the LSTD learning for LQR and derived the sample complexity to estimate the value function for a fixed policy. Subsequently, Krauth, Tu, and Recht (2019) established the convergence and sample complexity of the LSTD policy iteration method under the LQR setting. On the subject of adopting AC to solve LQR, Yang et al. (2019) provided the first finite-time analysis with convergence guarantee and sample complexity under the double-loop setting. For the more practical yet challenging single-sample two-timescale AC, there is no such theoretical guarantee so far, which is the focus of this paper.

Notation. For two sequences $\{x_n\}$ and $\{y_n\}$, we write $x_n = \mathcal{O}(y_n)$ if there exists an constant C such that $x_n \leq Cy_n$. We use $\|\omega\|$ to denote the ℓ_2-norm of a vector ω and use $\|A\|$ to denote the spectral norm of a matrix A. We also use $\|A\|_F$ to denote the Frobenius norm of a matrix A. We use $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$ to denote the minimum and maximum singular values of a matrix A respectively. We also use $\mathrm{Tr}(A)$ to denote the trace of a square matrix A. For any symmetric matrix $M \in \mathbb{R}^{n\times n}$, let $\mathrm{svec}(M) \in \mathbb{R}^{n(n+1)/2}$ denote the vectorization of the upper triangular part of M such that $\|M\|_F^2 = \langle \mathrm{svec}(M), \mathrm{svec}(M)\rangle$. Besides, let $\mathrm{smat}(\cdot)$ denote the inverse of $\mathrm{svec}(\cdot)$ so that $\mathrm{smat}(\mathrm{svec}(M)) = M$. We denote by $A \otimes_s B$ the symmetric Kronecker product of two matrices A and B.

Preliminaries

In this section, we introduce the AC algorithm and provide the theoretical background of LQR.

Actor-Critic Algorithms

Reinforcement learning problems can be formulated as a discrete-time Markov Decision Process (MDP), which is defined by $(\mathcal{X}, \mathcal{U}, \mathcal{P}, c)$. Here $\mathcal{X}$ and $\mathcal{U}$ denote the state and the action space, respectively. At each time step t, the agent selects an action $u_t \in \mathcal{U}$ according to its current state $x_t \in \mathcal{X}$. In return, the agent will transit into the next state x_{t+1} and receive a running cost $c(x_t, u_t)$. This transition kernel is defined by $\mathcal{P}$, which maps a state-action pair (x_t, u_t) to the probability distribution over x_{t+1}. The agent's behavior is defined by a policy $\pi_\theta(u|x)$ parameterized by θ, which maps a given state to a probability distribution over actions. In the following, we will use ρ_θ to denote the stationary state distribution induced by the policy π_θ.

The goal of the average RL is to learn a policy that minimizes the infinite-horizon time-average cost (Sutton et al. 1999; Yang et al. 2019), which is given by

$$J(\theta) := \lim_{T\to\infty} \mathbb{E}_\theta \frac{\sum_{t=0}^{T} c(x_t, u_t)}{T} = \mathop{\mathbb{E}}_{x\sim\rho_\theta, u\sim\pi_\theta}[c(x,u)], \tag{1}$$

where $\mathbb{E}_\theta$ denotes the expected value of a random variable whose state-action pair (x_t, u_t) is obtained from policy π_θ. Under this setting, the state-action value of policy π_θ can be calculated as

$$Q_\theta(x,u) = \sum_{t=0}^{\infty} \mathbb{E}_\theta[c(x_t,u_t) - J(\theta)|x_0 = x, u_0 = u]. \tag{2}$$

The typical AC consists of two alternate processes: (1) critic update, which estimates the Q-function $Q_\theta(x,u)$ of current policy π_θ using temporal difference (TD) learning (Sutton and Barto 2018), and (2) actor update, which improves the policy to reduce the cost function $J(\theta)$ via gradient descent. By the policy gradient theorem (Sutton et al. 1999), the gradient of $J(\theta)$ with respect to parameter θ is characterized by

$$\nabla_\theta J(\theta) = \mathbb{E}_{x\sim\rho_\theta, u\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(u|x) \cdot Q_\theta(x,u)].$$

One can also choose to update the policy using the natural policy gradient, which is the basic idea behind natural AC algorithms (Kakade 2001). The natural policy gradient is given by

$$\nabla_\theta^N J(\theta) = F(\theta)^\dagger \nabla_\theta J(\theta). \tag{3}$$

where

$$F(\theta) = \mathbb{E}_{x\sim\rho_\theta, u\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(u|x)(\nabla_\theta \log \pi_\theta(u|x))^\top]$$

is the Fisher information matrix and $F(\theta)^\dagger$ denotes its Moore Penrose inverse.

The Linear Quadratic Regulator Problem

As a canonical optimal control problem, the linear quadratic regulator (LQR) has become a convenient testbed for the optimization landscape analysis of various RL methods. In this paper, we aim to analyze the convergence performance of the AC algorithm applied to LQR. In particular, we consider a stochastic version of LQR (called noisy LQR), where the system dynamics and the running cost are specified by

$$x_{t+1} = Ax_t + Bu_t + \epsilon_t, \tag{4a}$$

$$c(x, u) = x^\top Qx + u^\top Ru. \tag{4b}$$

Here $x_t \in \mathbb{R}^d$ and $u_t \in \mathbb{R}^k$, $A \in \mathbb{R}^{d\times d}$ and $B \in \mathbb{R}^{d\times k}$ are system matrices, $Q \in \mathbb{S}^{d\times d}$ and $R \in \mathbb{S}^{k\times k}$ are performance matrices, and $\epsilon_t \sim \mathcal{N}(0, D_0)$ are i.i.d Gaussian random variables with $D_0 > 0$.

The goal of the noisy LQR problem is to find an action sequence $\{u_t\}$ that minimizes the following infinite-horizon time-average cost

$$\underset{\{u_t\}}{\text{minimize}} \quad J(\{u_t\}) := \lim_{T\to\infty} \mathbb{E}[\frac{1}{T}\sum_{t=1}^{T} x_t^\top Qx_t + u_t^\top Ru_t]$$

$$\text{subject to} \quad \text{(4a)}.$$

From the optimal control theory (Anderson and Moore 2007; Bertsekas 2011, 2019), the optimal policy is given by a linear feedback of the state

$$u_t = -K^* x_t, \tag{5}$$

where $K^* \in \mathbb{R}^{k\times d}$ can be calculated as

$$K^* = (B^\top P^* B)^{-1} B^\top P^* A$$

with P^* being the unique solution to the following Algebraic Riccati Equation (ARE) (Anderson and Moore 2007)

$$P^* = Q + A^\top P^* A - A^\top P^* B(B^\top P^* B + R)^{-1} B^\top P^* A.$$

Actor-critic for LQR

Although the optimal solution of LQR can be easily found by solving the corresponding ARE, its solution relies on the complete model knowledge. In this paper, we pursue finding the optimal policy in a *model-free* way by using the AC method, without knowing or estimating A, B, Q, R.

Based on the structure of the optimal policy in (5), we parameterize the policy as

$$\{\pi_K(\cdot|x) = \mathcal{N}(-Kx, \sigma^2 I_k), K \in \mathbb{R}^{k\times d}\}, \tag{6}$$

where K is the policy parameter to be solved and $\sigma > 0$ is the standard deviation of the exploration noise. In other words, given a state x_t, the agent will take an action u_t according to $u_t = -Kx_t + \sigma\zeta_t$, where $\zeta_t \sim \mathcal{N}(0, I_k)$. As a consequence, the closed-loop form of system (4a) under policy (6) is given by

$$x_{t+1} = (A - BK)x_t + \xi_t, \tag{7}$$

where

$$\xi_t = \epsilon_t + \sigma B\zeta_t \sim \mathcal{N}(0, D_\sigma)$$

with $D_\sigma = D_0 + \sigma^2 BB^\top$.

The set $\mathbb{K}$ of all stabilizing policies is given by

$$\mathbb{K} := \left\{K \in \mathbb{R}^{k\times d} : \rho(A - BK) < 1\right\}, \tag{8}$$

where $\rho(\cdot)$ denotes the spectral radius. Before adopting AC to find the optimal policy π_{K^*} that minimizes the corresponding cost $J(K)$ defined in (1), we first need to establish the analytical formula of the average cost $J(K)$, the Q-function $Q_K(x, u)$, and the policy gradient $\nabla_K J(K)$ for a given stabilizing policy.

It is well known that if $K \in \mathbb{K}$, the Markov chain in (7) has a stationary distribution $\mathcal{N}(0, D_K)$, where D_K satisfies the following Lyapunov equation

$$D_K = D_\sigma + (A - BK)D_K(A - BK)^\top. \tag{9}$$

Similarly, we define P_K as the unique positive definite solution to

$$P_K = Q + K^\top RK + (A - BK)^\top P_K(A - BK). \tag{10}$$

Based on D_K and P_K, the following lemma characterizes the explicit expression of $J(K)$ and its gradient $\nabla_K J(K)$.

Lemma 1. *(Yang et al. 2019) For any $K \in \mathbb{K}$, the cost function $J(K)$ and its gradient $\nabla_K J(K)$ take the following forms*

$$J(K) = \text{Tr}(P_K D_\sigma) + \sigma^2 \text{Tr}(R), \tag{11a}$$

$$\nabla_K J(K) = 2E_K D_K, \tag{11b}$$

where $E_K := (R + B^\top P_K B)K - B^\top P_K A$.

It can be shown that the natural gradient of $J(K)$ can be calculated as (Fazel et al. 2018; Yang et al. 2019)

$$\nabla_K^N J(K) = \nabla_K J(K) D_K^{-1} = E_K. \tag{12}$$

Note that we omit the constant coefficient since it can be absorbed by the stepsize.

The expression of $Q_K(x, u)$ will also play an important role in our analysis later on.

Lemma 2. *(Bradtke, Ydstie, and Barto 1994; Yang et al. 2019) For any $K \in \mathbb{K}$, the Q-function $Q_K(x, u)$ takes the following form*

$$\begin{aligned} Q_K(x, u) =& (x^\top, u^\top)\Omega_K \begin{pmatrix} x \\ u \end{pmatrix} \\ & - \sigma^2 \text{Tr}(R + P_K BB^\top) - \text{Tr}(P_K D_K), \end{aligned} \tag{13}$$

where

$$\Omega_K := \begin{bmatrix} \Omega_K^{11} & \Omega_K^{12} \\ \Omega_K^{21} & \Omega_K^{22} \end{bmatrix} := \begin{bmatrix} Q + A^\top P_K A & A^\top P_K B \\ B^\top P_K A & R + B^\top P_K B \end{bmatrix}. \tag{14}$$

Single-sample Natural Actor-Critic

The expressions of $\nabla J(K)$ and $Q_K(x,u)$ in Lemmas 1 and 2 depend on A, B, Q, and R. For model-free learning, we establish a single-sample two-timescale AC algorithm for LQR in the following.

In view of the structure of the Q-function given in (13), we define the following feature functions,

$$\phi(x,u) = \text{svec}\left[\begin{pmatrix} x \\ u \end{pmatrix}\begin{pmatrix} x \\ u \end{pmatrix}^\top\right].$$

Then we can parameterize $Q_K(x,u)$ by the following linear function

$$\hat{Q}_K(x,u;w) = \phi(x,u)^\top w + b.$$

To drive $\hat{Q}_K(x,u;w)$ towards its true value $Q_K(x,u)$ in a model-free way, the TD learning technique is applied to tune its parameters ω:

$$\begin{aligned}\omega_{t+1} &= \omega_t + \beta_t[(c_t - J(K) + \phi(x_{t+1},u_{t+1})^\top\omega_t \\ &\quad + b - \phi(x_t,u_t)^\top\omega_t - b)]\phi(x_t,u_t) \\ &= \omega_t + \beta_t[(c_t - J(K))\phi(x_t,u_t) \\ &\quad - \phi(x_t,u_t)(\phi(x_t,u_t) - \phi(x_{t+1},u_{t+1}))^\top)\omega_t],\end{aligned} \tag{15}$$

where β_t is the step size of the critic.

To further simplify the expression, we denote (x',u') as the subsequent state-action pair of (x,u) and abbreviate $\mathbb{E}_{x\sim\rho_K,u\sim\pi_K(\cdot|x)}$ as $\mathbb{E}_{(x,u)}$. By taking the expectation of ω_{t+1} in (15) with respect to the stationary distribution, for any given ω_t, the expected subsequent critic can be written as

$$\mathbb{E}[\omega_{t+1}|\omega_t] = \omega_t + \beta_t(b_K - A_K\omega_t), \tag{16}$$

where

$$\begin{aligned}A_K &= \mathbb{E}_{(x,u)}[\phi(x,u)(\phi(x,u) - \phi(x',u'))^\top)], \\ b_K &= \mathbb{E}_{(x,u)}[(c(x,u) - J(K))\phi(x,u)].\end{aligned} \tag{17}$$

Given a policy π_K, it is not hard to show that if the update in (16) has converged to some limiting point ω_K^*, i.e., $\lim_{t\to\infty}\omega_t = \omega_K^*$, ω_K^* must be the solution of

$$A_K\omega = b_K. \tag{18}$$

We characterize the uniqueness and the explicit expression of ω_K^* in Proposition 3.

Proposition 3. *Suppose $K \in \mathbb{K}$. Then the matrix A_K defined in* (17) *is invertible such that the linear equation* (18) *has an unique solution ω_K^*, which is in the form of*

$$\omega_K^* = \text{svec}(\Omega_K). \tag{19}$$

Combining (12), (14), and (19), we can express the natural gradient of $J(K)$ using only ω_K^*:

$$\nabla_K^N J(K) = \Omega_K^{22}K - \Omega_K^{21} = \text{smat}(\omega_K^*)^{22}K - \text{smat}(\omega_K^*)^{21}.$$

This enables us to estimate the natural policy gradient using the critic parameters ω_{t+1}, and then update the actor in a model-free manner

$$K_{t+1} = K_t - \alpha_t\widehat{\nabla_{K_t}^N J(K_t)}, \tag{20}$$

where α_t is the (actor) step size and $\widehat{\nabla_{K_t}^N J(K_t)}$ is the natural gradient estimation depending on ω_{t+1}:

$$\widehat{\nabla_{K_t}^N J(K_t)} = \text{smat}(\omega_{t+1})^{22}K_t - \text{smat}(\omega_{t+1})^{21}. \tag{21}$$

With the critic update rule (15) and the actor update rule (20) in place, we are ready to present the following single-sample two-timescale natural AC algorithm for LQR.

Algorithm 1: Single-Sample Two-timescale Natural Actor-Critic for Linear Quadratic Regulator

1: **Input** initialize actor parameter $K_0 \in \mathbb{K}$, critic parameter ω_0, average cost η_0, step sizes α_t, β_t, and γ_t.
2: **for** $t = 0, 1, 2, \cdots, T-1$ **do**
3: Sample x_t from the stationary distribution ρ_{K_t}.
4: Take action $u_t \sim \pi_{K_t}(\cdot|x_t)$ and receive $c_t = c(x_t,u_t)$ and the subsequent state x_t'.
5: Obtain $u_t' \sim \pi_{K_t}(\cdot|x_t')$.
6: TD error calculation
$$\delta_t = c_t - \eta_t + \phi(x_t',u_t')^\top\omega_t - \phi(x_t,u_t)^\top\omega_t$$
7: Average cost estimate
$$\eta_{t+1} = \Pi_U(\eta_t + \gamma_t(c_t - \eta_t))$$
8: Critic estimate
$$\omega_{t+1} = \Pi_{\bar{\omega}}(\omega_t + \beta_t\delta_t\phi(x_t,u_t))$$
9: Actor update
$$K_{t+1} = K_t - \alpha_t(\text{smat}(\omega_{t+1})^{22}K_t - \text{smat}(\omega_{t+1})^{21})$$
10: **end for**

We call this algorithm "single-sample" because we only use exactly one sample to update the critic and the actor at each step. Line 3 of Algorithm 1 samples from the stationary distribution corresponding to policy π_K, which is common in analysis of the LQR problem (Yang et al. 2019). Such a requirement is only made to simplify the theoretical analysis. As shown in Tu and Recht (2018), when $K \in \mathbb{K}$, the Markov chain in (7) is geometrically β-mixing and thus mixes quickly. Therefore, in practice, one can run the Markov chain in (7) for a sufficient time and sample from the last one.

Since the update of the critic parameter in (15) requires the knowledge of the average cost $J(K)$, Line 7 is to provide an estimate of the cost function $J(K)$. Besides, compared with (15), we introduce a projection operator in Line 8 to keep the critic norm-bounded, which is necessary to stabilize the algorithm and attain convergence guarantee. Similar operation has been commonly adopted in other literature (Wu et al. 2020; Yang et al. 2019; Xu, Wang, and Liang 2020b).

Main Theory

In this section, we establish the global convergence and analyze the finite-sample performance of Algorithm 1. All the corresponding proofs are provided in Chen et al. (2022).

Before preceding, the following assumptions are required, which are standard in the theoretical analysis of AC methods (Wu et al. 2020; Fu, Yang, and Wang 2020; Yang et al. 2019; Zhou and Lu 2022).

Assumption 4. *There exists a constant $\bar{K} > 0$ such that $\|K_t\| \leq \bar{K}$ for all t.*

The above assumes the uniform boundedness of the actor parameter. One can also ensure this by adding a projection operator to the actor. In this paper, we follow the previous works (Konda and Tsitsiklis 1999; Bhatnagar et al. 2009; Karmakar and Bhatnagar 2018; Barakat, Bianchi, and Lehmann 2022) to explicitly assume that our iterations remain bounded. As shown in our proof, it is only made to guarantee the uniform boundedness of the feature functions, which is a standard assumption in the literature of AC methods with linear function approximation (Xu, Wang, and Liang 2020b; Wu et al. 2020; Fu, Yang, and Wang 2020).

Assumption 5. *There exists a constant $\rho \in (0,1)$ such that $\rho(A - BK_t) \leq \rho$ for all t.*

Assumption 5 is made to ensure the stability of the closed loop systems induced in each iteration and thus ensure the existence of the stationary distribution corresponding to policy π_{K_t}. In the single-sample case, the estimation of the natural gradient of $J(K)$ is biased and the policy change is noisy. Therefore, it is difficult to obtain a theoretical guarantee for this condition. Nevertheless, we will present numerical examples to support this assumption. Moreover, the assumption for the existence of stationary distribution is common and has been widely used in Zhou and Lu (2022); Olshevsky and Gharesifard (2022).

Under these two assumptions, we can now prove the convergence of Algorithm 1.

We first establish the finite-time convergence of the critic learning.

Theorem 6. *Suppose that Assumptions 4 and 5 hold. Choose $\alpha_t = \frac{c_\alpha}{(1+t)^\delta}, \beta_t = \frac{1}{(1+t)^v}, \gamma_t = \frac{1}{(1+t)^v}$, where $0 < v < \delta < 1$, c_α is a small positive constant. We have*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\omega_t - \omega^*_{K_t}\|^2] = \mathcal{O}(\frac{1}{T^{1-v}}) + \mathcal{O}(\frac{1}{T^v}) + \mathcal{O}(\frac{1}{T^{2(\delta-v)}}).$$

Note that $\|\omega_t - \omega^*_{K_t}\|^2$ measures the difference between the estimated and true parameters of the corresponding Q-function under K_t. Despite the noisy single-sample critic update, this result establishes the convergence and characterizes the sample complexity of the critic for Algorithm 1. The complexity order depends on the selection of step sizes δ and v, which will be determined optimally later according to the finite-time bound of the actor.

Based on this finite-time convergence result of the critic, we further characterize the global convergence of Algorithm 1 below.

Theorem 7. *Suppose that Assumptions 4 and 5 hold. Choose $\alpha_t = \frac{c_\alpha}{(1+t)^\delta}, \beta_t = \frac{1}{(1+t)^v}, \gamma_t = \frac{1}{(1+t)^v}$, where $0 < v < \delta < 1$, c_α is a small positive constant. We have*

$$\min_{0\leq t<T}\mathbb{E}[J(K_t) - J(K^*)] = \mathcal{O}(\frac{1}{T^{1-\delta}}) + \mathcal{O}(\frac{1}{T^v}) + \mathcal{O}(\frac{1}{T^{2(\delta-v)}}).$$

The optimal convergence rate of the actor is attained at $\delta = \frac{3}{5}$ and $v = \frac{2}{5}$. In particular, to obtain an ϵ-optimal policy, the optimal complexity of Algorithm 1 is $\mathcal{O}(\epsilon^{-2.5})$. To our knowledge, this is the first convergence result for solving LQR using single-sample two-timescale AC method.

To see the merit of our proof framework, we sketch the main proof steps of Theorems 6 and 7 in the following. The supporting propositions and theorems mentioned below can be found in Chen et al. (2022). Note that since the critic and the actor are coupled together, we define the following notations to clarify their dependency:

$$A(T) = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[(\eta_t - J(K_t))^2],$$
$$B(T) = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\omega_t - \omega^*_{K_t}\|^2],$$
$$C(T) = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla^N_{K_t} J(K_t)\|^2],$$

where $\nabla^N_{K_t} J(K_t) = E_{K_t}$ is defined in (12).

Proof Sketch:

1. Prove the convergence of the average cost. Note that in Line 7 of Algorithm 1, the average cost estimator η_t is only coupled with actor K_t via the cost c_t. We bound η_t utilizing the local Lipschitz continuity of $J(K)$ shown in Chen et al. (2022, Proposition 12) and the boundedness of K_t. Then it can be proved that
$$A(T) \leq \sqrt{A(T)}\mathcal{O}(T^{\frac{1}{2}-2(\delta-v)}) + \mathcal{O}(\frac{1}{T^{1-v}}) + \mathcal{O}(\frac{1}{T^v}),$$
where $\mathcal{O}(T^{\frac{1}{2}-2(\delta-v)})$ comes from the ratio between actor step size and critic step size, which reveals how the merit of two-timescale method can contribute to the proof. The other two terms $\mathcal{O}(\frac{1}{T^{1-v}})$ and $\mathcal{O}(\frac{1}{T^v})$ are induced by the step size for the average cost, which is $\gamma_t = \frac{1}{(1+t)^v}$. Solving this inequality gives the convergence of η_t which we presented in Chen et al. (2022, Theorem 13).
2. Prove the convergence of the critic. Note that the critic is coupled with both η_t and actor K_t. We decouple the critic and the actor in a similar way to step 1 utilizing the Lipschitz continuity of ω^*_t as shown in Chen et al. (2022, Proposition 15). Then, the following inequality can be obtained,
$$B(T) \leq \sqrt{A(T)B(T)} + \sqrt{\mathcal{O}(\frac{1}{T^{2(\delta-v)}})B(T)} + \mathcal{O}(\frac{1}{T^{1-v}}) + \mathcal{O}(\frac{1}{T^v}),$$

where $\sqrt{A(T)B(T)}$ shows the coupling between the average cost estimator η_t and the critic ω_t. Terms $\mathcal{O}(\frac{1}{T^{2(\delta-v)}})$, $\mathcal{O}(\frac{1}{T^{1-v}})$ and $\mathcal{O}(\frac{1}{T^v})$ are induced by the stepsizes. Combining the bound for $A(T)$, we can conclude the convergence of critic detailed in Theorem 6.

3. Prove the convergence of the actor. We utilize the almost smoothness property of the cost function $J(K)$ to establish the relation between actor, critic, and the natural gradient. We first prove that

$$C(T) \leq \sqrt{B(T)C(T)} + \mathcal{O}(\frac{1}{T^{1-\delta}}) + \mathcal{O}(\frac{1}{T^{\delta}}),$$

where $\sqrt{B(T)C(T)}$ shows the coupling between critic and actor. The terms $\mathcal{O}(\frac{1}{T^{1-\delta}})$ and $\mathcal{O}(\frac{1}{T^{\delta}})$ are induced by the step size of actor. By the convergence of critic established in Theorem 6, we can show that $C(T)$ converges to zero, which means the convergence of natural gradient. Finally, using the following gradient domination conditions (see Chen et al. (2022, Lemma 17))

$$J(K) - J(K^*) \leq \frac{1}{\sigma_{\min}(R)} \|D_{K^*}\| \mathrm{Tr}(E_K^\top E_K),$$

we can further prove the global convergence of actor shown in Theorem 7.

Experiments

In this section, we provide two examples to validate our theoretical results.

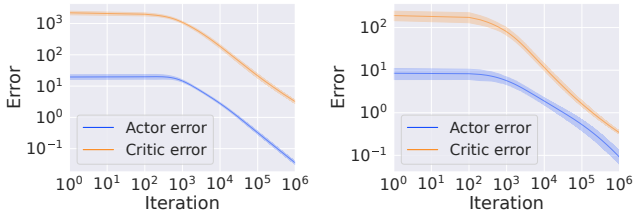
Example 1. *Consider a two-dimensional system with*

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, Q = \begin{bmatrix} 9 & 2 \\ 2 & 1 \end{bmatrix}, R = \begin{bmatrix} 1 & 2 \\ 2 & 8 \end{bmatrix}.$$

Example 2. *Consider a four-dimensional system with*

$$A = \begin{bmatrix} 0.2 & 0.1 & 1 & 0 \\ 0.2 & 0.1 & 0.1 & 0 \\ 0 & 0.1 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 \end{bmatrix}, B = \begin{bmatrix} 0.3 & 0 & 0 \\ 0.2 & 0 & 0.3 \\ 1 & 1 & 0.3 \\ 0.3 & 0.1 & 0.1 \end{bmatrix},$$
$$Q = \begin{bmatrix} 1 & 0 & 0.2 & 0 \\ 0 & 1 & 0.1 & 0 \\ 0.2 & 0.1 & 1 & 0.1 \\ 0 & 0 & 0.1 & 1 \end{bmatrix}, R = \begin{bmatrix} 1 & 0.1 & 1 \\ 0.1 & 1 & 0.5 \\ 1 & 0.5 & 2 \end{bmatrix}.$$

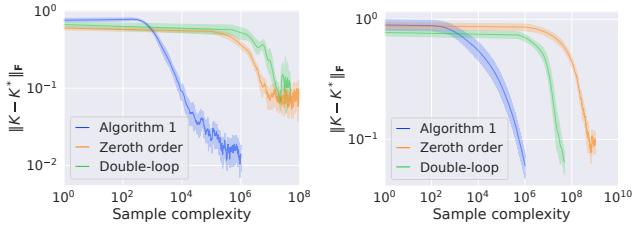
The learning results of Algorithm 1 for these two examples are shown in Figure 1. Consistent with our theoretical analysis, both the critic and the actor gradually converge to the optimal solution. Interested readers can refer to the Appendix of Chen et al. (2022) for experimental details.



(a) Example 1 (b) Example 2

Figure 1: Learning curves of the critic and the actor. Critic error refers to $\frac{1}{T}\sum_{t=0}^{T-1} \|\omega_t - \omega^*_{K_t}\|^2$ while actor error refers to $\frac{1}{T}\sum_{t=0}^{T-1}[J(K_t) - J(K^*)]$. The solid lines correspond to the mean and the shaded regions correspond to 95% confidence interval over 10 independent runs.

We also compare our algorithm with the double-loop AC algorithm proposed in Yang et al. (2019) and the zeroth-order method described in Fazel et al. (2018). We plotted the relative error of the actor parameters for all three methods in Figure 2. These simulation results show the superior sample-efficiency of Algorithm 1 empirically, confirming the practical wisdom of single sample two-timescale AC method.



(a) Example 1 (b) Example 2

Figure 2: Sample complexity comparison. The solid lines correspond to the mean and the shaded regions correspond to 95% confidence interval over 10 independent runs.

Conclusion and Discussion

In this paper, we establish the first finite-time global convergence analysis for the two-timescale AC method under LQR setting. We adopt a more practical single-sample two-timescale AC method and achieve an $\mathcal{O}(\epsilon^{-2.5})$ sample complexity. Our proof techniques of decoupling the actor and critic updates and controlling the accumulative estimate errors of the actor induced by the critic are novel and applicable to analyzing other AC methods where the actor and critic are updated simultaneously. Our future work includes further tightening the sample complexity bound under more relaxed settings and assumptions.

Acknowledgments

The work of X. Chen and L. Zhao was supported in part by the Singapore Ministry of Education Academic Research Fund Tier 1 (R-263-000-E60-133). The work of J. Duan is sponsored by NSF China with 52202487. The work of Y. Liang was supported in part by the U.S. National Science Foundation under the grant CCF-1761506.

References

Abbasi-Yadkori, Y.; Bartlett, P.; Bhatia, K.; Lazic, N.; Szepesvari, C.; and Weisz, G. 2019. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, 3692–3702. PMLR.

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

Anderson, B. D.; and Moore, J. B. 2007. *Optimal control: linear quadratic methods*. Courier Corporation.

Barakat, A.; Bianchi, P.; and Lehmann, J. 2022. Analysis of a Target-Based Actor-Critic Algorithm with Linear Function Approximation. In *International Conference on Artificial Intelligence and Statistics*, 991–1040. PMLR.

Bertsekas, D. 2019. *Reinforcement learning and optimal control*. Athena Scientific.

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

Bhandari, J.; and Russo, D. 2021. On the linear convergence of policy gradient methods for finite mdps. In *International Conference on Artificial Intelligence and Statistics*, 2386–2394. PMLR.

Bhandari, J.; Russo, D.; and Singal, R. 2018. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, 1691–1692. PMLR.

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

Bradtke, S. J.; Ydstie, B. E.; and Barto, A. G. 1994. Adaptive linear quadratic control using policy iteration. In *Proceedings of 1994 American Control Conference-ACC'94*, volume 3, 3475–3479. IEEE.

Castro, D. D.; and Meir, R. 2010. A convergent online single time scale actor critic algorithm. *The Journal of Machine Learning Research*, 11: 367–410.

Chen, X.; Duan, J.; Liang, Y.; and Zhao, L. 2022. Global Convergence of Two-timescale Actor-Critic for Solving Linear Quadratic Regulator. *arXiv preprint arXiv:2208.08744*.

Cohen, A.; Koren, T.; and Mansour, Y. 2019. Learning Linear-Quadratic Regulators Efficiently with only $\sqrt{T}$ Regret. In *International Conference on Machine Learning*, 1300–1309. PMLR.

Dann, C.; Neumann, G.; Peters, J.; et al. 2014. Policy evaluation with temporal differences: A survey and comparison. *Journal of Machine Learning Research*, 15: 809–883.

Dean, S.; Mania, H.; Matni, N.; Recht, B.; and Tu, S. 2018. Regret bounds for robust adaptive control of the linear quadratic regulator. *Advances in Neural Information Processing Systems*, 31.

Dean, S.; Mania, H.; Matni, N.; Recht, B.; and Tu, S. 2020. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4): 633–679.

Duan, J.; Li, J.; Li, S. E.; and Zhao, L. 2022. Optimization landscape of gradient descent for discrete-time static output feedback. In *2022 American Control Conference (ACC)*, 2932–2937. IEEE.

Fazel, M.; Ge, R.; Kakade, S.; and Mesbahi, M. 2018. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, 1467–1476. PMLR.

Fu, Z.; Yang, Z.; and Wang, Z. 2020. Single-Timescale Actor-Critic Provably Finds Globally Optimal Policy. In *International Conference on Learning Representations*.

Hu, Y.; Ji, Z.; and Telgarsky, M. 2021. Actor-critic is implicitly biased towards high entropy optimal policies. In *International Conference on Learning Representations*.

Kakade, S.; and Langford, J. 2002. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer.

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

Karmakar, P.; and Bhatnagar, S. 2018. Two time-scale stochastic approximation with controlled Markov noise and off-policy temporal-difference learning. *Mathematics of Operations Research*, 43(1): 130–151.

Khodadadian, S.; Doan, T. T.; Romberg, J.; and Maguluri, S. T. 2022. Finite sample analysis of two-time-scale natural actor-critic algorithm. *IEEE Transactions on Automatic Control*.

Konda, V.; and Tsitsiklis, J. 1999. Actor-critic algorithms. *Advances in neural information processing systems*, 12.

Krauth, K.; Tu, S.; and Recht, B. 2019. Finite-time analysis of approximate policy iteration for the linear quadratic regulator. *Advances in Neural Information Processing Systems*, 32.

Kumar, H.; Koppel, A.; and Ribeiro, A. 2019. On the sample complexity of actor-critic method for reinforcement learning with function approximation. arXiv:1910.08412.

Malik, D.; Pananjady, A.; Bhatia, K.; Khamaru, K.; Bartlett, P.; and Wainwright, M. 2019. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2916–2925. PMLR.

Mania, H.; Tu, S.; and Recht, B. 2019. Certainty equivalence is efficient for linear quadratic control. *Advances in Neural Information Processing Systems*, 32.

Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; and Kavukcuoglu, K. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, 1928–1937. PMLR.

Nesterov, Y.; and Polyak, B. T. 2006. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1): 177–205.

Olshevsky, A.; and Gharesifard, B. 2022. A Small Gain Analysis of Single Timescale Actor Critic. arXiv:2203.02591.

Polyak, B. T. 1963. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4): 864–878.

Qiu, S.; Yang, Z.; Ye, J.; and Wang, Z. 2021. On finite-time convergence of actor-critic algorithm. *IEEE Journal on Selected Areas in Information Theory*, 2(2): 652–664.

Ruder, S. 2016. An overview of gradient descent optimization algorithms. arXiv:1609.04747.

Shalev-Shwartz, S.; and Ben-David, S. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Silver, D.; Lever, G.; Heess, N.; Degris, T.; Wierstra, D.; and Riedmiller, M. 2014. Deterministic policy gradient algorithms. In *International conference on machine learning*, 387–395. PMLR.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. 2017. Mastering the game of go without human knowledge. *nature*, 550(7676): 354–359.

Sutton, R. S. 1984. *Temporal credit assignment in reinforcement learning*. University of Massachusetts Amherst.

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

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

Sutton, R. S.; McAllester, D.; Singh, S.; and Mansour, Y. 1999. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12.

Tu, S.; and Recht, B. 2018. Least-squares temporal difference learning for the linear quadratic regulator. In *International Conference on Machine Learning*, 5005–5014. PMLR.

Tu, S.; and Recht, B. 2019. The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. In *Conference on Learning Theory*, 3036–3083. PMLR.

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

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

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3): 229–256.

Witten, I. H. 1977. An adaptive optimal controller for discrete-time Markov environments. *Information and control*, 34(4): 286–295.

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

Xu, T.; Wang, Z.; and Liang, Y. 2020a. Improving sample complexity bounds for (natural) actor-critic algorithms. *Advances in Neural Information Processing Systems*, 33: 4358–4369.

Xu, T.; Wang, Z.; and Liang, Y. 2020b. Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms. arXiv:2005.03557.

Yang, Z.; Chen, Y.; Hong, M.; and Wang, Z. 2019. Provably global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. *Advances in neural information processing systems*, 32.

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

Zhou, M.; and Lu, J. 2022. Single Time-scale Actor-critic Method to Solve the Linear Quadratic Regulator with Convergence Guarantees. arXiv:2202.00048.